MAT137 Lecture Notes

Tyler Holden, ©2014-2015

Contents

1 Logic and Proofs

Many of you have already waded through the quagmire that is high-school calculus, endlessly berated with salvos of mindless computational questions asking you to find numbers with which you associate no meaning. This is what I call 'recipe mathematics,' wherein the student is provided with a recipe and the necessary ingredients, and is told to prepare a mathematical meal.

This is far removed from our goal in this class: our goal is to show the students how real mathematics is done. We are going to be problem solving and proving theorems, and this is a murky world which is hard to teach, and harder to learn.

Mathematics is often obsessed with *rigour*: the process of removing ambiguity and attaining the highest possible level of absolute, infallible deduction. To do this, the student must first understand the rigid framework in which mathematics is presented.

1.1 Sets and notation

Before we can begin to speak complicated sentences, we must first learn the words of a language. A *set* is any collection of well-defined and distinct objects. By this we mean that you can put as many things as you like into a set, so long as they are concrete and all different. We often surround the elements of a set by curly braces $\{,\}$, for example

$$\{1, 2, 3, ...\}, \qquad \{\text{cat}, \text{dog}, \text{bird}\}, \qquad \{☺, ☹\}, \qquad \{\heartsuit, \clubsuit, \diamondsuit, \spadesuit\}.$$

We can put anything we want into a set[1] so long as the object is a well-defined thing (for example, we cannot consider the set of all objects which I think are interesting. What objects are in this set? It is ambiguous), and all the elements of the set are distinct (so the object $\{1, 1, 2\}$ is not a set, because the element 1 appears multiple times).

We use the symbol '$\in$' (read as 'in') to talk about when an element is in a set; for example, $1 \in \{1, 2, 3\}$ but $☹ \notin \{\text{dog}, \text{cat}\}$. We can also talk about subsets, which are collections of items in a set and indicated with a '$\subseteq$' sign. For example,

$$\{2, 4, 6\} \subseteq \{1, 2, 3, 4, 5, 6\}$$

since every element on the left-hand-side is also present in the right-hand-side.

Since sets can have many objects within them, it is often impractical to list them all explicitly. Instead, we might use *set-builder* notation, which allows to say "the set of all things which satisfy some property." For example,

$$\{x : x > 0\}$$

is read as "the set of all x such that x is greater than 0," while

$$\{\text{month} : \text{month ends in 'ber'}\} = \{\text{September, October, November, December}\},$$

[1] This is actually a complete lie, but the reason why it is a lie is rather subtle. The quintessential example of this is something known as *Russell's paradox*. Let S be the set whose elements are the sets which do not contain themselves as subsets. Is S an element of itself? This is a self-referential paradox.

is the set of all months for which the end of the name of the month ends in 'ber'. Another way we could say this, is to let M be the set of all months, and consider the set

$$\{x \in M : x \text{ ends in 'ber'}\}.$$

More interesting than months are sets of numbers, since they tend to be quite big. Some sets that we will be involved with a lot are as follows:

- The **naturals**[2] $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$,
- The **integers** $\mathbb{Z} = \{..., -2, -1, 0, 1, 2, ...\}$,
- The **rationals** $\mathbb{Q} = \{p/q : p, q \in \mathbb{Z}, q \neq 0\}$,
- The **reals** $\mathbb{R}$ (the set of all infinite decimal expansion).

I have been somewhat sloppy in defining the real numbers here since their construction is actually rather involved. Nonetheless, I believe the average student is familiar with some of the basic properties[3] of $\mathbb{R}$.

1.1.1 Intervals

In addition to the "universes" listed above, we will often find ourselves concerned with subsets of $\mathbb{R}$, called intervals. Almost certainly the student has seen these in one form or another, but we re-introduce them here to ensure that everyone is on the same page. Effectively, intervals represent "connected" subsets of real numbers, such as

$$1 \leq x \leq 5, \qquad -2 < x < 4, \qquad -10 < x \leq -1, \qquad x > 10.$$

These can be a pain to write down though, so for the sake of compactness we introduce a new notation: If a, b are real numbers with $a < b$, the following hold[4]

$$\begin{aligned}
(a, b) &= \{x \in \mathbb{R} : a < x < b\} \\
[a, b) &= \{x \in \mathbb{R} : a \leq x < b\} \\
(a, b] &= \{x \in \mathbb{R} : a < x \leq b\} \\
[a, b] &= \{x \in \mathbb{R} : a \leq x \leq b\}
\end{aligned}$$

Intervals of the form (a, b) are said to be *open intervals*, while $[a, b]$ are said to be *closed intervals*. The other two $[a, b)$ and $(a, b]$ are *half-open intervals*. With a slight abuse of notation, we can also choose to let our sets be unbounded by writing infinities:

$$\begin{aligned}
(-\infty, a) &= \{x \in \mathbb{R} : x < a\} \\
(a, \infty) &= \{x \in \mathbb{R} : x > a\}
\end{aligned}$$

[2] Some mathematicians do not believe that 0 is a natural number.

[3] Strange things can happen if we are not careful with defining the real numbers. For example, consider the number $0.\bar{9}$, where our bar indicates that the number 9 is repeated infinitely often after the decimal place. It turns out that $0.\bar{9} = 1$, demonstrating that decimal expansions of real numbers are not unique.

[4] Some people prefer to use the notation $]a, b[$ for the interval (a, b), but we will avoid this relatively uncommon practice in this course.

We must always use open brackets when enclosing the infinity, since to use closed brackets $[,]$ would imply a point at infinity, and no such point exists in[5] $\mathbb{R}$. Combining these together, we may write the set of all real numbers are $\mathbb{R} = (-\infty, \infty)$ (noting that we always use open brackets, since there is no "number infinity" contained in the interval).

1.2 Fundamental Logic

The logic we will use will be similar to that found in computer science and philosophy classes, and while we will introduce it in a philosophical context, do not underestimate its importance throughout mathematics.

1.2.1 If, Then

We say that $P(x)$ is a proposition if it has a truth value associated to it; for example, we might have $P(x) =$ "x is a dog," in which case

$$P(\text{Lassie}) = \text{true}, \qquad P(\text{Shitzu}) = \text{true}, \qquad P(\text{Stephen Harper}) = \text{false}.$$

This allows us to string together logical sentences, possibly the most important of which is the "if,then" statement: Let $P(x), Q(x)$ be propositions, and consider

"If $P(x)$, then $Q(x)$."

This is a sentence, and we can always write it down regardless of what $P(x)$ and $Q(x)$ might mean. However, what we would really like to do is decide whether the sentence is true. For example, if $P(x) =$"x is a dog" and $Q(x) =$"x is an animal" then the statement

"If $P(x)$, then $Q(x)$" translates to "If x is a dog, then x is an animal."

This sentence is certainly true, since every dog is an animal. More generally (and confusingly), the statement "If $P(x)$, then $Q(x)$" is true if whenever $P(x)$ is true then $Q(x)$ is true.[6] In the interest of removing all words, we sometimes write $P(x) \Rightarrow Q(x)$ in lieu of the "If, then" sentence.

The *converse* of $P(x) \Rightarrow Q(x)$ is the statement $Q(x) \Rightarrow P(x)$. Note that *these are not logically equivalent*! For example, taking $P(x)$ and $Q(x)$ as before, the converse of "If x is a dog, then x is an animal" becomes "If x is an animal, then it is a dog" and this is certainly false, since not all animals are dogs.

The interesting thing happens when both $P(x) \Rightarrow Q(x)$ and the converse $Q(x) \Rightarrow P(x)$ are both true. In this case we write $P(x) \Leftrightarrow Q(x)$ (read as "$P(x)$ if and only if $Q(x)$"), and this means that $P(x)$ and $Q(x)$ are logically equivalent; namely, the truth of $P(x)$ is exactly the same as the truth of $Q(x)$. For example,

[5] When we formally add the points $\pm\infty$ to $\mathbb{R}$, we get what are called the *extended reals*. Weirder yet, if we make $+\infty = -\infty$, then what we get in the end is a circle! I may elaborate more on this later in these notes.

[6] $P(x) \Rightarrow Q(x)$ also evaluates to true whenever $P(x)$ is false, and this is called a *vacuous truth*. However, I do not want to confuse the student with why this is true right now, so you should ignore this fact if you find it confusing.

An integer is even *if and only if* it is divisible by 2.

These two statements are completely equivalent: if an integer is divisible by 2 it is even, while if it is not divisible by 2 then it is not even.

While we will explore it more in Section 1.3, we can 'prove' if-then statements, such as the following:

Example 1.1

If a and b are two odd integers, then their product ab is also an odd integer.

Solution. Recall that even integers can be written as $2n$ for some n (since they are divisible by 2) and so odd integers are written as $2n+1$. Since we have been told that a and b are odd, we know we can find m and n such that $a = 2n+1$ and $b = 2m+1$. Multiplying them together we get

$$ab = (2n+1)(2m+1) = 4nm + 2n + 2m + 1 = 2(2nm + n + m) + 1.$$

This number is again of the form 2(integer) + 1 and hence is odd, as we wanted to show. ■

1.2.2 Negation and the Contrapositive

If $P(x)$ is a proposition, we can *negate* its truth by creating the proposition "not $P(x)$", often written as $\neg P(x)$. This means that whenever $P(x)$ is true, $\neg P(x)$ is not true; namely, $P(x)$ is false. If $P(x)$ is as before

$$\neg P(\text{Lassie}) = \text{false}, \qquad \neg P(\text{Shitzu}) = \text{false}, \qquad \neg P(\text{Stephen Harper}) = \text{true}.$$

We mentioned in the last section that $P(x) \Rightarrow Q(x)$ is not equivalent to the converse, but what it is equivalent to is the *contrapositive*. The contrapositive of $P(x) \Rightarrow Q(x)$ is $\neg Q(x) \Rightarrow \neg P(x)$; in our usual example, this becomes

"If x is not an animal, then x is not a dog."

This is definitely still true. The idea is that if you're not even an animal, then you had no chance of being a dog.

The following is a (contrived) example:

Example 1.2

Let x be a positive integer. If $x \neq 1$, then $x \neq x^2$.

Solution. It is a big hassle to try and show this directly, since we would have to check every positive integer other than 1. Instead, we can try to show the contrapositive:

"Let x be a positive integer. If $x = x^2$ then $x = 1$.

Of course, this follows since

$$\begin{aligned} & x = x^2 \\ \Leftrightarrow \quad & x - x^2 = 0 \\ \Leftrightarrow \quad & x(1-x) = 0 \\ \Leftrightarrow \quad & x = 0 \text{ or } x = 1 \end{aligned}$$

and since we said x was a positive integer, x must be 1. ■

1.2.3 Quantifiers

Now we add some spice to our logical lives by introducing the notion of quantifiers. Quantifiers do exactly as the word implies: they denote some sort of quantity. We have two quantifiers in mathematics, the universal quantifier ($\forall$) and the existential quantifier ($\exists$). Remembering these as universal and existential would be silly, and so the student should instead read these as $\forall =$"for all" and $\exists =$"there exists". The easy way to remember this is that $\forall$ looks like an upside-down A corresponding to **A**ll , while $\exists$ looks like a backwards E, corresponding to **E**xists.[7] As an example, consider the sentence

$$\forall x \in \mathbb{R}, \exists y \in \mathbb{R}, y > x.$$

Right now this is just a bunch of symbols, and our job is to translate it into English. Using our newfound knowledge of quantifiers, we can directly translate this as

"For all x in the real numbers, there exists y in the real numbers, y is greater than x.

This is not particularly enlightening, so we take some linguistic liberties and re-arrange to say

"For every real number x, there is a real number y which is bigger than x."

That's much easier to read and to understand. In fact, we can even go one step further along the translation and write

"There is no largest real number,"

though that might have been a big conceptual jump from the last line (be sure to think about why those two sentences are the same!). This is true statement: there is no largest real number.

Order of Quantifiers: It is very important to realize that the order in which the quantifiers appears is important, as they do not "commute" with each other. More precisely, we always have that $\forall x \forall y = \forall y \forall x$ and $\exists x \exists y = \exists y \exists x$, but it may not be the case that $\forall x \exists y = \exists y \forall x$. The student is likely screaming about the equality sign here. What does it mean for these things to be equal!?

[7] Why is the A upside down and the E backwards? This is probably related to the mnemonic above. Unfortunately, the A and E exhibit symmetry relations: A is symmetric through a vertical axis while E is symmetric through a horizontal axis. Thus we must flip the A and reverse the E, necessitating the orientation we see today.

By equality, I simply mean that they evaluate to the same level of truth. An example should shed some light on this discussion.

Example 1.3

Consider the following logical statements:

$$\forall x \in \mathbb{Z}, \exists y \in \mathbb{Z}, \ x + y = 0 \tag{1.1}$$

and the statement

$$\exists x \in \mathbb{Z}, \forall y \in \mathbb{Z}, \ x + y = 0. \tag{1.2}$$

Compare these expressions by translating them as follows:

1. Convert the mathematical notation into English.
2. Turn the sentence derived above into a simple sentence, which does not involve any variables.

I would first like to comment that one of these statement is *true* and the other is *false*. Some students are resistant to idea of writing down false sentences mathematically, though this should not be an issue. Pigs can fly. See that? I just wrote down a false sentence in English. Why then is there any obstacle to writing false things mathematically?

Solution. We start with equation (1.1) for which a direct translation of the notation into English gives us

"For all x in the integers, there exists y in the integers, (such that) $x + y = 0$."

Unfortunately, this was the easy part. To see what this says in simple English, we proceed one step at a time. The statement $x + y = 0$ is equivalent to $x = -y$, so we can also read this as "for all x in the integers there exists y in the integers such that x is negative y." This is still pretty clumsy, so we now drop the variables and get

"Every integer has a negative."

I think it is pretty obvious that this is a true statement. If you give me an integer a, I can easily describe for you its negative. I simply throw a minus sign on it, giving you $-a$.

Let us now take a look at (1.2)

$$\exists y \in \mathbb{Z}, \forall x \in \mathbb{Z}, \ x + y = 0. \tag{1.3}$$

We can do our same translation trick, to determine that the corresponding English sentence is

"There is an element which is the negative of every integer."

Let us think about what this means. This says there is a number to which we can add any other number and always get zero. Certainly this is not true! If it were, then there would be a number n such that $n + a = 0$ and $n + b = 0$ for any integers a and b. Equating these expressions, we would find that $n + a = n + b$ which in turn implies that $a = b$. This would force all integers to be equal, and this is clearly nonsense. ■

By changing the order of the quantifiers, we have changed whether the statement is true or not, and hence we must be careful about the order in which quantifiers are presented.

Negating Quantifiers: Negating quantifiers turns out to be rather easy. Let $P(x)$ be a proposition, then negation of quantifiers works as follows:

Statement	Negation
$\forall x, P(x)$	$\exists x, \neg P(x)$
$\exists x, P(x)$	$\forall x, \neg P(x)$

As we can see, negating the universal quantifier gives the existential quantifier and vice versa. Whenever we go through a sentence, we just negate all of its parts.

Example 1.4

Negate the sentences given in (1.1) and (1.2).

Solution. Let's start with (1.1) which states

$$\forall x \in \mathbb{Z}, \exists y \in \mathbb{Z},\ x + y = 0.$$

Recall that in English this meant "Every integer has a corresponding negative integer." To negate this, we go step by step and negative each part of the sentence. Remember that $\forall$ will become $\exists$ and vice-versa. Hence we get

$$\exists x \in \mathbb{Z}, \forall y \in \mathbb{Z},\ x + y \neq 0.$$

Now intuitively, this should say the opposite of what we determined above; namely "There is some integer that does not have a corresponding negative number." Let's see if we get this using the same techniques we employed in Example **??**. Translating the English verbatim we get

"There exists an x in the integers, for all y in integers, $x + y \neq 0$."

This is a good start. Refining it further we get

"There is an integer such that no matter what we add to it, we can't get 0"

And this of course finally becomes the sentence we were expecting.

For (1.2) we move a bit quicker. The statement is

$$\exists x \in \mathbb{Z}, \forall y \in \mathbb{Z},\ x + y = 0$$

and its negation is therefore

$$\forall x \in \mathbb{Z}, \exists y \in \mathbb{Z},\ x + y \neq 0. \tag{1.4}$$

■

Exercise: Determine the English translation for Equation (1.4).

1.3 Definitions and Theorems

1.3.1 Definitions

Definitions are just what their name implies: they are a way of defining/creating a new notion by using older notions. For example, a random person off the street might walk up to you and ask

"What is a prime number?"

While we often have fuzzy, intuitive ideas about many notions, it does not do well to build a structure upon shaky ground. Since mathematics continuously builds upon itself, it is essential that we ensure our foundation is rock-solid. An answer to the above question might be something along the lines of

"A prime number is a number whose only factors are 1 and itself."

This is an acceptable definition in most cases, but what if I were to now ask you "Is 1 a prime number?" This is a somewhat tricky question: certainly the only factor of 1 is itself, but something feels dubious about this definition. It turns out that 1 *is not prime*, and the reason we were unable to deduce this is because our definition above is not sufficient.[8]

The additional course notes (located on the course website) has a great treatment of how to turn the intuitive idea of an increasing function into a proper definition. I will leave this to the student to read on his/her own. Instead, I'm going to introduce a different notion which we will use later in the course. I will first give you the intuitive idea, and then we will work towards building up a proper definition.

Injective Functions: Here's the idea that we want to define: we know that some functions are "many-to-one" in the sense that different inputs can yield the same output. A clear example of this is the square function $f(x) = x^2$. Notice that $f(-2) = 4 = f(2)$, and in general for any $a \neq 0$ we will have $f(-a) = a^2 = f(a)$. This function is thus two-to-one: there are two inputs which map to a single out. Similarly, the function $g(x) = \sin(x)$ is infinite-to-one, since for any $x \in \mathbb{R}$ we have

$$\sin(x) = \sin(x + 2\pi) = \sin(x + 4\pi) = \ldots.$$

We want to determine when a function is *one-to-one*, also known as an *injective* function.

There is a geometric way of determining whether a function is one-to-one, called the horizontal line test. Exactly akin to the vertical line test for determining whether a map is a function, the horizontal line test is roughly stated as

"The function $f(x)$ satisfies the horizontal line test if any horizontal line drawn in the plane meets the graph of $f(x)$ at most once."

[8]The real definition of a prime is somewhat complicated. The student should take it as "If $p \in \mathbb{Z}$ is a number not equal to 1, then p is prime if its only factors are 1 and itself". More generally, mathematicians define a prime as follows: "If $p \in \mathbb{Z}$ does not have a multiplicative inverse, then p is a prime if whenever p divides a product ab, then p divides either a or p divides b. This is a better definition because it extends nicely to sets other than just $\mathbb{Z}$.

Okay, so we are starting to get an idea for how these functions work, but we need to write down a formal definition. How do we start? There are two ways we can go, but I'm going to choose one and then I'll show you the other afterwards:

"A function f is injective on (a, b) if whenever we plug in two different numbers, the value of f at those two numbers is different."

This roughly captures what we are try to say: different inputs lead to different outputs. Clearly $f(x) = x^2$ does not satisfy this, since we can put in different numbers (2 and -2) and get the same output. Still, it's not a very rigorous definition: what are the inputs? What are the outputs? Let's spruce it up a bit more.

"A function f is injective on (a, b) if whenever $x, y \in (a, b)$ and $x \neq y$ then $f(x) \neq f(y)$."

This is actually a pretty good definition so far, but we could make it even better by throwing in a universal quantifier and mathematizing everything:

Definition 1.5

A function $f : (a, b) \to \mathbb{R}$ is injective on (a, b) if

$$\forall x \in (a, b), \forall y \in (a, b), \quad x \neq y \quad \Rightarrow \quad f(x) \neq f(y).$$

This is a nice definition in terms of understanding the concept, but it turns out to not be terribly practical to show that a function is injective using this definition. Far more useful is the contrapositive of this statement. Take a moment to see if you can determine the answer before reading on.

"A function $f : (a, b) \to \mathbb{R}$ is injective if for all $x, y \in (a, b)$, $f(x) = f(y)$ implies that $x = y$."

Example 1.6

If $c \neq 0$, show that the function $f(x) = cx$ is injective on all of $\mathbb{R}$.

Solution. Using the contrapositive definition of injective, we want to show that for all $x, y \in \mathbb{R}, f(x) = f(y)$ implies that $x = y$. Let x and y be arbitrarily chosen (choosing them arbitrarily is the same as letting them be any number, letting us treat all numbers at the same time!). If $f(x) = f(y)$ then by definition of f we must have $cx = cy$. Re-arranging, we can write this as

$$0 = cx - cy = c(x - y).$$

Now the product of two numbers is zero only if one of the factors is zero, but we know that $c \neq 0$ so it must be that $x - y = 0$ or equivalently, $x = y$. But this is exactly what we wanted to show! So $f(x) = cx$ is injective whenever $c \neq 0$. ■

Exercise: Show that the polynomial $f(x) = x^n$ is injective whenever n is odd, and is not injective whenever n is even.

Exercise: We say that a function $f : \mathbb{R} \to \mathbb{R}$ is *even* if $f(-x) = f(x)$. Can an even function ever be injective?

1.3.2 Theorems and Proofs

Theorems (or to a lesser degree, lemmas and propositions) are the major results of mathematics. Up to poetic licence, these are phrased in the form

Theorem: If HYPOTHESES, then RESULTS

When proving a theorem, one typically starts by assuming the hypotheses, then showing that the results are true. When applying a theorem, one must first check that the hypotheses are satisfied. Let's jump into examples to elucidate what we mean.

Theorem 1.7

If $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ are both injective functions, then their composition $f \circ g : \mathbb{R} \to \mathbb{R}$ is also injective.

This is clearly stated as an if-then theorem with

- **Hypothesis:** f and g are injective functions,
- **Result:** The composition $f \circ g$ is also injective; that is, $f(g(x)) = f(g(y))$ implies that $x = y$.

Proof. To prove this, we start with our hypotheses. Thus let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be injective functions. So that we have this information in front of us, let's recall what this means:

- Since f is injective, whenever $f(x) = f(y)$ we must have $x = y$.
- Since g is injective, whenever $g(x) = g(y)$ we must have $x = y$.

Okay, our goal is now to show that $f \circ g$ is injective. Looking at our definition of injective, we should start by assuming that $x, y \in \mathbb{R}$ satisfy $(f \circ g)(x) = (f \circ g)(y)$. We want to show that $x = y$.

At this point we get stuck, and when you get stuck it is usually a good idea to look back at your hypotheses and ask yourself "what haven't I used yet?" Since f is injective and $f(g(x)) = f(g(y))$, we must have that $g(x) = g(y)$. If this is not clear, try writing $n = g(x)$ and $m = g(y)$ so that

$$f(g(x)) = f(g(y)) \qquad \text{is equivalent to} \qquad f(m) = f(n).$$

Using the fact that f is injective implies that $m = n$, or equivalently, $g(x) = g(y)$.

Now we get stuck again since we can't seem to reduce any further. Looking back at our hypotheses, we have yet to use the fact that g is injective. But this is precisely what we need to finish our proof; namely, since g is injective, $g(x) = g(y)$ implies that $x = y$.

So let's quickly summarize what we found. We showed that if $f(g(x)) = f(g(y))$ then $g(x) = g(y)$, and this implied that $x = y$. This is the definition of what it means for $f \circ g$ to be injective, so we are done the proof! $\square$

Example 1.8

Assume that $g : \mathbb{R} \to \mathbb{R}$ is an injective function. Show that $g(x)^3$ is also an injective function.

Solution. If we think about this for a moment, we see that by defining $f(x) = x^3$ we get $g(x)^3 = f(g(x))$. Now we would like to apply Theorem 1.7, but first we must check that the hypotheses apply; that is, both f and g are injective.

We were told that $g(x)$ was injective to start with, so there's nothing to check there. The diligent student will have shown in Exercise 1.3.1 that $f(x) = x^3$ is injective. Hence both f and g are injective, and Theorem 1.7 applies. We conclude that the composition $f \circ g$ is also injective, hence $g(x)^3$ is injective as required. ■

Theorem 1.9

Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be two functions. If the function $f \circ g$ is injective, then g is injective.

Proof. This is a result that often confuses students, since it seems to go the wrong way compared to our last result. We are given two functions, and all we know is that when we combine them to form $f \circ g$ then this new function is injective. From that, we would like to deduce that g is injective.

- **Hypotheses:** $f \circ g$ is injective; that is, $f(g(x)) = f(g(y))$ implies that $x = y$.
- **Want to show:** g is injective; that is, $g(x) = g(y)$ implies that $x = y$.

Okay, so to show that g is injective, we should start with the expression we always start with when we want to show injectivity; namely, assume that $x, y \in \mathbb{R}$ satisfy $g(x) = g(y)$. At this point we are stuck and do not know what to do, so we go back to our hypotheses. By applying f to both sides of our equating, we can get $f(g(x)) = f(g(y))$. Now our hypothesis tells us that $x = y$, which is what we wanted to show. □

Example 1.10

Find an example of functions f and g such that $f \circ g$ is injective but f is not injective. Conclude that the converse of Theorem 1.7 does not hold.

Solution. Two possible solutions are as follows:

1. Let $f(x) = x^2$ and $g(x) = \sqrt{x}$. The student can easily check that f is not injective but g is injective. Now $f(g(x)) = (\sqrt{x})^2 = x$ is an injective function, as required.
2. Consider the functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ given by $f(x) = \tan(x)$ and $g(x) = \arctan(x)$. One can show that $g(x)$ is injective, and it is easy to see that $f(x)$ is not injective. However, their composition is

$$f(g(x)) = \tan(\arctan(x)) = x$$

which is again injective.

Now Theorem 1.7 said that if f and g are injective, so too is their composition. The converse of this statement would be

"If $f \circ g$ is injective, then both f and g are injective."

But we just showed in the above example that this is not true, so we conclude that the converse of Theorem 1.7 is not true. In fact, in general if $g(x)$ is injective but not surjective[9] then g will have a left-inverse f such that $f(g(x)) = x$ but f will not be injective itself. ■

1.4 Induction

Proof by mathematical induction is a proof technique used to show that a result holds for every natural number $\mathbb{N}$. For example, let's try and show that for every $n \in \mathbb{N}$, $2n \leq 2^n$.

Let's try out a few values of n and see what happens:

n	$2n$	2^n	$2n \leq 2^n$
0	0	1	true
1	2	2	true
2	4	4	true
3	6	8	true
4	8	16	true
5	10	32	true

The student can likely see the pattern, and it is apparent that 2^n is growing very quickly, so $2n$ will never catch up and the result will be true. Nonetheless, this is not a mathematical proof! It's not enough to wave your arms and say "it looks like it should be correct, so it is!"[10]

What mathematical induction brings to the table is a very succinct way of showing that something holds for all n, and it relies on the following principle:

> Let P be some proposition. If $P(1)$ is true, and $P(k) \Rightarrow P(k+1)$, then $P(n)$ is true *for all* $n \in \mathbb{N}$

The idea is that this principle acts like dominos. We know that $P(k) \Rightarrow P(k+1)$. Since $P(1)$ is true, $P(2)$ is therefore true. Since $P(2)$ is true, $P(3)$ is therefore true, etc. Hence $P(n)$ will be true for any n by just doing this chain long enough.

When proving something by mathematical induction, there is thus a two-step process that you should always observe:

[9] A function $g : \mathbb{R} \to \mathbb{R}$ is surjective if for every $y \in \mathbb{R}$ there exists an $x \in \mathbb{R}$ such that $f(x) = y$.

[10] For a long time, mathematicians have been trying to get bounds on the number of primes less than some number. There is a function $Li(x)$ which roughly counts the number of primes in the interval $[0, x]$, and for a long time it was thought that $Li(x)$ was always an *upper bound* for the number of primes. In 1914, we found a counterexample. The first known location of this counter-example occurs somewhere around 10^{316}.

1. **Base Case:** Show that $P(1)$ is true (or $P(k)$, where k is the smallest number for which the result holds).
2. **Induction Step:** Show that, under the assumption that $P(k)$ is true, $P(k+1)$ is also true.

Step (1) just confirms that $P(1)$ is true according to our induction principle, while step (3) is the $P(k) \Rightarrow P(k+1)$ step.

Example 1.11

Show, using mathematical induction, that $2n + 2 \leq 4n$ for all integers $n \geq 1$.

Solution.

1. **Base Case:** The smallest number for which this occurs is $n = 1$, and in this case we have $2n + 2 = 4$ and $4n = 4$, so the result holds in the base case.
2. **Induction Step:** Assume that $2k + 2 \leq 4k$ for some natural number k. We want to show that $2(k + 1) + 2 \leq 4(k + 1)$. Indeed, notice that

$$\begin{aligned} 4(k+1) &= 4k + 4 \\ &\geq (2k+2) + 2 && \text{using the induction hypothesis } 4k > 2k+2 \\ &= 2k + 4 = 2(k+1) + 2. \end{aligned}$$

We conclude from the induction principle that $2k + 2 \leq 4k$ for all $k \in \mathbb{N}$. ■

Exercise: Show that $2n \leq 2^n$ for all $n \in \mathbb{N}$. [Hint: You will need to use Example **??**.]

Example 1.12

Let k be some fixed positive integer such that $k \geq 1$. Show that

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

Solution. As always, we follow the program enumerated above.

1. **Base Case:** Here we check the easiest possible case, which corresponds to $k = 1$. When $k = 1$ the left-hand-side becomes

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

which the right-hand-side is $\frac{1}{1+1} = \frac{1}{2}$. Clearly both sides agree, so the base case is satisfied.

2. **Induction Step:** Let k be some fixed by arbitrary number and assume that

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

We would like to show that the result holds for $k+1$. It makes most sense to start by working with the left-hand-side, since it will give us the most "flexibility." Notice that

$$\begin{aligned}
\sum_{n=1}^{k+1} \frac{1}{n(n+1)} &= \left.\frac{1}{n(n+1)}\right|_{n=k+1} + \sum_{n=1}^{k} \frac{1}{n(n+1)} \\
&= \frac{1}{(k+1)(k+2)} + \frac{k}{k+1} && \text{via the induction hypothesis} \\
&= \frac{1+k(k+2)}{(k+1)(k+2)} = \frac{k^2+2k+1}{(k+1)(k+2)} && \text{common denominator} \\
&= \frac{(k+1)^2}{(k+1)(k+2)} = \frac{k+1}{k+2} && \text{factoring and cancelling} \\
&= \frac{(k+1)}{(k+1)+1}.
\end{aligned}$$

This is precisely what we wanted to show, and so we are done. ■

The point I would like to make is that while the above may have looked somewhat messy, there was really only one thing we could have done. By intelligently setting up the problem, the rest of the work simply "fell out.

Definition 1.13

Let m, n be integers. We say that $m|n$ (read as m *divides* n) if there exists some integer k such that $mk = n$.

To see what this means intuitively, notice that in general when we divide two integers, we get a rational number. For example, $\frac{3}{2}$ is rational and not an integer. However, we say that $m|n$ if $\frac{n}{m}$ is actually an integer. Thus $2|4$ since $\frac{4}{2} = 2$.

Example 1.14

Prove that for all positive integers k, $5|6^k - 1$.

Solution.

1. **Base Case:** The simplest case is $k = 1$, for which we see that $6^k - 1 = 5$. Clearly $5|5$ since $\frac{5}{5} = 1$, so the base case is satisfied.

2. **Induction Step:** For some positive integer k, assume that $5|6^k - 1$. Since by hypothesis, we know that $5|6^k - 1$ we know there is some integer d such that $\frac{6^k-1}{5} = d$. Consider $6^{k+1} - 1$

which we may write as

$$\begin{aligned} 6^{k+1} - 1 &= 6(6^k) - 1 \\ &= (1+5)(6^k) - 1 \\ &= 5(6^k) + (6^k - 1) \end{aligned}$$

We claim that 5 divides this number. To see that this is the case, let us divide by 5 and see what we get.

$$\begin{aligned} \frac{6^{k+1} - 1}{5} &= \frac{5(6^k) + (6^k - 1)}{5} \\ &= \frac{5 \cdot 6^k}{5} + \frac{6^k - 1}{5} \\ &= 6^k + d && \text{by induction hypothesis.} \end{aligned}$$

This is clearly an integer, so $5 | 6^{k+1} - 1$ as required. ■

2 Limits

2.1 A Quick Primer on Absolute Values

Absolute values often give students trouble, but will be absolutely crucial for the upcoming section. Our goal is thus twofold: the first is to ensure that the student understand the intuition for absolute values and why we use them; the second objective is to show students how to deal with problems that involve absolute values.

2.1.1 The Definition

The absolute value of a number n is denoted $|n|$, and many students often interpret the absolute value as being "the thing which makes n positive." This is correct to a degree, as evidenced by the proper definition of the absolute value.

Definition 2.1

The absolute value is a function $|\cdot| : \mathbb{R} \to \mathbb{R}$ defined as

$$|x| = \begin{cases} x & x \geq 0 \\ -x & x < 0 \end{cases}. \tag{2.1}$$

This definition is a little weird: we know that the absolute value of a number should always be positive, so why does this negative sign appear in the definition? Let's try computing a few absolute values using this definition, and hopefully that will make things clear:

Example 2.2

Compute the absolute values of 3 and -4.

Solution. We start by computing $|3|$. By looking at (2.1) we see that since $3 > 0$ the absolute value is given by just writing down the argument; namely, $|3| = 3$.

To compute $|-4|$ we must be a bit more careful. Clearly $-4 < 0$ which means we fall into the bottom-half of the piecewise definition, so the absolute value is the *negative* of the argument; namely $|-4| = -(-4) = 4$. ■

The absolute value performed as expected, when we put a negative number into it, we got its positive counterpart out. The fact is that when $x < 0$ we know that x is a negative number. The definition of the absolute value then tells us to set $|x| = -x$, which in turn guarantees that $-x > 0$.

More than just a technical definition, (2.1) is actually the key tool to apply when dealing with problems that involve absolute values: We break the problem into a piecewise problem using the definition.

Example 2.3

Determine the values of x which satisfy $|x+3| > 2x+4$

Solution. To get rid of the absolute value, we use (2.1). Indeed,

$$|x+3| = \begin{cases} x+3 & x+3 \geq 0 \\ -x-3 & x+3 < 0 \end{cases} = \begin{cases} x+3 & x \geq -3 \\ -x-3 & x < -3 \end{cases}.$$

This allows us to break the problem into two smaller problems, based on whether $x \geq -3$ or $x < -3$.

- **Case** $x \geq -3$: When $x \geq -3$ we have $|x+3| = x+3$, so our problem becomes $x+3 > 2x+4$. This is now easy to solve and we get that $x < -1$. Since we must have both $x \geq -3$ and $x < -1$, this will only be satisfied for $x \in [-3, -1)$.
- **Case** $x < -3$: When $x < -3$ we have $|x+3| = -x-3$, so our problem becomes $-x-3 > 2x+4$ which we can solve to get $x < -7/3$. Hence we must have both $x < -3$ and $x < -7/3$, which means that our equation is satisfied when $x \in (-\infty, -env3)$.

Combining both solutions, we get that $|x+3| > 2x+4$ when $x \in (-\infty, -3) \cup [-3, -1) = (-\infty, -1)$. ■

To re-iterate: when solving a problem which involves an absolute value, break the problem into cases to get rid of the absolute value, then solve as normal.

Very quickly, here are the important properties of the absolute value:

Proposition 2.4

The absolute value function satisfies the following properties:

1. **Multiplicative:** If $x, y \in \mathbb{R}$ then $|xy| = |x||y|$.
2. **Non-degenerate:** $|x| = 0$ if and only if $x = 0$
3. **Triangle Inequality:** For any $x, y \in \mathbb{R}$, $|x+y| \leq |x| + |y|$.

Exercise: Prove these properties of the absolute value using the definition.

2.1.2 A Measure of Distance

This tells us how to deal with absolute values when they arise and we will see more example shortly, but let's stop for a moment and think about what an absolute value means. An absolute value is a measure of *distance* or *length*, in particular, from the number 0. For example, $|3|$ is the distance from the number 0 to the number 3: unsurprising this is 3-units. On the other hand, $|-4|$ is the distance from 0 to the number -4, and that is 4-units (since distances are never negative).

Of far greater interest is that we can use the absolute value to measure distances between two non-zero numbers. For any $x, y \in \mathbb{R}$, the distance between x and y is $|x - y|$. For example, what do we intuitively expect the distance to be between the numbers -1 and 5? I think it's going to be 6 units, and indeed $|(-1) - 5| = |-6| = 6$ as we expected. Of course, distance does not care about which element 'came first' and we could have just as easily written $|5 - (-1)| = |6| = 6$ and gotten the same answer.

An incredibly useful thing that we should check is the following: Assume that $r > 0$ is some real number. What is the set $\{x \in \mathbb{R} : |x| < r\}$? If we read this properly, this should be the set of all elements that are within r of 0; namely, $(-r, r)$. Let's see if this is true:

First, we break the absolute value down into cases. We know that if $x \geq 0$ then $|x| = x$, and so if $x \geq 0$ and $x < r$ then $x \in [0, r)$. Similarly, if $x < 0$ then $|x| = -x$ and we have $-x < r$ which is the same as $x > -r$. In this case, we get $x \in (-r, 0)$. Combining the two solutions, we get

$$x \in (-r, 0) \cup [0, r) = (-r, r)$$

as required. We don't always want to write down intervals though, and we know that $x \in (-r, r)$ is the same thing as $-r < x < r$. Hence we can often make the following transformation:

$$|x| < r \quad \Leftrightarrow \quad -r < x < r. \tag{2.2}$$

Example 2.5

Write the interval $(1, 3)$ using an absolute value, thought of as a distance.

Solution. Equation (2.2) show us that writing $|x| < r$ is the same thing as $-r < x < r$, and this is very symmetric. To write $(1, 3)$ in the same way, we notice that the centre of $(1, 3)$ is the number 2, and that all points are within a distance of 1 from the center. We hypothesize that

$$(1, 3) = \{x \in \mathbb{R} : |x - 2| < 1\}.$$

Let's check to see if this is correct. Using a modified version of Equation (2.2) we get

$$\begin{aligned} |x - 2| < 1 &\Leftrightarrow -1 < x - 2 < 1 \\ &\Leftrightarrow -1 + 2 < x - 2 + 2 < 1 + 2 && \text{adding 2 to both sides} \\ &\Leftrightarrow 1 < x < 3 \\ &\Leftrightarrow x \in (1, 3). \end{aligned}$$

■

Example 2.6

If $|x + 1| < 2$, find an upper bound on the function $|2x + 4|$.

Solution. There are many ways to solve this, and since we were only asked to find an upper bound, there are infinitely many different solutions. One could probably guess that if we chose a very large number, it would probably bound $|2x+4|$, but we 1) want to prove it mathematically, and 2) often want as good a bound as possible.

Now certainly the function $|2x+4|$ is not bounded in general, but the fact that $|x+1| < 2$ means that we are only looking at a very small segment of the real line. The first method for solving this is to unravel the definition of $|x+1| < 2$ and see if we can transform it into something that looks like $|2x+4|$. Indeed, we know that

$$\begin{aligned} |x+1| < 2 \quad &\Leftrightarrow \quad -3 < x < 1 \\ &\Leftrightarrow \quad -6 < 2x < 2 && \text{multiply by 2} \\ &\Leftrightarrow \quad -2 < 2x+4 < 6 && \text{add 4} \end{aligned}$$

So we now have upper and lower bounds on the function $2x+4$, but we want to turn this into bounds on the function $|2x+4|$. If we think about this, we can see that since $2x+4$ can only take numbers in the interval $(-2, 6)$, then the largest it can be in absolute value is 6, hence $|2x+4| < 6$.

Alternatively, if we are clever we can use the triangle inequality. Since $|x+1| < 2$, we have that

$$|2x+4| = |2(x+1)+2| \leq 2|x+1| + 2 < 2(2) + 2 = 6$$

which is the same result we had before. ■

> **Exercise:** Let $f(x)$ be a function such that $a < f(x) < b$. Convince yourself that $|f(x)| < \max\{|a|, |b|\}$.

2.2 Limited Intuition

Limits are the method why which we, as manifestly finite being, deal with concepts of infinities and infinitesimals. The key goal towards which we are looking is the description of an instantaneous rate of change, so let us ponder what this means.

The majority of us have been in a car at some point or another, and have afforded a casual glance at the speedometer. Let us say that at the instant we look down, the speedometer reads 90 km/hr. Have you ever thought about what it means, at that single instant in time, to be travelling at that speed? As suggested by its units, speed is an object which requires both distance and time to measure, but at a single moment, neither any time nor any distance has passed, so what does this mysterious quantity mean?

Despite my claims that the previous example should get you thinking about how the word "instantaneous" really affects a quantity, many of you will simply shrug aside my suggestions. In anticipation of this reaction, what if we change the associated quantities around and instead of the instantaneous speed of a car, we discuss shopping! At any given point of time, somebody on this planet is making a purchase. Let us assume that we were able to measure the rate at which people were spending money, and I told you that at this moment in time the human species was globally spending $140 million dollars and hour? What does this mean!

Now on the other hand, what if you were asked to determine the instantaneous speed of a race car at the instant its front bumper passes a finish line? Being clever students, you decide to measure how far the car has travelled in the minute before it hits the finish line, and get a result of 1500 meters. Hence the car was travelling

$$\frac{1500 \text{ metres}}{1 \text{ minute}} \times \frac{1 \text{ kilometre}}{1000 \text{ metres}} \times \frac{60 \text{ seconds}}{1 \text{ hour}} = \frac{90\text{kilometres}}{1 \text{ hour}}.$$

But what if the cars speed was not constant during that minute? What if the driver accelerated at the end? You decide that you can get a better estimate of the speed at the finish line by instead just looking at how far the car travelled in the single second before the car hit the finish line. This time the car travelled 30 metres, so you calculate

$$\frac{30 \text{ metres}}{1 \text{ second}} = \frac{108 \text{ kilometres}}{1 \text{ hour}}.$$

But still, this does not account for any change in acceleration which occurred in the last second. Your guess of 108km/hr is probably close, but close is not good enough in mathematics! So you try again by measuring the distance after 0.1 seconds, then 0.01 seconds, and so on (does this remind you of anything?), but no matter how hard you try you cannot get the exact speed because there is always the chance that the car was not travelling at a constant speed during your measurements. Nonetheless, we know there must be an answer: the car was travelling at some speed, so what is it? Limits provide the solution.

2.2.1 Hand Waving Arguments

Limits are the mathematical device which allow us to infer information about a (possibly misbehaved) point by analyzing information about well-behaved points nearby. The way this is done in a formal environment can be a little bit messy, so let us take a moment to get some simple examples under our belts before diving headfirst into the chaos.

Let $f(x)$ be an arbitrary function and $c \in \mathbb{R}$. We say that "the limit of $f(x)$ as x approaches c is equal to L" if, whenever we let x get arbitrarily close to c then $f(x)$ gets arbitrarily close to L. This is written as

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

The best way to gain an intuitive understanding of limits is to see a few examples. We warn the student that this first example is rather nicely behaved and fails to capture why we use limits. Nonetheless, simple examples are often the best for getting a grasp as to how something works.

Example 2.7

Consider the function $f(x) = 4x + 2$. Determine the limits

$$\lim_{x \to 0} f(x), \qquad \lim_{x \to -4} f(x), \qquad \lim_{x \to 5} f(x).$$

Form a hypothesis as to what the limit is as $x \to c$ for any value of c.

Solution. We implore the student to keep in mind that this solution is purely heuristic and is only presented in a way to show the student how to think about these problems. A limit should *never* be computed in the following manner.

The first example asks us to consider what happens when $x = 0$, so we would like to see what happens for values of x which are close (but not equal to zero). Hopefully the student can guess that as x gets close to zero, $4x + 2$ gets close to $4 \cdot 0 + 2 = 2$. Similarly, as $x \to -4$ then $4x + 2$ approaches $4 \cdot (-4) + 2 = -14$. The following table corroborates this idea:

$x \to 0$				$x \to -4$			
$x < 0$		$x > 0$		$x < -4$		$x > -4$	
x	$f(x)$	x	$f(x)$	x	$f(x)$	x	$f(x)$
-0.1	1.6	0.1	2.4	-4.1	-14.4	-3.9	-13.6
-0.05	1.8	0.05	2.2	-4.05	-14.2	-3.95	-13.8
-0.01	1.96	0.01	2.04	-4.01	-14.04	-3.99	-13.96
-0.005	1.98	0.005	2.02	-4.005	-14.02	-3.995	-13.98
-0.001	1.9996	0.001	2.0004	-4.001	-14.004	-3.999	-13.996
-0.0005	1.9998	0.0005	2.0002	-4.0005	-14.002	-3.9995	-13.998

.

As a matter of fact, it looks as though

$$\lim_{x\to 0} f(x) = f(0) = 2, \qquad \lim_{x\to -4} f(x) = f(-4) = -14$$

so we guess that in general,

$$\lim_{x\to c} f(x) = f(c) = 4c + 2.$$

■

In example 2.7 we guessed that the limit as $x \to c$ could be determined by evaluating $f(c)$, and it turns out that in this example that is correct. However, one must be very careful about just freely plugging in numbers into equations as the function might not always be defined at that point.

Example 2.8

Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = \frac{x^2+x-6}{x-2}$. Determine the limit

$$\lim_{x\to 2} f(x)$$

Solution. Unlike the previous example, attempting the substitute $x = 2$ into $f(x)$ will result in division-by-zero, which we know is never permitted. However, we can evaluate $f(x)$ at any number other than 2 and the hope is that this will tell us what the function looks like at $x = 2$. Indeed, creating the following table we once again find that

x	$f(x)$
2.1	5.1
2.05	5.05
2.01	5.01
2.005	5.005
2.001	5.001
2.0005	5.0005

so it certainly appears as though $f(x)$ is approaching 5. Indeed, if $x \neq 2$ then we may factor $f(x)$ as

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

and the behaviour of this function as $x \to 2$ agrees with our observations. ■

The previous example demonstrates that a function does not need to be defined at a point for the limit at that point to exist. In fact, this is an excellent opportunity to point out that the functions $f(x) = \frac{x^2+x-6}{x-2}$ and $g(x) = x+3$ are similar but *are not equal*: the distinction being that the domain of $f(x)$ is $\mathbb{R} \setminus \{2\}$ while the domain of $g(x)$ is $\mathbb{R}$. If two functions have different domains, they certainly cannot be equal! Of course, $x = 2$ is the only point where the functions do not agree, and their graphs are even identical with the exception that the graph of $f(x)$ will have a hole at $x = 2$. Luckily, this does not matter when we are taking limits, and we have the equality

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

The reason is that while the functions differ at the point $x = 2$, the limit only looks at what the functions do at points close to *but not equal* to 2. Thus the limits see them as the same function (cf Figure 1).

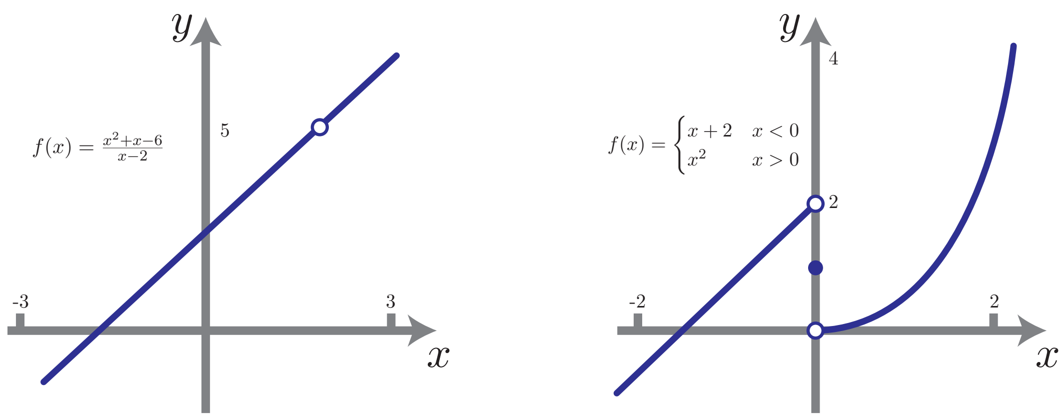


Figure 1: **Left:** The function $\frac{x^2+x-6}{x-2}$ is identical to the function $x+3$ except for the presence of a hole at $x = 2$. This does not affect the limit though, as the limit is only concerned with the behaviour of the function *near* $x = 2$. **Right:** A graph of a piecewise function whose limit at zero is dependent upon the direction of approach. Notice that in either case, the limit disagrees with the value of the function at zero.

Example 2.9

Compute the limit

$$\lim_{x\to 2} \frac{x-2}{\sqrt{x+7}-3}.$$

Solution. This is actually identical to the previous example, but it may not be obvious why. Instead, the usual thing to do in such situations where one summand is a square root is to multiply by the conjugate. In this case, $\sqrt{x+7}+3$. In that case we have

$$\begin{aligned}\lim_{x\to 2}\frac{x-2}{\sqrt{x+7}-3}\frac{\sqrt{x+7}+3}{\sqrt{x+7}+3} &= \lim_{x\to 2}\frac{(x-2)(\sqrt{x+7}+3)}{(x+7)-9}\\ &= \lim_{x\to 2}\frac{x-2}{x-2}\left[\sqrt{x+7}+3\right]\\ &= \lim_{x\to 2}\left[\sqrt{x+7}+3\right] = 6.\end{aligned}$$

The reason this is identical to Example 2.8 is that

$$x-2 = (x+7)-9 = \left[\sqrt{x+7}-3\right]\left[\sqrt{x+7}+3\right]$$

and so the steps we just iterated may be replaced by a much simpler cancellation argument. ■

One Sided Limits: In each of the above examples, our tables were typically biased by choosing to approach our limiting number in one direction. For example, when looking at the limit as $x \to 0$ we approached 0 by considering successively smaller positive numbers, but we could have equivalently looked at successively smaller negative numbers. We say that the limit at 0 exists if and only if the limit of from the positive numbers is equal to the limit given by the negative numbers.

More generally, we will say that "the limit of $f(x)$ as x approaches c from the *right* is L" if whenever $x > c$ gets arbitrarily close to c, $f(x)$ gets arbitrarily close to L. This is written in symbols as

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

Similarly, we say that "the limit of $f(x)$ as x approaches c from the *left* is L" if whenever $x < c$ gets arbitrarily close to c, $f(x)$ gets arbitrarily close to L, and in this case we write

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

We then define the two-sided limit to be L if both limits exist and are equal. There are plenty of examples where the two-sided limit does not exist, as our following examples demonstrate.

Example 2.10

Consider the function

$$f(x) = \begin{cases} x+2 & x<0 \\ 1 & x=0 \\ x^2 & x>0 \end{cases}.$$

Compute the limit of $f(x)$ as $x \to 0^-$ and as $x \to 0^+$. Does the two-sided limit exist?

Solution. We first look at the limit as $x \to 0^-$. In this case, we know that x is always less than 0, and so in this regime $f(x)$ effectively looks like the function $x+2$. As $x \to 0^-$ we see that $x+2 \to 2$ and so we conclude that

$$\lim_{x\to 0^-} f(x) = 2.$$

On the other hand, the limit $x \to 0^+$ guarantees that x is always positive. Here, $f(x)$ looks like the function x^2 and as x approaches 0, x^2 approaches 0 as well, so

$$\lim_{x\to 0^+} f(x) = 0.$$

Each one-sided limit exists, but they are not equal. Hence the two-sided limit does not exist. The graph of $f(x)$ is given in Figure 1. ■

The other way that a two-sided limit can fail to exist is if the one-sided limits do not exist either. This can happen in one of two ways: The first is that it is impossible to find a number L to which the function gets close. This can happen, for example, if our function oscillates infinitely in any small interval around a point. Alternatively, the limits can *diverge*, meaning that the function goes to either positive or negative infinity. This next example demonstrates both phenomena.

Example 2.11

Let $f(x)$ be the function

$$f(x) = \begin{cases} \sin\left(\frac{1}{x}\right) & x < 0 \\ \frac{1}{1-x} & 0 \leq x < 1 \\ \frac{x^2}{x-1} & x > 1 \end{cases}$$

Examine the limits as $x \to 0^{\pm}$ and $x \to 1^{\pm}$.

Solution. Take a look at Figure 2 which plots $f(x)$. In the limit as $x \to 0^-$ we want to examine how the function $\sin({}^1\!/_x)$ behaves when x is a small negative number. With any luck, the student will recognize that the function begins to oscillate infinitely often as x gets small, making it impossible for the function to ever converge to a single point. This tells us that

$$\lim_{x\to 0^-} f(x) \qquad \text{does not exist.}$$

In the case $x \to 0^+$, we are only interested in what happens in a small positive neighbourhood of 0 and hence we can assume that $0 < x < 1$ so that $f(x)$ looks like $\frac{1}{1-x}$. We encounter no problem by substituting small positive numbers into this function, and one can quickly deduce that

$$\lim_{x\to 0^+} f(x) = 1.$$

Now $\frac{1}{1-x}$ is not defined at $x = 1$, so we must be careful in looking at the limit as $x \to 1^-$. As $x < 1$ gets arbitrarily close to 1, the number $1 - x$ is a very small *positive* number. This means that $\frac{1}{1-x}$ is a very large positive number. As we can make $1 - x$ arbitrarily small, $\frac{1}{1-x}$ can be arbitrarily large so

$$\lim_{x\to 1^-} f(x) = \infty.$$

Finally, as $x \to 1^+$ we again run into troubles with the function $\frac{x^2}{x-1}$. But we can see that the number is always positive and tends to the number 1, while the denominator $x - 1$ gets arbitrarily small, but positive since $x > 1$. Making x arbitrarily close to 1 implies that $\frac{x^2}{x-1}$ can get arbitrarily large, so that

$$\lim_{x\to 1^+} f(x) = \infty.$$

We conclude that neither of the two sided limits $x \to 0$ nor $x \to 1$ exist. ■

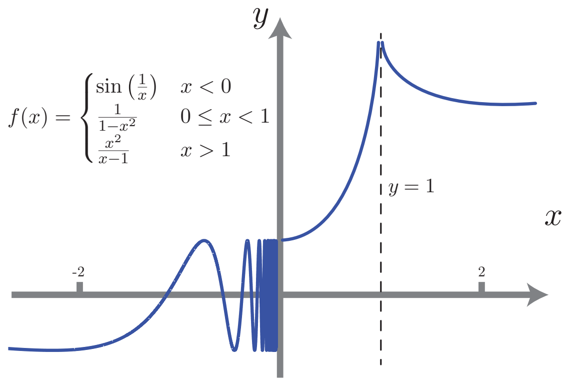


Figure 2: A pathological function for which no two sided limit exists at the points $x = 0$ and $x = 1$. Note that as $x \to 0^-$ the limit fails to converge, while as $x \to 1^\pm$ the function diverges to infinity.

2.3 Of epsilons and deltas

The ϵ-δ definition of the limit is the bane of every calculus student and will not cease haunting you until your time in this course has finished. However, it is so crucially important in developing a rigorous theory of limits that the sooner the student grasps this concept the better.

The unfortunate reality of the situation is that the definition of a limit is awkward and clumsy, though is absolutely necessary for reasons I will make clear shortly. There are multiple ways to understand ϵ-δ concepts and I will try to convey the most important to you.

Definition 2.12

Let f be a function and p be a point. If f is defined in an open neighbourhood[a] of p then we say that "The limit of $f(x)$ as x approaches c is L" precisely if

$$\forall \epsilon > 0, \exists \delta > 0,\ 0 < |x - c| < \delta \quad \Rightarrow \quad |f(x) - L| < \epsilon. \tag{2.3}$$

[a] That is, there exists a $\gamma > 0$ such that f is defined on $(c - \gamma, c) \cup (c, c + \gamma)$

With our study of logical quantifiers in hand, let us see if we can translate this into a plain sentence. Reading verbatim, (2.3) says

"For every positive ϵ, there is a positive δ such that if $|x - c| < \delta$ then $|f(x) - L| < \epsilon$."

Well, this has failed to clear anything up, so we must make a very small detour to clarify things. Recall that aside from the technical definition of the absolute value, the quantity $|a - b|$ is just the *distance* for a to b. The absolute value is necessary to ensure that the distance is always positive since it makes no sense to talk about negative distances.

Hence $|a - b|$ can be read as *the distance of a from b* (or of b from a) and $|x - c| < \delta$ means "the distance from x to c is at most δ, and $|f(x) - L| < \epsilon$ means "the distance from $f(x)$ to L is at most ϵ. Our translation of (2.3) now becomes

"No matter how close I want f to get to L, I can find an x that is sufficiently close to c such that $f(x)$ is closer."

The textbook uses a diagram similar to Figure 3 to explain this situation. The idea is that given L and a prescribed ϵ, we may draw the lines corresponding to $L \pm \epsilon$. Where these lines meet the function $f(x)$, we then draw vertical lines that meet the x-axis. We then take the largest symmetric interval about c and define δ to be the distance from c to these intersecting lines.

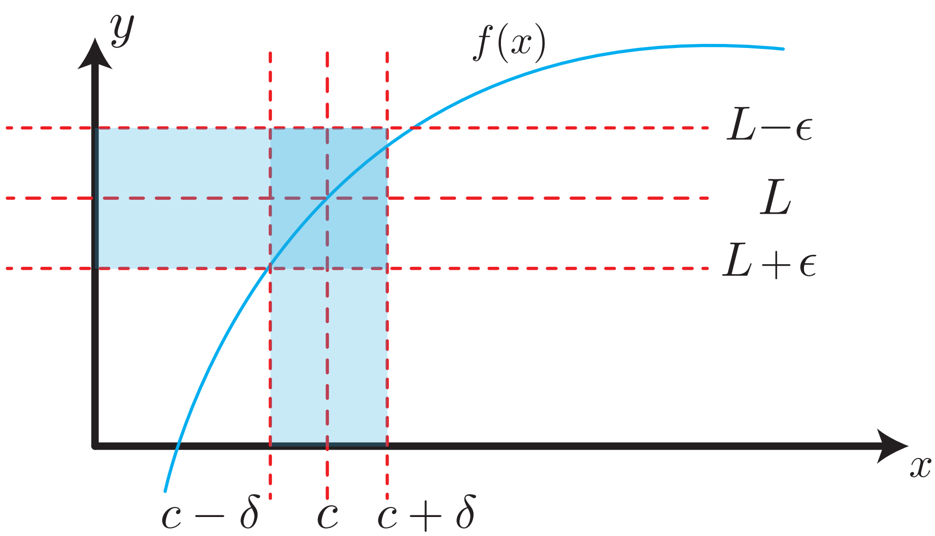


Figure 3: The geometric interpretation of the ϵ and δ "bands".

This describes what is happening geometrically, but hopefully many of you are wondering why we walk around with this lanky definition in our bags. Certainly, this seems like those crazy mathematicians just want to make life really complicated for the average student, but let me assure you there is a good reason. In order to elaborate, allow me to tell you the story of two young mathematicians named Newton and Leibniz.

A long time ago there were two young mathematicians, one named Newton and one named Leibniz. They lived in a small city, and attended the local university. One day, a geology professor approached our young protagonists and asked them to do him a favor. There had recently been an earthquake and the professor was interested in a crevasse that had formed on a hillside, two kilometres from the base. In particular, the professor was interested in the altitude at which the tear in the hill had occurred. A young and ambitious Newton volunteered immediately, and set forth climbing the hill. To determine the altitude, Newton took his altimeter and stepped into the crevasse. He immediately fell to his death, and so ended the story of young Newton (see Figure 4).

On the other hand, the young Leibniz witnessed Newton's untimely death and devised an alternative method. With his altimeter in hand, he approached the crevasse, but did not try to stand on it. Instead, he told the professor that he could determine the altitude of the crevasse by getting arbitrarily close to it. In fact, if the professor specified what level of accuracy he would like in the measurement, Leibniz would be able to approach sufficiently close in order to make the

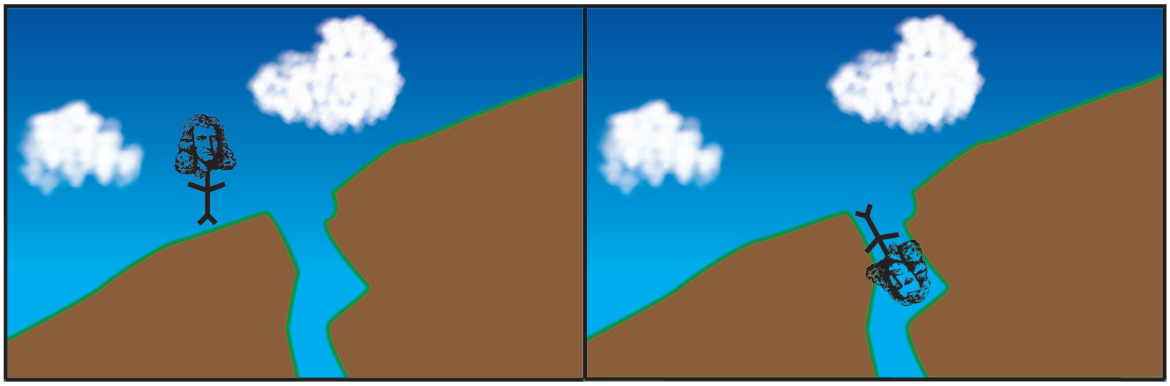


Figure 4: The tragic death of young Newton.

desired measurement. Thus Leibniz was able to determine the height of the crevasse without falling to his death.

If we look at Figure 3, we question why we need consider all these complicated epsilons and deltas. The reason is that the function in question is what we will later learn is called *continuous*, in which case it is very simple to determine limits. The problem arises when our function is not continuous, which intuitively corresponds to the function containing holes, instantaneous jumps, or diverging to infinity. An attempt to evaluate a limit directly at such points will result in less dire consequences than our dear friend Newton, but not by much. Hence we endeavour to always use the technique of young Leibniz when calculating limits.

One may have all the geometric intuition in the world, but this is as useful as a waterproof teabag if we cannot figure out how to actually perform ϵ-δ proofs. Allow me to show you a **very** simple proof, and then we may go through the steps of how to write such things.

Example 2.13

Show that

$$\lim_{x\to 3} 2x + 1 = 7. \tag{2.4}$$

Solution. Let $\epsilon > 0$ be given and choose $\delta = \frac{\epsilon}{2}$. We would like to show that if $|x-3| < \delta$ then $|(2x+1)-7| < \epsilon$. Indeed this is the case since

$$|(2x+1)-7| = |2x-6| = 2|x-3| < 2\frac{\epsilon}{2} = \epsilon \tag{2.5}$$

as required. ■

The most important question you should all have is "how did he know to choose $\delta = \epsilon/2$?" I know because I did a bunch of rough work ahead of time, then wrote up the proper solution already knowing that this choice of δ would work. So now let me explain how you should do your "rough work."

It may help to think of this as some sort of curious game which, speaking from experience, is a great way to break the ice at parties. In particular, let us assume that some dodgy mysterious fellow approaches you on the street, and gives you an $\epsilon > 0$. He tells you that you can do whatever you want with the quantity $|x - c|$, and that you win the game if you can find a $\delta > 0$ such that $|f(x) - L| < \epsilon$ whenever $|x - c| < \delta$. What should our strategy be? Our only tool is that we can make $|x - c|$ do whatever we want, so we should try to manipulate $|f(x) - L|$ to look $K|x - c|$ for some constant K. If we can do this, then no matter what ϵ we are given, we can automatically just choose $\delta = \epsilon/K$ and win the game. Let me illustrate this with a simple but more general example than that given above:

Example 2.14

Show that for any $a, b, c \in \mathbb{R}$ with $a \neq 0$ then

$$\lim_{x \to c} ax + b = ac + b. \tag{2.6}$$

Solution. We begin by doing the rough work. Let $\epsilon > 0$ be given. Per our strategy discussed above, we should try to make $|(ax + b) - (ac + b)|$ look like $K|x - c|$ for some constant K. Indeed,

$$|(ax + b) - (ac + b)| = |ax + b - ac - b| = |ax - ac| = |a||x - c|. \tag{2.7}$$

Now that was not that hard was it? For equation (2.7) to be less than ϵ, we want $|a||x - c| < \epsilon$, or rather $|x - c| < \epsilon/a$. Since we can let $|x - c|$ be as small as we want, we choose $\delta = \epsilon/|a|$.

Having done the rough-work, we may now give the proper proof, which we would write as follows:

Let $\epsilon > 0$ be given and choose $\delta = \epsilon/|a|$. If $|x - c| < \delta$ then we have

$$\begin{aligned} |f(x) - L| &= |(ax + b) - (ax - c)| = |ax + b - ax - b| \\ &= |ax - ac| = |a||x - c| \\ &< |a| \frac{\epsilon}{|a|} = \epsilon \end{aligned}$$

which is precisely what we wanted to show. ■

Let's finish with some slightly harder examples:

Example 2.15

Show that

$$\lim_{x \to 5} x^2 - 9 = 16. \tag{2.8}$$

Solution. As always, let $\epsilon > 0$ be given and consider $|f(x) - L|$ which we can manipulate as

$$|(x^2 - 9) - 16| = |x^2 - 25| = |(x - 5)(x + 5)| = |x - 5||x + 5|. \tag{2.9}$$

We recall that our goal is to get this to look like something of the form $K|x - 5|$, but we have this pesky $|x + 5|$ term hanging around. How do we deal with this? Since we only want $|f(x) - L|$ to be *less than* epsilon, we are allowed to make some approximations.

Let us assume for the moment that $|x-5|<1$ (which roughly corresponds to choosing $\delta = 1$), though we can choose any number that we like. Under this assumption we would like to determine an upper bound for $|x+5|$. But notice that

$$|x+5| = |(x-5)+10| \leq |x-5| + 10 < 11. \tag{2.10}$$

Thus if $\delta = 1$ we have

$$|f(x) - L| \leq 11|x-5| \tag{2.11}$$

and we should choose $\delta = \epsilon/11$. We have almost completed our rough work, but notice that in order to get this bound, we needed to assume that $|x-5|<1$. In order to take care of this, we simply set $\delta = \min(1, \epsilon/11)$, and then our proof will work. Indeed, the proof is as follows:

Let $\epsilon > 0$ be given and set $\delta = \min(1,\epsilon/11)$. Notice that

$$|(x^2-9) - 16| = |x^2 - 25| = |x+5||x-5|. \tag{2.12}$$

Since we chose $\delta = \min(1, \epsilon/11)$ it is certainly the case that $|x-5| < \delta \leq 1$. Hence (2.10) holds and so $|x+5| < 11$. Thus

$$|(x^2-9)-16| = |x+5||x-5| \leq 11|x-5| < 11\frac{\epsilon}{11} = \epsilon \tag{2.13}$$

as required. ■

As I noted above, in approximating $|x+5|$ we could have assumed that $|x-5|$ was less than **any** positive number. For example, if we had used 2 then we would have found that $|x+5| < 12$. The only change we then need to make in the proof is to set $\delta = \min(2, \epsilon/12)$.

Exercise: If $a,b,c,d \in \mathbb{R}$ with $a \neq 0$ show that

$$\lim_{x\to d} ax^2 + bx + c = ad^2 + bd + c. \tag{2.14}$$

In particular, if we make the approximation that $|x-d| < n$ for some $n > 0$, show that choosing

$$\delta = \min\left(n, \frac{\epsilon}{|a|n + 2|ad| + |b|}\right)$$

is sufficient for the proof.

In the following examples, I have only done the rough work required for each question. I have left the "proof" part as a simple exercise for the student.

This being said, there are occasions in which it can be a little tricky to start from $|f(x)-L|$ and get $|x-c|$. In fact, sometimes it is actually easier to go the other direction!

Example 2.16

Show that

$$\lim_{x\to 9} \sqrt{x} = 3. \tag{2.15}$$

Solution. Let $\epsilon > 0$ be given. In this case, we realize that $|x-9| = |\sqrt{x}-3||\sqrt{x}+3|$ so that we can write

$$|f(x) - L| = |\sqrt{x} - 3| = \frac{|x-9|}{|\sqrt{x}+3|}. \tag{2.16}$$

Following the procedure outlined above, our goal should then be to find a number K such that $|\sqrt{x}+3| \geq K$. We must use the greater-than sign here, since when we take the reciprocal the greater-than sign will switch to the less-than sign that we desire. There are two ways to do this: the easy way and the hard but more general way.

The first way is to realize that $\sqrt{x}+3 > 3$ for all x and so $|\sqrt{x}+3| = \sqrt{x}+3 > 3$. In this case, we now have

$$|\sqrt{x} - 3| = \frac{|x-9|}{|\sqrt{x}+3|} \leq \frac{1}{3}|x-9| \tag{2.17}$$

so choosing $\delta = 3\epsilon$ will do the trick.

On the other hand, we may also proceed with the same trick as before and hypothesize that $|x-9| < 1$. In this case, we then have

$$\begin{gathered} -1 < x - 9 < 1 \\ 8 < x < 10 \\ \sqrt{8} < \sqrt{x} < \sqrt{10} \\ \sqrt{8} + 3 < \sqrt{x} + 3 < \sqrt{10} + 3. \end{gathered}$$

Now when we take absolute values, we can only guarantee the the function will be greater than whichever of these is smaller. Thus $|\sqrt{x}+3| \geq \min\left\{\sqrt{8}+3, \sqrt{10}+3\right\} = \sqrt{8}+3$ and we get

$$|\sqrt{x} - 3| = \frac{|x-9|}{|\sqrt{x}+3|} \leq \frac{1}{\sqrt{8}+3}|x-9| \tag{2.18}$$

so choosing $\delta = \min\left\{1, (\sqrt{8}+3)\epsilon\right\}$ will do the trick. ∎

Finally, we can use this same idea to treat rational functions.

Example 2.17

Show that

$$\lim_{x \to 5} \frac{x+5}{x-7} = -5. \tag{2.19}$$

Solution. Let $\epsilon > 0$ be given. Starting with $|f(x) - L|$ we have

$$\begin{aligned} \left|\frac{x+5}{x-7} + 5\right| &= \left|\frac{x+5+5x-35}{x-7}\right| \\ &= \left|\frac{6(x-5)}{x-7}\right| \\ &= 6\left|\frac{x-5}{x-7}\right|. \end{aligned}$$

We want to try to bound $|x-7|$ from below so that its reciprocal is bounded from above. Assume that $|x-5|<1$ and notice that

$$\begin{gathered} -1 < x-5 < 1 \\ -1-2 < x-5-2 < 1-2 \\ -3 < x-7 < -1 \end{gathered}$$

and so $|x-7| \geq \min\{|-3|, |-1|\} = 1$. We thus have

$$\left| \frac{x+5}{x-7} + 5 \right| \leq 6\,|x-5| \tag{2.20}$$

so taking $\delta = \min\{1, \epsilon/6\}$ will work. ■

2.3.1 Negation of Limits

In the previous section we determined how to use the ϵ-δ definition of a limit to show that a limit exists. In this section, we demonstrate the (much harder) task of showing that a limit does not exist. Having spent a great deal of time learning how to negate logical sentences, we can just apply our know-how to directly negate the limit:

$$\lim_{x\to c} f(x) \neq L \text{precisely if } \exists \epsilon > 0, \forall \delta > 0, \exists x, 0 < |x-c| < \delta \text{ and } |f(x) - L| \geq \epsilon.$$

In particular, we can translate this into English and read it as

> There is some positive distance such that, no matter what interval we take around c, there is always a point in that interval which makes $f(x)$ further away from L than the prescribed distance.

Example 2.18

Show that

$$\lim_{x\to 0} \frac{x}{|x|} \neq 0.$$

Solution. A quick look at the graph of $x/|x|$ indicates that indeed, the limit should not be zero, but we can now rigorously show this using our negation above. We have to find an $\epsilon > 0$ which works. If $\epsilon > 1$ then our function will always live within the given ϵ-band, so we should choose some $0 < \epsilon < 1$. Let's take $\epsilon = 1/2$.

Let $\delta > 0$ be given and arbitrary, so that we are looking at the interval $(-\delta, 0) \cup (0, \delta)$. We have to choose a point in here to make $|f(x) - 0| = |f(x)| > \epsilon$. Let's try taking $x_0 = \delta/2$. Notice that $|x_0 - 0| = |x_0| = \delta/2 < \delta$ so this x_0 lies within the desired interval. Furthermore, $|f(x)| = |1| = 1$, and so

$$|f(x)| = 1 > \frac{1}{2} = \epsilon,$$

as required. ■

This shows that the limit of a function is not a single limit point, but it doesn't say that it can't be some other point. To show that a limit does not exist at all, we have to first realize that implicit in the ϵ-δ definition of a limit is another existential quantifier; namely, there exists a limit L. Hence to show that a limit does not exist for all L, our negation should be

$$\forall L \in \mathbb{R}, \exists \epsilon > 0, \forall \delta > 0, \quad 0 < |x - c| < \delta \text{ and } |f(x) - L| \geq \epsilon.$$

Take careful note of how L, ϵ, and δ are permitted to depend on one another.

1. The ϵ is allowed to change depending upon the choice of L.
2. The ϵ must work for all choices of δ.

Example 2.19

Show that $\lim_{x \to 0} \frac{x}{|x|}$ does not exist (for any L).

Solution. If $|L| \neq 1$ then the above proof can be adapted to work, so we break our proof into two cases.

Case 1 ($|L| \neq 1$) Let's assume that $L > 0$ since precisely the same argument holds if $L < 0$. What's the idea here? Try choosing a random limit $L \neq 1$. How do we choose ϵ to ensure that the function will not lie in the given ϵ-band? We make ϵ smaller than the distance from L to 1. Hence set $d = |L - 1|$ and take $\epsilon = \frac{d}{2}$. Let $\delta > 0$ be given, and choose x to be any positive number $0 < x < \delta$, so $x_0 = \delta/2$. In this case we have

$$|f(x) - L| = |1 - L| > \frac{d}{2} = \epsilon.$$

Case 2 ($|L| = 1$) Once again, take $L = 1$ and use a similar argument for $L = -1$. If $L = 1$ then our previous arugment will not work (try it!). The trick is now to take x to be negative. Indeed, set $\epsilon = 1$ and take $x_0 = -\delta/2$, so that $f(x_0) = -1$ and

$$|f(x) - L| = |(-1) - 1| = |-2| = 2 > 1 = \epsilon. \qquad \blacksquare$$

Rather than do everything with the ϵ-δ definition, we can use the powerful but unsophisticated tools to build an infrastructure for problem solving. Of course, such infrastructures are created through bootstrapping, meaning that we are going to be using the ϵ-δ definition a lot in this section.

Our first theorem tells us that once we have found a limit, we can be confident that it is the only one:

Theorem 2.20

If f is a function defined in a neighbourhood of c and $\lim_{x \to c} f(x)$ exists, then the limit is unique.

Proof. We will proceed by contradiction. What would go wrong if we had two different limits $L_1 \neq L_2$? If we draw a picture and think about ϵ-bands, we see that it should be impossible for two different limits to be contained in the same band. In particular, if we let ϵ be less than the distance between L_1 and L_2, something should go terribly wrong.

Let $d = |L_1 - L_2|$ which we know is not zero by assumption. Since we assume that

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

then by setting $\epsilon = d/4$ we know there exists $\delta_1, \delta_2 > 0$ such that

$$0 < |x - c| < \delta_1 \Rightarrow |f(x) - L_1| < \frac{d}{4}$$
$$0 < |x - c| < \delta_2 \Rightarrow |f(x) - L_2| < \frac{d}{4}.$$

That is, f gets arbitrarily close to *both* L_1 and L_2. Let $\delta = \min\{\delta_1, \delta_2\}$ so that if $0 < |x - c| < \delta$ then

$$\begin{aligned} d = |L_1 - L_2| &= |L_1 - f(x) + f(x) - L_2| \\ &\leq |f(x) - L_1| + |f(x) - L_2| \\ &= \frac{d}{4} + \frac{d}{4} = \frac{d}{2}. \end{aligned}$$

Of course, since $d > 0$, this is impossible and we have arrived at a contradiction. Hence we have shown that $L_1 = L_2$ and we conclude that limits are unique. □

We will not put this theorem to much use in this course, but it is a load off our shoulders knowing that nothing terribly weird can happen with limits.

On to our next subject: Mathematicians love to be lazy, in the sense that if we have already performed a calculation, why should we repeat it ever again? Similarly, we like to build complicated examples from simple examples. To this end, we formulate the following collection of limit laws, which are intended to dramatically simplify our life:

Theorem 2.21: Limit Laws

Let f and g be two functions defined in a neighbourhood of $c \in \mathbb{R}$. Furthermore, if

$$\lim_{x \to c} f(x) = L, \qquad \lim_{x \to c} g(x) = M$$

both exist, then

1. $\lim_{x \to c} [\alpha f(x)] = \alpha \lim_{x \to c} f(x) = \alpha L,$
2. $\lim_{x \to c} [f(x) \pm g(x)] = \left[\lim_{x \to c} f(x)\right] \pm \left[\lim_{x \to c} g(x)\right] = L \pm M,$
3. $\lim_{x \to c} [f(x)g(x)] = \left[\lim_{x \to c} f(x)\right]\left[\lim_{x \to c} g(x)\right] = LM,$
4. $\lim_{x \to c} \left[\frac{f(x)}{g(x)}\right] = \frac{\lim_{x \to c} f(x)}{\lim_{x \to c} g(x)} = \frac{L}{M}$ if $M \neq 0$.

It is crucial that both limits must exist. It is a common mistake for students to gleefully attempt to apply the above limits laws in instances in which it is not permitted. For example, the following **IS NOT CORRECT**:

$$\lim_{x\to 0} x^2 \sin\left(\frac{1}{x}\right) = \left[\lim_{x\to 0} x^2\right]\left[\lim_{x\to 0}\sin\left(\frac{1}{x}\right)\right] = 0 \times \left[\lim_{x\to 0}\sin\left(\frac{1}{x}\right)\right] = 0.$$

Interestingly, this is the correct answer, but for all the wrong reasons. Hence this solution is absolutely incorrect.

Proof. We will only do the proof of number 2, though we will do it in great detail. The remaining proofs are left as exercises for the student.

We are given a new function which is the sum $(f+g)$ of our given functions, and we would like to show that their limit is the corresponding sum $L+M$. This means that we have to show that, given any $\epsilon > 0$, there exist some $\delta > 0$ such that $0 < |x-c| < \delta$ implies that

$$\Big|[f(x)+g(x)] - [L+M]\Big| < \epsilon. \tag{2.21}$$

Let's take a moment to think about what hypotheses we have been given. First, we are told that both limits exist; hence for any $\epsilon > 0$ we know there exist $\delta_f, \delta_g > 0$ such that by taking $|x-c|$ smaller than δ_f (respectively, δ_g) then we can make $|f(x)-L|$ (respectively $|g(x)-M|$) smaller than ϵ. Since this is the only information we have been given, we should then try to reduce (2.21) to an expression which involves $|f(x)-L|$ and $|g(x)-M|$. By staring at (2.21) we realize we can re-arrange it as follows:

$$\begin{aligned}\Big|[f(x)+g(x)] - [L+M]\Big| &= \Big|[f(x)-L] + [g(x)-M]\Big| \\ &\le |f(x)-L| + |g(x)-M| \qquad \text{triangle inequality.}\end{aligned}$$

This is perfect! We can control how small $|f(x)-L|$ and $|g(x)-M|$ become, so in particular, let's make them less than $\epsilon/2$. Let $\delta_f, \delta_g > 0$ be given such that

$$\begin{aligned} 0 < |x-c| < \delta_f \quad &\Rightarrow \quad |f(x)-L| < \frac{\epsilon}{2} \\ 0 < |x-c| < \delta_g \quad &\Rightarrow \quad |g(x)-M| < \frac{\epsilon}{2}. \end{aligned} \tag{2.22}$$

The problem now is that we have two deltas, and the definition of the limit requires that we give a single delta. The way to solve this is to recall that if we have found a delta which works, any smaller delta will also work. Hence by setting $\delta = \min\{\delta_f, \delta_g\}$ we can guarantee that $\delta \le \delta_f$ and $\delta \le \delta_g$ so that both equations in (2.22) hold. Hence if $0 < |x-c| < \delta$ then

$$\Big|[f(x)+g(x)] - [L+M]\Big| \le |f(x)-L| + |g(x)-M| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

This is what we wanted to show, and so the proof is complete. □

Exercise: Complete the proofs for Theorem 2.21.

We can immediately use this to prove some very powerful and useful results.

Corollary 2.22

If $f(x) = \frac{p(x)}{q(x)}$ is any rational functions (so that $p(x)$ and $q(x)$ are polynomials), and $c \in \mathbb{R}$ is such that $q(c) \neq 0$ then

$$\lim_{x\to c} \frac{p(x)}{q(x)} = \frac{p(c)}{q(c)}.$$

Proof. The key is to first show this for polynomials and apply the limit laws. It is easy to see (the student should check!) that

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

and so by induction, (the student can show that)

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

Let $p(x) = a_n x^n + a_{n-1}x^{n_1} + \cdots + c_1 x + c_0$ be an arbitrary polynomial. Our discussion tells us that we know how to deal with every occurrence of the function $x \mapsto x$, and so using the Theorem 2.21 we have

$$\begin{aligned}
\lim_{x\to c} p(x) &= \lim_{x\to c} \left[a_n x^n + \cdots + a_1 x + a_0\right] \\
&= a_n \left[\lim_{x\to c} x\right]^n + a_{n-1}\left[\lim_{x\to c} x\right]^{n-1} + \cdots + a_1 \left[\lim_{x\to c} x\right] + a_0 \\
&= a_n c^n + a_{n-1}c^{n-1} + \cdots + a_1 c + a_0 \\
&= p(c).
\end{aligned}$$

Thus the result holds for any polynomial. Now if $p(x)$ and $q(x)$ are two polynomials and $q(c) \neq 0$, then the limit laws for quotients implies

$$\lim_{x\to c} \frac{p(x)}{q(x)} = \frac{\lim_{x\to c} p(x)}{\lim_{x\to c} q(x)} = \frac{p(c)}{q(c)},$$

as required. □

2.4 Limits at Infinity

There are two notions of infinity that we will be interested in tackling, both with definitions similar to the ϵ-δ definition we saw before.

Definition 2.23

- We say that $\lim_{x\to\infty} f(x) = L$ if for every $\epsilon > 0$ there exists an $M \in \mathbb{R}$ such that whenever $x > M$ then $|f(x) - L| < \epsilon$.
- We say that $\lim_{x\to c} f(x) = \infty$ if for every $M \in \mathbb{R}$ there exists a $\delta > 0$ such that if $0 < |x - c| < \delta$ then $f(x) > M$.

In both cases, to consider $-\infty$ we replace $x > M, f(x) > M$ with $x < M, f(x) < M$.

Exercise: Determine an appropriate statement for $\lim_{x\to\infty} f(x) = \infty$.

Example 2.24

Show that $\lim_{x\to\infty} \frac{x}{x+1} = 1$.

Solution. Let $\epsilon > 0$ be given and take $M = \max\{-1, \epsilon^{-1} - 1\}$. If $x > M$ then since $x \geq M \geq -1$ we know that $|x+1| = x+1$. Furthermore, since $x > \epsilon^{-1} - 1$ we know that $x + 1 > \frac{1}{\epsilon}$, thus

$$\left|\frac{x}{x+1} - 1\right| = \left|\frac{-1}{x+1}\right| = \frac{1}{|x+1|} < \epsilon$$

as required. ■

Example 2.25

Show that $\lim_{x\to 0} \frac{1}{x^2} = \infty$.

Solution. Let $M \in \mathbb{R}$ be given and let $\delta = \frac{1}{\sqrt{|M|}}$. If $|x| < \delta = \frac{1}{\sqrt{|M|}}$ then $0 < x^2 < \frac{1}{|M|}$ and so

$$f(x) = \frac{1}{x^2} > |M| > M.$$ ■

Just as it was possible to use the ϵ-δ definition to show that finite limits do not exist, we can use our rigorous definitions to show that limits at infinity do not exist. Indeed, we say that

$$\lim_{x\to\infty} f(x) \text{ does not exist}$$

if for every $L \in \mathbb{R}$ there exists an $\epsilon > 0$ such that for every $M \in \mathbb{R}$ there exists an x such that $x > M$ and $|f(x) - L| \geq \epsilon$.

Example 2.26

Show that $\lim_{x\to\infty} \sin(x)$ does not exist.

Solution. We will consider the case where $L > 0$, noting that the $L < 0$ case follows precisely the same argument. Assume that $L > 1$ and let $\epsilon = \frac{L-1}{2}$. Let x be any number such that $x > M$ and notice that

$$\begin{aligned} |\sin(x) - L| &= L - 1 && \text{since } \sin(x) \leq 1 < L \\ &\geq \frac{L-1}{2} = \epsilon. \end{aligned}$$

Thus assume that $0 < L < 1$ and set $\epsilon = \frac{L}{2}$. Let $M \in \mathbb{R}$ be given, and let $n \in \mathbb{N}$ satisfy $n > \frac{M}{\pi}$. Set $x = n\pi$ so that $x > M$ and

$$|\sin(x) - L| = |\sin(n\pi) - L| = L > \frac{L}{2} = \epsilon.$$

We conclude that the limit does not exist. ■

2.5 Continuity

In many of the cases that we have hitherto examined our functions have been nicely behaved, in the sense that the limits were often 'obvious' (if not so obvious to prove). We give such functions a very nice name:

Definition 2.27

We say that a function f, defined in a neighbourhood of c, is *continuous at the point* $c \in \mathbb{R}$ if

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

If the function $f(x)$ is continuous at all points in $\mathbb{R}$, we say that it is *continuous*.

Of course, we can write this in terms of ϵ-δ: *$f(x)$ is continuous at c if*

$$\forall \epsilon > 0, \exists \delta > 0, |x - c| < \delta \ \Rightarrow \ |f(x) - f(c)| < \epsilon.$$

The important point to take away from either definition is that the limit of the function may be evaluated simply by substituting the value of the function at that point.

Example 2.28

If $\exists C > 0$ such that $\forall x, y \in \mathbb{R}$

$$|f(x) - f(y)| \leq C|x - y| \tag{2.23}$$

show that f is a continuous function.

Solution. Here it is quite natural to use the ϵ-δ definition of continuity. Indeed, let $\epsilon > 0$ be given and set $\delta = \epsilon/C$. If $p \in \mathbb{R}$ is arbitrary but fixed and $|x - p| < \delta$, then

$$|f(x) - f(p)| \leq C|x - p| < C\frac{\epsilon}{C} = \epsilon \tag{2.24}$$

so f is continuous at p. Since p was arbitrary, this must hold for every point, and hence f is everywhere continuous. ■

We have already seen an entirely family of continuous functions; namely, recall that from Corollary 2.22 we know that if $p(x)$ is any polynomial then

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

showing that all polynomial are continuous. Similarly, if $p(x)$ and $q(x)$ are polynomials and c is in the domain of $p(x)/q(x)$ (which implies that $q(c) \neq 0$) then

$$\lim_{x\to c} \frac{p(x)}{q(x)} = \frac{p(c)}{q(c)}$$

showing that rational functions are continuous everywhere on their domain.

Proposition 2.29

The following functions are continuous on the given domains:

1. $\sqrt[n]{x}$ on $(0,\infty)$ if n is an even integer and $\mathbb{R}$ if n is an odd integer,
2. $|x|$ on all of $\mathbb{R}$

Solution. The majority of these proofs often reduce to doing as we did in the previous section, but instead of fixing a limit point we allow that limit point to vary in general.

1. We shall do the example for $\sqrt{x}$ and leave the other roots as an exercise (albeit a tricky one). We want to show that

$$\lim_{x\to c} \sqrt{x} = \sqrt{c}, \qquad \forall c \in (0,\infty). \tag{2.25}$$

   Indeed, let $\epsilon > 0$ be given and take $\delta = \sqrt{c}\epsilon$. We know that $\delta > 0$ since $\sqrt{c} \neq 0$. Now

$$\begin{aligned} |f(x) - f(c)| = |\sqrt{x} - \sqrt{c}| &= \frac{|x-c|}{|\sqrt{x}+\sqrt{c}|} \\ &\leq \frac{|x-c|}{\sqrt{c}} && \text{since } |\sqrt{x}+\sqrt{c}| \geq \sqrt{c} \\ &< \frac{\sqrt{c}}{\sqrt{c}}\epsilon = \epsilon. \end{aligned}$$

   Hence 2.25 holds as required.

2. Let $\epsilon > 0$ be given. If $c = 0$ it is easy to check that $\delta = \epsilon$ will do the trick. Thus let $c \neq 0$ and without loss of generality, assume that $c > 0$: a similar argument holds if $c < 0$. Let $\delta = \min\{c, \epsilon\}$. Hence if $|x - c| < \delta \leq c$ then $0 < x < 2c$ so in particular, $x > 0$. Thus

$$|f(x) - f(c)| = \Big||x| - |c|\Big| = |x - c| < \delta \leq \epsilon. \qquad \blacksquare$$

In fact, we can use the Limit Laws to say very powerful things about continuous functions:

Theorem 2.30

If f and g be functions which are continuous at c, then

1. For every $\alpha \in \mathbb{R}$, αf is continuous at c,
2. $f \pm g$ is continuous at c,
3. fg is continuous at c,
4. If $g(c) \neq 0$ then f/g is continuous at c.

Proof. As mentioned above, these all follow immediately from the limit laws. Indeed, let's show (iii) in more detail. Since f and g are both continuous at c, we know that the following limits exist

$$\lim_{x\to c} f(x) = f(c), \qquad \lim_{x\to c} g(x) = g(c).$$

Since the limits exists, we know that the limit of the product is the product of the limits, and hence

$$\lim_{x\to c}[fg](x) = \lim_{x\to c}[f(x)g(x)] = \left[\lim_{x\to c} f(x)\right]\left[\lim_{x\to c} g(x)\right] = f(c)g(c) = [fg](c),$$

showing that $[fg](x)$ is continuous at c, as required. □

Another very useful way of thinking about continuity is that we can "take limits inside the function." Indeed, since $\lim_{x\to a} x = a$ (this is clear I hope), then we say that a function is continuous if

$$\lim_{x\to a} f(x) = f\left(\lim_{x\to a} x\right). \tag{2.26}$$

This is a little contrived at the moment, but really turns out to be (an alternative to) the defining characteristic of continuous functions:

Theorem 2.31

If f, g are functions such that $\lim_{x\to c} g(x) = L$ exists and f is continuous, then

$$\lim_{x\to c} f(g(x)) = f\left(\lim_{x\to c} g(x)\right) = f(L). \tag{2.27}$$

Proof. Let $\epsilon > 0$ be given. Since we know $f(x)$ is continuous, there exists some $\eta > 0$ such that if $|y - L| < \eta$ then $|f(y) - f(L)| < \epsilon$. Furthermore, since $g(x) \xrightarrow{x\to c} L$ we know there exist some $\delta > 0$ such that if $0 < |x - c| < \delta$ then $|g(x) - L| < \eta$. By setting $y = g(x)$, we thus have that

$$0 < |x - c| < \delta \quad \Rightarrow \quad |g(x) - L| = |y - L| < \eta \quad \Rightarrow \quad |f(g(x)) - f(L)| < \epsilon$$

and hence $\lim_{x\to c} f(g(x)) = f(L)$ as required. □

As an immediate corollary, we get the following exceptionally useful corollary:

Corollary 2.32

If f, g are functions such that g is continuous at c and f is continuous at $g(c)$, then $f \circ g$ is continuous at c.

Proof. By using Theorem 2.31, there is almost nothing to show. Indeed,

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

as required. □

Definition 2.33

If f is a function defined in a neighbourhood of c, denote by $L_\pm$ the one-sided limits

$$L_+ = \lim_{x\to c^+} f(x), \qquad L_- = \lim_{x\to c} f(x).$$

If $f(x)$ fails to be continuous at c, we say that c is

1. A *removable discontinuity* if both L_+ and L_- exist and $L_+ = L_-$.
2. A *jump discontinuity* if L_+, L_- exist but $L_+ \neq L_-$.
3. An *essential/other discontinuity* if one of $L_\pm$ does not exist or is infinite.

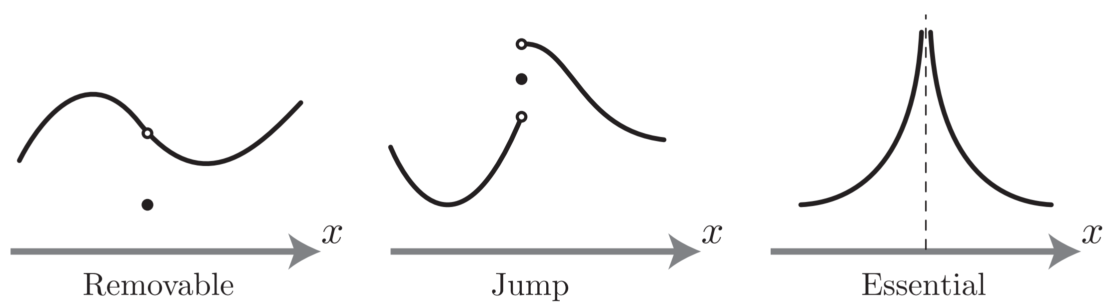


Figure 5: Examples of the types of discontinuity that can occur.

Exercise: Let $\mathbb{Q}$ denote the rational numbers; that is,

$$\mathbb{Q} = \left\{\frac{p}{q} : p, q \in \mathbb{Z}, q \neq 0\right\}.$$

If we define the function

$$f(x) = \begin{cases} 0 & x \in \mathbb{R} \setminus \mathbb{Q} \\ \frac{1}{q} & x \in \mathbb{Q}, x = \frac{p}{q}, \gcd(p,q) = 1 \end{cases} \tag{2.28}$$

then at what points is f continuous? Show that this is the case.

2.6 The Squeeze Theorem

Some limits can be rather tricky to determine, simply because they behave in curious ways around the limit point. In this section we will introduce the Squeeze Theorem, sometimes also referred to as the Pinching Theorem, Sandwich Theorem, or sometimes the Police Theorem. The Squeeze Theorem (as I will likely call it, from here on out) allows us to determine troublesome limits by bounding one function in terms of two other functions which converge to the same point.

Theorem 2.34: The Squeeze Theorem

If $f(x), g(x)$, and $h(x)$ are all defined in an interval around the point c, satisfying $f(x) \leq g(x) \leq h(x)$ on that interval, and

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

then the limit of $g(x)$ as $x \to c$ also exists and is equal to L; that is,

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

Proof. Let $\epsilon > 0$ be given. Since the limits of $f(x)$ and $h(x)$ both exist at the point c, we know there exist $\delta_f, \delta_h > 0$ such that

$$\begin{aligned} 0 < |x-c| < \delta_f \ &\Rightarrow\ |f(x) - L| < \epsilon \\ 0 < |x-c| < \delta_h \ &\Rightarrow\ |h(x) - L| < \epsilon \end{aligned} \tag{2.29}$$

Set $\delta = \min\{\delta_f, \delta_h\}$, so that if $0 < |x-c| < \delta$ then our conditions in (2.29) become

$$\begin{aligned} L - \epsilon <&f(x) < L + \epsilon \\ L - \epsilon <&h(x) < L + \epsilon. \end{aligned} \tag{2.30}$$

Combining this information together with the fact that $f(x)$ and $h(x)$ bound $g(x)$, we get

$$L - \epsilon < f(x) \leq g(x) \leq h(x) < L + \epsilon$$

showing that $|g(x) - L| < \epsilon$ as required. □

As is demonstrated by Figure 6, the idea is that the functions $f(x)$ and $h(x)$ squeeze $g(x)$ into having the same limit.

Example 2.35

Show that

$$\lim_{x\to 0} x^2 \sin\left(\frac{1}{x}\right) = 0.$$

Solution. Regardless of its argument, the functions $\sin(x)$ is always bounded above by $y = 1$ and below by -1; that is,

$$-1 \leq \sin\left(\frac{1}{x}\right) \leq 1.$$

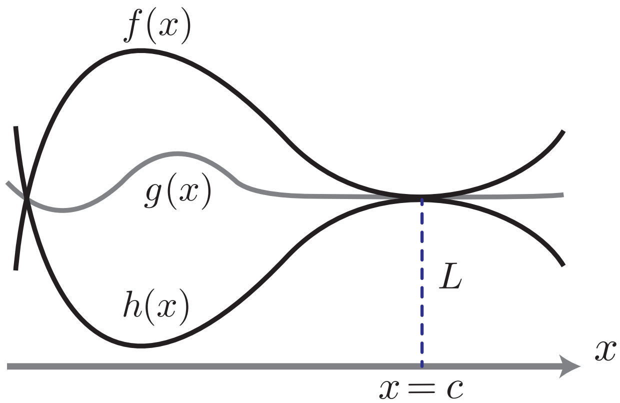


Figure 6: A visualization of a typical Squeeze Theorem example.

As $x^2 > 0$ everywhere, we can multiply through to see that

$$-x^2 \leq x^2 \sin\left(\frac{1}{x}\right) \leq x^2.$$

We are interested in the behaviour as $x \to 0$, and we notice that

$$\lim_{x\to 0} x^2 = \lim_{x\to 0} -x^2 = 0.$$

The Squeeze Theorem thus applies and we get

$$\lim_{x\to 0} x^2 \sin\left(\frac{1}{x}\right) = 0$$

as required. ■

Exercise: Using the Squeeze Theorem, show the following two results:

1. If $\lim_{x\to 0} |f(x)| = 0$ then $\lim_{x\to 0} f(x) = 0$. [Hint: Convince yourself that $-|f(x)| \leq f(x) \leq |f(x)|$.]
2. If $f(x)$ is bounded (so that there is some $M > 0$ with $|f(x)| < M$ for all x) and $\lim_{x\to 0} g(x) = 0$ then $\lim_{x\to 0} f(x)g(x) = 0$.

Two useful limits that appears in this section are the following

$$\lim_{x\to 0} \frac{\sin(x)}{x} = 1, \qquad \lim_{x\to 0} \frac{\cos(x) - 1}{x} = 0. \tag{2.31}$$

We will not present the proofs of these facts, but will instead defer to the textbook for the boring and technical details as to how these are shown. Instead, I would like to emphasize several points here that may be pertinent in other courses. Looking at (2.31), we see that these heuristically say that $\sin(x) \approx x$ and $\cos(x) \approx 1 + x$ as gets gets very close to 0. In physics, these are often referred to as the *small angle approximations*, and are responsible for a plethora of dodgy (yet somehow accurate) physical results. The results of (2.31) may be generalized even further by realizing the following result, which we leave as an exercise for the students:

Proposition 2.36

If $\lim_{x\to 0} f(x) = L$ then for any $c \neq 0$ we have $\lim_{x\to 0} f(cx) = L$.

Applying this result to (2.31) we get that for any $c \neq 0$,

$$\lim_{x\to 0} \frac{\sin(ax)}{ax} = 1, \qquad \lim_{x\to 0} \frac{\cos(ax) - 1}{ax} = 0.$$

Example 2.37

Determine the limit

$$\lim_{x\to 0} \frac{x^2 \sin(4x)}{\sin^3(7x)}.$$

Solution. The first important point to realize is that we can evaluate a limit as follows:

$$\lim_{x\to 0} \frac{\sin(ax)}{x} = \lim_{x\to 0} \frac{\sin(ax)}{x} \frac{a}{a} = a \lim_{x\to 0} \frac{\sin(ax)}{ax} = a.$$

The next key point is to try to match up our x components with our sine components. However, we notice that we have four sines and only two x components. Luckily, we can artificially introduce additional x's as follows:

$$\begin{aligned}
\lim_{x\to 0} \frac{x^2 \sin(4x)}{\sin^3(7x)} &= \lim_{x\to 0} \frac{x}{\sin(7x)} \frac{x}{\sin(7x)} \frac{\sin(4x)}{\sin(7x)} \\
&= \lim_{x\to 0} \frac{x}{\sin(7x)} \frac{x}{\sin(7x)} \frac{\sin(4x)}{\sin(7x)} \frac{x}{x} \\
&= \lim_{x\to 0} \frac{x}{\sin(7x)} \frac{x}{\sin(7x)} \frac{\sin(4x)}{x} \frac{x}{\sin(7x)} \\
&= \left[\lim_{x\to 0} \frac{x}{\sin(7x)}\right] \left[\lim_{x\to 0} \frac{x}{\sin(7x)}\right] \left[\lim_{x\to 0} \frac{\sin(4x)}{x}\right] \left[\lim_{x\to 0} \frac{x}{\sin(7x)}\right] \\
&= \frac{1}{7} \times \frac{1}{7} \times 4 \times \frac{1}{7} = \frac{4}{7^3}.
\end{aligned}$$

■

2.7 The Value Theorems

With just continuity under our belts, we can take a look at two exceptionally powerful theorems: the Intermediate Value Theorem (IVT) and the Extreme Value Theorem (EVT).

2.7.1 Intermediate Value Theorem

Theorem 2.38: Intermediate Value Theorem

Let f be a continuous function on $[a, b]$ and assume that $f(a) < f(b)$. For every $v \in (f(a), f(b))$ there exists an $x \in (a, b)$ such that $f(x) = v$.

This theorem also holds if $f(b) < f(a)$ so long as we take $v \in (f(b), f(a))$. Once we wipe away the mysterious mathematical jargon, the idea of this theorem is actually rather obvious: It says that if a function is continuous on an interval, it cannot 'jump' over any point. Additionally, notice that the theorem implies the existence of an x such that $f(x) = v$, but in no way indicates how to find that x. This is an example of an *non-constructive theorem*. The proof of this theorem is rather hard[11]

Remark: There is a version of the Intermediate Value Theorem that uses open intervals, though it is more complicated to write down. The most important parts of the hypotheses are that the function be continuous, and the interval be "connected."

Example 2.39

Show that there exists a solution to the equation

$$e^x = \sin(\pi x) + 2$$

on the interval $[0, 1]$.

Solution. I have deliberately underlined the trigger words for this question: notice that I have not asked the student to find the solution. Indeed, one would have a very difficult time solving for x without the use of a computer. The trick here is the use the Intermediate Value Theorem (which is perhaps unsurprising given that this is the topic of our discussion). The key to doing most questions like this are to define a new function, say

$$f(x) = e^x - \sin(\pi x) - 2$$

and notice that if we can find a point $p \in [0, 1]$ such that $f(p) = 0$ then we will be done, since

$$f(p) = 0 = e^p - \sin(\pi p) - 2 \quad \Rightarrow \quad e^p = \sin(\pi p) + 2.$$

To invoke the intermediate value theorem, we first must show that $f(x)$ is a continuous function. We know that this is true though since $f(x)$ is a sum of continuous functions and hence is continuous. Next, we must show that there are values in $[0, 1]$ such that $f(x) < 0$ and $f(x) > 0$. This problem has been rigged so that these are very easy to find. Indeed

$$f(0) = e^0 - \sin(0) - 2 = -1, \qquad f(1) = e - \sin(\pi) - 2 > 0$$

where in calculating $f(1)$ we have noted that $e \approx 2.7$ so that $e - 2 > 0$. Thus $f(0) < 0$ and $f(1) > 0$ implies there is some $p \in [0, 1]$ such that $f(p) = 0$, and this is precisely what we wanted to show. ■

[11] Using first principles of the real numbers, it requires a 'forcing' argument. There is a much simpler proof using the notion of connectedness, but this is too advanced for this course. In fact, the IVT is a special case of a far more powerful theorem: The continuous image of a connected set is always connected.

Now that we have seen an example, let us talk about the general methodology for proving problems with the intermediate value theorem. The questions will almost always say something along the lines of "Here are two continuous functions $f(x)$ and $g(x)$. Show there is a point such c that $f(c) = g(c)$." It might not be this clear in the problem statement, but you can always re-arrange it so that this is the problem you are asked to solve.

In such instances, your strategy should be as follows: Define a new function $h(x) = f(x) - g(x)$, which is continuous as it is the sum of continuous functions. Find a point a such that $h(a) < 0$ and b such that $h(b) > 0$. By the intermediate value theorem, there is a point c between a and b such that $h(c) = 0$. This is precisely the point we want to find, since

$$0 = h(c) = f(c) - g(c)$$

so re-arranging we have $f(c) = g(c)$. Let's see some more examples, this time a little more theoretical.

Example 2.40

If $f(x)$ is a continuous function such that $\forall x \in [0,1]$ we have $f(x) \in (0,1)$ show that $f(x)$ has a fixed point; that is, there is a point p such that $f(p) = p$.

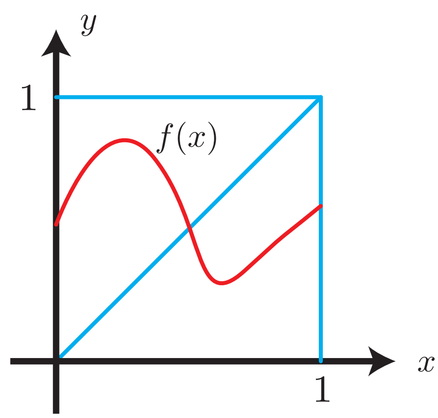


Figure 7: The geometric idea behind Exercise **??**.

Solution. We refer to Figure 7 for geometric intuition. The question is essentially asking us to show that any function which lies strictly in the box $[0,1] \times [0,1]$ must cross the diagonal $y = x$ at some point.

This question is asking us to find a solution to the equation $f(x) = x$. If we proceed in the same fashion as Exercise **??** and our discussion above, we should define a new function

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

and show that there is a point p such that $g(p) = 0$. To apply the Intermediate Value Theorem, we must first confirm that $g(x)$ is continuous. Indeed, since $f(x)$ is continuous and x is continuous,

$g(x)$ is a difference of continuous functions and hence is continuous itself. Next, we want to show that g goes below zero and above zero. Indeed, this is the case since

$$g(0) = f(0) - 0 = f(0)$$

and since $f(0) \in (0,1)$ we know that $g(0) = f(0) > 0$. On the other hand,

$$g(1) = f(1) - 1$$

and since $f(1) \in (0,1)$ it must be that $g(1) < 0$. Thus by the intermediate value theorem, there exists $p \in [0,1]$ such that $g(p) = 0$ and so $f(p) = p$ as required. ■

Exercise: I personally hate the above stated version of the Intermediate Value Theorem. Here is an (equivalent) and much better statement:

> "If f is continuous on the interval $[a,b]$ and $f(a)f(b) < 0$, then there exists some x such that $f(x) = 0$."

Convince yourself that this statement is equivalent to Theorem 2.38.

2.7.2 Extreme Value Theorem

Theorem 2.41: Extreme Value Theorem

If f is continuous on $[a,b]$ then f attains its maximum and minimum values on $[a,b]$.

Remark: The important parts of this theorem are the 'opposite' of the IVT. Indeed, it is *crucial* for this theorem that we look at the closed interval $[a,b]$, and we can actually put multiple closed intervals into this theorem and still have it be true.

Being a named theorem, the EVT finds a tonne of applications in mathematics[12]. Unfortunately (or perhaps fortunately for the student), the most interesting examples are beyond the scope of this course. For this reason, the variety of questions regarding the EVT is quite sparse. Nonetheless, let us look at some examples:

Example 2.42

Determine on which of the following intervals the function $\frac{1}{x}$ attains a global maximum and minimum.

$$[-1,1], \qquad (0,1), \qquad [1,2].$$

Solution. According to the Extreme Value Theorem, we can be guaranteed that $\frac{1}{x}$ attains its max and min if it is continuous on a closed interval. The first interval $[-1,1]$ is closed but $\frac{1}{x}$ is not continuous at 0 which is a point in $[-1,1]$. Hence we cannot guarantee that $\frac{1}{x}$ attains a max and

[12]Just as in the case of the intermediate value theorem, the EVT is a special case of a far more powerful theorem: "The continuous image of a compact set is compact."

min on $[-1, 1]$ (though note that it does attain global minima at $x = \pm 1$). Similarly, the interval $(0, 1)$ is not closed so we cannot guarantee that the maximum or minimum is attained. In fact, there is no max or min of $\frac{1}{x}$ in $(0, 1)$. Finally, $\frac{1}{x}$ *is* continuous on $[1, 2]$ and so by the Extreme Value Theorem the max and min are attained. ■

3 Derivatives

Derivatives are the tools which allow us to dynamically analyze the properties of functions, and are intimately connected with limits (in so much as we require the limit's ability to capture the infinitesimal behaviour of a function).

3.1 First Principles

The speed of the car is measured as the distance travelled in a given amount of time, but limiting ourselves to a single instant means that neither any time nor distance has passed. How then can we ascribe a sensible meaning to this instantaneous velocity? One approach was to approximate the speed by measuring the distance travelled by the car in the minute before (or after) crossing the finish line, and taking the average speed in that duration. Of course, we can get a better approximation by measuring the distance in a single second before/after crossing the finish line, or any time smaller than a second. If $f(t)$ represents the position of the car at time t and the car passes the finish line at time t_0, then the average speed of the car in the second after passing the finish line is given by

$$\text{average speed} \quad = \frac{\text{distance}}{\text{time}} = \frac{f(t_0 + 1\text{ sec}) - f(t_0)}{1\text{ sec}}.$$

We can do this more generally by letting h denote an arbitrary quantity of time, so that

$$\text{average speed} \quad = \frac{\text{distance}}{\text{time}} = \frac{f(t_0 + h) - f(t_0)}{h}$$

where if $h > 0$ then we are measuring after crossing the finish line, and if $h < 0$ we are measuring prior to crossing the finish line. The instantaneous speed of the car may then be determined by taking the limit as our time interval becomes arbitrarily small; that is, by taking $h \to 0$.

While this works well for cars, there is no reason that we cannot generalize this discussion to a more abstract setting or to other examples. The key point is that having precise knowledge about an object at any given point allows us to determine how quickly some attribute is changing. This can be realized geometrically as follows:

Definition 3.1

Consider the graph of a function $f(x)$ and let $a < b$ be distinct real numbers. The *secant line from a to b* is the unique straight line which passes through the points $(a, f(a))$ and $(b, f(b))$.

Of interest to use is the slope of this secant line, given by

$$m_{ab} = \frac{f(b) - f(a)}{b - a}.$$

In fact, this slope describes the average value of the function $f(x)$ between $f(a)$ and $f(b)$. We can get successively better approximations to the instantaneous rate of change of a function by taking the distance between a and b to be successively smaller, and the instantaneous rate of change will be given by taking a limit as $a \to b$. Geometrically, in the limit we get the tangent line.

Definition 3.2

If $f(x)$ is a function and a is in the domain of $f(x)$, we define the *derivative of* $f(x)$ *at* a to be the value
$$f'(a) = \lim_{x\to a} \frac{f(x) - f(a)}{x - a}$$
if this limit exists. In such an instance we say that $f(x)$ is *differentiable at* a, and when $f(x)$ is differentiable everywhere we say it is simply *differentiable*.

Example 3.3

Compute the derivative of $f(x) = x^2$ at the points $x = -1$ and $x = 4$.

Solution. We proceed by applying the definition of the derivative. For $x = -1$ we have
$$\begin{aligned} f'(-1) &= \lim_{x\to -1} \frac{f(x) - f(-1))}{x - (-1)} \\ &= \lim_{x\to -1} \frac{x^2 - (-1)^2}{x+1} = \lim_{x\to -1} \frac{x^2 - 1}{x+1} \\ &= \lim_{x\to -1} \frac{(x+1)(x-1)}{x+1} \\ &= \lim_{x\to -1} (x-1) = -2. \end{aligned}$$
Similarly, at the point $x = 4$ we get
$$\begin{aligned} f'(-4) &= \lim_{x\to 4} \frac{f(x) - f(4)}{x - 4} \\ &= \lim_{x\to 4} \frac{x^2 - 16}{x - 4} = \lim_{x\to 4} \frac{(x-4)(x+4)}{x-4} \\ &= \lim_{x\to 4} (x+4) = 8. \end{aligned}$$
■

One of the things that the student may have noticed in Example 3.3 was that the computation of the derivative at each point is rather redundant. Indeed, notice that in our computation of the derivative of $f(x) = x^2$ at $x = -1$ and $x = 4$, what if we used the number $x = a$ in general? We would find that
$$\begin{aligned} f'(a) &= \lim_{x\to a} \frac{f(x) - f(a)}{x - a} \\ &= \lim_{h\to 0} \frac{x^2 - a^2}{x - a} \\ &= \lim_{x\to a} \frac{(x+a)(x-a)}{x-a} = \lim_{x\to a} (x+a) \\ &= 2a. \end{aligned}$$
This agrees with what we found when $a = -1$ and $a = 4$, but gives us the derivative at any point a. Hence one can view the derivative of a function as a function itself, with
$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{y\to x} \frac{f(y) - f(x)}{y - x}.$$

In fact, this is where the word *derivative* comes from: the value of the function $f'(x)$ is derived from that of $f(x)$.

Remark 3.4 If f is a function and we compute its derivative f', what is the domain of f' and how is it related to that of f? The way this is usually handled often depends on the type of mathematics being done. For example, some people insist that the domain of f and f' must be the same. This is typically done for convenience, but can sometimes be important to the study of your space. On the other hand, some mathematicians define the domain of f' to be the largest subset of f such that the derivative exists. In this case, the domain of f' is a subset of the domain of f. For the most part unless otherwise specified, this class will always use the latter definition.

A different parameterization: An alternative way of writing the derivative comes from changing how we parameterize the limit. Instead of taking two distinct points a and b, let us set $b = a+h$, so that the limit $b \to a$ is equivalent to the limit $h \to 0$. Thus an equivalent definition of the derivative is given by

$$f'(a) = \lim_{h\to 0} \frac{f(a+h) - f(a)}{(a+h) - a} = \lim_{h\to 0} \frac{f(a+h) - f(a)}{h}. \tag{3.1}$$

Exercise: Convince yourself of the equality of the two definitions for the derivative given in (3.1).

Example 3.5

Compute the derivative of the function $f(x) = 9/x$. Compare the domain of f and f'.

Solution. Let us apply our alternative definition of the derivative to this example (though I encourage the student to also compute the derivative using the former definition and see that they are the same).

$$\begin{aligned} f'(x) &= \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{9/(x+h) - 9/x}{h} \\ &= \lim_{h\to 0} \frac{\frac{9x - 9(x+h)}{x(x+h)}}{h} = \lim_{h\to 0} \frac{-9h}{xh(x+h)} \\ &= \lim_{h\to 0} \frac{-9}{x(x+h)} = -\frac{9}{x^2}. \end{aligned}$$

The domain of f is $\mathbb{R} \setminus \{0\}$ and coincides with the domain of f'. ■

Tangent Lines: We know that $f'(a)$ thus represents the slope of the tangent line to the graph of $f(x)$ at the point a. Since we have a slope and a point on in the plane $(a, f(a))$, we can form the equation of the tangent line through $f(x)$ at a using the point-slope formula:

$$y - f(a) = f'(a)(x - a).$$

Example 3.6

Determine the equation of the tangent line through $f(x) = \sqrt{x}$ at the point $x = 1$.

Solution. While we are at it, we might as well determine the derivative of the function $f(x) = \sqrt{x}$.

$$\begin{aligned}
f'(x) &= \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = \lim_{h\to 0} \frac{\sqrt{x+h}-\sqrt{x}}{h} \\
&= \lim_{h\to 0} \frac{\sqrt{x+h}-\sqrt{x}}{h}\left[\frac{\sqrt{x+h}+\sqrt{x}}{\sqrt{x+h}+\sqrt{x}}\right] && \text{Multiply by conjugate} \\
&= \lim_{h\to 0} \frac{x+h-x}{h(\sqrt{x+h}+\sqrt{x})} \\
&= \lim_{h\to 0} \frac{x}{\sqrt{x+h}+\sqrt{x}} = \frac{1}{2\sqrt{x}}.
\end{aligned}$$

At $x = 1$ we have $f(1) = \sqrt{1} = 1$, so we know our line passes through the point $(1, 1)$. Our derivative formula tells us that $f'(1) = \frac{1}{2}$, so using our point slope formula, we thus deduce that the equation of the tangent line is $y - 1 = \frac{1}{2}(x - 1)$ or $y = \frac{1}{2}x + \frac{1}{2}$. ■

Example 3.7

Compute the derivative of $f(x) = 1/x$ for all $x \neq 0$.

Solution. Using the definition of the derivative, we get

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

■

There are occasions when in lieu of using a function to find a derivative, we would instead to prefer to define a relationship amongst variables, with one variable relying explicitly upon the other variable. Such a relationship might be written as $y = x^2$. In the event that the relationship is given by a function $f(x)$, we often write $y = f(x)$. This alternative method of thinking about the relationship between the input and output of an assignment has a correspondingly alternative method for writing derivative: If $y = f(x)$ then

$$f'(x) = \frac{\mathrm{d}y}{\mathrm{d}x}.$$

One should understand the notation $\frac{\mathrm{d}y}{\mathrm{d}x}$ as meaning a quantity which describes how y is changing with respect to x. The motivation for this definition comes from the definition of the derivative as the limit of secant lines. If $a < b$ are two distinct real numbers, the change in the y-value of the function $f(x)$ between a and b may be written as $\Delta y = f(b) - f(a)$ (this is still often used amongst physicists), with the change in the x-value being written as $\Delta x = b - a$. The secant line between a and b is then

$$m_{ab}^f = \frac{\Delta y}{\Delta x} \qquad \text{“} \xrightarrow{\Delta x \to 0} \text{”} \qquad \frac{\mathrm{d}y}{\mathrm{d}x}$$

and in the limit as $\Delta x \to 0$ these delta's "transform" into d's.

Example 3.8

Determine $\frac{\mathrm{d}y}{\mathrm{d}x}$ if $y = (x^2+1)/(2x)$.

Solution. As all examples previous, this amounts to an exercise in algebraic manipulation of the definition of the derivative. Indeed, one finds that

$$\begin{aligned}
\frac{\mathrm{d}y}{\mathrm{d}x} &= \lim_{h\to 0} \frac{\frac{(x+h)^2+1}{2(x+h)} - \frac{x^2+1}{2x}}{h} \\
&= \lim_{h\to 0} \frac{\frac{(x^2+2xh+h^2)(2x)-(x^2+1)(2x+2h)}{(2x)(2x+2h)}}{h} \\
&= \lim_{h\to 0} \frac{(2x^3+4x^2h+2x)-(2x^3-2x^2h-2x-2h)}{h(2x)(2x+2h)} \\
&= \lim_{h\to 0} \frac{2x^2h-2h}{h(2x)(2x+h)} = \frac{2x^2-2}{4x^2} \\
&= \frac{x^2-1}{2x^2}.
\end{aligned}$$

■

3.1.1 Failures of Differentiability

A posteriori, there are three ways in which a function which fail to be differentiable.

1. The function may have a cusp.
2. The function may fail to be continuous.
3. The function may have a vertical tangent line.

With any luck, the student may be able to convince his/herself as to why a cusp will quickly fail to be differentiable. If one were to visualize the tangent lines to the graph in a neighbourhood around a cusp, they would see that the tangents have distinctly different values at the cusp point. This implies that the two-sided limit

$$\lim_{h\to 0} \frac{f(x+h)-f(x)}{h} \tag{3.2}$$

does not exist, despite the fact that the one-sided limits might. A similar argument holds for discontinuity. If $f(x)$ is not continuous at the point x_0 then either $f(x_0)$ does not exist (making Equation 3.2 read as non-sense in the case of removable or essential discontinuities) or $f(x+h)-f(x)$

converges to a finite number as $h \to 0$, so that the limit in Equation 3.2 is not defined. Finally, a vertical tangent line has slope "infinity," meaning that the derivative diverges at that point.

Example 3.9

Consider the functions $f(x) = |x|, g(x) = x/|x|$ and $h(x) = \sqrt[3]{x}$. Determine the points where each function is not differentiable, and classify the manner in which the function fails to be differentiable.

Solution. We begin by analyzing the function $f(x) = |x|$. A glance at the graph of $f(x)$ shows that $f(x)$ has a cusp at the point $x = 0$, so let us check whether $f(x)$ is differentiable at 0. Each one sided limit may be computed as

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

and as the limit disagree, the function is not differentiable.

The function $g(x)$ is not defined at $x = 0$, so it does not even make sense to write down the equation

$$\lim_{h\to 0} \frac{g(h) - g(0)}{h}.$$

It follows that $g(x)$ cannot be differentiable at 0, as it fails to be continuous.

Finally, let us look at the function $h(x) = \sqrt[3]{x}$. If we recall that we may factor a difference of cubes as

$$a^3 - b^3 = (a - b)(a^2 + ab + b^2)$$

we may compute the derivative of $h(x)$ in general to be

$$\begin{aligned} h'(x) &= \lim_{y\to x} \frac{\sqrt[3]{x} - \sqrt[3]{y}}{x - y} \\ &= \lim_{y\to x} \frac{\sqrt[3]{x} - \sqrt[3]{y}}{(\sqrt[3]{x} - \sqrt[3]{y})(\sqrt[3]{x^2} + \sqrt[3]{xy} + \sqrt[3]{y^2})} \\ &= \lim_{y\to x} \frac{1}{\sqrt[3]{x^2} + \sqrt[3]{xy} + \sqrt[3]{y^2}} \\ &= \frac{1}{3\sqrt[3]{x^2}}. \end{aligned}$$

This limit will thus hold true for any $x \neq 0$, but at $x = 0$ the slope of the tangent line is infinite. ■

I had mentioned previously that continuity measures the well-behavedness of a function; in particular, a function $f(x)$ is continuous at a if, in a neighbourhood of a, $f(x)$ is bounded, does not oscillate infinitely quickly, and does not jump. The above discussion shows that a function $f(x)$ is differentiable at a if it is behaved in many of the same ways as a continuous function, with the added assumption that $f(x)$ has no cusps. We may glean this as an indication that differentiability measures the "smoothness" of a function, and in fact the word 'smooth' has a formal mathematical definition which I will allude to shortly. Of more immediate importance is the fact that being differentiable is *stronger* than being continuous; that is, **all differentiable functions are continuous**, though the converse is not true (not all continuous functions are differentiable).

Proposition 3.10

All differentiable functions are continuous.

Proof. Let $f(x)$ be differentiable at the point a, so that

$$\lim_{x\to a} \frac{f(x)-f(a)}{x-a} = L \text{ exists and is finite.} \tag{3.3}$$

To show continuity, we wish to show that

$$\lim_{x\to a} f(x) = f(a), \qquad \text{or equivalently} \qquad \lim_{x\to a} [f(x)-f(a)] = 0.$$

At this point we say to ourselves, "what part of the hypothesis have we failed to use?" Examination will show that we know something about the quantity $(f(x)-f(a))/(x-a)$ and now wish to say something about $f(x)-f(a)$, so we should multiply and divide by the quantity $(x-a)$ to find

$$\begin{aligned}
\lim_{x\to a} [f(x)-f(a)] &= \lim_{x\to a} \frac{f(x)-f(a)}{x-a}(x-a) \\
&= \left[\lim_{x\to a} \frac{f(x)-f(a)}{x-a}\right] \left[\lim_{x\to a} x-a\right] && \text{By the limit laws since both limits exist} \\
&= L \times 0 = 0
\end{aligned}$$

as desired. □

3.2 Some Derivative Formulas

Introducing the Limit Laws meant that we could avoid having to create annoying value-tables each time we wished to evaluate a limit. Similarly, it is rather cumbersome lugging around the awkward definition of the derivative. In this section, we give a few formulas to simplify some calculations, and examine how to differentiate polynomials and exponential functions.

Proposition 3.11

If $f(x)$ and $g(x)$ are differentiable at c, then

1. $\dfrac{\mathrm{d}}{\mathrm{d}x}[\alpha f(x)] = \alpha f'(x)$ for any real number α,
2. $\dfrac{\mathrm{d}}{\mathrm{d}x}[f(x)+g(x)] = f'(x)+g'(x)$

Proof. 1. By the definition of the derivative, we have

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[\alpha f(x)\right] = \lim_{h\to 0}\frac{\alpha f(x+h)-\alpha f(x)}{h} = \alpha\lim_{h\to 0}\frac{f(x+h)-f(x)}{h} = \alpha f'(x).$$

2. Again, using the definition of the derivative we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\left[f(x)+g(x)\right] &= \lim_{h\to 0}\frac{[f(x+h)+g(x+h)]-[f(x)+g(x)]}{h} \\ &= \lim_{h\to 0}\frac{[f(x+h)-f(x)]+[g(x+h)-g(x)]}{h} && \text{re-arranging terms} \\ &= \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}+\lim_{h\to 0}\frac{g(x+h)-g(x)}{h} && \text{by the limit laws} \\ &= f'(x)+g'(x).\end{aligned}$$

□

The previous theorem tells us that $\frac{\mathrm{d}}{\mathrm{d}x}$ is what is called a *linear operator*, in that it preserves scalar multiplication and addition. As an immediate corollary, we see that

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[f(x)-g(x)\right] = \frac{\mathrm{d}}{\mathrm{d}x}\left[f(x)+(-1)\times g(x)\right] = f'(x)+(-1)\times g'(x) = f'(x)-g'(x)$$

so that the derivative of the difference is also the difference of the derivatives.

These facts are ultimately important, but are often very easy to remember and are not particularly enlightening. However, the are extremely useful in that they tell us exactly how to differentiate polynomials, once we have the following result in hand.

Proposition 3.12

For any positive integer n, $\dfrac{\mathrm{d}}{\mathrm{d}x}x^n = nx^{n-1}$.

Proof. Recall that one may always factor the n^{th}-powered difference of two elements as

$$(a^n - b^n) = (a-b)(a^{n-1}+a^{n-2}b^1+\cdots+a^1b^{n-2}+b^{n-1}).$$

Applying this to the definition of the derivative, we see that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}x^n &= \lim_{y\to x}\frac{y^n-x^n}{y-x} = \lim_{y\to x}\frac{(y-x)(y^{n-1}+y^{n-2}x+\cdots+yx^{n-2}+x^{n-1})}{y-x} \\ &\lim_{y\to x}\underbrace{y^{n-1}+y^{n-2}x+\cdots+yx^{n-2}+x^{n-1}}_{n\text{-times}} \\ &= nx^{n-1}.\end{aligned}$$

□

Now every polynomial function is built for scalar multiplication and addition of monomials of the form a_nx^n, hence we may now differentiate any polynomial quite easily as follows:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}x}\left[a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2 x^2 + a_1 x + a + 0\right] \\
&= a_n\left[\frac{\mathrm{d}}{\mathrm{d}x}x^n\right] + a_{n-1}\left[\frac{\mathrm{d}}{\mathrm{d}x}x^{n-1}\right] + \cdots + a_1\left[\frac{\mathrm{d}}{\mathrm{d}x}x\right] + 0 \\
&= na_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \cdots + 2a_2 x + a_1.
\end{aligned}$$

Example 3.13

Compute the derivative of $x^{654} + 13x^{45} - 20x^5 + 6$.

Solution. Using our template above, we see that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x}\left[x^{654} + 13x^{45} - 20x^5 + 6\right] &= (654)x^{653} + 13(45)x^{44} - 250(5)x^4 \\
&= 654x^{653} + 585x^{44} - 1250x^4.
\end{aligned}$$

■

While the above proposition only showed that $\frac{\mathrm{d}}{\mathrm{d}x}^n = nx^{n-1}$ for any *positive integer* n, it turns out that the formula works for all real numbers. There is some work involved in showing that this is the case, so for the moment the student should take this fact for granted.

Exercise: Show that $\frac{\mathrm{d}}{\mathrm{d}x}^n = nx^{n-1}$ for all integers $n \in \mathbb{Z}$.

Exercise: Recall that if n is a positive integer, we define $n! = n(n-1)(n-2)\cdots(3)(2)$. For example, $3! = 3 \times 2 = 6$ and $4! = 4 \times 3 \times 2 = 24$. Consider the function

$$f(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots .$$

Compute the derivative of $f(x)$. If you are given that derivatives are unique up to additive constants, what function must $f(x)$ be?

3.2.1 The Product Rule

Computing the derivative of sums of functions was ultimately rather simple. However, it turns out that computing the derivative of a product is a far more complicated affair.

Theorem 3.14

If $f(x)$ and $g(x)$ are differentiable at c, then $f(x)g(x)$ is differentiable at c and

$$\frac{\mathrm{d}}{\mathrm{d}x}f(x)g(x) = f'(x)g(x) + g'(x)f(x).$$

Proof. By the limit definition of the derivative, we have

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

□

Example 3.15

Let $f(x)$ be a differentiable function such that $f'(x) = 1/f(x)$. Compute $\frac{\mathrm{d}}{\mathrm{d}x}[f(x)]^2$.

Solution. Applying the product rule, we have

$$\frac{\mathrm{d}}{\mathrm{d}x}[f(x)]^2 = f'(x)f(x) + f'(x)f(x) = 2f'(x)f(x).$$

Now since we were told that $f'(x) = 1/f(x)$ we may substitute this to find that

$$\frac{\mathrm{d}}{\mathrm{d}x}[f(x)]^2 = 2f'(x)f(x) = 2\frac{f(x)}{f(x)} = 2.$$

■

Exercise: Find a differentiable function that satisfies $f'(x) = [f(x)]^{-1}$ as in Example 3.15.

Example 3.16

Let $f(x)$ be differentiable and satisfy $f(1) = 1$ and $f'(1) = 2$. Compute $g'(1)$ where $g(x) = f(x)/x$.

Solution. We have already seen that $\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{x} = -\frac{1}{x^2}$, so the product rule tells us that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x} g(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\left[f(x)\times\frac{1}{x}\right] = f'(x)\left(\frac{1}{x}\right) + f(x)\left(\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{x}\right) \\
&= \frac{f'(x)}{x} + f(x)\left(-\frac{1}{x^2}\right) \\
&= \frac{f'(x)}{x} - \frac{f(x)}{x^2}.
\end{aligned}$$

If we now substitute $x = 1$ into this equation we find

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

■

In fact, there is no reason to limit ourselves to considering the product of only two functions. If $f(x), g(x)$, and $h(x)$ are all differentiable then

$$\frac{\mathrm{d}}{\mathrm{d}x} f(x)g(x)h(x) = f'(x)g(x)h(x) + f(x)g'(x)h(x) + f(x)g(x)h'(x).$$

The way to see this is simple: define a new function $n(x) = g(x)h(x)$ so that $n'(x) = g'(x)h(x) + g(x)h'(x)$ and $f(x)g(x)h(x) = f(x)n(x)$. Since the right-hand-side is a product of two functions, the product rule again gives us

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x} f(x)n(x) &= f'(x)n(x) + f(x)n'(x) = f'(x)g(x)h(x) + f(x)\left[g'(x)h(x) + g(x)h'(x)\right] \\ &= f'(x)g(x)h(x) + f(x)g'(x)h(x) + f(x)g(x)h'(x)\end{aligned}$$

and this process is easily generalized to any number of functions.

Theorem 3.17: The Quotient Rule

If $f(x)$ and $g(x)$ are differentiable at c and $g(c) \neq 0$ then $f(x)/g(x)$ is differentiable at c and

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{f(x)}{g(x)} = \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2}.$$

Exercise:

1. Let $f(x)$ be a differentiable function. Show that for any $n \in \mathbb{Z}$ we have $\frac{\mathrm{d}}{\mathrm{d}x} f(x)^n = nf(x)^{n-1}f'(x)$. [Hint: Use induction and the product rule].
2. Use part (a) to show that the Product Rule and the Quotient Rule are equivalent.

Example 3.18

Confirm the computation of the derivative of $f(x)/x$ given in Example 3.16.

Solution. Applying the quotient rule to $f(x)/x$ we find that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\frac{f(x)}{x} &= \frac{f'(x)x - \left(\frac{\mathrm{d}}{\mathrm{d}x}x\right)f(x)}{[x^2]} \\ &= \frac{xf'(x) - f(x)}{x^2}.\end{aligned}$$

In Example 3.16 we found that

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{f(x)}{x} = \frac{f'(x)}{x} - \frac{f(x)}{x^2} = \frac{xf'(x) - f(x)}{x^2}$$

exactly as expected. ■

3.2.2 Higher Derivatives

Differentiating a function $f(x)$ resulted in another function $f'(x)$. If $f'(x)$ is nice enough, we should be able to differentiate it again to get $f''(x)$, the *second derivative of* $f(x)$. Once again, $f''(x)$ is a function and we may differentiate it to get a third derivative $f'''(x)$. After three primes, one usually uses the notation $f^{(n)}(x)$ to denote the n^{th} derivative of $f(x)$. These have important interpretations in both mathematics and science which we shall explore later. In Leibniz notation, we use the operator $\frac{\mathrm{d}}{\mathrm{d}x}$ to take subsequent derivatives, hence if $y = f(x)$ then the first derivative is $\frac{\mathrm{d}y}{\mathrm{d}x}$, while the second, third, and fourth derivatives are

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}^2 y}{\mathrm{d}x^2}, \qquad \frac{\mathrm{d}}{\mathrm{d}x}\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} = \frac{\mathrm{d}^3 y}{\mathrm{d}x^3}, \qquad \frac{\mathrm{d}}{\mathrm{d}x}\frac{\mathrm{d}^3 y}{\mathrm{d}x^3} = \frac{\mathrm{d}^4 y}{\mathrm{d}x^3}$$

respectively, with the patter continuing *ad infinitum*.

Example 3.19

Compute the second derivative of the function $f(x) = \frac{1}{x}$. Try to determine a formula for the n^{th} derivative $f^{(n)}(x)$, and prove that this is the case using induction.

Solution. In Example 3.7 we showed that $f'(x) = -1/x^2$. To compute $f''(x)$ we take the derivate of $f'(x)$ and find that

$$\begin{aligned} f''(x) = \frac{\mathrm{d}}{\mathrm{d}x} f'(x) &= \lim_{h\to 0} \frac{f'(x+h) - f'(x)}{h} \\ &= \lim_{h\to 0} \frac{-\frac{1}{x+h}^2 + \frac{1}{x^2}}{h} = \lim_{h\to 0} \frac{\frac{-x^2 + (x+h)^2}{x^2(x+h)^2}}{h} \\ &= \lim_{h\to 0} \frac{2xh + h^2}{hx^2(x+h)^2} = \lim_{h\to 0} \frac{2x+h}{x^2(x+h)^2} \\ &= \frac{2x}{x^4} = \frac{2}{x^3}. \end{aligned}$$

Of course, having completed exercise 3.2 we can also confirm this using the power-rule.

If we were to continue on in this fashion, the third derivative would be computed to be $\frac{-6}{x^4}$, and so on. We hypothesize that in general:

$$f^{(n)}(x) = \frac{(-1)^n n!}{x^{n+1}}.$$

We have done the base case already. In the induction step, we thus have

$$\begin{aligned} f^{(n+1)}(x) = \frac{\mathrm{d}}{\mathrm{d}x} f^{(n)}(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{(-1)^n n!}{x^{n+1}}\right] \\ &= (-1)^n n! \frac{\mathrm{d}}{\mathrm{d}x}\left[x^{-n+1}\right] \\ &= (-1)^n n! \left[-(n+1)x^{-(n+2)}\right] \\ &= \frac{(-1)^{n+1}(n+1)!}{x^{n+2}}, \end{aligned}$$

as required. ■

3.2.3 Trigonometric Derivatives

We would now like to explore the derivatives of the trigonometric functions $\sin(x)$ and $\cos(x)$. Since all other functions (such as $\tan(x)$ and $\sec(x)$) are built from quotients of these two functions, our knowledge of the product and quotient rules will allow us to compute the derivatives of all the remaining fundamental trigonometric functions. For the sake of clarity, recall the following two limits from Section 2.6.

$$\lim_{\theta\to 0}\frac{\sin(\theta)}{\theta} = 1, \qquad \lim_{\theta\to 0}\frac{\cos(\theta)-1}{\theta} = 0. \tag{3.4}$$

Also, recall the angle sum identities:

$$\sin(x\pm y) = \sin(x)\cos(y)\pm\cos(x)\sin(y), \quad \cos(x\pm y) = \cos(x)\cos(y)\mp\sin(x)\sin(y).$$

Theorem 3.20

The functions $\sin(x)$ and $\cos(x)$ are everywhere differentiable, and in particular

$$\frac{\mathrm{d}}{\mathrm{d}x}\sin(x) = \cos(x), \qquad \frac{\mathrm{d}}{\mathrm{d}x}\cos(x) = -\sin(x).$$

Proof. I will demonstrate the proof for the derivative of $\sin(x)$ and leave the proof for $\cos(x)$ as an exercise for the student. Applying the definition of the derivative we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x}\sin(x) &= \lim_{h\to 0}\frac{\sin(x+h)-\sin(x)}{h} \\
&= \lim_{h\to 0}\frac{\sin(x)\cos(h)+\cos(x)\sin(h)-\sin(x)}{h} && \text{by the double angle identity} \\
&= \lim_{h\to 0}\sin(x)\frac{\cos(h)-1}{h}+\cos(x)\frac{\sin(h)}{h} \\
&= \sin(x)\underbrace{\left[\lim_{h\to 0}\frac{\cos(h)-1}{h}\right]}_{=0}+\cos(x)\underbrace{\left[\lim_{h\to 0}\frac{\sin(h)}{h}\right]}_{=1} && \text{by the limit laws} \\
&= \cos(x).
\end{aligned}$$

□

Example 3.21

Using the quotient rule, compute the derivative of $\tan(x)$.

Solution. Since $\tan(x) = \sin(x)/\cos(x)$ we can apply the quotient rule to find

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x}\tan(x) &= \frac{\left[\frac{\mathrm{d}}{\mathrm{d}x}\sin(x)\right]\cos(x)-\sin(x)\left[\frac{\mathrm{d}}{\mathrm{d}x}\cos(x)\right]}{\cos^2(x)} \\
&= \frac{\cos^2(x)+\sin^2(x)}{\cos^2(x)} = \frac{1}{\cos^2(x)} \\
&= \sec^2(x).
\end{aligned}$$

■

Computing the remaining trigonometric derivatives is an excellent exercise in the use of the quotient rule, but I shall provide here a table of the trigonometric derivatives:

$\frac{\mathrm{d}}{\mathrm{d}x}\sin(x) = \cos(x)$	$\frac{\mathrm{d}}{\mathrm{d}x}\csc(x) = -\csc(x)\cot(x)$
$\frac{\mathrm{d}}{\mathrm{d}x}\cos(x) = -\sin(x)$	$\frac{\mathrm{d}}{\mathrm{d}x}\sec(x) = \sec(x)\tan(x)$
$\frac{\mathrm{d}}{\mathrm{d}x}\tan(x) = \sec^2(x)$	$\frac{\mathrm{d}}{\mathrm{d}x}\cot(x) = -\csc^2(x)$

Table 1: List of Trigonometric Derivatives.

As a general rule, many of these identities are not worth remembering. Instead, the student should focus on how to *derive* these relationships using the quotient rule, so that the formulae can be called upon when needed.

Periodicity in the differentiation process: It is not too hard to see that by iteratively differentiating sine/cosine, we eventually cycle back on ourselves:

$$\frac{\mathrm{d}}{\mathrm{d}x}\sin(x) = \cos(x), \quad \frac{\mathrm{d}^2}{\mathrm{d}x^2}\sin(x) = -\sin(x), \quad \frac{\mathrm{d}^3}{\mathrm{d}x^2} = -\cos(x), \quad \frac{\mathrm{d}^4}{\mathrm{d}x^4}\sin(x) = \sin(x),$$

so differentiation the sine function four times gives us back the function back (or alternatively, differentiate the function twice gives us back the negative of the function). The student should check that precisely the same argument works with the cosine function.[13]

Degrees vs Radians: It was essential that we used radians in lieu of degrees when doing all of our trigonometric limits. The reason is precisely because the identity $\lim_{x\to 0}\frac{\sin(x)}{x} = 1$ would no longer be true, and would instead given a value of $\frac{180}{\pi}$. One can then see how this would affect our calculations of the derivative by either imitating the proof of Theorem 3.20 or using the fact that if r is radians and θ is degrees, then $\theta = \frac{180}{\pi}r$. Indeed,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\sin(r) = \sin\left(\frac{180}{\pi}\theta\right) = \frac{180}{\pi}\cos\left(\frac{180}{\pi}\theta\right).$$

This destroys the nice symmetry between the derivatives of sine and cosine, and introduces a pesky scaling factor every time we differentiate.

Even more importantly, since we learn calculus from the point of view of radians, if the student is ever asked to do a computation, using degrees will actually give the wrong answer. The moral of the story: Always use radians.

[13] Akin to this is the following idea: We say that the *order* of a number a is n if $a^n = 1$. Notice that the order of -1 is 2, since $(-1)^2 = 1$, while the order of $i = \sqrt{-1}$ is four, since $i^4 = 1$. It turns out that one way of seeing that sine/cosine repeat every four-derivatives is precisely because $i^4 = 1$.

3.3 Rates of Change

We already saw at the beginning of this section how derivatives can be used to deduce the *instantaneous rate of change* of one quantity with respect to another. The power of this is that it allows us to take boring, static relationships, and differentiate them to discover the dynamic interplay between variables.

Example 3.22

The volume of a sphere of radius r is $V = \frac{4}{3}\pi r^3$. Determine how the volume changes as r is allowed to vary.

Solution. All we need to do is differentiate the given expression with respect to r to find that

$$\frac{\mathrm{d}V}{\mathrm{d}r} = 4\pi r^2.$$

This says that the rate of change of volume V with respect to the radius r is $4\pi r^2$. As an example, this means that double the radius of a sphere will quadruple its volume, while tripling the radius will increase the volume 9-fold! ■

One of the more utilized relationships is that of position, velocity, acceleration, jerk, etc. If $p(t)$ describes the position of an object with respect to time, then $p'(t)$ is its velocity and $p''(t)$ is its acceleration with respect to time. This can be used to model physical situations which can then be solved by mathematical methods:

3.3.1 Some Physics

Gravity: We have learned at this point that if the trajectory of an object is described as $x(t)$, we may think of $\frac{\mathrm{d}x}{\mathrm{d}t}$ as the instantaneous velocity and $\frac{\mathrm{d}^2x}{\mathrm{d}t^2}$ as the instantaneous acceleration of that object.

Example 3.23

Consider a child standing on the surface of the earth. If the child has a ball of mass m and throws the ball from an initial height y_0 with an initial velocity v_0, describe the trajectory of the ball as a function of time.

Solution. The common theme in these types of questions is to exploit Newton's second law, which say that the net force F_{net} exerted on a body is equal to simply the object's inertial mass times its acceleration; that is, $F_{\text{net}} = ma$. In our case, we shall ignore air-friction and assume that the force of gravity is a constant, denoted by g. A quick look at Figure 8 shows us that the net force acting on the ball is $F = -mg$, where the negative sign is included to indicate that the force of gravity is acting in the opposite direction of initial motion. Newton's second law then implies that

$$F = ma = -mg.$$

Now since the mass is the same in either case (see the note at the bottom) we may cancel to deduce that $a = -g$. But we mentioned earlier that if y is the ball's vertical displacement, then $a = \frac{d^2y}{dt^2}$, yielding the ordinary differential equation

$$\frac{d^2y}{dt^2} = -g.$$

We are not yet at a stage wherein we will properly be able to treat differential equations. Nonetheless, this is not too difficult to handle since g is a constant. In particular, we ask ourselves "What kind of function may we differentiate twice to get a constant?" The immediate answer is that quadratic functions will suffice. To see this, recall that

$$\frac{d^2}{dt^2}\left(at^2 + bt + c\right) = 2a \tag{3.5}$$

which is a constant. We thus make the *ansatz* (educated guess) that we may describe the equation of motion as $y(t) = at^2 + bt + c$ for which we then need to determine the coefficients $a, b, c \in \mathbb{R}$. Equation (3.5) implies that $y''(t) = 2a = -g$ so that $a = -g/2$. The other coefficients are determined by the two pieces of information that we have yet to use: the initial height and initial velocity. The initial velocity (at time $t = 0$) is v_0 which means that

$$v_0 = \frac{dy}{dt}(0) = [2at + b]_{t=0} = b$$

and the initial height is

$$y_0 = y(0) = \left[at^2 + bt + c\right]_{t=0} = c.$$

Putting this all together we conclude that the equation of motion is given by

$$y(t) = -\frac{g}{2}t^2 + v_0 t + y_0. \tag{3.6}$$

Note that this makes sense intuitively (Figure 8): Indeed, the leading term of t^2 should be negative since the ball is being projected upwards and then is returning to the ground.

The derivative yields $y'(t) = -gt + v_0$ which has a zero at $t = v_0/g$. This is clearly a maximum since the second derivative test is $y''(t) = -g < 0$. The height at this time is given by

$$y\left(\frac{v_0}{g}\right) = -\frac{g}{2}\left(\frac{v_0}{g}\right)^2 + v_0\left(\frac{v_0}{g}\right) + y_0 = \frac{v_0^2}{2g} + y_0 \tag{3.7}$$

■

Remark 3.24 In the above derivation of our equation of motion, we commented that we could simply "cancel" the masses on each side of the equation. While this is something that we often do without thinking, I would ask the student to be very thoughtful about this procedure. Philosophically, these are in fact *different* masses: one is the object's inertial mass and the other is the object's gravitational mass. The fact that they are essentially the same is something called the *Equivalence Principle* and is actually one of the foundational postulates of Einstein's theory of general relativity.

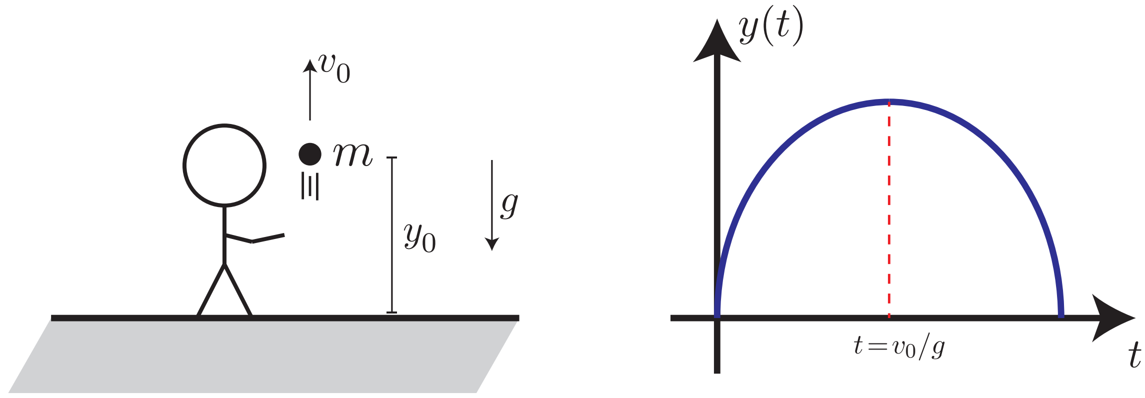


Figure 8: **Left:** A child throws a ball of mass m into the air from an initial height of y_0. It is pulled back to Earth via the force of gravity. **Right:** A plot of the height of the ball as a function of time.

A good way of checking the feasibility of your answers in physics is to ensure that things are dimensionally correct. For example, the dimension of y_0, denoted $[y_0]$, is length. We write this as $[y_0] = L$. Since v_0 is velocity its units are length per time, written as $[v_0] = LT^{-1}$. Acceleration is length per time per time, so $[g] = LT^{-2}$. Finally, time is clearly just $[t] = T$. We can multiply any collection of dimensions, but we may only add dimensions when they are the same. Looking at equation (3.6) it is thus easy to see that

$$\left[\frac{g}{2}t^2\right] = [g][t]^2 = (LT^{-2})T^2 = L, \quad [v_0 t] = [v_0][t] = (LT^{-1})T = L, \quad [y_0] = L$$

so that each component actually has the correct dimensions. Similarly, in (3.7) we have

$$\left[\frac{v_0^2}{g}\right] = \frac{L^2T^{-2}}{LT^{-2}} = L$$

so again our solution makes sense. Notice then that all of this implies that $\left[\frac{\mathrm{d}}{\mathrm{d}t}\right] = T^{-1}$. In general, differentiating by a quantity with units U results in an object whose units are affected by U^{-1}.

The Simple Harmonic Oscillator The simple harmonic oscillator is easily one of the most important physical systems that can be described. It appears constantly throughout all of physics and much of mathematics, and so warrants some of our attention.

Example 3.25

Consider an object of mass m attached to a spring with spring constant k. If the object is initially a distance of x_0 from its equilibrium and is given an initial velocity of v_0, what is the equation of motion of the spring?

Solution. Let the equilibrium point of the spring be given by $x = 0$, in which case Hooke's law tells us that the net force experienced by the object is given by $F = -kx$. Again a simple application

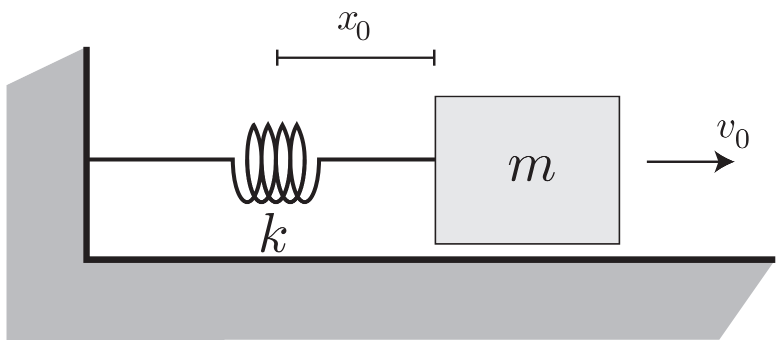


Figure 9: A spring-mass system, wherein the spring coefficient is given by the quantity k.

of Newton's second law tells us that

$$F = ma = -kx. \tag{3.8}$$

Writing $a = \frac{d^2x}{dt^2}$, dividing by m, and defining a new variable $\omega = \sqrt{k/m}$, equation (3.8) is equivalent to

$$\frac{d^2x}{dt^2} + \omega^2 x = 0. \tag{3.9}$$

Again, a proper treatment of this question warrants the use of some ODE theory, but we can formulate another ansatz as to the solutions of this equation. Essentially, this equation is asking us "What function, when you differentiate it twice, gives you back the negative of the function?" The only such functions (of which you are currently aware) are the sine and cosine functions. In fact, the student should convince themselves that equation (3.9) is satisfied by

$$x(t) = A\sin(\omega t) + B\cos(\omega t)$$

for as-of-yet determined coefficients $A, B \in \mathbb{R}$. Just as we did before, we may calculate A, B using the unused information given to us in the problem. The initial displacement is y_0 implying that

$$y_0 = x(0) = A\sin(0) + B\cos(0) = B$$

and the initial velocity is

$$v_0 = x'(0) = A\omega\cos(0) - B\omega\sin(0) = A\omega.$$

Putting this all together, the equation of motion for the spring is

$$x(t) = \frac{v_0}{\omega}\sin(\omega t) + y_0\cos(\omega t). \qquad \blacksquare$$

Remark 3.26 The ω introduced above is known as the *resonance frequency* and is the natural frequency of the system. Many systems have natural frequencies and much technology has been developed by exploiting this frequency (example: Magnetic *Resonance* Imaging). On the other hand, resonance is also responsible for some catastrophes. Look for videos of the original Tacoma Narrows Bridge. When engineers designed this bridge, they failed to consider what the bridge's natural frequency would be. There would be some days wherein high-winds would match this frequency causing the bridge to oscillate. The bridge eventually collapsed.

The Pendulum: So far I have given you some very simple problems whose solutions could easily be found. However, the reality is that while we can often easily formulate differential equations to describe a system, it is very rare that these equations can be properly solved (we will see more on this when we get to integration theory). Nonetheless, there is an entire field of mathematical study dedicated to solving differential equations which do not have closed form solutions. The moral of the story: we can use calculus to describe the motion of a great deal many things.

Here is an example which does not have a closed form solution, but whose equation of motion is simple to derive.

Example 3.27

Consider a bob of mass m hanging from a rigid rod of length L. If the rod is not in an equilibrium state, describe its equation of motion.

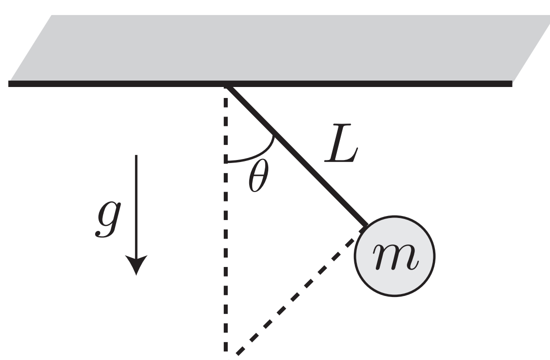


Figure 10: A pendulum system.

Solution. This is best done by considering Figure 10. The net force acting on the bob will correspond to $F = mg\sin\theta$ and so Newton's second law tells us that $ma = mg\sin\theta$. If we try our similar trick of writing $a = \frac{d^2x}{dt^2}$ we will have an equation in both the variable x and θ, which is not what we desire. Instead, we must find a way to relate x and θ. This is not too hard once we realize that x describes the arclength swept out by the rod, meaning that $x = L\theta$ (this is just the arclength formula). Twice differentiating tells us that $\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = L\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2}$ so substituting we get

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} - \frac{g}{L}\sin\theta = 0.$$

It may not be obvious at this point that this equation cannot by solved analytically, but let me assure you that it cannot. However, there is one extenuating circumstance in which we may find *approximate* solutions. If θ is a very small angle then $\sin\theta \approx \theta$ in which case this becomes

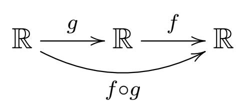
$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} - \frac{g}{L}\theta = 0$$

which we recognize as the equation for the simple harmonic oscillator. ■

3.4 The Chain Rule

We have seen how to take the derivatives of sums, products, and quotients of functions. The only major operation left is to look at function composition.

The reason that this rule is called the chain rule is because we can think of composition as a chain: Given functions f and g the compositions $f \circ g$ means that we first apply g then apply f afterwards.



So how do we differentiate the composition? The answer is perhaps more insightful if we look at Leibniz notation. Let $y = g(x)$ and $w = f(y)$ so that $w = f(y) = f(g(x))$. We should first remark that some students find the above equation disconcerting. What does it mean that f is a function of y, but that $f(g(x))$ is a function of x?

The idea behind this is easier to comprehend if we think of a physical situation. Let's say that you're a weatherman, and you want to determine the outside temperature over the course of the day. You realize that the temperature H depends on the amount of sunshine s, and they are related through a function $H = f(s)$. This makes perfect sense, and you can even use calculus to determine $\frac{\mathrm{d}H}{\mathrm{d}s}$, the instantaneous rate of change of temperature with respect to sunshine.

But maybe you find it difficult to measure the amount of sunshine directly. Instead, you realize that sunshine itself is a (periodic) function of time t; that is, you can write $s = g(t)$ for some function g. This allows you to determine $\frac{\mathrm{d}s}{\mathrm{d}t}$, the instantaneous rate of sunshine with respect to time.

But what if you now want to determine how the temperature changes with respect to time? It seems like you should be able to do this: after all, you know how temperature changes with sunshine, and how sunshine changes with time:

$$\text{time} \longrightarrow \text{sunshine} \longrightarrow \text{temperature}.$$

Here, sunshine just acts as an intermediary for getting from time to sunshine, and we can write $T = f(s) = f(g(t))$.

Now back to the question of differentiation. To differentiate temperature with respect to time in this way, we are definitely going to have to go through time. We have already seen how the differentials can act like fractions, and keeping this in mind it is tempting to say that

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \frac{\mathrm{d}H}{\mathrm{d}s}\frac{\mathrm{d}s}{\mathrm{d}t}.$$

In fact, it turns out that this is correct:

Theorem 3.28: Chain Rule

Let f, g be function such that g is differentiable at c and f is differentiable at $g(c)$. Then the composition $f \circ g$ is differentiable at c and $(f \circ g)'(c) = f'(g(c))g'(c)$. In Leibniz notation, if $w = f(y)$ and $y = g(x)$ then

$$\left.\frac{\mathrm{d}w}{\mathrm{d}x}\right|_a = \left.\frac{\mathrm{d}w}{\mathrm{d}y}\right|_{g(a)} \left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_a. \tag{3.10}$$

Example 3.29

Compute the derivative of $\sqrt{x + \sqrt{x}}$.

Solution. The key to differentiation this function is to realize it as the composition of two functions. In particular, let $f(x) = \sqrt{x}$ and $g(x) = x + \sqrt{x}$ so that

$$f(g(x)) = \sqrt{g(x)} = \sqrt{x + \sqrt{x}}.$$

Now we know that $f'(x) = \frac{1}{2\sqrt{x}}$ and $g'(x) = 1 + \frac{1}{2\sqrt{x}}$ so using the chain rule we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\sqrt{x+\sqrt{x}} &= \frac{\mathrm{d}}{\mathrm{d}x}f(g(x)) = f'(g(x))g'(x) \\ &= \frac{1}{2\sqrt{g(x)}}\left(1 + \frac{1}{2\sqrt{x}}\right) \\ &= \frac{2\sqrt{x}+1}{4\sqrt{x^2 + x\sqrt{x}}}.\end{aligned}$$

■

Example 3.30

Recall that in Exercise 3.2.1 we show that for any $n \in \mathbb{Z}$,

$$\frac{\mathrm{d}}{\mathrm{d}x}f(x)^n = nf(x)^{n-1}f'(x).$$

Confirm this using the chain rule.

Solution. By setting $g(x) = x^n$ we have that $f(x)^n = g(f(x))$. Furthermore, $g'(x) = nx^{n-1}$, so applying the chain rule we have

$$\frac{\mathrm{d}}{\mathrm{d}x}f(x)^n = g'(f(x))f'(x) = nf(x)^{n-1}f'(x)$$

as required. ■

Example 3.31

Denote by $f^{\circ n}(x)$ the n-fold composition of f. For example, $f^{\circ 3}(x) = f(f(f(x)))$. Assuming that $f(1) = 1$ and $f'(1) = 2$ find the derivative of $f^{\circ n}$ evaluated at $x = 1$.

Solution. This is just a repeated exercise of the chain rule requiring a small bit of trickery. First, we notice that since $f(1) = 1$ then no matter how many times we compose by f, we will always get 1. More explicitly, notice that $f(f(1)) = f(1) = 1$, and in general $f^{\circ n}(1) = 1$. For the sake of intuition, let us try this when $n = 3$. Notice in this case that

$$(f^{\circ 3})'(x) = f'(f(f(x))) \cdot f'(f(x)) \cdot f'(x)$$

so that

$$(f^{\circ 3})'(1) = f'(f(f(1))) \cdot f'(f(1) \cdot f'(1) = f'(1) \cdot f'(1) \cdot f'(1) = \left[f'(1)\right]^3 = 8.$$

In fact, precisely the same procedure will work for general n, and we get

$$(f^{\circ n})'(x) = f'(f^{\circ(n-1)}(x))f'(f^{\circ(n-2)}(x)) \cdots f'(f(x))f'(x) = \prod_{i=0}^{n-1} f'(f^{\circ i}(x))$$

where $f^{\circ 0} = x$. Hence we get

$$(f^{\circ n})'(1) = \underbrace{f'(1) \cdots f'(1)}_{n-\text{times}} = 2^n.$$ ■

Just as we were able to use the product rule to extend the power rule from $n \in \mathbb{N}$ to $n \in \mathbb{Z}$, we can use the product rule to tell us something about inverse functions (we will discuss this more in a subsequent section).

Theorem 3.32: Inverse Function Theorem

Let f be continuously differentiable at the point c, and assume further that $f'(c) \neq 0$. Then there is a neighbourhood of c on which f is injective with continuously differentiable inverse, and the derivative of the inverse is given as

$$\left(f^{-1}\right)'(x) = \frac{1}{f'(f^{-1}(x))}. \tag{3.11}$$

The proof of this theorem is not terribly difficult (until you generalize it to higher dimensions). The key is to show that since $f'(c) \neq 0$, one can find an inverse function f^{-1} which is actually continuous is a neighbourhood of c. The fact that the function is differentiable is not too much more work, so we leave it as an exercise for the student[14]. Instead, let's look at how (3.11) is derived.

[14] This question looks daunting and is requires some concentrated thinking, but is not impossibly hard. Try drawing a prototypical picture with ϵ and δ bands and think about what is happening for a long time, tilting your head to the side as necessary.

Let us begin by assuming that we have been given f^{-1} and told that it is differentiable. By definition of how the inverse function behaves, we know that $f(f^{-1}(x)) = x$ for all x in the range of f. Differentiating both sides (applying the chain rule to the composition) we then get

$$1 = \frac{\mathrm{d}}{\mathrm{d}x} f(f^{-1}(x)) = f'(f^{-1}(x))(f^{-1})'(x).$$

We can solve for $(f^{-1})'(x)$ by re-arranging, to get $(f^{-1})'(x) = \left[f(f^{-1}(x))\right]^{-1}$ as required. Again, this is an instance in which the derivation of the formula is so easy that it would be wasteful to memorize (3.11). Instead, we emphasize that the student should focus on the derivation itself.

Exercise: Prove Theorem 3.32.

3.5 The Inverse Trigonometric Functions

3.5.1 The Inverse of What?

The word "inverse" has many different meanings depending on the context in which it is used. For example, what if we were to ask the student to find the inverse of the number 2? What does this mean? To what are we taking the inverse? To properly understand this, we need to understand the following: Given a binary operator (an operator which takes in two things and produces a single thing in return, such as addition and multiplication), we say that a number id is the *identity* of that operator if operating against it does nothing to the input. For example, in the case of addition, the operator will satisfy $x + \mathrm{id}_+ = x$ for all possible x; for example,

$$2 + \mathrm{id}_+ = 2, \qquad -5 + \mathrm{id}_+ = -5.$$

Our experience tells us that $\mathrm{id}_+ = 0$. Similarly, for multiplication the identity $\mathrm{id}_\times$ will satisfy $x \times \mathrm{id}_\times = x$ for all x; for example,

$$3 \times \mathrm{id}_\times = 3, \qquad \pi \times \mathrm{id}_\times = \pi.$$

Again our experience tells us that $\mathrm{id}_\times = 1$. We thus say that 0 is the additive identity and 1 is the multiplicative identity. We say that the *inverse of* x is an element which, when paired against x, gives the identity. Hence the additive inverse of 2 is the number y such that $2 + y = \mathrm{id}_+ = 0$, or rather -2. In general, the additive inverse of n is $-n$, and this always exists! For multiplication, it is not too hard to convince ourselves that the multiplicative inverse of x is $\frac{1}{x}$; for example, $2 \times \frac{1}{2} = 1 = \mathrm{id}_\times$. Notice that there is no multiplicative inverse for the number 0, so in this case the inverse does not always exist.

Function composition $f \circ g$ is another example of a binary operator: we put two functions f, g in, and we get one function $f \circ g$ out. What is the identity for this operation? Well, we would like a function $\mathrm{id}_\circ$ such that

$$\begin{aligned} f(\mathrm{id}_\circ(x)) &= f(x) \\ &= \mathrm{id}_\circ(f(x)). \end{aligned}$$

If we think about this for a moment, the identity function is the function $\text{id}_\circ(x) = x$, the function which does nothing to the argument! Now what is the inverse of a function? The inverse of a function $f(x)$ is a function $f^{-1}(x)$ such that $f \circ f^{-1} = f^{-1} \circ f = \text{id}_\circ$.

The underlined word above is important: We need our inverse map to also be a function so that we can apply our usual ideas to it. In particular, let's think about this in the context of the vertical line test. Recall from Section 1 that a function f is said to be injective if

$$f(x) = f(y) \quad \Rightarrow \quad x = y;$$

that is, the function is one-to-one, or satisfies the horizontal line test. Since geometrically, the inverse of f arises by reflecting the graph of f about the line $y = x$, the horizontal line test will afterwards become a vertical line test. We conclude that for the inverse map f^{-1} to be a function, f must be injective.

3.5.2 The Inverse Trigonometric Functions

As suggested by this section name, our goal is to differentiate the inverse trigonometric functions. But wait a second! We just said that for the inverse of a function to even exist, our original function had to be injective. When we think about the trigonometric functions like $\sin(x)$ and $\cos(x)$, they are certainly not injective, so what does it even meant to talk about their inverse?

The solution is that we technically cannot define their inverse. What we do instead is create new functions which are restrictions of the old functions to domains where they are injective. Moreover, we need to ensure that we capture as much information as possible (namely, that our range is $[-1, 1]$), suggesting that we take in interval of length π. In the case of $\sin(x)$, there are many possible intervals that we could choose. For the sake of simplicity, we choose the maximal injective interval at the origin, namely $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. In the case of $\cos(x)$ we take $[0, \pi]$. On these intervals, $\sin(x)$ has an inverse $\arcsin(x)$ and $\cos(x)$ has an inverse $\arccos(x)$. We can perform a similar analysis of all the remaining trigonometric functions and deduce the following table:

Name	Definition	Notation	Maps		
arcsine	$\theta = \arcsin(x)$	$\sin(\theta) = x$	$[-1, 1]$	$\longrightarrow$	$\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$
arccosine	$\theta = \arccos(x)$	$\cos(\theta) = x$	$[-1, 1]$	$\longrightarrow$	$[0, \pi]$
arctangent	$\theta = \arctan(x)$	$\tan(\theta) = x$	$\mathbb{R}$	$\longrightarrow$	$\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$
arccosecant	$\theta = \text{arccsc}(x)$	$\csc(\theta) = x$	$\mathbb{R} \setminus (-1, 1)$	$\longrightarrow$	$\left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \setminus \{0\}$
arcsecant	$\theta = \text{arcsec}(x)$	$\sec(\theta) = x$	$\mathbb{R} \setminus (-1, 1)$	$\longrightarrow$	$[0, \pi] \setminus \left\{\frac{\pi}{2}\right\}$
arccotangent	$\theta = \text{arccot}(x)$	$\cot(\theta) = x$	$\mathbb{R}$	$\longrightarrow$	$(0, \pi)$

3.5.3 The Derivatives

We can use the Inverse Function Theorem (Theorem 3.32) to compute the derivatives of the inverse trigonometric functions quite easily. Let us proceed with an example on the sine-function.

Theorem 3.33

The function $\arcsin : [-1,1] \to \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ is differentiable on $(-1,1)$ and its derivative is given by

$$\frac{\mathrm{d}}{\mathrm{d}x} \arcsin(x) = \frac{1}{\sqrt{1-x^2}}. \tag{3.12}$$

Proof. Set $f(x) = \sin(x)$ so that $f^{-1}(x) = \arcsin(x)$. Note that $\frac{\mathrm{d}}{\mathrm{d}x}\sin(x) = \cos(x)$ is never zero on the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$ and hence by the Inverse Function Theorem, we know that $f^{-1}(x) = \arcsin(x)$ is differentiable on $(-1,1)$. Now we know that on $(-1,1)$ we have $\sin(\arcsin(x)) = x$ and hence differentiating both sides we have[15]

$$1 = \cos(\arcsin(x)) \left[\frac{\mathrm{d}}{\mathrm{d}x} \arcsin(x)\right] \qquad \Rightarrow \qquad \frac{\mathrm{d}}{\mathrm{d}x} \arcsin(x) = \frac{1}{\cos(\arcsin(x))}.$$

This is the derivative, but does not agree with the statement given in (3.12). To deduce $\cos(\arcsin(x))$, set $y = \arcsin(x)$ so that $\sin(y) = x$ (see Figure 11). From the figure, we immediately read off that $\cos(y) = \cos(\arcsin(x)) = \sqrt{1-x^2}$ which gives us

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

as required. □

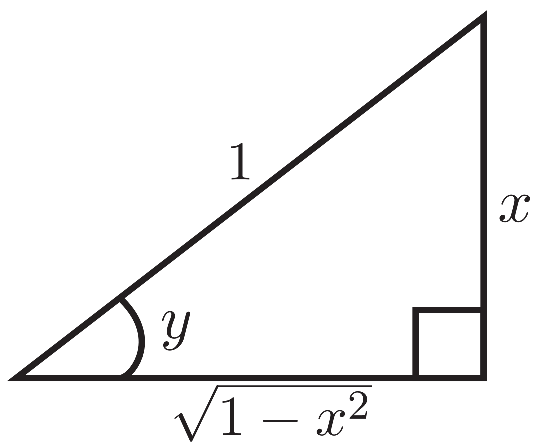


Figure 11: The right-triangle corresponding to setting $y = \arcsin(x)$.

The student is encouraged to attempt deriving the remaining functions, which I will provide below in an easily readable table[16]:

[15] This is of course equivalent to using (3.11).

[16] Be careful with $\operatorname{arccsc}(x)$, since it can be quite tricky depending on how you do it.

$$\frac{\mathrm{d}}{\mathrm{d}x}\arcsin(x) = \frac{1}{\sqrt{1-x^2}} \qquad \frac{\mathrm{d}}{\mathrm{d}x}\operatorname{arccsc}(x) = -\frac{1}{|x|\sqrt{x^2-1}}$$
$$\frac{\mathrm{d}}{\mathrm{d}x}\arccos(x) = -\frac{1}{\sqrt{1-x^2}} \qquad \frac{\mathrm{d}}{\mathrm{d}x}\operatorname{arcsec}(x) = \frac{1}{|x|\sqrt{x^2-1}}$$
$$\frac{\mathrm{d}}{\mathrm{d}x}\arctan(x) = \frac{1}{1+x^2} \qquad \frac{\mathrm{d}}{\mathrm{d}x}\operatorname{arccot}(x) = -\frac{1}{1+x^2}.$$

3.6 Exponentials and Logarithms

Exponentials and logarithms are two exceptionally important types of functions that, along with inverse trigonometric functions, are referred to as transcendental functions. Unfortunately, they are tricky to define, despite the fact that the average student has probably seen them before. Indeed, while there are several possible ways to define them, to be completely rigorous we must either show the student the Monotone Convergence Theorem, start with Taylor Series, or develop the theory of integration. As we have not yet covered any of these topics, we are certainly in a sticky situation!! However, the ubiquity of these functions throughout much of the natural sciences necessitates that we introduce them earlier rather than later.

3.6.1 The Functions

The Exponential: For now, let us just naively assume that if $a > 0$, the object a^x makes sense, for all real numbers x. We then have the following properties:

1. $a^0 = 1$
2. $a^1 = a$
3. $a^x a^y = a^{x+y}$
4. $(a^x)^y = a^{xy}$
5. $(ab)^x = a^x b^x$.

From this list of properties we can derive other properties. For example, combining the fact that $a^x a^y = a^{x+y}$ we must then have $a^x a^{-x} = a^0 = 1$ which implies that $a^x = a^{-x}$. Furthermore, $1^x = 1$ for all x and $a^0 = 1$ for all $a > 0$. Notice that for every $a \neq 1$, a^x has domain $\mathbb{R}$ and range $(0, \infty)$.

Logarithms: One can show that the non-constant exponential functions are monotonic; that is, they are either increasing or decreasing. Since such functions are automatically injective, we can be guaranteed that the exponential function has an inverse, if we choose the inverse function to have appropriate domain. Hence if $a > 0$, we define the map $\log_a : (0, \infty) \to \mathbb{R}$ as $\log_a(x) = y$ if $a^y = x$; that is, the logarithm is the inverse function to a^x:

$$a^{\log_a(x)} = x, \qquad \log_a(a^x) = x.$$

A few easy properties of $\log_a$ that come from the properties of the exponential are as follows:

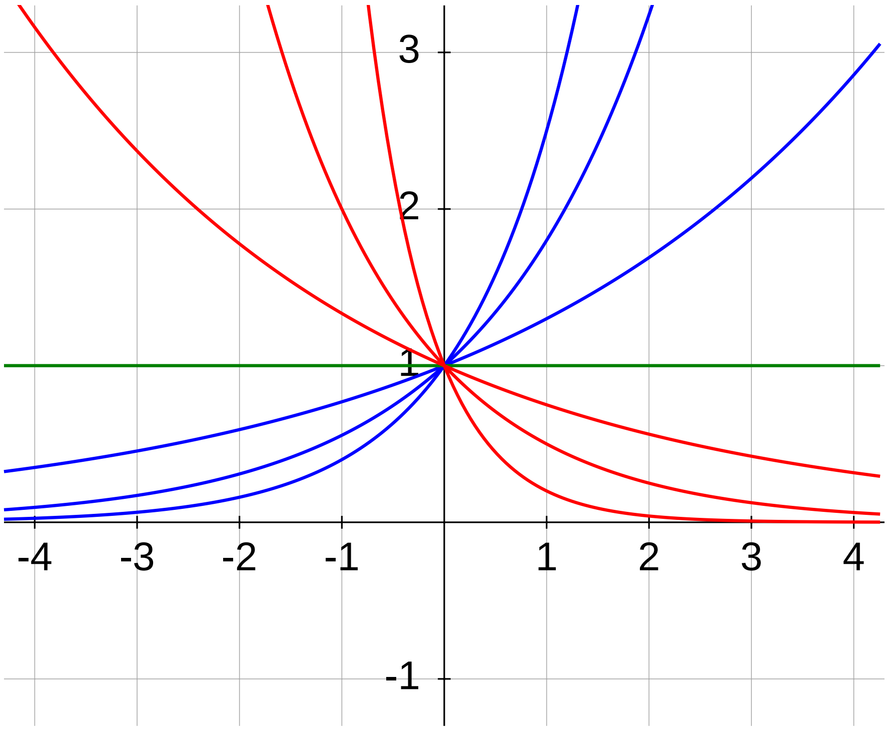


Figure 12: A graph of some of the exponential functions. Functions in blue are a^x for $a > 1$, while those in red are for $0 < a < 1$. The green line corresponds to $f(x) = 1^x \equiv 1$.

Proposition 3.34

For any $a, b, x, y > 0$ the logarithm has the following properties

1. $\log_a(x^d) = d\log_a(x)$
2. $\log_a(xy) = \log_a(x) + \log_a(y)$
3. $\log_a(x/y) = \log_a(x) - \log_a(y)$
4. $\log_x(y) = \frac{\log_a(y)}{\log_a(x)}$

3.6.2 Their Derivatives

Using the properties of the exponential function as well as the definition of the limit, we can start to get an idea about the derivative of the exponential. Indeed,

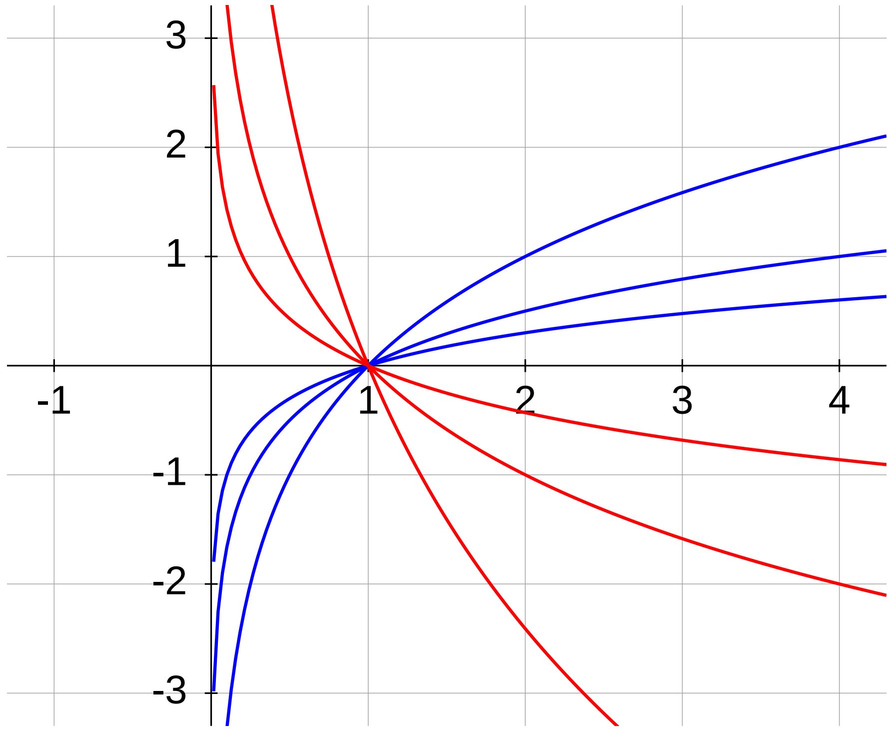


Figure 13: A graph of some of the logarithmic functions. Functions in blue are $\log_a x$ for $a > 1$, while those in red are for $0 < a < 1$.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x} a^x &= \lim_{h\to 0} \frac{a^{x+h} - a^x}{h} \\ &= \lim_{h\to 0} \frac{a^x a^h - a^x}{h} \\ &= a^x \lim_{h\to 0} \frac{a^h - 1}{h} \\ &= C_a a^x.\end{aligned}$$

where $C_a = \lim_{h\to 0} \frac{a^h-1}{h}$. We shall blindly assume that this constant exists, and we shall define *Euler's number* e to be the (blindly assumed) unique number such that $C_e = 1$. Hence by definition,

$$\frac{\mathrm{d}}{\mathrm{d}x} e^x = e^x.$$

For this number e, we shall denote the inverse of e^x as simply[17] $\log(x)$.

To differentiate $\log(x)$, we recall that $e^{\log(x)} = x$, so differentiating both sides we get

$$1 = e^{\log(x)} \left[\frac{\mathrm{d}}{\mathrm{d}x} \log(x) \right]$$

[17]Some people in the sciences use the notation $\ln(x)$ for the inverse of e^x, but this is not common throughout mathematics and does not generalize well. It should be taken that if no base to the logarithm is supplied, it is always base e.

which is easily solved for to get

$$\frac{\mathrm{d}}{\mathrm{d}x}\log(x) = \frac{1}{x}.$$

We can use our new found knowledge of the derivatives of e^x and $\log(x)$ to determine the remaining C_a. Indeed, notice that for any $a > 0$ we can write

$$a^x = e^{\log(a^x)} = e^{x\log(a)}.$$

Differentiating we get

$$\frac{\mathrm{d}}{\mathrm{d}x}a^x = \frac{\mathrm{d}}{\mathrm{d}x}e^{x\log(a)} = \log(a)e^{x\log(a)} = \log(a)a^x.$$

To determine the remaining derivatives for the other bases of the logarithm, convince yourself that we can always write $\log_a(x) = \log(x)/\log(a)$, in which case we simply get

$$\frac{\mathrm{d}}{\mathrm{d}x}\log_a(x) = \frac{\mathrm{d}}{\mathrm{d}x}\frac{\log(x)}{\log(a)} = \frac{1}{x\log(a)}.$$

Corollary 3.35: Generalized Power Rule

For all $n \in \mathbb{R}$ we have $\frac{\mathrm{d}}{\mathrm{d}x}x^n = nx^{n-1}$

Proof. We can write $x^n = e^{n\log x}$, so that differentiating yields

$$\frac{\mathrm{d}}{\mathrm{d}x}x^n = \frac{\mathrm{d}}{\mathrm{d}x}e^{n\log x} = \frac{n}{x}e^{n\log x} = \frac{n}{x}x^n = nx^{n-1}. \qquad \square$$

3.6.3 Logarithmic Differentiation

One of the great things about logarithms is there is a sense in which they decrease the complexity of an operation. For example, we often think of addition as being easier than multiplication, and multiplication being easier than exponents:

$$\log(xy) = \log(x) + \log(y), \qquad \log(x^y) = y\log x.$$

At the cost of introducing a logarithm, we are thus able to convert product to sums, and powers to products! Since the logarithm is not very hard to differentiate, this does not seem like such a terrible cost.

This idea in general is known as *logarithmic differentiation*. Where it can be particularly useful is when we have a product/quotient of many objects which are individually simple to differentiate, but which will become complicated when nested with the product rule. For example, given a collection of functions $f_1(x), \ldots, f_n(x)$ and $g_1(x), \ldots, g_m(x)$, notice that we can write

$$\log\left[\frac{f_1(x)\cdots f_n(x)}{g_1(x)\cdots g_m(x)}\right] = \log f_1(x) + \cdots + \log f_n(x) - \log g_1(x) - \cdots - \log g_m(x).$$

Hence implicit differentiation yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\frac{f_1(x)\cdots f_n(x)}{g_1(x)\cdots g_m(x)} &= \frac{f_1(x)\cdots f_n(x)}{g_1(x)\cdots g_m(x)}\frac{\mathrm{d}}{\mathrm{d}x}\left[\log f_1(x) + \cdots + \log f_n(x) - \log g_1(x) - \cdots - \log g_m(x)\right] \\ &= \frac{f_1(x)\cdots f_n(x)}{g_1(x)\cdots g_m(x)}\left[\frac{f_1'(x)}{f_1(x)} + \cdots + \frac{f_n'(x)}{f_n(x)} - \frac{g_1'(x)}{g_1(x)} - \cdots - \frac{g_m'(x)}{g_m(x)}\right]\end{aligned}$$

Example 3.36

Compute the derivative of $f(x) = \dfrac{(x-1)^2(x^2+2)\sqrt{x}}{x^4+5}$.

Solution. This would be an absolute nightmare to compute using the quotient rule, so instead we use logarithm differentiation. Taking the logarithm of both sides yields:

$$\log f(x) = 2\log(x-1) + \log(x^2+2) + \frac{1}{2}\log(x) - \log(x^4+5).$$

Differentiating implicitly gives

$$\begin{aligned}f'(x) &= f(x)\frac{\mathrm{d}}{\mathrm{d}x}\left[2\log(x-1) + \log(x^2+2) + \frac{1}{2}\log(x) - \log(x^4+5)\right] \\ &= \frac{(x-1)^2(x^3+2)\sqrt{x}}{x^4+5}\left[\frac{2}{x-1} + \frac{2x}{x^2+2} + \frac{1}{2x} - \frac{4x^3}{x^4+5}\right].\end{aligned}$$

■

This takes care of converting products to sums, but now what about powers to products? Given two functions $f(x)$ and $g(x)$, let's try to differentiate $f(x)^{g(x)}$. The problem here is that neither the power rule, nor the rules for differentiating exponents can apply (in both of those cases, the function should only occur in the power or the base, but not both). To deal with this, we set $y = f(x)^{g(x)}$ so that $\log y = g(x)\log f(x)$. We can now differentiate implicitly:

$$\frac{1}{y}\frac{\mathrm{d}y}{\mathrm{d}x} = g'(x)\log f(x) + g(x)\frac{f'(x)}{f(x)}$$

which we may then solve for $\frac{\mathrm{d}y}{\mathrm{d}x}$ to get

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f(x)^{g(x)}\left[g'(x)\log f(x) + g(x)\frac{f'(x)}{f(x)}\right]. \tag{3.13}$$

Like many of the formulae that we've derived, Equation (3.13) is not worth remembering on its own. Rather, what is important is remembering how this derivation was performed so that it can be repeated when necessary.

Example 3.37

Compute the derivative of $x^{\sin(x)}$.

Solution. Setting $y = x^{\sin(x)}$ we have $\log y = \sin(x)\log x$. Differentiating implicitly we get

$$\frac{1}{y}\frac{\mathrm{d}y}{\mathrm{d}x} = \cos(x)\log(x) + \frac{\sin(x)}{x}$$

which we may solve for $\frac{\mathrm{d}y}{\mathrm{d}x}$ to get

$$\frac{\mathrm{d}y}{\mathrm{d}x} = x^{\sin(x)}\left[\cos(x)\log(x) + \frac{\sin(x)}{x}\right].$$

■

4 Applications Of Derivatives

We now have an extensive collection of tools at our disposal, but we did not create this toolbox simply to admire the lustre of the implements, but rather to use them to solve problems. Here we will begin to see how we can use the tools of calculus.

4.1 Implicit Differentiation

4.1.1 The Idea of Implicit Functions

The idea of implicit differentiation is that we may be given variables in which it is *implied* that those variables depend on other variables, though we may not be able to explicitly write that relationship down. For example, to this point we have typically seen examples where we might write $y = f(x)$, in which case it is clear that that changes in x affect changes in y, as prescribed by the function x. We were then able to determine the rate of change of y with respect to x by computing $\frac{\mathrm{d}y}{\mathrm{d}x}$.

However, we can sometimes write relationships down without being able to solve for one variable as a function of the other; for example

$$\begin{aligned} x^2 + y^2 &= 25 \\ e^x + x\cos(y) + y &= 1 \\ f(x,y) &= k \quad \text{for some constant } k \end{aligned}$$

We can convince ourselves that the variables x and y above depend on one another. For example, consider the equation of the circle $x^2 + y^2 = 25$. If we set $y = 5$ then x is forced to be 0, while if we were to set $x = 3$ then y would have to be one of $y = \pm 4$. However, there is no function which makes the relationship between x and y explicit, since as our above example indicates, a single x-value may correspond to two possible y-values, and hence the relationship is not one given by a function.[18]

As an alternative example, consider the volume V of a cylinder as a function of its radius r and height h:

$$V = \pi r^2 h.$$

This equation defines an explicit relationship between the three entities V, r, and h. However, what if our cylinder were made of metal, and we were told that as temperature T increase, both the radius and the height increase, while when temperatures decrease, the radius and the height decrease[19]? In that case, we are *implicitly* assuming that r and h are functions of temperature T. This means that V is also implicitly a function of temperature, and we have

$$V(T) = \pi r(T)^2 h(T). \tag{4.1}$$

[18] Some students might be distressed at the fact that I have written $x^2 + y^2 = 25$ as an implicit equation, since certainly we could solve to find

$$y = \sqrt{25 - x^2},$$

but I claim that this actually not an explicit representation of this function. The reason is that, for example, both $(0, 5)$ and $(0, -5)$ are solutions to this equation, but we are unable to recover $(0, -5)$ from the expression $y = \sqrt{25 - x^2}$.

[19] Metals typically expand with increased heat

This implicit understanding that all the variables now depend on temperature allows us to determine how the volume of our cylinder is changing with temperature. Certainly, one could imagine a scenario in which this could be important!

In both of the aforementioned cases, it seems as though we should still be able to discuss the rate of change of one variable with respect to another, even if we are unable to explicitly describe the relationship between the variables using a function. This leads us to a process known as implicit differentiation.

4.1.2 How Implicit Differentiation Works

Let us say that we know a variable y implicitly depends upon another variable x. The idea of implicit differentiation is to differentiate as though the exact nature of the relationship were known. The best way to understand this is to see an example.

Example 4.1

Consider the equation of the circle centered at the origin with radius 1, $x^2+y^2=1$. Determine the rate of change of y with respect to x.

Solution. Our goal is to compute $\frac{\mathrm{d}y}{\mathrm{d}x}$. We saw earlier that the equation of a circle is an implicit relationship as there is no function which describes how y changes with respect to x or vice versa. Nonetheless, we are going to differentiate the equation $x^2+y^2=1$, but we keep in mind always that we are assuming that y is a function of x. To make this more clear, let's actually write $y=f(x)$, so that our equation of the circle is

$$1 = x^2 + y^2 = x^2 + f(x)^2.$$

Now differentiating both sides, we have

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}x}\left(x^2 + f(x)^2\right) \\ &= 2x + 2f(x)f'(x). \end{aligned}$$

Remember that we are trying to solve for $\frac{\mathrm{d}y}{\mathrm{d}x}$, which under our choice of $y=f(x)$ is just $\frac{\mathrm{d}y}{\mathrm{d}x}=f'(x)$, hence

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f'(x) = -\frac{x}{y}. \tag{4.2}$$

Note: Many people do not like to use this $y=f(x)$ notation, and instead will just write down

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(x^2+y^2\right) = 2x + 2y\frac{\mathrm{d}y}{\mathrm{d}x}.$$

This is acceptable and if you are comfortable using it, then you should feel free to do that. However, writing this down often hides what is really happening, so we have used the $y=f(x)$ notation just to be clear.

On the other hand, let me write $\tilde{y} = \sqrt{1-x^2}$, where the fact that I have used the tilde is to indicate that $\tilde{y}$ is not actually the same thing as y. When we differentiate we get

$$\frac{\mathrm{d}\tilde{y}}{\mathrm{d}x} = \frac{-x}{\sqrt{1-x^2}} = -\frac{x}{\tilde{y}}. \tag{4.3}$$

The inquisitive student may realize looks very similar to (4.2), but I claim is not quite the same.

Indeed, let us try to find the slope of the tangent line to the circle at the point $x = \frac{1}{\sqrt{2}}$. Notice on the circle that there are two possible y values, corresponding to $y_+ = +\frac{1}{\sqrt{2}}$ and $y_- = -\frac{1}{\sqrt{2}}$. Using (4.2) we find that the slope of the tangent lines at $y_\pm$ are

$$\left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{\left(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right)} = -\frac{1/\sqrt{2}}{1/\sqrt{2}} = -1, \qquad \left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{\left(\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}}\right)} = -\frac{1/\sqrt{2}}{-1/\sqrt{2}} = 1$$

which is in fact what we would expect. On the other hand, using (4.3) we find that at $x = \frac{1}{\sqrt{2}}$ there is only one possible $\tilde{y}$ value, corresponding to $\tilde{y} = \frac{1}{\sqrt{2}}$ in which case the slope of the tangent line is the same as that found above, namely

$$\left.\frac{\mathrm{d}\tilde{y}}{\mathrm{d}x}\right|_{\left(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right)} = -\frac{1/\sqrt{2}}{1/\sqrt{2}} = -1.$$

We have in fact lost the other tangent line! This is because when we took the square root of $y^2 = 1 - x^2$ we needed to make a choice as to whether to take the positive or negative root. In doing so, we actually lost information. ■

Example 4.2

Consider the volume of a cylinder as a function of temperature, as given in (4.1). Determine the rate of change of V with respect to temperature, written in terms of how r and h vary with respect to temperature.

Solution. We already know that $V = \pi r^2 h$. Although (4.1) has the temperature dependence written in, we will ignore it for this exercise to show the student how the notation is typically conveyed. Differentiating, we get

$$\begin{aligned}\frac{\mathrm{d}V}{\mathrm{d}T} &= \frac{\mathrm{d}}{\mathrm{d}T}\left[\pi r^2 h\right] \\ &= \pi\left[2r\frac{\mathrm{d}r}{\mathrm{d}T}h + r^2\frac{\mathrm{d}h}{\mathrm{d}T}\right]\end{aligned}$$ ■

The power of implicit differentiation can be even greater though if one is given a transcendental equation such as $e^x + x\cos(y) + y = 1$ in cases where y has no change of being isolated. In such instances, one is indeed *forced* to use implicit differentiation.

Example 4.3

Compute the derivative $\frac{\mathrm{d}y}{\mathrm{d}x}$ of y in the equation $e^x + x\cos(y) + y = 1$.

Solution. Keeping in mind that y is a function of x, we apply $\frac{\mathrm{d}}{\mathrm{d}x}$ to both sides of our equation to find:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\left(e^x + x\cos(y) + y\right) &= \frac{\mathrm{d}}{\mathrm{d}x}1 \\ e^x + \cos(y) - x\sin(y)\frac{\mathrm{d}y}{\mathrm{d}x} + \frac{\mathrm{d}y}{\mathrm{d}x} &= 0 \\ \frac{e^x + \cos(y)}{x\sin(y) - 1} &= \frac{\mathrm{d}y}{\mathrm{d}x}.\end{aligned}$$

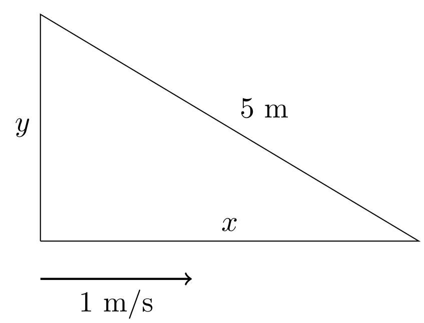
■

4.2 Related Rates

Much as we saw in our example relating the volume of a cylinder to its radius and height, related rates are way of using the relationship between objects to infer how they change with respect to an implicit variable. The trick when doing related rates questions is to find an equation that involves the quantities you are interested in, assume everything varies with time, then differentiate implicitly. The best way to get the hang of related rates questions is to do examples, so let's get started.

Example 4.4

Assume that a custodian is standing on a latter 5 metres tall and a pesky student pulls that ladder out from the wall at a rate of 1 metre/second. Determine how quickly the custodian is falling when he is 3 metres from the ground.



Solution. If one begins by drawing a simple diagram of the situation, we see that we get a triangle whose side lengths are $(x, y, 5)$, where x is the distance of the ladder from the wall, y is the height of the ladder above the ground, and 5 is length of the ladder (the hypotenuse). The obvious relationship between the variables is to use the Pythagorean theorem, $x^2 + y^2 = 25$. Since x and y are both changing as a function of time, we differentiate both implicitly to find

$$2x\frac{\mathrm{d}x}{\mathrm{d}t} + 2y\frac{\mathrm{d}y}{\mathrm{d}t} = 0.$$

Since we want to know how fast the custodian is falling we would like to evaluate $\frac{\mathrm{d}y}{\mathrm{d}t}$, which is solved as $\frac{\mathrm{d}y}{\mathrm{d}t} = -\frac{x}{y}\frac{\mathrm{d}x}{\mathrm{d}t}$. We already know that $\frac{\mathrm{d}x}{\mathrm{d}t} = 1$ (as this is the rate at which the latter is being pulled from the wall), and we are told to evaluate when $y = 3$. Hence we need only find the x-value which corresponds to a y-value of 3. The Pythagorean theorem again implies that if $y = 3$ then $x = \sqrt{5^2 - 3^2} = 4$, so that

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -\frac{4}{3}(1) = -\frac{4}{3}.$$

Thus the custodian is falling at 4/3 metres per second. ■

Example 4.5

A sphere is shrinking in such a way that its surface area changes at a constant rate of 1 unit^2 per time. Find the rate at which the diameter of the sphere is decreasing when the diameter is 5 units.

Solution. First, we look at the quantities involved. We are given information about the surface area A of the sphere, and its diameter δ. We know that the formula for the surface area of a sphere is given by $A = 4\pi r^2$ where r is the radius of the sphere. This is not quite what we want though, since we are interested in the diameter. Luckily, we also know that the radius is half the diameter, so that $r = \frac{1}{2}\delta$. Substituting this we find that

$$A = 4\pi\left(\frac{1}{2}\delta\right)^2 = \pi\delta^2.$$

Now it is clear that both the area and the diameter are changing with time, so when we differentiate we will have to do so implicitly. Indeed, one gets that

$$\frac{\mathrm{d}A}{\mathrm{d}t} = 2\pi\delta\frac{\mathrm{d}\delta}{\mathrm{d}t}.$$

We want to know $\frac{\mathrm{d}\delta}{\mathrm{d}t}$ as this represents the change in the diameter with respect to time, so we solve to find that

$$\frac{\mathrm{d}\delta}{\mathrm{d}t} = \frac{1}{2\pi\delta}\frac{\mathrm{d}A}{\mathrm{d}t}.$$

Since A is decreasing at a rate of 1 unit^2 per time, we set $\frac{\mathrm{d}A}{\mathrm{d}t} = -1$ (the negative sign indicates that we are decreasing). Thus we know every quantity on the right-hand-side, so we may substitute to find that

$$\frac{\mathrm{d}\delta}{\mathrm{d}t} = \frac{1}{2\pi \cdot 5}(-1) = -\frac{1}{10\pi}$$

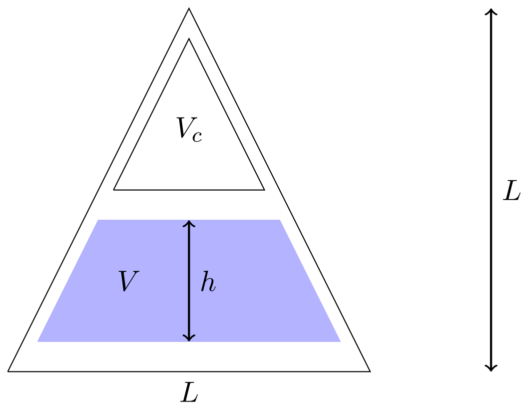
and so we are done. ■

Example 4.6

Consider a square pyramid which is being filled with some mysterious physically pertinent quantity. The base square has length and width both equal to $L > 0$ and the height of the apex of the pyramid is also a distance L. At what rate must we poor the mysterious quantity into the pyramid such that the rate of change of the height of the material does not depend on the size of the base square. Is this rate dependent on the height of the pyramid?

Solution. Let V be the volume of the liquid contained in the pyramid, which we know is an increasing function in time. We have the capability of adjusting the flow rate, which corresponds to the quantity $\frac{\mathrm{d}V}{\mathrm{d}t}$. We are interested in the determining how the height of the water h is changing as a function of time, so that we might choose an appropriate $\frac{\mathrm{d}V}{\mathrm{d}t}$ to ensure that $\frac{\mathrm{d}h}{\mathrm{d}t}$ has no dependence upon L.

If we project the pyramid into two dimensions, we have the following figure:



Where V_c is the volume not occupied by water. Notice that the height of the pyramid bounding V_c is $L-h$, so if its base has dimension ℓ, its volume is $V_c = \frac{1}{3}\ell^2(L-h)$. In order to relate this to the volume of the water, notice that the total volume of the pyramid is

$$V_{\text{tot}} = \frac{1}{3}Ah = \frac{1}{3}L^3 = V + V_c$$

so we can write $V = \frac{1}{3}L^3 - V_c$ (which implies that $\frac{\mathrm{d}V}{\mathrm{d}t} = -\frac{\mathrm{d}V_c}{\mathrm{d}t}$). Now the triangle which bounds the projected pyramid is similar to the triangle which bounds V_c, and this implies that the ratio of their side lengths will be equal. In particular, if we let ℓ be the length of the base of the triangle bounding V_c, then we must have

$$\frac{\ell}{L-h} = \frac{L}{L} = 1$$

so that $\ell = L - h$. Hence we have

$$V = \frac{1}{3}L^3 - V_c = \frac{1}{3}L^3 - \frac{1}{3}(L-h)^3.$$

Differentiating, we thus get

$$\frac{\mathrm{d}V}{\mathrm{d}t} = (L-h)^2\frac{\mathrm{d}h}{\mathrm{d}t}, \qquad \text{or equivalently} \qquad \frac{\mathrm{d}h}{\mathrm{d}t} = -\frac{1}{(L-h)^2}\frac{\mathrm{d}V}{\mathrm{d}t}.$$

Thus by setting $\frac{\mathrm{d}V}{\mathrm{d}t} = C(L-h)^2$ for some constant C, we can guarantee that $\frac{\mathrm{d}h}{\mathrm{d}t}$ does not depend on L, and changes at a constant rate C. ■

Example 4.7

A cat is sitting on a ledge a distance d from a door when suddenly a frantic math professor bursts through, chased by a mob of angry students. The cat, having watched many Olympic races, figures that the teacher is running at a rate of s metres per second. Assuming that the angry mob of students is unable to catch the spry professor, at what rate must the cat's head move in order to watch the chase once the professor has run r metres. Your answer should be expressed purely in terms of r, s and d.

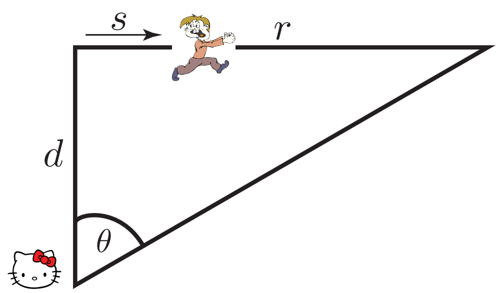


Figure 14: The professor running, while the cat patiently observes.

Solution. Let us begin by drawing an appropriate diagram, labeling the angle the cats head makes with the door θ and the distance from the professor to the door by y. Notice we will be interested in the case when $y = r$ (signifying that the professor has run a distance r) and when $\frac{dy}{dt} = s$ (since this is the speed of the professor). This is illustrated in Figure 14.

The triangle tells us that $\tan(\theta) = \frac{y}{d}$ and so implicit differentiation yields

$$\sec^2(\theta)\theta' = \frac{y'}{d}.$$

Since all units are consistent and the professor is running at s metres per second, we have that $y' = s$ giving

$$\theta' = \frac{s}{d\sec^2(\theta)}$$

and we need only remove the dependency on θ to be finished. The triangle illustrated in Figure 14 tells us that the hypotenuse is given by $\sqrt{r^2 + d^2}$ so that $\cos(\theta) = \frac{d}{\sqrt{r^2+d^2}}$. Since $\sec^2(\theta) = \frac{1}{\cos^2(\theta)} = \frac{r^2+d^2}{d^2}$ we conclude that the cat's head is moving at a rate of

$$\theta' = \frac{sd}{r^2 + d^2} \quad \text{radians per second.}$$ ■

4.3 The Mean Value Theorem

The Mean Value Theorem, being another of these named theorems, is an exceptionally important result in analysis. However, first year students often find it difficult to use. Unlike the other named theorems in this class, we will actually prove the Mean Value Theorem, so we present it below now (despite the fact that it will take us a while to prove it).

Theorem 4.8: Mean Value Theorem

If f is a function and $[a, b]$ is an interval such that f continuous on $[a, b]$ and differentiable on (a, b), then there exists $c \in (a, b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}. \tag{4.4}$$

Remarks:

1. There are two hypotheses; namely the continuity and differentiability of f. Many students like to ignore these hypotheses and instead focus on Equation (4.7). While the equation is critical, it is simply false if we ignore the hypotheses, so be mindful of them.
2. Like the Intermediate and Extreme Value Theorems, this theorem is not constructive; that is, we are guaranteed that such a c exists, but have no way of finding it.
3. The quantity $\frac{f(b)-f(a)}{b-a}$ is the slope of the secant line between the points $(a, f(a))$ and $(b, f(b))$ (see Figure 15).

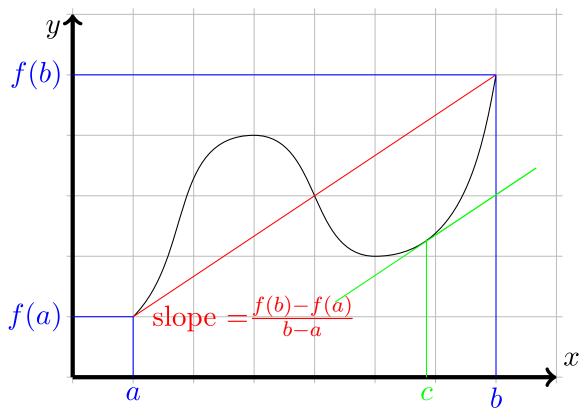


Figure 15: The Mean Value Theorem says that there is a point on this graph such that the tangent line has the same slope as the secant between $(a, f(a))$ and $(b, f(b))$.

4.3.1 Extrema and Rolle's Theorem

To prove the Mean Value Theorem, we start with a seemingly more simple (yet actually equivalent) result, known as Rolle's theorem. However, in order to prove Rolle's theorem, we will need to discuss something of maxima and minima.

Definition 4.9

If f is a (not necessarily continuous) function, we say that a point c in the domain of f is a *local minimum* of f is there is a neighbourhood $I = (c - \gamma, c + \gamma)$ such that, for all $x \in I$ we have $f(x) \geq f(c)$. Similarly, c is a *local maximum* if for all $x \in I$ we have $f(x) \leq f(c)$.

The point, which we shall expound on in more detail later, is that local minima and maxima which occur on the interior necessarily correspond to points where the derivative is zero. Indeed, if we look at Figure 15 we see that the function has a local minimum and local maximum which occur on the interior of the domain of f. At these points, it certainly appears as though the tangent line is horizontal, and this is indeed the case.

Proposition 4.10

If f is differentiable in a neighbourhood of a point c in the interior of its domain, and c is either a local maximum or local minimum, then necessarily $f'(c) = 0$.

Proof. We shall do the proof for the case when c corresponds to a local maximum and leave the proof of the other case to the student. Since c is a local maximum, we know there is some neighbourhood $I \subseteq D$ of c such that for all $x \in I, f(x) \leq f(c)$.

Since c corresponds to a maximum of f, for all $h > 0$ sufficiently small so that $c + h \in I$, we have that $f(c+h) \leq f(c)$. Hence $f(c+h) - f(c) \leq 0$, and since h is positive, the difference quotient satisfies $\frac{f(c+h)-f(c)}{h} \leq 0$. In the limit as $h \to 0^+$ we thus have

$$\lim_{h \to 0^+} \frac{f(c+h) - f(c)}{h} \leq 0 \tag{4.5}$$

Similarly, if $h < 0$ we still have $f(c + h) - f(c) \leq 0$ but now with a negative denominator our difference quotient is non-negative and

$$\lim_{h \to 0^-} \frac{f(c+h) - f(c)}{h} \geq 0. \tag{4.6}$$

Combining (4.5) and (4.6) and using the fact that f is differentiable at c, we have

$$0 \leq \lim_{h \to 0^-} \frac{f(c+h) - f(c)}{h} = f'(c) = \lim_{h \to 0^+} \frac{f(c+h) - f(c)}{h} \leq 0$$

which implies that $f'(c) = 0$. □

Theorem 4.11: Rolle's Theorem

Let f be continuous on $[a, b]$ and differentiable on (a, b). If $f(a) = f(b) = 0$ then there exists a point $c \in (a, b)$ such that $f'(c) = 0$.

The intuition for Rolle's theorem is quite clear. If our function is continuous (and differentiable), then in order for the function to start and 0 and end at 0, at some point it must have been increasing and then decreasing (or vice-versa). This corresponds to a derivative which positive and then negative, which we suspect will imply that at some point the derivative was in fact zero.

Proof. If f is constant (and hence the zero function) then we are done, so assume that f is a non-constant function. Since f is continuous on $[a, b]$, by the Extreme Value Theorem we know that it achieves its maximum and minimum on $[a, b]$. Since the function is non-constant, at least one of the maximum or minimum is non-zero. Without loss of generality, assume that f takes on a maximum at $c \in (a,b)$ and that $f(c) > 0$. Since $f(c) > 0$ then c cannot be an endpoint (since $f(a) = f(b) = 0$) and hence must be an interior point. By Proposition 4.10, it then follows that $f'(c) = 0$ as required. □

Rolle's Theorem on its own can be applied to give some fairly useful results. One example is that we can use it to say something about the number of roots of a function.

Example 4.12

Let $f(x)$ be a differentiable function, and assume that $f'(x) \neq 0$ for all x. Show that f has at most one root.

Solution. As we have discussed before, the fact that $f(x)$ is continuously differentiable means that $f'(x)$ is continuous. Since $f'(x) \neq 0$ for all $x \in (a,b)$ it must follow that $f'(x)$ is either strictly positive or strictly negative (otherwise there are points c_1, c_2 such that $f(c_1) > 0$ and $f(c_2) < 0$ and the Intermediate Value Theorem would then imply a point between c_1 and c_2 such that $f'(c) = 0$).

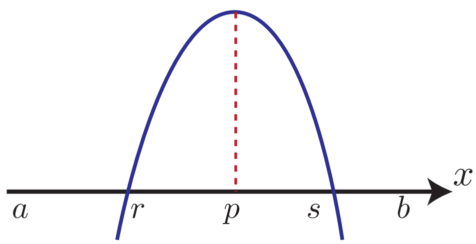


Figure 16: Plot of a decreasing (left) and increasing (right) function. No matter how we choose to draw such a function, it may only have at most one root.

Think about any nice (read: differentiable) graph with two roots: it typically looks something along the lines of Figure 16 not occur? The reason is that the derivative must switch signs at some point. In particular, there is a point $p \in (a,b)$ where $f'(p) = 0$. This is the contradiction at which we would like to arrive, and Rolle's Theorem will give it to us.

For the sake of contradiction, assume that the function has more than one root and choose any two roots, say $r, s \in (a,b)$, so that $f(r) = f(s) = 0$. Since our function is continuously differentiable, it certainly satisfies the hypotheses of Rolle's theorem, so there exists $c \in (a,b)$ such that

$$f'(c) = 0.$$

This is a contradiction to the fact that $f'(x) \neq 0$ for all $x \in (a,b)$. Hence the function cannot have two roots, so we conclude it has at most one root. ■

Example 4.13

Show that the function $f(x) = x^3 - 6x^2 - 12x + e^x - 4$ has exactly one root.

Solution. We will first use the Intermediate Value Theorem to show that there is a root, then we will argue that this can be the only root.

Indeed, our function is certainly continuous, and we notice that $\lim_{x\to -infty} f(x) = -\infty$ and $\lim_{x\to\infty} f(x) = \infty$. This certainly suggests there is a root. If we want to be more precisely, one can show that $f(-10) < 0$ and $f(10) > 0$ without too much trouble. Hence by the Intermediate Value Theorem, there is a root in $[-10, 10]$.

Now we claim that this is the only root. Since f is differentiable everywhere, by Example 4.12 it suffices to show that $f'(x) \neq 0$ for all x. Differentiating we get

$$f'(x) = 3x^2 - 12x - 12 + e^x = 3(x-2)^2 + e^x.$$

The summands here satisfy $3(x-2)^2 \geq 0$ and $e^x > 0$, which means that between them we have $f'(x) > 0$. We conclude that the function has at most one root. Combining our data, we find that f has exactly one root, as required. ■

4.3.2 The Theorem Proper

We implied earlier that the Mean Value Theorem and Rolle's Theorem are in fact equivalent. It's not too hard to convince ourselves that Rolle's Theorem is actually just a special case of the Mean Value Theorem, but the fact that we use Rolle's theorem to prove the Mean Value Theorem is why they are equivalent.

The idea of why Rolle's theorem is so useful is that the Mean Value Theorem effectively reduces to 'turning your head to the side' and applying Rolle's Theorem.

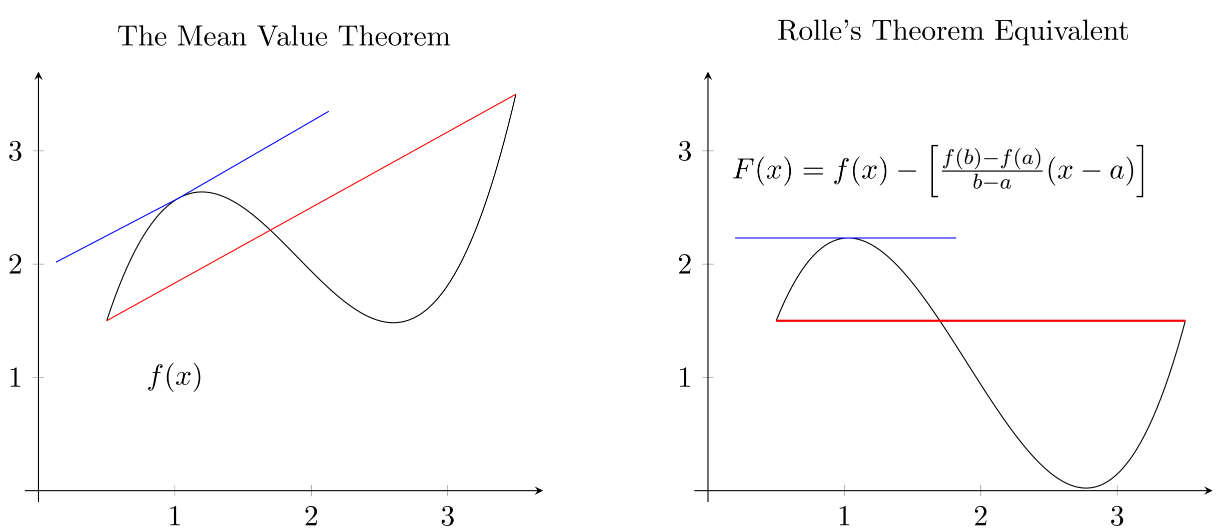


Figure 17: We can turn the statement of the Mean Value Theorem (left) into an equivalent problem which can be solved using Rolle's Theorem (right).

Theorem 4.14: The Mean Value Theorem

Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Then there exist a $c \in (a, b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}. \tag{4.7}$$

Proof. As mentioned prior to the theorem statement, the trick is to define a new function by pivoting $f(x)$ about the point $f(a)$ until $f(a) = f(b)$. At this point we will be able to use Rolle's theorem, which will complete the result.

Define a function $F : [a, b] \to \mathbb{R}$ by

$$F(x) = f(x) - \frac{f(b) - f(a)}{b - a}(x - a)$$

and notice that F is continuous on $[a, b]$ and differentiable on (a, b). Furthermore,

$$F(a) = f(a) - \frac{f(b) - f(a)}{b - a}(a - a) = f(a), \quad F(b) = f(b) - \frac{f(b) - f(a)}{b - a}(b - a) = f(a)$$

so that $F(a) = F(b)$. Applying Rolle's Theorem to the function F, we know there exists some $c \in (a, b)$ such that $F'(c) = 0$. This in turn implies that

$$0 = F'(c) = f'(c) - \frac{f(b) - f(a)}{b - a}, \quad \Rightarrow \quad f'(c) = \frac{f(b) - f(a)}{b - a}$$

which is precisely what we wanted to show. □

There are many "intuitive" results that can now be proven using the Mean Value Theorem, such as the following corollary:

Corollary 4.15

If $f : [a, b] \to \mathbb{R}$ is differentiable and $f'(x) = 0$ for all $x \in [a, b]$, then f is a constant function on $[a, b]$.

Proof. Since we want to show that f is constant, it is sufficient to show that $f(x) = f(a)$ for all x (since $f(a)$ is a constant). Let $p \in [a, b]$ be some arbitrary point and consider the interval $[a, p] \subseteq [a, b]$. Since f is differentiable on the larger of these two, its restriction is continuous on $[a, p]$ and differentiable on (a, p).

Recalling that $f'(x) = 0$ for all x, we can now apply the Mean Value Theorem to find some $c \in (a, p)$ such that

$$0 = f'(c) = \frac{f(p) - f(a)}{p - a}.$$

This in turn implies that $f(p) - f(a) = 0$ so that $f(p) = f(a)$. Since p was arbitrary, we must have that $f(x) = f(a)$ for all $x \in [a, b]$, and so f is the constant function. □

4.4 Maxima and Minima of Functions

We introduced the idea of maxima and minima in Section 4.3 as it was necessary to prove the Mean Value Theorem. Here now we develop tools from the Mean Value Theorem for determining where maxima and minima occur. Intuitively, we saw that a function is a maximum if our function is increasing to the left and decreasing to the right (with an analogous notion for minima). In order to make this idea more formal, we introduce the notion of increasing/decreasing functions.

Definition 4.16

A function $f : [a, b] \to \mathbb{R}$ is said to be

1. *increasing on* $[a, b]$ if whenever $x_1, x_2 \in [a, b]$ satisfy $x_1 < x_2$, then $f(x_1) \leq f(x_2)$. We say that f is *strictly increasing on* $[a, b]$ if $x_1 < x_2$ implies that $f(x_1) < f(x_2)$.

2. *decreasing on* $[a, b]$ if whenever $x_1, x_2 \in [a, b]$ satisfy $x_1 < x_2$, then $f(x_1) \geq f(x_2)$. We say that f is *strictly decreasing on* $[a, b]$ if $x_1 < x_2$ implies that $f(x_1) > f(x_2)$.

The Mean Value Theorem gives us our first tool for determining the relationship between increasing/decreasing and the derivative of a function:

Proposition 4.17

Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$, and differentiable on (a, b).

1. If $f'(x) < 0$ for all $x \in [a, b]$, then f is a decreasing function on $[a, b]$,

2. If $f'(x) > 0$ for all $x \in [a, b]$, then f is an increasing function on $[a, b]$.

Proof. We first consider the case when $f'(x) < 0$ for all $x \in [a, b]$. Let $x_1 < x_2$ be two points in $[a, b]$, and consider the function f restricted to the interval $[x_1, x_2] \subseteq [a, b]$. Since f is differentiable on the larger of these two, it is certainly continuous on $[x_1, x_2]$ and differentiable on (x_1, x_2). Applying the Mean Value Theorem, we find a $c \in (x_1, x_2)$ such that

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

But $x_2 - x_1 > 0$ and $f'(c) < 0$ implies that $f(x_2) - f(x_1) < 0$, or rather, $f(x_1) > f(x_2)$. Since x_1 and x_2 were chosen arbitrarily, it then follows that whenever $x_1 < x_2$ that $f(x_1) > f(x_2)$, which is precisely the definition of decreasing.

For the case when $f'(x) > 0$, define a new function $g(x) = -f(x)$ and notice that $g'(x) = -f'(x) < 0$. By our previous case, we know that $g'(x)$ is decreasing on $[a, b]$; that is, if $x_1 < x_2$ then $g(x_1) > g(x_2)$. This in turn implies that

$$-f(x_1) > -f(x_2) \quad \Rightarrow \quad f(x_1) < f(x_2)$$

showing that f is increasing as required. □

Example 4.18

Determine the intervals on which the function $f(x) = \cos(\pi x^2)$ is decreasing.

Solution. In light of the previous corollary, our goal is to find intervals on which the derivative is negative. Computing the derivative we get

$$f'(x) = -2\pi x \sin(\pi x^2).$$

If $x > 0$ then the sign of $f'(x)$ is entirely determined by $\sin(\pi x^2)$. Now we know that if $y > 0$ then $\sin(y) < 0$ for $y \in [(2k+1)\pi, (2k+2)\pi]$, $k \in \mathbb{N}$. Hence $\sin(\pi x^2) < 0$ whenever $x \in [\sqrt{2k+1}, \sqrt{2k+2}]$ or $x \in [-\sqrt{2k+2}, -\sqrt{2k+1}]$. But since we wanted $x > 0$, we can throw away the negative solutions.

For $x < 0$ we want $\sin(\pi x^2) > 0$ which happens to coincide with $x \in [-\sqrt{2k+2}, -\sqrt{2k+1}]$, and so our final solution is

$$\bigcup_{-k\in\mathbb{N}} [-\sqrt{-2k+2}, \sqrt{-2k+1}] \cup \bigcup_{k\in\mathbb{N}} [\sqrt{2k+1}, \sqrt{2k+2}]. \qquad \blacksquare$$

Via Proposition 4.10 we saw that extreme points which occured on the interior of the domain of a function necessarily had zero derivative. These values are in fact so special that they have a name:

Definition 4.19

If $f : D \to \mathbb{R}$ is a differentiable function and $f'(c) = 0$, we say that c is a *critical point* of f, and $f(c)$ is a *critical value* of f.[a] If f is not differentiable, we also say that c such that $f'(c)$ does not exist are critical points.

[a] Critical points and values are exceptionally interesting in mathematics, though our treatment of them in this course will be very limited. One example is Sard's theorem, which says that if f is differentiable, then there are "very few" critical values. Alternatively, there is an entire study called Morse Theory, which is interested in reconstructing spaces by using information about critical values of functions on that space.

Proposition 4.10 tolds us that if c is a max/min on the interior of the domain of a function, the c is a critical point. Does the converse need to be true? Unfortunately the answer is no: A simple example is the function $f(x) = x^3$. This function has a critical point at $x = 0$ since $f'(x) = 2x^2$. However, $f(x)$ is actually an increasing function, so the point 0 certainly cannot be a maximum or minimum!

Since it is only necessary for extrema to be critical points, we need to develop a strategy for determining whether a critical point actually corresponds to a max/min. Our first such test is primitive yet intuitive:

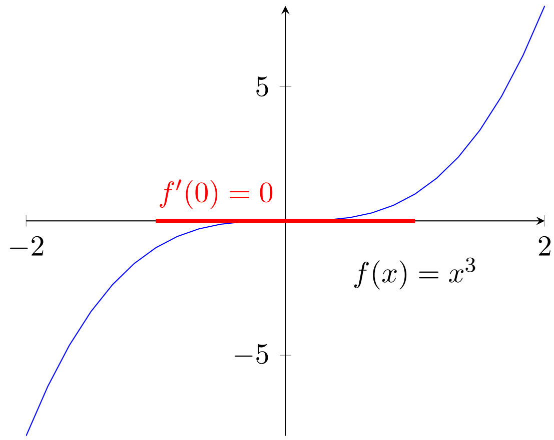


Figure 18: The function $f(x) = x^3$ has a critical point at $x = 0$ despite not having a max or min at $x = 0$.

Theorem 4.20: First Derivative Test

Let $f : [a, b] \to \mathbb{R}$ be differentiable with critical point $c \in (a, b)$.

1. If there exists $\delta > 0$ such that $f'(x) > 0$ on $(c - \delta, c)$ and $f'(x) < 0$ on $(c, c + \delta)$ then c is a maximum.
2. If there exists $\delta > 0$ such that $f'(x) < 0$ on $(c - \delta, c)$ and $f'(x) > 0$ on $(c, c + \delta)$ then c is a minimum.

Proof. Effectively, the theorem says that if our function is increasing to the left of the critical point and decreasing to the right, then we have a maximum, and vice versa for a minimum. Indeed, this is what we suspect.

We will do the proof for the maximum only, and leave the other as an exercise for the student. Hence assume that such a δ exists. By Proposition 4.17, since $f'(x) > 0$ for all $x \in (c - \delta, c)$ we know that f is an increasing function on this interval, and since f is continuous it then follows that $f(x) \leq f(c)$ for all $x \in (c - \delta, c)$. Similarly, since $f'(x) < 0$ for all $x \in (c, c + \delta)$ we have that $f(x) \leq f(c)$ for all $x \in (c, c + \delta)$, hence c is a maximum. □

Example 4.21

Find the (local) maxima and minima of the function $f(x) = 6x^4 - 3x^2 + 2$.

Solution. We begin by finding the critical points, so we differentiate to get $f'(x) = 24x^3 - 6x = 6x(4x^2 - 1) = 6x(2x - 1)(2x + 1)$. Setting $f'(x) = 0$ and solving for x, we get $x = 0, \pm\frac{1}{2}$. Now since $f'(x)$ splits into linear factors, and each linear factor can only switch sign once, we can determine

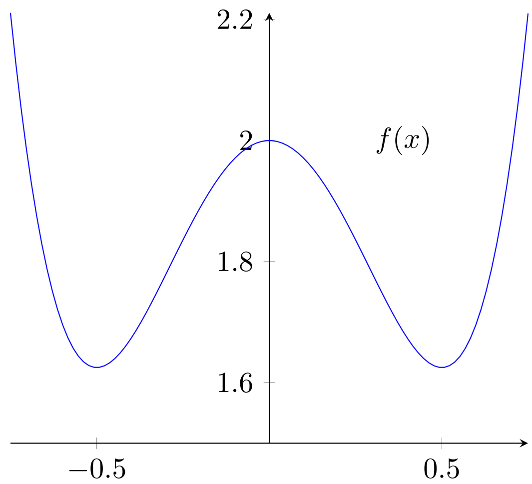


Figure 19: The function $f(x) = 6x^4 - 3x^2 + 2$ from Example 4.21
.

whether f is increasing or decreasing on each interval by creating the following chart:

	$x < -\frac{1}{2}$	$-\frac{1}{2} < x < 0$	$0 < x < \frac{1}{2}$	$x > \frac{1}{2}$
$2x+1$	$-$	$+$	$+$	$+$
x	$-$	$-$	$+$	$+$
$2x-1$	$-$	$-$	$-$	$+$
$f(x)$	$-$	$+$	$-$	$+$

From our chart we see that $-1/2$ is a min, 0 is a max, and $\frac{1}{2}$ is a min. This is confirmed by the graph of our function. ■

Ths chart from Example 4.21 was non-trivial to produce. It turns out there is a much more simple test:

Theorem 4.22: Second Derivative Test

Let $f(x)$ be a function which is twice continuously differentiable in a neighbourhood of a critical point c.

1. If $f''(c) < 0$ then c is a local maximum.
2. If $f''(c) > 0$ then c is a local minimum.
3. If $f''(c) = 0$ let k be the smallest positive integer such that $f^{(k)}(c) \neq 0$. If k is odd then c is neither a max nor a min. If k is even and $f^{(k)}(c) > 0$ then c is a local min, while if $f^{(k)}(c) < 0$ then c is a local max.

Notice that this takes care of the pesky $f(x) = x^3$ example, since $f^{(3)}(0) \neq 0$ the theorem implies that 0 is neither a max nor a min. As this appears on your next problem set, I will omit the proof.

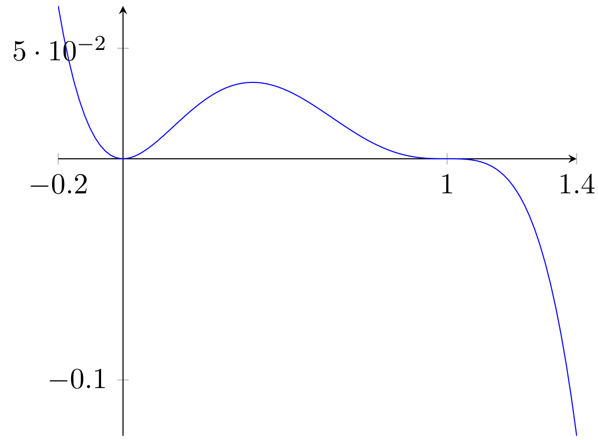


Figure 20: The function $f(x) = x^2(1-x)^3$ from Example 4.23

.

Example 4.23

Let $f(x) = x^2(1-x)^3$. Find the critical points of f and classify them.

Solution. To find the critical points, we begin by differentiating:

$$f'(x) = 2x(1-x)^3 - 3x^2(1-x)^2$$

Setting $f'(x) = 0$ and solving for x, we immediately see that $x = 0$ and $x = 1$ are solutions. Furthermore,

$$\begin{aligned} & 2x(1-x)^3 = 3x^2(1-x)^2 \\ \Leftrightarrow \quad & 2(1-x) = 3x \\ \Leftrightarrow \quad & x = \frac{2}{5} \end{aligned}$$

To classify the critical points, we take another derivative.

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

At $x = 0$ we get $f''(0) = 2 > 0$, implying that $x = 0$ is a min. At $x = \frac{2}{5}$ we get $f''(2/5) = -\frac{149}{125} < 0$ implying that $x = 2/5$ is a max. Finally, at $x = 1$ we get $f''(1) = 0$, so our test is indeterminate. Differentiating one more time gives

$$f'''(x) = -18(1-x)^2 + 36x(1-x) - 6x^2$$

and $f'''(1) = -6$ so $x = 1$ is neither a max nor a min. Figure 20 confirms our findings. ∎

We have seen how to determine local minima/maxima which occur on the interior of the domain of a function, but if our function is defined on a closed and bounded domain (like $[a,b]$) then it could have endpoints which are maxima and minima. The definition for an endpoint to be a local max or min is the same as that given in Definition 4.9, except that we only look at one-sided neighbourhoods.

Rather than being interested in local minima and maxima, we are more often interested in finding *global* extrema. The global maximum of a function f is a value c such that for all x in the domain of f, $f(x) \leq f(c)$: similarly for the global minimum. Of course, global extrema can occur at interior points, endpoints, or might not exist at all! However, since we know that extrema can only occur at critical points or at endpoints, all we need to do is check all possible (hopefully finite) points and determine which amongst them is the largest/smallest.

Example 4.24

Compute the global maximum and minimum of the function $f(x) = x - \arctan(x)$ on the interval $[-1, 1]$.

Solution. We recall that $\arctan(\pm 1) = \pm\frac{\pi}{4}$, so that

$$f(-1) = -1 - \frac{\pi}{4}, \qquad f(1) = 1 - \frac{\pi}{4}.$$

Next we compute the critical points. The derivative of $f(x)$ is

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

which has a single zero at $x = 0$. Plugging this back into our function we get $f(0) = 0 - \arctan(0) = 0$. A quick comparison reveals that

$$f(-1) < f(0) < f(1)$$

so that the global min is at -1 and the global max is at $+1$. ■

Example 4.25

Find the area of the largest rectangle which can be inscribed in a circle of radius R.

Solution. The trick to doing this question is to choose an intelligent parameterization of the rectangles. Assume that the circle is centred at the origin and notice that every inscribed rectangle can be uniquely prescribed by the angle one of its vertices makes with the positive x axis (see Figure 21). In this way one may construct the area function

$$A(x) = 4R^2 \sin\theta \cos\theta = 2R^2 \sin(2\theta)$$

for[20] $\theta \in [0, \pi/2]$. We differentiate to get $A'(\theta) = 4R^2 \cos(2\theta)$, which has a zero a $\theta = \pi/4$. It is not too hard to see that this is a max, since

$$A''(\theta) = -8R^2 \sin\left(\frac{\pi}{2}\right) < 0.$$

Furthermore, the two endpoints actually have zero area, so this is the only such maximum. ■

[20] Strictly speaking, we need only take $\theta \in [0, \pi/4]$ but this requires arguing about the symmetry of the problem which will just serve to complicate the solution.

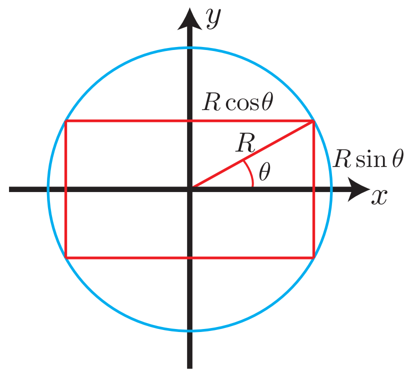


Figure 21: The construction of the parameterization of inscribed rectangles of a circle of radius R.

4.5 L'Hôpital's Rule

Just when we thought we had left limits behind forever, they find a way to creep back up to us. This time however, we are wielding a powerful weapon: the Mean Value Theorem. Recall that when evaluating limits which were not continuous, we often had to perform some algebraic trickery to manipulate the function into a form more amenable to substitution. It was a task which often required an intimate knowledge of the behaviour of the function. L'Hôpital's Rule shall make this much easier.

Definition 4.26

Let f, g be two functions,

1. If $\lim_{x\to c} f(x) = \lim_{x\to c} g(x) = 0$ then we say that $\lim_{x\to c} \frac{f(x)}{g(x)}$ is *indeterminate of type* $0/0$.
2. If $\lim_{x\to c} |f(x)| = \lim_{x\to c} |g(x)| = \infty$ then we say that $\lim_{x\to c} \frac{f(x)}{g(x)}$ is *indeterminate of type* ∞/∞.

Note: If we take the limit as $x \to \infty$ or $x \to -\infty$, or either of the one sided limits $x \to c^{\pm}$ we will still say that the limits are of the above indeterminate types.

Indeterminate forms are precisely the aforementioned pesky functions. Here is how we attack them.

Theorem 4.27: L'Hôpital's Rule

Suppose that f and g are differentiable in a neighbourhood of c, and that the limit $\lim_{x\to c} \frac{f(x)}{g(x)}$ is indeterminate of either type $0/0$ or type ∞/∞. If one of the following conditions holds:

1. The limit $\lim_{x\to c} \frac{f'(x)}{g'(x)}$ exists,
2. There is a neighbourhood of c on which g' is never zero,
3. There is a neighroubhood of c in which g' does not switch signs infinitely often,

then $\lim_{x\to c} \frac{f(x)}{g(x)}$ exists and $\lim_{x\to c} \frac{f(x)}{g(x)} = \lim_{x\to c} \frac{f'(x)}{g'(x)}$.

Proof. We shall do the proof for the 0/0-indeterminate type and with condition (1). The others are left as an exercise for the student. Let $[c-\delta, c+\delta]$ be the neighbourhood on which f, g are differentiable. Define if necessary the auxiliary functions

$$F(x) = \begin{cases} f(x) & x \neq c \\ 0 & x = c \end{cases}, \quad G(x) = \begin{cases} g(x) & x \neq c \\ 0 & x = c \end{cases}$$

so that both f and g are continuous at 0. Let $h > 0$ be sufficiently small so that $[c, c+h] \subseteq (c-\delta, c+\delta)$, and notice that the function $h(x) = F(x)G(c+h) - G(x)F(c+h)$ is continuous on $[c, c+h]$ and differentiable on $(c, c+h)$. Furthermore, $h(c) = 0$ and $h(c+h) = 0$, so by Rolle's Theorem, there exists some $\theta \in (c, c+h)$ such that

$$0 = h'(\theta) = F'(\theta)G(c+h) - G'(\theta)F(c+h).$$

In particular, this means that

$$\frac{F'(\theta)}{G'(\theta)} = \frac{F(c+h)}{G(c+h)}.$$

Since $c < \theta < c+h$, then as a function of h we have that $\theta \xrightarrow{h\to 0^+} c$, and so

$$\lim_{x\to c^+} \frac{f(x)}{g(x)} = \lim_{h\to 0^+} \frac{F(x)}{G(x)} = \lim_{\theta\to c} \frac{F'(\theta)}{G'(\theta)}.$$

Hence the result holds for the one sided limit. The two sided limit can be done by repeating the argument for $h < 0$. □

Example 4.28

Determine the limit $\lim_{x\to 0} \frac{xe^{nx} - x}{1 - \cos(nx)}$ for $n \neq 0$.

Solution. It is easy to check that this is of type $0/0$, so we can apply L'Hôpital to get

$$\begin{aligned}\lim_{x\to 0}\frac{xe^{nx}-x}{1-\cos(nx)} &\overset{\langle H\rangle}{=} \lim_{x\to 0}\frac{e^{nx}+nxe^{nx}-1}{n\sin(nx)}\\ &\overset{\langle H\rangle}{=} \lim_{x\to 0}\frac{ne^{nx}+ne^{nx}+n^2xe^{nx}}{n^2\cos(nx)}\\ &= \frac{2n}{n^2}=\frac{2}{n}.\end{aligned}$$ ■

This is an example where we actually need to apply L'Hôpital's rule more than once. There are also lots of interesting results we can get from this theorem.

Example 4.29

If f is twice continuously differentiable function, compute the limits

$$\lim_{h\to 0}\frac{f(x+h)-f(x-h)}{2h}, \qquad \lim_{h\to 0}\frac{f(x+h)-2f(x)+f(x-h)}{h^2}. \tag{4.8}$$

Solution. Contextually, it should be obvious that we are going to proceed via L'Hôpital's rule, though the student might recall that I pointed out the first limit earlier in the course.

We recognize that the limit in question is indeterminate of type $0/0$ and so we have

$$\lim_{h\to 0}\frac{f(x+h)-f(x-h)}{2h} \overset{\langle H\rangle}{=} \lim_{h\to 0}\frac{f'(x+h)-(-1)f'(x-h)}{2} = f'(x) \tag{4.9}$$

wherein again we stress that we must *differentiate with respect to h*.

The second limit in (4.8) is a little bit harder to see without using L'Hôpital, so let's just go ahead and compute to find

$$\begin{aligned}\lim_{h\to 0}\frac{f(x+h)-2f(x)+f(x-h)}{h^2} &\overset{\langle H\rangle}{=} \lim_{h\to 0}\frac{f'(x+h)-f'(x-h)}{2h} && \text{by L'Hôpital with type } 0/0\\ &= (f')'(x) && \text{by (4.9)}\\ &= f''(x).\end{aligned}$$ ■

Corollary 4.30

Let f be continuous at a point $p\in\mathbb{R}$ and assume f' exists everywhere in a neighbourhood of p. If $\lim_{x\to p} f'(x)$ exists, then f is continuously differentiable at p; that is, $f'(p)$ exists and

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

Solution. Define the function $F(x)=f(x)-f(p)$ and let $G(x)=x-p$. We claim that we can apply L'Hôpital's rule to the quotient $F(x)/G(x)$ at the point 0. Since f is continuous, we clearly have that

$$\lim_{x\to p} F(x) = \lim_{x\to p}\left[f(x)-f(p)\right] = f(p)-f(p)=0$$

and similarly $G(x) \xrightarrow{p\to 0} 0$. Furthermore, $F'(x) = f'(x)$ and $G'(x) = 1$ so that

$$\lim_{x\to p} \frac{F'(x)}{G'(x)} = \lim_{x\to p} f'(x), \tag{4.10}$$

which exists by assumption. Hence

$$\begin{aligned} f'(p) &= \lim_{x\to p} \frac{f(x)-f(p)}{x-p} \\ &= \lim_{x\to p} \frac{F(x)}{G(x)} \\ &= \lim_{x\to p} \frac{F'(x)}{G'(x)} && \text{L'Hôpital's Rule} \\ &= \lim_{x\to p} f'(x) && \text{Equation (4.10)} \end{aligned}$$

and this is what we wanted to show. ■

4.5.1 Other Indeterminate Types

There are several other indeterminate types, which we approach by converting them into the type 0/0 case.

1. $0 \times \infty$: Possibly the easiest case to deal with, assume that we are given a limit of the form

$$\lim_{x\to a} f(x)g(x), \qquad \text{where } \lim_{x\to a} f(x) = 0, \qquad \text{and} \qquad \lim_{x\to a} g(x) = \infty.$$

   We can thus turn this into $0/0$ or ∞/∞ as

$$\lim_{x\to a} \frac{f(x)}{1/g(x)} \quad \text{type } 0/0, \qquad \lim_{x\to a} \frac{g(x)}{1/f(x)} \quad \text{type } \infty/\infty.$$

2. $\infty - \infty$: In contrast to the previous cases, these are possibly the hardest to transform into one of $0/0$ or ∞/∞. This case arises when we have a limit of the form:

$$\lim_{x\to a} [f(x) - g(x)], \qquad \text{where } \lim_{x\to a} f(x) = \infty, \qquad \text{and} \qquad \lim_{x\to a} g(x) = \infty.$$

   There is no obvious way to deal with these problems, and often need to be done on a case-by-case basis.

3. $1^\infty, 0^0, \infty^0$: These cases are all handled identically, so we shall treat them together. For the sake of concreteness, assume we are given a limit of the form

$$\lim_{x\to a} f(x)^{g(x)}, \qquad \text{where } \lim_{x\to a} f(x) = \infty, \qquad \text{and} \qquad \lim_{x\to a} g(x) = 0$$

   This case is handled by calling upon our old friend, implicit differentiation. Let $y = f(x)^{g(x)}$ so that $\log(y) = g(x) \log f(x)$. As we have seen previously, continuity of e^x means that it is sufficient to compute

$$\lim_{x\to a} \log y, \qquad \text{since} \qquad \lim_{x\to a} y = \lim_{x\to a} e^{\log(y)} = e^{\lim_{x\to a} \log y}.$$

Now in the limit, $\log(y) = g(x)\log f(x)$ is of type $0 \times \infty$, so we resort to case (1) which tells us that we compute this limit by either

$$\lim_{x\to a} \frac{g(x)}{1/\log f(x)}, \qquad \text{or} \qquad \lim_{x\to a} \frac{\log f(x)}{1/g(x)}.$$

In the case where $f(x) \to 1$ and $g(x) \to \infty$, or $f(x) \to 0$ and $g(x) \to 0$ we also get $0 \times \infty$ and so these cases are treated the same as above.

Note that there is no type 0^∞ despite the fact that this seems like it could potentially be indeterminate. The intuitive reason for this is that as soon as the base becomes less than the number 1, taking powers will actually drive the limit closer to 0. Hence 0^∞ and is not an indeterminate form.

Example 4.31

Compute the limit $\lim_{x\to 0^+} x\log(x)$.

Solution. In the limit as $x \to 0$ the function $\log(x) \to -\infty$, so this limit is indeterminate of type $0 \times \infty$. We now have to make a choice as to which component to invert. Let us choose to first invert the x, giving

$$\begin{aligned} \lim_{x\to 0^+} x\log(x) &= \lim_{x\to 0} \frac{\log(x)}{1/x} \overset{\langle H\rangle}{=} \lim_{x\to 0} \frac{1/x}{-1/x^2} \\ &= \lim_{x\to 0} -x = 0. \end{aligned}$$ ■

Example 4.32

Determine the limit $\lim_{x\to 0}[\csc(x) - \cot(x)]$

Solution. It is easy to see that this limit is indeterminate of type $\infty - \infty$, but as everything is trigonometric, there is likely some simplification that can be performed. Writing everything in terms of $\sin(x)$ and $\cos(x)$ we get

$$\begin{aligned} \lim_{x\to 0}[\csc(x) - \cot(x)] = \lim_{x\to 0}\left[\frac{1}{\sin(x)} - \frac{\cos(x)}{\sin(x)}\right] &= \lim_{x\to 0} \frac{1-\cos(x)}{\sin(x)} \\ &\overset{\langle H\rangle}{=} \lim_{x\to 0} \frac{\sin(x)}{\cos(x)} = 0. \end{aligned}$$

Note that alternatively in this last step, we could have avoided using L'Hôpital's Rule by writing

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

Example 4.33

Determine the limit $\lim_{x\to\infty}\left(1+\frac{a}{x}\right)^x$.

Solution. It is not too hard to see that this is indeterminate of type 1^∞, so we must take logarithms in order to simplify our lives. Indeed, set $y = \left(1+\frac{a}{x}\right)^x$ so that $\log y = x\log\left(1+\frac{a}{x}\right)$. In the limit as $x\to\infty$, this is of type $0\times\infty$, so we take the reciprocal of one of the elements and get

$$\begin{aligned}\lim_{x\to\infty} x\log\left(1+\frac{a}{x}\right) &= \lim_{x\to\infty}\frac{\log\left(1+\frac{a}{x}\right)}{1/x}\\ &\overset{\langle H\rangle}{=} \lim_{x\to\infty}\frac{(-a/x^2)/\left(1+\frac{a}{x}\right)}{-1/x^2}\\ &= \lim_{x\to\infty}\frac{a}{1+\frac{a}{x}} = a.\end{aligned}$$

Hence re-exponentiating, we get

$$\lim_{x\to\infty}\left(1+\frac{a}{x}\right)^x = e^{\lim\limits_{x\to a}\log y} = e^a. \qquad \blacksquare$$

4.5.2 Little 'o'-notation

In many areas of computer science and number theory, one is interested in the growth rate of functions relative to other functions. There are many ways of characterizing these growth rates, but one way is with little 'o' notation.

Definition 4.34

Let f, g be functions such that g is eventually non-zero. We say that $f = o(g)$ if

$$\lim_{x\to\infty}\frac{f(x)}{g(x)} = 0.$$

Example 4.35

Show that for any $n\in\mathbb{N}$, $x^n = o(e^x)$.

Solution. It suffices to show that $\lim_{x\to\infty}\frac{x^n}{e^x} = 0$ for any $n\in\mathbb{N}$. We shall proceed by induction. As a base case, notice that $\lim_{x\to\infty} xe^{-x}$ is of type $0/0$ and hence we may apply L'Hôpital to get

$$\lim_{x\to\infty}\frac{x}{e^x}\overset{\langle H\rangle}{=}\lim_{x\to\infty}\frac{1}{e^x} = 0.$$

Now assume that the result holds for some $k\geq 1$, and notice that

$$\lim_{x\to\infty}\frac{x^{k+1}}{e^x}\overset{\langle H\rangle}{=} k\lim_{x\to\infty}\frac{x^k}{e^x} = 0$$

where the last equality is via the induction hypothesis. We thus conclude that $x^n = o(e^x)$ for all $n \in \mathbb{N}$. ■

Exercise: Show that $\log(\log(x)) = o(\log(x))$.

4.6 Curve Sketching

Without the assistance of computers, it can be difficult to assess the qualitative properties of a function. Luckily, we can use our differentiation tools to begin to analyze how functions behave. The first notion we would like to examine is concavity. The true definition of concavity is rather subtle and not terribly enlightening. Since we often deal with very nice (read: differentiable) functions, we shall instead define concavity as follows:

Definition 4.36

Let f be twice a differentiable function on the interval (a, b). We say that the graph of f is *concave up* at $c \in (a, b)$ if f' is increasing in a neighbourhood of c, and *concave down* if f' is decreasing in a neighbourhood of c. We say that $c \in (a, b)$ is a *inflection point* if the concavity of the function changes in any neighbourhood of c.

If our function is twice differentiable, concavity may be ascertained by examining the sign of $f''(c)$ (since this describes whether f' is increasing or decreasing), and it will be necessary that points of inflection satisfy $f''(c) = 0$. On the other hand, if f is not twice differentiable, points of inflection could corresponds to places where $f''(x)$ does not exist. The notion of concavity is intimately related to that of maxima and minima. Recall that from the second derivative test, if $f'(c) = 0$ and $f''(c) \neq 0$ then c is a min if $f''(c) > 0$ and a max if $f''(c) < 0$.

Exercise: Show that if f is concave up on an interval (a, b) then for any $x_1, x_2 \in (a, b)$ and any $t \in [x_1, x_2]$, we have

$$f(x_1 + t(x_2 - x_1)) \leq f(x_1) + t\left[f(x_2) - f(x_1)\right].$$

Determine a similar condition for when f is concave down.

Exercise 4.6 is the actual notion of concavity, and as the student can see it leaves much to be desired. In effect, the idea is that a function is concave up on an interval (a, b) if, between any two points in the graph of f on (a, b), the graph of the function lies beneath the secant line joining those two points.

4.6.1 Boring Curve Sketching

Our goal is thus to combine all of our information into a system which allows us to analyze the qualitative behaviour of a function without knowing the nitty-gritty details of its exact value at every point. There are approximately seven pieces of information that we need to compute to ascertain the general behaviour.

1. Domain (and range of if possible),
2. Intercepts (x- and y-),
3. Symmetry (even/odd/none),
4. Asymptotes and asymptotics.
5. Increasing/Decreasing,
6. Maxima (and minima),
7. Concavity.

To be more precise about asymptotics, we have the following definition:

Definition 4.37

Let $f(x)$ and $g(x)$ be continuous functions. We say that $f(x)$ *behaves like* $g(x)$ *asymptotically* if

$$\lim_{x\to\pm\infty} [f(x) - g(x)] = 0.$$

We say that $f(x)$ has an *oblique asymptote* if $f(x)$ behaves asymptotically like $g(x) = mx + b$ for some $m \neq 0$.

Example 4.38

Compute any asymptotics of the function $f(x) = \frac{2x^3-x^2+2x}{x^2+1}$.

Solution. The easiest way to proceed is to try to write $f(x)$ as an improper rational function. Performing long division, we see that

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

The idea is that in the limit as $x \to \infty$, then $1/(x^2 + 1)$ term will die off and contribute very little to the behaviour of $f(x)$, so that $f(x)$ looks like the function $2x - 1$. To see that this satisfies the definition above, set $g(x) = 2x - 1$ so that

$$\begin{aligned}\lim_{x\to\infty} [f(x) - g(x)] &= \lim_{x\to\infty} \left[2x - 1 + \frac{1}{x^2+1} - (2x-1)\right] \\ &= \lim_{x\to\infty} \frac{1}{x^2+1} = 0,\end{aligned}$$

which is precisely what we wanted to show. Similarly, the limit as $x \to -\infty$ shows that $f(x)$ is also asymptotically like $g(x)$ in that limit as well. Thus $2x - 1$ is an oblique asymptote for $f(x)$ at both $\pm\infty$. ■

Now that we know how to compute all terms involved in this computation, we shall proceed with some examples.

Example 4.39

Plot the function $f(x) = \dfrac{x^3}{(x+1)^2}$.

Solution. **Domain:** The only point which could possibly give us trouble is $x = -1$. Hence our domain is simply $\mathbb{R} \setminus \{-1\}$.

Intercepts: The y-intercept occurs when $x = 0$, so namely $f(0) = 0$. Similarly the x-intercept comes when $y = 0$, for which we see that

$$\frac{x^3}{(x+1)^2} = 0, \qquad \Leftrightarrow \qquad x = 0.$$

Thus the x- and y-intercepts both occur at the origin.

Symmetry: There is no symmetry involved: Since the functions are polynomial they have no periodicity. The student may check that $f(-x)$ has no relation to $f(x)$, so that the function is neither even nor odd.

Asymptotes: We begin with the horizontal asymptotes. It is easy to see that since the numerator dominates the denominator, the limit will go to infinity (check this by dividing top and bottom by $1/x^3$). Further, since the denominator is always positive, the sign is determined entirely by the x^3 factor, so

$$\lim_{x\to\infty} \frac{x^3}{(x+1)^2} = \infty, \qquad \lim_{x\to-\infty} \frac{x^3}{(x+1)^2} = -\infty.$$

We conclude there are no horizontal asymptotes.

The only candidate for a vertical asymptote occurs at $x = -1$. Again the denominator $(x+1)^2$ is always positive, so the sign of the "infinity" is entirely determined by the behaviour of x^3 around $x = -1$, which is negative. It is then clear that

$$\lim_{x\to 1^\pm} \frac{x^3}{(x+1)^2} = -\infty.$$

Finally, we want to check for oblique asymptotes. Using long polynomial division we may easily find that

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

so we claim that $y = x - 2$ is an oblique asymptote. Indeed, notice that

$$\begin{aligned}
\lim_{x\to\pm\infty}[f(x)-(x-2)] &= \lim_{x\to\pm\infty}\left[\left((x-2)-\frac{3x+2}{x^2+2x+1}\right)-(x-2)\right] \\
&= -\lim_{x\to\pm\infty}\frac{3x+2}{x^2+2x+1} \\
&= -\lim_{x\to\pm\infty}\frac{3/x+2/x^2}{1+2/x+1/x^2} && \text{multiply and divide by } 1/x^2. \\
&= 0.
\end{aligned}$$

First Derivative: Computing the first derivative can be a chore, but we find that

$$\begin{aligned}
f'(x) &= \frac{3x^2(x+1)^2-2(x+1)x^3}{(x+1)^4} \\
&= \frac{3x^4+6x^3+3x^2-2x^4-2x^3}{(x+1)^4} \\
&= \frac{x^2(x+3)(x+1)}{(x+1)^4} \\
&= \frac{x^2(x+3)}{(x+1)^3}
\end{aligned}$$

so that the critical points correspond to $x = 0$ and $x = -3$. The y-values for these points will be useful when we plot, so we substitute to find that $f(0) = 0$ and $f(-3) = 27/4$. To determine where the function is increasing and decreasing, we consider the following table:

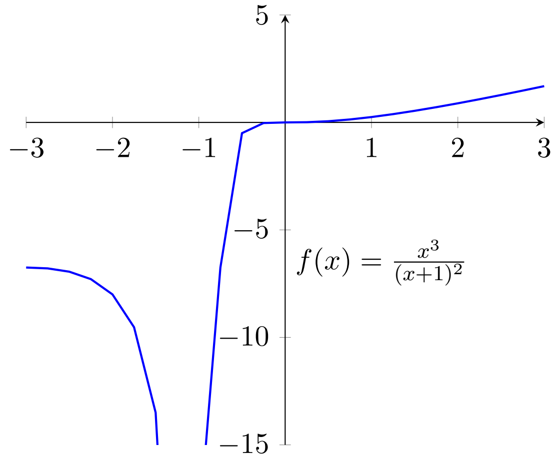
	$x < -3$	$-3 < x < -1$	$-1 < x < 0$	$0 < x$
$x+3$	$-$	$+$	$+$	$+$
$(x+1)^3$	$-$	$-$	$+$	$+$
x^2	$+$	$+$	$+$	$+$
$f'(x)$	$+$	$-$	$+$	$+$

Second Derivative: The second derivative is a little messy, but simplifies if done correctly.

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x}\frac{x^2(x+3)}{(x+1)^3} &= \frac{(3x^2+6x)(x+1)^3-3(x+1)^2(x^3+3x^2)}{(x+1)^6} \\
&= \frac{3x^3+3x^2+6x^2+6x-3x^3-9x^2}{(x+1)^4} \\
&= \frac{6x}{(x+1)^4}
\end{aligned}$$

so there is an inflection point at $(0, 0)$ (telling us that one of the critical points is an inflection point. Furthermore, $f''(-3) = -9/8 < 0$ so the point $(-3, 27/4)$ is a max. Since the denominator is a quartic it is always positive, and we can see that concavity is entirely determined by the numerator $6x$. Hence $f(x)$ is concave up when $f''(x) > 0$, corresponding to $x > 0$; and $f(x)$ is concave down when $f''(x) < 0$, corresponding to $x < 0$.

Plotting: Putting all of this information together, the student should get the following plot: ■



Example 4.40

Plot the function $f(x) = \dfrac{x^3}{1-x^2}$.

Solution. Following our new six-step program we set to work:

Domain: By this point, this should not be too hard to see. In particular, our function will not be defined whenever the denominator is zero. This happens at the points $x = \pm 1$ and so our domain is $\mathbb{R} \setminus \{\pm 1\}$.

Intercepts: The y-intercept occurs when $x = 0$, so namely $f(0) = 0$. Similarly the x-intercept comes when $y = 0$, for which we see that

$$\frac{x^3}{1-x^2} = 0, \qquad \Leftrightarrow \qquad x = 0.$$

Thus the x- and y-intercepts both occur at the origin.

Symmetry: Since we are dealing with polynomials, there is no obvious periodicity to worry about. It's not too hard to see that this is actually an odd function, since

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

Asymptotes: The vertical asymptotes will clearly occur at $x = \pm 1$. Typically, one would calculate the limits

$$\lim_{x\to 1^\pm} \frac{x^3}{1-x^2}, \quad \lim_{x\to -1^\pm} \frac{x^3}{1-x^2}$$

but this is laborious and is redundant once we have information on the first derivative. For the interested student who would like to see how to do this all the same, we have the following table

	x^3	$1-x^2$	$x^3/1-x^2$
$x \to 1^+$	$+$	$-$	$-$
$x \to 1^-$	$+$	$+$	$+$
$x \to -1^+$	$-$	$+$	$-$
$x \to -1^-$	$-$	$-$	$+$

so that

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

Because the degree of the numerator is strictly greater than the degree of the denominator, there are no horizontal asymptotes:

$$\lim_{x\to\pm\infty} \frac{x^3}{1-x^2} = \mp\infty.$$

Finally, we want to check for oblique asymptotes. Using long polynomial division we may easily find that

$$\frac{x^3}{1-x^2} = -x + \frac{x}{1-x^2}$$

so we claim that $y = -x$ is an oblique asymptote. Indeed, notice that

$$\begin{aligned} \lim_{x\to\pm\infty} [f(x) - (-x)] &= \lim_{x\to\pm\infty} \left[\frac{x^3}{1-x^2} + x\right] \\ &= \lim_{x\to\pm\infty} \frac{x}{1-x^2} \\ &= \lim_{x\to\pm\infty} \frac{1/x}{1/x^2 - 1} \\ &= 0. \end{aligned}$$

First Derivative: This step allows us to determine where the function is increasing, decreasing, the critical points, and when combined with the second derivative, maxima and minima. The first derivative is computed to be

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x} \frac{x^3}{1-x^2} &= \frac{(3x^2)(1-x^2) - (-2x)(x^3)}{(1-x^2)^2} \\ &= \frac{x^2(3-x^2)}{(1-x^2)^2}. \end{aligned}$$

We may simply read off the critical points as $\left\{\pm 1, \pm\sqrt{3}, 0\right\}$ with potential extrema at $\left\{0, \pm\sqrt{3}\right\}$. Setting up a quick table for increasing and decreasing we have

	$x < -\sqrt{3}$	$-\sqrt{3} < x < -1$	$-1 < x < 0$	$0 < x < -1$	$1 < x < \sqrt{3}$	$x > \sqrt{3}$
$f'(x)$	$-$	$+$	$+$	$+$	$+$	$-$

This table is easy to deduce once we realize that the $\frac{x^2}{(1-x^2)^2}$ portion is always positive, so the sign of $f'(x)$ is entirely determined by the sign of $3 - x^2$, which is negative whenever $|x| > \sqrt{3}$. It will likely be useful to know the function values corresponding to our critical points. We already know that $f(0) = 0$ and we find that

$$f(\pm\sqrt{3}) = \frac{\pm 3\sqrt{3}}{-2} = \frac{\mp 3\sqrt{3}}{2}.$$

Second Derivative: The second derivative is a little messy, but simplifies if done correctly.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\frac{x^2(3-x^2)}{(1-x^2)^2} &= \frac{(6x-4x^3)(1-x)^2 - 2(1-x^2)(-2x)(3x^3-x^4)}{(1-x^2)^4}\\ &= \frac{2x(x^2+3)}{(1-x^2)^3}.\end{aligned}$$

The inflection points will occur when $f''(x) = 0$ or does not exist, which we can again read off as being $\{0, \pm 1\}$. We form a table to check for concavity and find

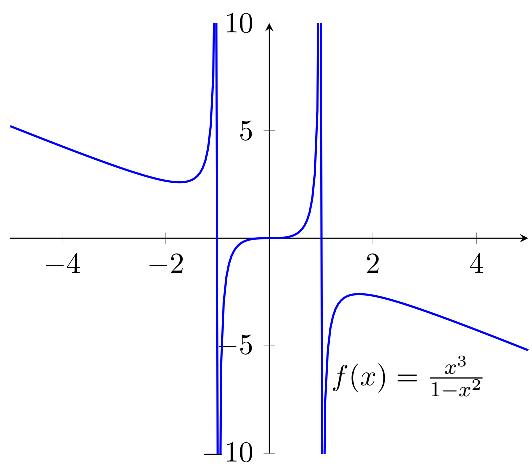
	$x < -1$	$-1 < x < 0$	$0 < x < 1$	$x > 1$
$f''(x)$	$+$	$-$	$+$	$-$

Finally, recalling that we have extrema candidate at ± 3 we check to find that

$$f(\pm\sqrt{3}) = \frac{\pm 2\sqrt{3}(3+3)}{(1-3)^2} = \pm\frac{12\sqrt{3}}{-8} = \mp\frac{3\sqrt{3}}{2}.$$

Thus $\left(\sqrt{3}, -\frac{3\sqrt{3}}{2}\right)$ is a local maximum and $\left(-\sqrt{3}, \frac{3\sqrt{3}}{2}\right)$ is a local minimum. Since $f''(0) = 0$ we cannot infer any information about this critical point. If we continue to take derivatives, we will find that $f^{(3)}(0) = 6$ and so by the generalized second derivative test, 0 is an inflection point.

Plotting: Putting all of this information together, the student should get the following plot: ■



4.6.2 Curves with Parameters

An interesting question comes when we ask ourselves "How much can we change the above examples and still end up with something which is qualitatively similar?" For example, what if we consider the function

$$f_c(x) = \frac{x^3}{c - x^2}$$

for some $c \in \mathbb{R}$? What values of c change the qualitative behaviour?

Our function still passes through the origin, is odd, and has an oblique asymptote of $-x$ (since polynomial long division yields $f_c(x) = -x + \frac{cx}{c-x^2}$), but we notice that whether or not $f_c(x)$ has vertical asymptotes depends on the sign of c. In particular, if $c > 0$ then there are asymptotes are $\pm c$, while if $c = 0$ our function becomes $f_c(x) = -x$ for all $x \neq 0$, so has no asymptotes but is still not defined at a point. Finally, if $c < 0$ then our function has no asymptotes, and has domain $\mathbb{R}$.

We can compute our derivative to be

$$f_c'(x) = \frac{x^2(3c - x^2)}{(c - x^2)^2}.$$

Again here, the sign of c is important. The critical point at 0 is unchanged: if $c > 0$ then there are also critical points at $\pm\sqrt{3c}$ and $\sqrt{c}$, while if $c < 0$ then there are no additional critical points.

	$x < -\sqrt{3c}$	$-\sqrt{3c} < x < -\sqrt{c}$	$-\sqrt{c} < x < 0$	$0 < x < -\sqrt{c}$	$\sqrt{c} < x < \sqrt{3c}$	$x > \sqrt{3c}$
$f_c'(x)$ $c>0$	$-$	$+$	$+$	$+$	$+$	$-$

	$x \in \mathbb{R}$
$f_c'(x)$ $c=0$	$-$

	$x < 0$	$x > 0$
$f_c'(x)$ $c<0$	$-$	$-$

This tells us that if $c \leq 0$ then there this function has no maxima/minima, and is always decreasing.

The second derivative can be computed to be

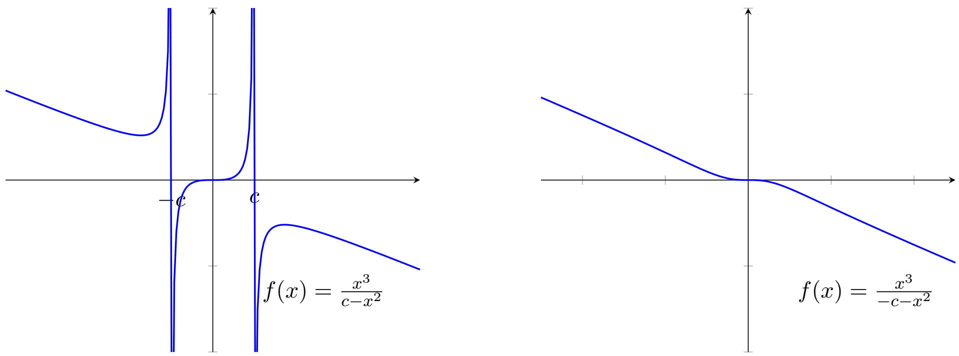
$$f_c''(x) = \frac{2cx(x^2 + 3c)}{(c - x^2)^3}.$$

If $c > 0$ then potential inflection points are $x = 0, x = \pm\sqrt{c}$. If $c = 0$ then $f_c''(x) = 0$. If $c < 0$ then the potential inflection points are $x = 0$ and $x = \pm\sqrt{-3c}$.

	$x < -\sqrt{c}$	$-\sqrt{c} < x < 0$	$0 < \sqrt{c}$	$x > \sqrt{c}$
$f_c''(x)$ $c>0$	$+$	$-$	$+$	$-$

	$x \in \mathbb{R}$
$f_c''(x)$ $c=0$	0

	$x < -\sqrt{-3c}$	$-\sqrt{-3c} < x < 0$	$0 < \sqrt{-3c}$	$x > \sqrt{-3c}$
$f_c''(x)$ $c<0$	$-$	$+$	$-$	$+$



5 Integration

The theory of integration will consume much of the remainder of this course, and is in many ways dual to differentiation. Consequently, the theory of integration is usually more difficult to establish, despite the fact that it is much more natural mathematically.

The over-arching goal of integration is to add things together in a continuous fashion. This manifests itself in applications as finding the area under a curve, the volume of an object, or even as calculating physical quantities such as work, flux, or voltage potentials. The fact that this is even remotely related to the process of differentiation is not at all obvious, though we will see shortly that there is in fact an intimate relationship.

There are many different techniques for defining the integral. Any student who has seen the integral in his/her secondary school days likely saw the method of Riemann sums, done in the typical hand-waving fashion. We are going to digress from this approach and introduce an alternative technique which generalizes much better to higher dimensions. This will allow for a more natural segue for students who have to take the follow up course, MAT237.

5.1 Infima and Suprema

We have to take a small digression before introducing the integral. In this section, we will introduce the notions of infima and suprema. While these could have been introduced at the beginning of the course, it is a difficult notion which can add complexity to an already chaotic first few weeks. Now that we are more comfortable with abstract thinking, we shall begin to probe the subject.

As motivation to the subject, recall the following problem. One would often be given a function f and told that its range is the interval $[0, 1)$. When asked "What is the maximum of this function," the student would often be tempted to say "The maximum is at 1," despite the fact that the correct answer is "There is no maximum." The reason there is no maximum is that the function never actually attains the value 1, despite getting arbitrarily close to it (see Figure 22).

Here we feel cheated! Morally, the number 1 does everything that we want a maximum to do, we just get bogged down in the pedantry of whether or not 1 actually belongs to the set or not. However, the precision of mathematics means that pedantry is essential, so we have to introduce a new way of discussing what this 'moral maximum' should be.

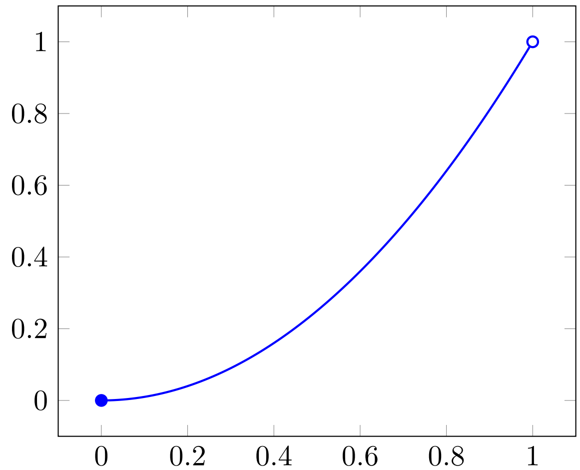


Figure 22: A function whose range is $[0, 1)$. This function does not have a maximum, despite the fact that it gets arbitrarily close to 1.

5.1.1 Least Upper Bound and Greatest Lower Bound

Definition 5.1

Let $S \subseteq \mathbb{R}$ be a set. We say that $M \in \mathbb{R}$ is an *upper bound for* S if for every $x \in S$ we have $x \leq M$. In this case we say that S *is bounded from above*. We say that M is the *supremum* or *least upper bound* of S if whenever M' is an upper bound, then $M \leq M'$. We will write $M = \sup S$ to indicate the M is the supremum of S. If $M \in S$ then we say that M is the *maximum* of S.

Just as the name suggests, the least upper bound of a set is the smallest of the upper bounds, though the word supremum is more commonly used by mathematicians. Naturally, there is a dual notion for lower bounds:

Definition 5.2

Let $S \subseteq \mathbb{R}$ be a set. We say that $m \in \mathbb{R}$ is an *lower bound for* S if for every $x \in S$ we have $x \geq M$. In this case we say that S *is bounded from below*. We say that m is the *infimum* or *greatest lower bound* of S if whenever m' is an lower bound, then $m \geq m'$. We will write $m = \inf S$ to indicate the m is the infimum of S. If $m \in S$ then we say that m is the *minimum* of S.

Hence we have that the infimum and supremum play the role of the moral minimum and maximum discussed earlier, but are only bona-fide minima and maxima when the infimum and supremum actually belong to the set.

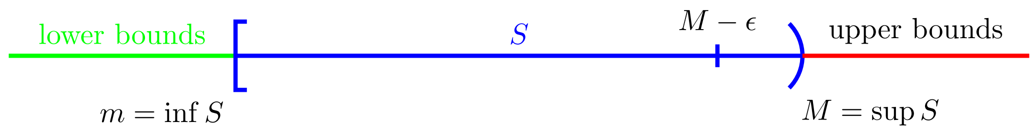


Figure 23: The blue represents the set S, green the set of lower bounds, and red the set of upper bounds for S. The number m is the largest of the green while M is the smallest of the red.

Example 5.3

1. The set $S = (-3, -1)$ is bounded both above and below. The infimum is -3 and supremum is -1. Since neither -3 nor -1 belong to S, the set S does not have a minimum or a maximum.
2. Let $S = \left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \ldots\right\}$. This set is bounded above and below. It supremum is 1, and since 1 is in the set this is actually a maximum. On the other hand, its infimum is 0, which is not in the set.
3. The empty set $\emptyset$ is bounded above and below. Every real number is both an upper bound and a lower bound. Since there is no smallest/largest real number, this set does not have an infimum or supremum, and so certainly does not have a maximum or a minimum.
4. Every finite set has a maximum and a minimum.
5. Let $S = \left\{x \in \mathbb{Q} : x^2 > 2, x > 0\right\}$. This set is bounded below but not above. The infimum of this set is $\sqrt{2}$ (convince yourself that this is true), and so cannot have a minimum since S is a subset of the rational numbers, and $\sqrt{2}$ is not rational.[a]

[a] In this course we have glossed over the construction/definition of the real numbers, and asked you to believe us that such a thing exists. One way of constructing the real numbers from the rational numbers is via a method called *Dedekind cuts*. The set $S = \left\{x \in \mathbb{Q} : x^2 \geq 2, x > 0\right\}$ is such a cut. One then *identifies* S with the number $\sqrt{2}$.

5.1.2 Properties and Results

An absolutely crucial property of the real numbers that we will take for granted is the following:

The Completeness Axiom: Every non-empty set of real numbers which is bounded above has a supremum. Equivalently, every non-empty set of real numbers which is bounded below has an infimum.

The necessity that the set be non-empty is exemplified by Example 5.3 (3), while the fact that the set must be bounded is demonstrated by Example 5.3 (5). There will be many instances in the sections which follow in which we would like to work with the supremum of a set without needing to know its value explicitly. The completeness axiom gives us the ability to guarantee that the supremum exists, allowing us to work with it.

Proposition 5.4

If $S \subseteq \mathbb{R}$ and $M = \sup S$ then for any $\epsilon > 0$ there exists an $x \in S$ such that $M - \epsilon < x \leq M$.

Proof. The proof of this fact is actually fairly obvious if we think about it, and I would encourage the student to give it a try before reading onwards.

Since M is an upper bound for S we have that $x \leq M$ for all $x \in S$, so this part is always true. Thus we need only show that there exists an $x \in S$ with $M - \epsilon < x$. For the geometric insight, we refer the reader back to Figure 23. The idea is that if there is no such x, then $M - \epsilon$ is also an upper bound and this is not possible.

More rigorously, for the sake of contradiction assume that no such x exists. Consequently, we have that $x \leq M - \epsilon$ for all $x \in S$. This is precisely what it means for $M - \epsilon$ to be an upper bound for S. However, $M - \epsilon < M$ which contradicts the fact that M is the *least* upper bound. Hence our original assumption must have been false, and we conclude that there must be an x satisfying $M - \epsilon < x$. □

Exercise: Determine the equivalent statement to Proposition 5.4 for the infimum of a set. Prove your statement.

The real reason we have introduced suprema and infima is that we which to determine these properties on the image of a function restricted to a subset of the real. To elaborate further, assume that f is a function and $S \subseteq \mathbb{R}$ is some set. Recall that the *image of S under f* is the set

$$f(S) = \{f(x) : x \in S\} = \{y : \exists x \in S, f(x) = y\}.$$

One can think of $f(S)$ as the 'range' of f when we restrict f to the set S. We are interested in the supremum and infimum of these sets. This is commonly written in one of the two following ways:

$$M = \sup f(S) = \sup_{x \in S} f(x).$$

Example 5.5

If $f(x) = \dfrac{x \sin(x)}{x+1}$ and $S = (0, \infty)$ determine $\sup f(S)$ and $\inf f(S)$.

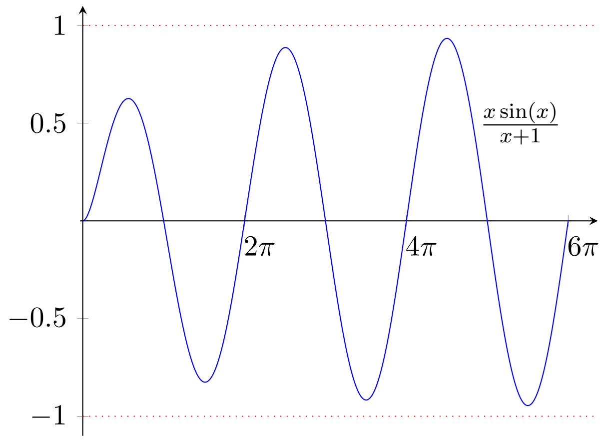
Solution. We know that $-1 \leq \sin(x) \leq 1$ for all $x \in \mathbb{R}$. Furthermore, since $x \in S$ we have that $x > 0$ so that $\frac{x}{x+1} > 0$, yielding

$$-\frac{x}{x+1} \leq \frac{x \sin(x)}{x+1} \leq \frac{x}{x+1}.$$

Now $x/(x+1)$ can never be greater than 1 (check this) but we have

$$\lim_{x \to \infty} \frac{x}{x+1} = 1.$$

Hence $f(S) = (-1, 1)$ and so $\sup f(S) = 1$ while $\inf f(S) = -1$. ■



5.2 Sigma Notation

Sigma notation is used to make complicated sums much easier to write down. In particular, we use a summation index to iterate through elements of a list and then sum them together. Let us take a moment to dissect the elements of the notation itself. Consider the expression

$$\sum_{i=n}^{m} r_i \tag{5.1}$$

which is read as "the sum from $i = n$ to m of r_i." The element i is known as the *dummy* or *summation* index, n and m are known as the *summation bounds*, and r_i is the *summand*. In order to decipher this rather cryptic notation, we adhere to the following algorithm:

1. Set $i = n$ and write down r_i;
2. Add 1 to the index i and add r_i to the current sum;
3. If i is equal to m then stop, otherwise go to step 2 and repeat.

For those computer savvy students out there, this is nothing more than a for-loop. Interpreting (5.1) we thus have

$$\sum_{i=n}^{m} r_i = r_n + r_{n+1} + r_{n+2} + \cdots r_m.$$

Example 5.6

Set $r_1 = 5, r_2 = -8, r_3 = 4$. Compute $\sum_{i=1}^{3} r_i$.

Solution. Via our discussion above, we may write the summation out explicitly as

$$\sum_{i=1}^{3} r_i = r_1 + r_2 + r_3 = 5 + (-8) + 4 = 1.$$

■

Now the r_i could be a collection of unrelated numbers as in Exercise 5.6, but they could also be a "function" of the index variable as follows:

Example 5.7

Compute $\sum_{i=1}^{4}(2i+1)$.

Solution. Following our algorithm, we start by setting $i = 1$ and then evaluating the summand. I will write out the steps in slightly more detail than usual to illustrate the process:

$$\begin{aligned}\sum_{i=1}^{4}(2i+1) &= (2i+1)_{i=1} + (2i+1)_{i=2} + (2i+1)_{i=3} + (2i+1)_{i=4} \\ &= (2\cdot 1+1) + (2\cdot 2+1) + (2\cdot 3+1) + (2\cdot 4+1) \\ &= 3+5+7+9 \\ &= 24.\end{aligned}$$

■

Sometimes we get lucky and can find closed form expressions for certain abstract summations. The following two examples are incredibly useful identities that are used throughout many fields of mathematics.

Example 5.8

For $n \geq 1$ compute $\sum_{j=1}^{n} 1$.

Solution. Notice that our summand (the number 1) does not depend on our index variable j. Nonetheless, we must proceed with our algorithm to find that

$$\begin{aligned}\sum_{j=1}^{n} 1 &= 1_{j=1} + 1_{j=2} + \cdots + 1_{j=n} \\ &= \underbrace{1+1+\cdots+1}_{n\text{-times}} \\ &= n.\end{aligned}$$

Since the summand did not depend on the index, we just repeated the index n-times with absolutely no changes between the iterations. Of course, we can be hyper-rigorous and prove this by induction if we so choose. ■

Example 5.9

For $n \geq 1$ compute $\sum_{j=1}^{n} j$.

Solution. This is another common example that we saw when we were doing induction, and indeed we know that the solution is

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

The student can prove this using induction.

There are many ways of determining this result, without guessing and using induction. Try to find some of these methods. ■

Notice now that we can combine Examples **??** and **??** in order to derive a general closed from solution for the sum considered in Example **??**. Indeed,

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

Plugging in $n = 4$ we get 24, just as we expected.

Remark 5.10 For any positive integer p, there is a closed form expression for

$$\sum_{i=1}^{n} i^p$$

but these expressions become successively more difficult to compute from a naive standpoint as done above. For example, can you think of a clever way to even compute the case when $p = 2$? Luckily, there is a standard way of deriving the closed form for any p using the *Bernoulli polynomials*, which are popular objects in the study of number theory but are tricky to define.

5.3 The Definite Integral

As mentioned in the beginning of this section, the goal of integrals is to sum up a continuous collection of objects. One great motivational example for this is computing the area under a curve, which can be thought of as 'summing' the values $f(x)$ as x ranges in some interval. The way this is done is by first breaking our curve into rectangles whose area is easily computed, and effectively analyze what happens as we allow the width of the rectangles to get arbitrarily small.

5.3.1 Partitions

Definition 5.11

Consider a bounded function f and an interval $[a, b]$. A *partition* of $[a, b]$ is a finite set of points, $P = \{x_1, \ldots, x_n\}$ such that

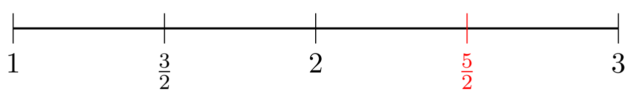
$$a = x_1 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b.$$

We will denote by $|P| = n$ the number of points in the partition.

As defined, a partition is nothing more than a finite collection of ordered points. However, that is not how you should think about a partition. Instead, you should have in mind that a partition P defines the endpoints of a collection of subintervals. So for example, consider the partition $P = \left\{1, \frac{3}{2}, 2, 3\right\}$ of the interval $[1, 3]$. Again, the idea is that each element of P represents a point at which to cut the interval $[1, 3]$. After doing this, we get the result:

$$[1, 3/2], \quad [3/2, 2], \quad [2, 3].$$

If P and Q are partitions such that $P \subseteq Q$, we say that Q is a *refinement* of P. In our example above, if $Q = \{1, 3/2, 2, 5/2, 3\}$ then $P \subseteq Q$. The reason why this is said to be a refinement is because the addition of the point $\frac{5}{2}$ means that we split the interval $[2, 3]$ into two more intervals $[2, 5/2], [5/2, 3]$. In a very real sense, the partition has been refined.



5.3.2 The Upper and Lower Integrals

Now that we have partitions in hand, we want to start computing areas under curves. We are going to use partitions to break our intervals into smaller intervals, and use rectangles to make coarse approximations to our curve.

Definition 5.12

Let f be a bounded function on $[a, b]$ and $P = \{x_1, \ldots, x_n\}$ be a partition. We define the *upper Riemann sum* $U_f(P)$ and *lower Riemann sum* $L_f(P)$ as

$$U_f(P) = \sum_{i=1}^{|P|-1} \left[\sup_{x \in [x_i, x_{i+1}]} f(x)\right](x_{i+1} - x_i)$$

$$L_f(P) = \sum_{i=1}^{|P|-1} \left[\inf_{x \in [x_i, x_{i+1}]} f(x)\right](x_{i+1} - x_i)$$

Note that since f is bounded on $[a, b]$, it is certainly bounded on subintervals of $[a, b]$, guaranteeing that the supremum and infimum actually exist.

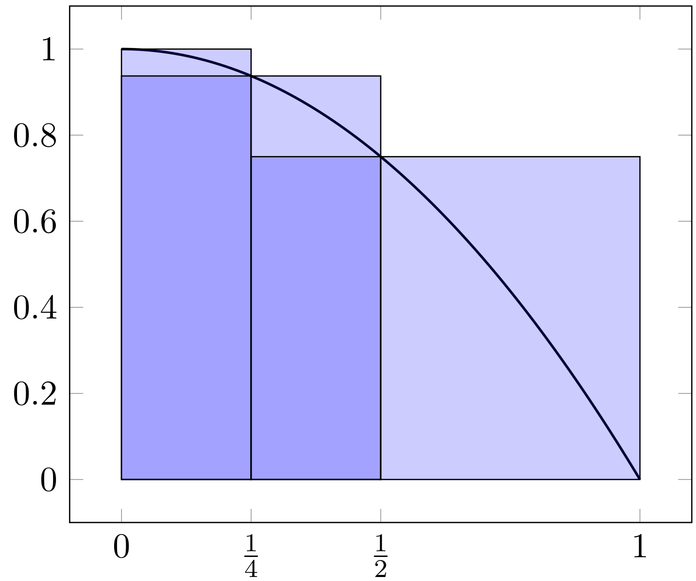


Figure 24: The upper (light blue) and lower (dark blue) Riemann sums for the function $f(x) = 1-x^2$ on the interval $[0, 1]$.

Example 5.13

Determine $U_f(P)$ and $L_f(P)$ if $f(x) = 1 - x^2$ on $[0, 1]$ and $P = \{0, 1/4, 1/2, 1\}$.

Solution. Conveniently, our function $f(x) = 1 - x^2$ is a decreasing function and is actually continuous. This means that not only do the supremum and infimum exist, but they are in fact maxima and minima. Furthermore, a decreasing function will attain its maximum at the left-endpoint, while attaining its min at the right end-point. This will make computing the upper and lower Riemann sums much simpler.

We have three subintervals: $[0, 1/4], [1/4, 1/2]$, and $[1/2, 1]$. To facilitate our computations, I have made the following table:

	$[0, 1/4]$	$[1/4, 1/2]$	$[1/2, 1]$
$x_{i+1} - x_i$	$1/4$	$1/4$	$1/2$
$\sup f(x)$	1	$15/16$	$3/4$
$\inf f(x)$	$15/16$	$3/4$	0

We can use this information to compute our upper and lower sums as follows:

$$\begin{aligned} U_f(P) &= \sum_{i=1}^{|P|-1} \left[\sup_{x \in [x_i, x_{i+1}]} f(x) \right] (x_i - x_{i+1}) \\ &= \left(1 \times \frac{1}{4}\right) + \left(\frac{15}{16} \times \frac{1}{4}\right) + \left(\frac{3}{4} \times \frac{1}{2}\right) = \frac{55}{64} \\ L_f(P) &= \sum_{i=1}^{|P|-1} \left[\inf_{x \in [x_i, x_{i+1}]} f(x) \right] (x_i - x_{i+1}) \\ &= \left(\frac{15}{16} \times \frac{1}{4}\right) + \left(\frac{3}{4} \times \frac{1}{4}\right) + \left(0 \times \frac{1}{2}\right) = \frac{27}{64} \end{aligned}$$

These numbers are quite different. ■

A rather important result is the following:

Lemma 5.14

Let f be a bounded function on $[a, b]$.

1. If P is a partition of $[a, b]$, then $L_f(P) \leq U_f(P)$,
2. If $P \subseteq Q$ are partitions (so Q is a refinement of P), then

$$L_f(P) \leq L_f(Q) \leq U_f(Q) \leq U_f(P),$$

3. For any partitions P, Q (not necessarily related) we have $L_f(P) \leq U_f(Q)$.

What Lemma 5.14 tells us is that:

- Refining a partition will cause the lower sum to get larger, while the upper sum to get smaller,
- The lower sum is *always* smaller than the upper sum.

The proof of this is a simple exercise that we leave for the reader.

What we are interested in is whether we can make the upper and lower sums of a function get really close to each other by choosing appropriate partitions. These seems like a limit question! Can we make the limit of $L_f(P)$ approach that of $U_f(P)$ as we change partitions? The problem is that the notion of defining limits in this case turns out to be extraordinarily tricky. As a result, we will put the idea of limits temporarily out of our minds and proceed as follows:

Definition 5.15

Let f be a bounded function on the interval $[a, b]$. We define the *lower integral* $\underline{I}_a^b(f)$ and *upper integral* $\overline{I}_a^b(f)$ as

$$\underline{I}_a^b(f) = \sup_P L_f(P)$$
$$\overline{I}_a^b(f) = \inf_P U_f(P)$$

If $\underline{I}_a^b(f) = \overline{I}_a^b(f)$ then we say that f is *integrable* on $[a, b]$. In that case, we write

$$\int_a^b f = \underline{I}_a^b(f) = \overline{I}_a^b(f),$$

called the *definite integral of* f *on* $[a, b]$.

Note that by Lemma 5.14 we always have that $\underline{I}_a^b(f) \leq \overline{I}_a^b(f)$, so if the upper and lower integrals are equal, their value is unique. Furthermore, when it is important to emphasize the particular variable, we might sometimes write

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

Of course, any other choice of variable other than x would also have served.

Theorem 5.16

1. **Additivity of Domain:** If f is integrable on $[a,b]$ and $[b,c]$ then f is integrable on $[a,c]$ and
$$\int_a^c f(x)dx = \int_a^b f(x)dx + \int_b^c f(x)dx.$$
2. **Additivity of Integral:** If f,g are integral on $[a,b]$ then $f+g$ is integrable on $[a,b]$ and
$$\int_a^b [f(x)+g(x)]\,dx = \int_a^b f(x)dx + \int_a^b g(x)dx.$$
3. **Scalar Multiplication:** If f is integrable on $[a,b]$ and $c \in \mathbb{R}$, then cf is integrable on $[a,b]$ and
$$\int_a^b cf(x)dx = c\int_a^b f(x)dx.$$
4. **Inherited Integrability:** If f is integrable on $[a,b]$ then f is integrable on any subinterval $[c,d] \subseteq [a,b]$.
5. **Monotonicity of Integral:** If f,g are integrable on $[a,b]$ and $f(x) \leq g(x)$ for all $x \in [a,b]$ then
$$\int_a^b f(x)dx \leq \int_a^b g(x)dx.$$
6. **Subnormality:** If f is integrable on $[a,b]$ then $|f|$ is integrable on $[a,b]$ and satisfies
$$\left|\int_a^b f(x)dx\right| \leq \int_a^b |f(x)|dx.$$

These are all left to the student as an exercise.

Theorem 5.17

All continuous functions are integrable.

Again, this result is beyond the scope of this course so we will not attempt it.

5.3.3 Other Riemann Sums

One approach often taken in defining the integral is to take a 'limit' as the length of partitions goes to zero, or alternatively as the number of partition points goes to infinity. We have avoided this technique because properly understanding what such a limit does turns out to be exceptionally complicated. There is special case where one might only consider partitions of uniform length, which might be something like the following: [21]

[21] If you are interested in the buzzwords, the limit can be computed because the set of partitions on an interval $[a,b]$ form a *directed set*. The fact that uniform partitions can be used follows form the fact that uniform partitions are

Assume *a priori* that f is integrable on the interval $[a,b]$. Since f is integrable, we know that the infimum of its upper sums is precisely the value $\int_a^b f$. For each $n \in \mathbb{N}$, define a partition P_n with 2^n+1 elements (hence containing 2^n subintervals) defined by

$$\begin{aligned} P_n &= \left\{a, a+\left(\frac{b-a}{2^n}\right), a+2\left(\frac{b-a}{2^n}\right), \ldots, a+(2^n-1)\left(\frac{b-a}{2^n}\right), b\right\} \\ &= \left\{a+i\left(\frac{b-a}{n}\right) : i=0,\ldots,2^n\right\}. \end{aligned}$$

Notice that[22] $P_{n-1} \subseteq P_n$ so that at each step we get a refinement of the previous partition. If $S_{i,n} = \left[a+i\left(\frac{b-a}{2^n}\right), a+(i+1)\left(\frac{b-a}{2^n}\right)\right]$ is a subinterval of P_n, let $M_{i,n} = \sup_{x\in S_{i,n}} f(x)$, so that the upper Riemann sum is just

$$U_f(P_n) = \sum_{i=0}^{2^n-1} M_{i,n}\left(\frac{b-a}{2^n}\right) = \left(\frac{b-a}{2^n}\right)\sum_{i=1}^{2^n-1} M_{i,n}.$$

Recall from Lemma 5.14 that since $P_{n-1} \subseteq P_n$ then $U_f(P_{n-1}) \leq U_f(P_n)$. Since the infimum of $U_f(P)$ is the integral and the $U_f(P_n)$ are decreasing, then[23]

$$\int_a^b f = \lim_{n\to\infty} \frac{b-a}{2^n}\sum_{i=0}^{2^n-1} M_{i,n}.$$

Example 5.18

If $f(x)=x$ on $[a,b]$, determine $\int_a^b f$.

Proof. Since f is continuous, we can use a Riemann sum to compute the definite integral. Furthermore, since $f(x)=x$ is monotonically increasing, it maximum on an interval will always occur at

cofinal amongst all partitions. Alternatively, if one assigns 'heights' to each partition (corresponding to the length of a maximal subinterval), one can always choose a uniform representative from the collection, which turns out to be the coarsest such partition amongst those partitions of the same height.

[22] The reason why we broke $[a,b]$ into 2^n subintervals is so that $P_{n-1} \subseteq P_n$. Notice that if we had only broken $[a,b]$ into n subintervals, there would be no relationship between partitions for subsequent n.

[23] We are cheating a bit here. We need to argue that the $U_f(P_n)$ do not 'get stuck' before reaching the infimum, but it's not too hard to convince ourselves that this is true.

the right endpoint. The upper Riemann sum for P_n is thus

$$\begin{aligned} U_f(P_n) &= \frac{b-a}{2^n}\sum_{i=0}^{2^n-1} f\left(a+(i+1)\left(\frac{b-a}{2^n}\right)\right) \\ &= \frac{b-a}{2^n}\sum_{i=0}^{2^n-1}\left[a+(i+1)\left(\frac{b-a}{2^n}\right)\right] \\ &= \frac{b-a}{2^n}\left[a\left(\sum_{i=0}^{2^n-1}1\right)+\frac{b-a}{2^n}\left(\sum_{i=0}^{2^n-1}(i+1)\right)\right] \\ &= \frac{b-a}{2^n}\left[a2^n+(b-a)\frac{2^n+1}{2}\right] \\ &= (ab-a^2)+(b-a)^2\left(\frac{2^n+1}{2^n}\right). \end{aligned}$$

It is easy to check that $\lim_{n\to\infty}\frac{2^n+1}{2^{n+1}}=\frac{1}{2}$, so

$$\begin{aligned} \int_a^b x\,\mathrm{d}x = \lim_{n\to\infty}U_f(P_n) &= (ab-a^2)+(b-a)^2\lim_{n\to\infty}\frac{2^n+1}{2^n} \\ &= ab-a^2+\frac{1}{2}(b-a)^2 = ab-a^2+\frac{1}{2}(b^2-2ab+a^2) \\ &= \frac{1}{2}(b^2-a^2). \end{aligned}$$

□

5.4 Anti-Derivatives

Anti-differentiation is the reverse process of differentiation; that is, if I give you a function $f(x)$ then our goal is to find a function $F(x)$ such that $F'(x)=f(x)$. To this end, we have the formal definition:

Definition 5.19

Given a function f on $[a,b]$, we say that a function F is an *anti-derivative* of f if $F'(x)=f(x)$ for all $x\in[a,b]$.

A decent part of this course will be dedicated to determining anti-derivatives of functions (though this is not yet obvious). In fact, just using the properties of differentiation, we can immediately infer a few results about anti-derivatives. Since the derivative is linear, we have

$$\frac{\mathrm{d}}{\mathrm{d}x}cf(x)=cf'(x), \qquad \frac{\mathrm{d}}{\mathrm{d}x}\left[f(x)+g(x)\right]=f'(x)+g'(x)$$

and this tells us that anti-differentiation will also be linear. To see this, let $F(x)$ and $G(x)$ be anti-derivatives of $f(x)$ and $g(x)$, so that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}\left[cF(x)\right] &= cF'(x)=cf(x)=c\frac{\mathrm{d}}{\mathrm{d}x}F(x) \\ \frac{\mathrm{d}}{\mathrm{d}x}\left[F(x)+G(x)\right] &= F'(x)+G'(x)=f(x)+g(x)=\frac{\mathrm{d}}{\mathrm{d}x}F(x)+\frac{\mathrm{d}}{\mathrm{d}x}G(x). \end{aligned}$$

Example 5.20

Compute the anti-derivative of $f(x) = 5x^4$.

Solution. We know that polynomials differentiate to give polynomials, so let's assume that $F(x) = x^n$ for some n. For $F(x)$ to be an anti-derivative of $f(x)$ it must be that $F'(x) = nx^{n-1} = f(x) = 5x^4$. It is not too hard to see that $n = 5$ works, so that the anti-derivative of $f(x) = 5x^4$ is $F(x) = x^5$. ■

The previous example was exceptionally easy to solve because of the coefficient 5 in the monomial term. If that term had not been there, then we would just artificially add it. For example, the anti-derivative of $3x^4$ may be computed by realizing that

$$3x^4 = 3 \cdot \frac{5}{5} x^4 = \frac{3}{5} 5x^4.$$

Since scalar multiples pass through derivatives, we hypothesize that the anti-derivative of $3x^4$ is $\frac{3}{5}x^5$ and a quick computation confirms this.

Note that the anti-derivative of a function is not unique, since we may add any constant to a function to find a new anti-derivative. For example, assume that $F(x)$ is an anti-derivative for $f(x)$ so that $F'(x) = f(x)$. Define a new function $F_c(x) = F(x) + c$ for any constant $c \in \mathbb{R}$. We then have that

$$\frac{\mathrm{d}}{\mathrm{d}x} F_c(x) = \frac{\mathrm{d}}{\mathrm{d}x} [F(x) + c] = F'(x) = f(x).$$

so that $F_c(x)$ is also an anti-derivative. This implies that there are an entire real number's worth of functions which are the anti-derivative of a function. More concretely, Example 5.20 shows that x^5 is the anti-derivative of $5x^4$, but a quick computation easily shows that $x^5 + c$ also differentiates to $5x^4$ for any constant c.

Of course, our knowledge of the Mean Value Theorem immediately tells us the following:

Corollary 5.21

If $f(x)$ is a function with an anti-derivative $F(x)$, then $F(x)$ is unique up to an additive constant; that is, if $\tilde{F}(x)$ is any other anti-derivative of $f(x)$, then there exists some constant c such that $F(x) = \tilde{F}(x) + c$.

Example 5.22

Compute $f(x)$ if $f''(x) = \sqrt{x} + \sin(x) - e^x$.

Solution. Notice that the second derivative is given, so we will have to compute the anti-derivative twice. Here in particular it is essential to recall that anti-derivatives are only defined up to additive constants. According to our table above, we have the following derivative - anti-derivative pairs:

$$\sqrt{x} = \frac{\mathrm{d}}{\mathrm{d}x} \frac{2}{3} x^{3/2}, \qquad \sin(x) = \frac{\mathrm{d}}{\mathrm{d}x} (-\cos(x)), \qquad e^x = \frac{\mathrm{d}}{\mathrm{d}e}{}^x$$

so that the first derivative (given by the anti-derivative of $f''(x)$ is

$$f'(x) = \frac{2}{3}x^{3/2} - \cos(x) + e^x + c$$

for some constant c. It is important to include the c here since when we take another anti-derivative, it will contribute to the solution. Once again, the anti-derivatives are given by

$$x^{3/2} = \frac{\mathrm{d}}{\mathrm{d}x}\frac{2}{5}x^{5/2}, \qquad \cos(x) = \frac{\mathrm{d}}{\mathrm{d}x}\sin(x), \qquad e^x = \frac{\mathrm{d}}{\mathrm{d}e}^x, \qquad c = \frac{\mathrm{d}}{\mathrm{d}c}x,$$

so that $f(x)$ is

$$f(x) = \frac{4}{15}x^{5/2} - \sin(x) + e^x + cx + d$$

where c, d are constants. ■

If additional criteria are supplied, such as the value of $f(x)$ (or its derivatives) at particular points, then a truly unique solution may be identified.

Example 5.23

Using your solution to Example 5.22, compute the unique anti-derivative which satisfies $f(0) = 10$ and $f'(0) = 0$.

Solution. Our above example showed that $f'(x) = \frac{2}{3}x^{3/2} - \cos(x) + e^x + c$. By substituting $x = 0$ into this we get

$$0 = f'(0) = \frac{2}{3}0^{3/2} - \cos(0) + e^0 + c = c$$

so that $c = 0$. Thus $f(x) = \frac{4}{15}x^{5/2} - \sin(x) + e^x + d$. Substituting $x = 0$ into this gives

$$10 = f(0) = \frac{4}{15}0^{5/2} - \sin(0) + e^0 + d = 1 + d$$

so that $d = 9$. In conclusion, the corresponding $f(x)$ is

$$f(x) = \frac{4}{15}x^{5/2} - \sin(x) + e^x + 9.$$ ■

Notice that Example 5.23 required two conditions to specify the number of constants. In general, if one is given the n^{th} derivative of a function, one needs to specify n-conditions to uniquely determine the function.

5.5 The Fundamental Theorem of Calculus

In this section, we will make the connection between the theory of integration and the theory of differentiation, by means of the *Fundamental Theorem of Calculus*. Let f be an integrable function on $[a, b]$. By Theorem 5.16 we know that for any subinterval $[c, d] \subseteq [a, b]$, f is also integrable on $[c, d]$. In particular, let's fix the left endpoint at a. Now for each $x \in [a, b]$, we have an integrable function on $[a, x]$ and hence the definite integral exists and produces a number. Thus we have a function

$$F(x) = \int_a^x f(s)ds$$

which assigns to each point x the value of the definite integral on $[a, x]$. Analogous to differentiation, wherein we had a function f on $[a, b]$ and created a function f' on $[a, b]$ with interesting properties, we now have the function F on $[a, b]$, and we are interested in its properties.

Theorem 5.24: Fundamental Theorem of Calculus

1. If f is integrable on $[a, b]$ then $F(x) = \int_a^x f(s)ds$ is continuous on $[a, b]$. Moreover, F is differentiable at any point where f is continuous, and in this case F is an anti-derivative of f.
2. If f is integrable on $[a, b]$, and F be a continuous anti-derivative of f which is differentiable at all but finitely many points, then

$$\int_a^b f(s)ds = F(b) - F(a).$$

Proof. 1. We begin by showing continuity, for which it is sufficient to show that

$$\lim_{x \to y} [F(x) - F(y)] = 0.$$

By Theorem 5.16 we know that

$$F(x) + \int_x^y f(s)ds = \int_a^x f(s)ds + \int_x^y f(s)ds = \int_a^y f(s)ds = F(y)$$

which we can re-arrange to get $F(y) - F(x) = \int_x^y f(s)ds$. Since f is necessarily bounded, there exists $M > 0$ such that $|f(s)| \leq M$ for all $s \in [a, b]$. Applying Theorem 5.16 we see that

$$|F(y) - F(x)| = \left| \int_x^y f(s)ds \right| \leq \int_x^y |f(s)|ds \leq \int_x^y M ds = M|y - x|.$$

If $\epsilon > 0$ is given, continuity can thus be demonstrated by choosing $\delta = \frac{\epsilon}{M}$.

We now wish to show that wherever f is continuous, F is differentiable and $F' = f$. Assume then that f is continuous at x, for which we would like to show that

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

We will proceed by using the ϵ-δ definition of the limit. Let $\epsilon > 0$ be given. Since f is continuous, there exists a δ such that for all $|x - y| < \delta$ we have $|f(x) - f(y)| < \epsilon$. If $|h| < \delta$ then

$$\frac{F(x+h) - F(x)}{h} - f(x) = \frac{1}{h} \int_x^{x+h} f(s)ds - f(x) = \frac{1}{h} \int_x^{x+h} [f(s) - f(x)]\, ds.$$

In the last equality, we have used the fact that we are integrating with respect to s, so $f(x)$ is a constant, and $\int_x^{x+h} f(x)ds = f(x)h$. Applying Theorem 5.16 we have

$$\left|\frac{F(x+h)-F(x)}{h} - f(x)\right| = \frac{1}{|h|}\left|\int_x^{x+h} [f(s)-f(x)]\right| ds \leq \frac{1}{|h|}\int_x^{x+h} |f(s)-f(x)|dx.$$

Now since $s \in [x, x+h]$ (or $[x+h, x]$ if $h < 0$) and $|h| < \delta$, then by continuity we know $|f(s) - f(x)| < \epsilon$, and so

$$\left|\frac{F(x+h)-F(x)}{h} - f(x)\right| \leq \frac{1}{|h|}\int_x^{x+h} |f(s)-f(x)|dx \leq \frac{\epsilon}{|h|}\int_x^{x+h} ds = \epsilon.$$

We thus conclude that

$$\lim_{h\to 0}\left|\frac{F(x+h)-F(x)}{h} - f(x)\right| = 0.$$

By the Squeeze theorem, this limit is zero if and only if the object within the absolute values is zero, and so differentiability follows.

2. Let $\{y_1, \ldots, y_n\}$ be the points where F fails to be differentiable. Let P' be an arbitrary partition, and consider the partition

$$P = P' \cup \{y_1, \ldots, y_n\} = \{x_1, \ldots, x_n\}.$$

Notice that F is continuous on each subinterval $[x_i, x_{i+1}]$ and differentiable on (x_i, x_{i+1}) and so by the Mean Value Theorem, there exists $\theta_i \in [x_i, x_{i+1}]$ such that

$$F(x_{i+1}) - F(x_i) = F'(\theta_i)(x_{i+1} - x_i) = f(\theta_i)(x_{i+1} - x_i)$$

where in the last inequality, we have used the fact that we know $F'(\theta_i) = f(\theta_i)$ by Part 1. Now

$$\begin{aligned} F(b) - F(a) &= [F(x_n) - F(x_{n-1})] + [F(x_{n-1}) - F(x_{n-2})] + \cdots \\ &\quad + [F(x_3) - F(x_2)] + [F(x_2) - F(x_1)] \\ &= \sum_{i=1}^{n-1} f(\theta_i)(x_{i+1} - x_i). \end{aligned}$$

But certainly we have $\inf_{x\in[x_i,x_{i+1}]} f(x) \leq f(\theta_i) \leq \sup_{x\in[x_i,x_{i+1}]} f(x)$ which implies that

$$L_f(P) \leq F(b) - F(a) \leq U_f(P).$$

This inequality is true regardless of the partition, and since the function is integrable, the resulting number $F(b) - F(a)$ must be the value of the integral, as required.

□

Remark: Effectively, the fundamental theorem of calculus indicates that differentiation and integration are 'inverses' of one another. This is not exactly true, as the following example demonstrates:

Example 5.25

Let f be a continuous function on $\mathbb{R}$. Evaluate
$$\frac{\mathrm{d}}{\mathrm{d}x}\int_0^x f(t)dt - \int_0^x \frac{\mathrm{d}}{\mathrm{d}t} f(t)dt.$$

Solution. If integration and differentiation were truly inverses, then this would simply evaluate to zero. However, let us be a bit more prudent in our evaluation. By the Fundamental Theorem of Calculus, $F(x) = \int_0^x f(t)dt$ is an anti-derivative of $f(x)$, and hence
$$\frac{\mathrm{d}}{\mathrm{d}x}\int_0^x f(t)dt = f(x).$$
On the other hand, $f(x)$ is clearly an anti-derivative of $f'(x)$, and so
$$\int_0^x \frac{\mathrm{d}}{\mathrm{d}t} f(t)dt = \int_0^x f'(t)dt = f(x) - f(0).$$
Hence the difference between these two terms comes out to $f(0)$; that is, they differ up to a constant. ∎

Example 5.26

Compute $\displaystyle\int_{-\pi}^{\pi} \sin(t)dt$, and $\displaystyle\int_0^1 \frac{1}{1+x^2}dx$.

Solution. We know that $F(x) = -\cos(x)$ is an anti-derivative of $\sin(t)$, so by the Fundamental Theorem of Calculus we have
$$\int_{-\pi}^{\pi} \sin(t)dt = F(\pi) - F(-\pi) = -\cos(\pi) - (-\cos(-\pi)) = 0.$$
Similarly, we know that $G(x) = \arctan(x)$ is an anti-derivative of $\frac{1}{1+x^2}$, and hence
$$\int_{-1}^{0} \frac{1}{1+x^2}dx = \arctan(0) - \arctan(-1) = -\frac{\pi}{4}.$$ ∎

5.6 Computing Areas with Integrals

The whole motivation for introducing integrals was to compute the areas under curves. Of course, Example 5.26 demonstrates a problem: It is possible for our areas to be 0, or even negative! The reason for this is that the integral as defined computes the *signed area* of the function, and hence assigns negative area to any area that occurs beneath the x-axis. The way to fix this is to use absolute values:

Example 5.27

Determine the unsigned (that is, total) area under the graph of $\sin(x)$ on the interval $[-\pi, \pi]$.

Solution. Example 5.26 told us that the signed area was exactly zero despite the fact that $\sin(x)$ is not identically the zero function. To compute the total area, we shall compute the integral $\int_{-\pi}^{\pi} |\sin(x)| dx$. We deal with the absolute value in precisely the same manner that we always deal with absolute values: We break it into cases:

$$|\sin(x)| = \begin{cases} \sin(x) & 0 \leq x \leq \pi \\ -\sin(x) & -\pi \leq x < 0 \end{cases}.$$

By additivity of domain, we thus get

$$\begin{aligned} \int_{-\pi}^{\pi} |\sin(x)| dx &= \int_{-\pi}^{0} |\sin(x)| dx + \int_{0}^{\pi} |\sin(x) dx \\ &= -\int_{-\pi}^{0} \sin(x) dx + \int_{0}^{\pi} \sin(x) dx \\ &= -[-\cos(x)]_{x=-\pi}^{0} + [-\cos(x)]_{x=0}^{\pi} \\ &= \cos(0) - \cos(\pi) - \cos(\pi) + \cos(0) = 4. \end{aligned}$$

■

We can also use the integral to find the area between two curves. Given two curves f, g on $[a, b]$, the area between f and g can be computed as

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

We note though that this is still a *signed area*. In particular, anywhere where $f > g$ will be assigned a positive area, while area where $f < g$ will be given a negative area. Of course, the total area can be computed by $\int_a^b |f(x) - g(x)| dx$.

Example 5.28

Find the area bounded between the functions $f(x) = \sqrt{x}$ and $g(x) = x^2$.

Solution. Notice that we were not explicitly given an interval over which to integrate. The reason is that if one sketches the graphs of f and g, there is only one area that can be said to be enclosed by the two functions. Consequently, we need to determine where the appropriate intersections occur. This can be done by equation $f(x) = g(x)$.

Setting $x^2 = \sqrt{x}$, one can easily to solve to find that the intercepts occur at $x = 0$ and $x = 1$. Furthermore, on $[0, 1]$ we have that $\sqrt{x} > x^2$, and hence our area is given as

$$\begin{aligned} \int_0^1 \left[\sqrt{x} - x^2\right] dx &= \left[\frac{2}{3} x^{3/2} - \frac{1}{3} x^3\right]_0^1 \\ &= \frac{2}{3} - \frac{1}{3} = \frac{1}{3}. \end{aligned}$$

■

5.7 Indefinite Integrals

We have seen that integration and differentiation are spiritual inverses of one another, up to an additive constant. In particular, for an integral f on $[a, b]$ we saw that $F(x) = \int_a^x f(x) dx$ was an anti-derivative of f, and for any anti-derivative G of f we have

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

Anti-derivatives are unique up to constants, so there exists some constant C such that $F = G + C$, with F playing a particularly nice representative. However, the constant seems rather artificial: we know that the anti-derivative of x^3 is $\frac{1}{4}x^4 + C$, but the meat-and-bones lies with the $\frac{1}{4}x^4$ term, not the constant. Hence our goal for this section is to represent the entire class of anti-derivatives, something called the *indefinite integral*.

The indefinite integral does not concern itself with upper and lower bounds of integration: our goal is to encompass an entire class of functions and imposing bounds forces us to look at particular representatives. Consequently, we denote the indefinite integral as the usual integral, albeit with the bounds omitted:

$$\int f(x)dx.$$

Example 5.29

Determine the following indefinite integrals:

1. $\displaystyle\int \left(\frac{x^4 + 2x^2 + 1}{x^3}\right) dx,$
2. $\displaystyle\int \sin(2x)\cos(2x)dx,$
3. $\displaystyle\int f(x)f'(x)dx.$

Solution. In time, we will learn more systematic ways of determining these integrals, but for now we will need to use the clever part of our brains to find appropriate classes of anti-derivatives.

1. Notice that we can re-write the integrand as

$$\frac{x^4 + 2x^2 + 1}{x^3} = x + \frac{2}{x} + \frac{1}{x^3}.$$

We are well acquainted with the functions which yield these as derivatives, and we get

$$\int \frac{x^4 + 2x^2 + 1}{x^3} = \int \left(x + \frac{2}{x} + \frac{1}{x^3}\right) dx = \frac{1}{2}x^2 + 2\log(x) - \frac{1}{2x^2} + C.$$

This yields to use our first curious case: Integrating a rational function resulted in a logarithm.

2. The important step here is to realize that $\frac{1}{2}\sin(4x) = \sin(2x)\cos(2x)$, hence our integral becomes

$$\int \sin(2x)\cos(2x)dx = \frac{1}{2}\int \sin(4x)dx = -\frac{1}{8}\cos(4x) + C.$$

3. This problem is a little more abstract: We need to find a function which differentiates to $f(x)f'(x)$. If we think hard, we see that $\frac{d}{dx}[f(x)]^2 = 2f(x)f'(x)$, so by dividing by 2 we will get the desired integrand. Applying the Fundamental Theorem of Calculus, we thus get
$$\int f(x)f'(x)dx = \int \frac{d}{dx}f(x)^2 dx = f(x)^2 + C. \tag{5.2}$$
■

Remark:

1. Part 3 can lead to rather terrible notation, not seen in mathematics so much as physics. Remember that physicists love to treat $\frac{dy}{dx}$ as a proper fraction, treating dy and dx as indivisible atoms which form the whole of the derivative. If one examines Equation (5.2) we see that treating dx as a fraction could cause them to cancel: consequently, some physicists might write
$$\int \frac{d}{dx}g(x)dx = \int dg(x) = g(x).$$
There is a sense in which this is valid, and it leads to an incredible world of complex and beautiful mathematics, but not one that is easily explained. For now, let us take this as a sign of physicist abuse and disregard it as little more than such.

2. While we are on the topic of bad notation, another convention sometimes seen is to write $\int dx f(x)$. While I cannot think of anything which is technically wrong with this notation, I find it distasteful.

Indefinite integrals, in one form or another, are essential in solving differential equations. These are equations of the form
$$\frac{dy}{dx} = f(y, x)$$
and are an exceptionally useful study in their own right. We will not invest much time on them in this course, but their importance is so great that they warrant many courses dedicated explicitly to their study. As a very simple example, consider an equation that one often sees in high-school physics. A particle thrown upwards from the surface of the Earth with initial height y_0 and velocity v_0 has its height $y(t)$ dictated as a function of time by the equation
$$y = -\frac{1}{2}gt^2 + v_0 t + y_0,$$
where g is the gravitational acceleration of the Earth. More generally, a particle given an acceleration a will travel a distance $y(t) = \frac{1}{2}at^2 + v_0 t + y_0$. How do we arrive at this equation? If we look back at Section 3.3.1, one can show that $\frac{d^2y}{dt^2} = a$. Using anti-derivatives, we can compute that $\frac{dy}{dt} = at + C_1$ and $y(t) = \frac{1}{2}at^2 + C_1 t + C_2$. This is not quite what we were after, but we can solve for the constants C_1, C_2 by using our *initial conditions*. Knowing that at time 0 our height was y_0, we have
$$y_0 = y(0) = 0 + 0 + C_2$$
so that $C_2 = y_0$. Similarly, at time 0 our speed was v_0, giving
$$v_0 = y'(0) = 0 + C_1,$$
so that $C_1 = v_0$. Combining these together yields $y(t) = \frac{1}{2}at^2 + v_0 t + y_0$ as required.

Notes on Notation: At this point, it is probably worth pointing out that we have been dramatically overloading our use of the integral sign. We have thus far seen three different objects: If f is integrable on $[a,b]$ then

$$\int_a^b f(s)ds, \qquad F(x) = \int_a^x f(s)ds, \qquad \int f(s)ds.$$

The first is simply a *number* which represents the signed area under the function f; the second is a *function* which assigns to each x the area under the function from a to x; the third is an infinite *family of functions*, all representing anti-derivatives of f. They are all intimately related to be certain, but each has a very different lifestyle. One must be careful not to confuse the relationships.

6 Integration Techniques

6.1 Integration by Substitution

Having seen that integration and differentiation are essentially inverses, we would like to develop some techniques and rules to compute integrals. It should be unsurprising that those rules will arise as the "inverse" operations of the rules obtained from differential calculus. We recall the chain rule of differential calculus tells us that

$$\frac{\mathrm{d}}{\mathrm{d}x} f(g(x)) = f'(g(x))g'(x)$$

hence by applying the fundamental theorem of calculus, we see that

$$\int f'(g(x))g'(x)dx = f(g(x)) + C. \tag{6.1}$$

Unfortunately, the majority of times nature will conspire against us and not write our integrand so plainly as $f'(g(x))g'(x)$. Hence we develop some techniques to make life simpler. Our strategy is as follows:

1. Look to see if we can find the occurrence of a function and its derivative. In (6.1) above, we are looking for the function $g(x)$, since it occurs in the argument of $f(x)$ and its derivative appears as $g'(x)$.

2. Define a new variable, $u = g(x)$ so that $\frac{\mathrm{d}u}{\mathrm{d}x} = g'(x)$. We often write this second equation as $du = g'(x)dx$.

3. Now replace everything until we only have dependencies on u instead of on x. Namely, recognize that

$$\underbrace{f'(g(x))}_{f(u)}\underbrace{g'(x)dx}_{du} = f'(u)du$$

4. Using the fundamental theorem of calculus, evaluate our new integral:

$$\int f'(u)du = f(u) + C.$$

5. We now have our solution, but it is in terms of the variable u. This is not a problem since we know that $u = g(x)$, so we just make this substitution to get our final solution

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

Let's try a simple example:

Example 6.1

Determine $\int \sin(2x)\cos(2x)dx$.

Solution. We can choose either $u = \sin(2x)$ or $u = \cos(2x)$. Indeed, let's see what happens if we choose either. If $u = \sin(2x)$ then $du = 2\cos(2x)dx$ and our integral becomes

$$\int \sin(2x)\cos(2x)dx = \frac{1}{2}\int u du = \frac{1}{4}u^2 + C = \frac{1}{2}\sin^2(2x) + C.$$

On the other hand, if we choose $u = \cos(2x)$ then $du = -\frac{1}{2}\sin(2x)$ and we get

$$\int \sin(2x)\cos(2x)dx = -\frac{1}{2}\int u du = -\frac{1}{4}\cos^2(2x) + C.$$

Of course, we know that $\sin^2(2x)$ and $\cos^2(2x)$ are very different functions, and here we see the importance of the constants.

To add to this bit of noise, notice that in Example 5.29 we calculated $\int \sin(x)\cos(x) = -\frac{1}{8}\cos(4x) + C$. ■

Example 6.2

For $a, b \neq 0$, compute $\displaystyle\int \frac{x^{n-1}}{\sqrt{a + bx^n}}dx$.

Solution. Following the above program, our first step should be to identify a function and its derivative. The fact that there is an x^n and an x^{n-1} is a pretty good sign. Note that since constants do not affect the integration, we can make our lives *even easier* if we define $u = a + bx^n$ so that $du = bnx^{n-1}dx$. Unfortunately, there is no $bnx^{n-1}dx$ in the integrand, but there is an $x^{n-1}dx$. Since these are related only up to a constant, we can divide both sides to find that $x^{n-1}dx = \frac{1}{bn}du$. Adding our substitutions we then get

$$\int \frac{x^{n-1}}{\sqrt{a + bx^n}}dx = \int \frac{\frac{1}{bn}du}{\sqrt{u}} = \frac{1}{bn}\int \frac{1}{\sqrt{u}}du.$$

This is now a very simple integral to calculate, and indeed we find that

$$\frac{1}{bn}\int \frac{1}{\sqrt{u}}du = \frac{2}{bn}\sqrt{u} + C.$$

We need this to be in terms of x rather than u, so we recall that $u = a + bx^n$ to finally find that

$$\int \frac{x^{n-1}}{\sqrt{a + bx^n}}dx = \frac{2}{bn}\sqrt{a + bx^n} + C.$$ ■

Substitution is not just handy for applying the chain rule. It also allows us to "change variables."

Example 6.3

Compute $\displaystyle\int x\sqrt{x + 1}dx$.

Solution. Notice that if we could somehow switch the x and the $x+1$, this integral would actually be really easy, since then $(x+1)\sqrt{x} = x^{\frac{3}{2}} + x$. Normally in mathematics, if we want to do such a thing, we just define a new variable $u = x+1$ so that $x = (u-1)$ and then a similar trick to the one above will work.

Since we are working with an integral though, we must be a bit more careful. We shall still define $u = x+1$ with $x = (u-1)$, but we must also track the differentials. Luckily, in this case $du = dx$ and there is nothing to do. We thus get

$$\begin{aligned}\int x\sqrt{x+1}dx &= \int (u-1)\sqrt{u}du \\ &= \int (u^{\frac{3}{2}} - u)du \\ &= \frac{2}{5}u^{\frac{5}{2}} - \frac{1}{2}u^2 + C\end{aligned}$$ ■

Very quickly, let us remark how we must change definite integrals when we do integration. Consider the following integral

$$\int_0^1 \cos\left(\frac{\pi}{2}x\right) dx.$$

One can easily see compute this as

$$\int_0^1 \cos\left(\frac{\pi}{2}x\right) dx = \frac{2}{\pi}\sin\left(\frac{\pi}{2}x\right)\Big|_0^1 = \frac{2}{\pi}.$$

Alternatively though, one should recognize that integrating $\cos\left(\frac{\pi}{2}\right)$ from 0 to 1 is that same as integrating $\cos(x)$ from 0 to $\frac{\pi}{2}$. How does this reflect in the integral if we make such a substitution? Let $u = \frac{\pi}{2}x$ so that $du = \frac{\pi}{2}dx$. If we were to naively carry though with the integration we would get

$$\int_0^1 \cos\left(\frac{\pi}{2}x\right) dx = \frac{2}{\pi}\int_0^1 \cos(x)dx = \frac{2}{\pi}\sin(x)\Big|_0^1 = \frac{2}{\pi}\sin(1).$$

We actually got a different answer. The reason can be seen in the first equality above. We said that integrating $\cos\left(\frac{\pi}{2}x\right)$ should be the same as integrating $\cos(x)$ from 0 to $\frac{\pi}{2}$, but we failed to change the upper and lower limits of integration. To do this correct, we think of the upper and lower limits as corresponding to $x_L = 0$ and $x_U = 1$. Having made the substitution to u, we need to find the corresponding u_L and u_U. Since u is just a function of x, we then get that $u_L = \frac{\pi}{2}x_L = 0$ and $u_U = \frac{\pi}{2}x_U = \frac{\pi}{2}$. Now if we do our work, we get the correct answer.

Exercise: Let $n \geq 3$ be a positive *integer*. Using only substitution, compute

$$n\int_0^\infty \frac{r^{n-1}}{(r^2+1)^{\frac{n+2}{2}}}dr.$$

6.2 Integration by Parts

Just as Integration by Substitution was the inverse of the chain rule, Integration by Parts is the analog of the product rule. Namely, we know that if $u(x)$ and $v(x)$ are functions, then

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[u(x)\cdot v(x)\right] = u'(x)v(x) + u(x)v'(x).$$

Integrating and applying the fundamental theorem of calculus, we then find that

$$\begin{aligned}\int \frac{\mathrm{d}}{\mathrm{d}x}(u\cdot v)dx &= u(x)\cdot v(x)\\ &= \int v\frac{\mathrm{d}u}{\mathrm{d}x}dx + \int u\frac{\mathrm{d}v}{\mathrm{d}x}dx\\ &= \int vdu + \int udv\end{aligned}$$

for which we may re-arrange to find

$$\int udv = uv - \int vdu. \tag{6.2}$$

In the event of the definite integral, the bounds of integration can be carried throughout; that is,

$$\int_a^b udv = uv\Big|_a^b - \int_a^b vdu.$$

Our strategy should be as follows: Assume we are integrating a product $\int f(x)g(x)dx$. We want to choose a candidate for dv and for u. Since we will need to integrate dv, it is often best to choose a function which is easy to integrate:

1. First look at the integrand and see if we can apply substitution. If so, do not worry about integration by parts.
2. Choose dv and u (I often choose dv to be whichever function is easiest to integrate),
3. Compute that v by integrating $\int dv = \int f(x)dx$. Compute du by differentiating u.
4. Substitute all appropriate variables into (6.2).

This is just a general idea of how you should proceed. To give some insight as to what is happening, consider Equation (6.2) by omitting the uv-term:

$$\int u(x)v'(x)dx = -\int u'(x)v(x)dx.$$

The is the power of integration by parts: It effectively allows us to transfer the derivative from one function to another![24]

[24] For a higher class of 'generalized functions,' this is actually how one defines the derivative. As an example, the *Dirac delta function* $\delta(x)$ is defined to satisfy $\int f(x)\delta(x)dx = f(0)$. The definition of $\delta'(x)$ is done using integration by parts; that is $\int \delta'(x)f(x)dx = -\int \delta(x)f'(x)dx = -f'(0)$.

Example 6.4

Evaluate the integral $\int x \sin(x) dx$.

Solution. Following our program, I personally find that $\sin(x)$ is easier to integrate than just x so we set $dv = \sin(x)dx$. Furthermore, my note above suggests that we should set $u = x$ so everything works out. Computing du and v we find that

$$\begin{array}{ll} u = x & dv = \sin(x)dx \\ du = dx & v = -\cos(x). \end{array}$$

Plugging these into (6.2) we find that

$$\begin{aligned} \int x \sin(x) dx &= -x\cos(x) + \int \cos(x) dx \\ &= -x\cos(x) + \sin(x) + C. \end{aligned}$$

The best thing about integration is that you can always check your answers by differentiating. Try it! ■

Alternatively, there are time when the integrand does not look like a product, but we may still apply integration by parts.

Example 6.5

Compute $\displaystyle\int \log(x) dx$.

Solution. Looking at the integrand, there do not immediately appear to be two functions, so how can we apply integration by parts? The solution is to realize that $\log(x) = \log(x) \cdot 1$, so that the constant function $1(x) = 1$ is actually our second function. I think that 1 is really easy to integrate, so let us set $dv = 1 \cdot dx$ and $u = \log(x)$ to find that

$$\begin{array}{ll} u = \log(x) & dv = 1 \cdot dx \\ du = \frac{dx}{x} & v = x \end{array}$$

Substituting these values into (6.2) we find

$$\begin{aligned} \int \log(x) dx &= x \log(x) - \int \frac{x}{x} dx \\ &= x \log x - x + C. \end{aligned}$$

Again, try differentiating this to ensure that it works! ■

I believe it is worth noting at this point that while we have never come across a function which could not be differentiated, there are *lots* of functions which cannot be analytically integrated. Perhaps the most famous example is the function e^{x^2}. It can be shown (using a field called Differential Galois Theory) that this function has no elementary anti-derivative.

Let's conclude with an example that will require all of our skills thus far.

Example 6.6

Compute the integral $\int \cos(\log x) dx$.

Solution. The tricky thing to notice here is that it is not clear that either integration by parts or substitution will actually work. Nonetheless, a change of variables is the appropriate thing to do, as it will let us "pull" the function $\log(x)$ outside of the $\cos(x)$ as follows. Let $y = \log(x)$ so that $dy = \frac{dx}{x}$. Alternatively, we can write $dx = xdy$. This is silly though, since if we were to substitute this into our integral we would have variables in both x and y and that would be horrible. Instead, realize that since $y = \log x$ then $x = e^y$ and so $dx = e^y dy$. Putting this all together, we have

$$\int \cos(\log x) dx = \int e^y \cos(y) dy.$$

Now we can proceed via integration by parts. It seems to be that e^y is easiest to integrate, so let us set $dv = e^y dy$ and $u = \cos(y)$ yielding

$$\begin{array}{ll} u = \cos(y) & dv = e^y dy \\ du = -\sin(y) dy & v = e^y. \end{array}$$

Plugging these into (6.2) we get

$$\int e^y \cos(y) dy = e^y \cos(y) + \int e^y \sin(y) dy. \tag{6.3}$$

Looking at the integral $\int e^y \sin(y) dy$, we have failed to simplify the integral. However, let's just see what happens if we integrate by parts once more. Set $u = \sin(y)$ and $dv = e^y$ so that

$$\begin{array}{ll} u = \sin(y) & dv = e^y dy \\ du = \cos(y) dy & v = e^y. \end{array}$$

Equation (6.2) then tells us that

$$\int e^y \sin(y) dy = e^y \sin(y) - \int e^y \cos(y) dy. \tag{6.4}$$

Putting (6.3) and (6.4) together we have

$$\begin{aligned}
\int e^y \cos(y)dy &= e^y \cos(y) + \int e^y \sin(y) \\
&= e^y \cos(y) + e^y \sin(y) - \int e^y \cos(y)dy. \\
2\int e^y \cos(y)dy &= e^y \Big(\cos(y) + \sin(y)\Big) && \text{adding } \int e^y \cos(y) \text{ to both sides} \\
\int e^y \cos(y)dy &= \frac{1}{2}e^y \Big(\cos(y) + \sin(y)\Big) && \text{dividing by 2.}
\end{aligned}$$

Now we have our solution in terms of y, but need to revert to a solution in terms of x. Since $y = \log(x)$ we conclude

$$\int \cos(\log x)dx = \frac{1}{2}e^{\log x}\Big(\cos(\log x) + \sin(\log x)\Big) = \frac{x}{2}\Big(\cos(\log x) + \sin(\log x)\Big). \qquad \blacksquare$$

6.3 Integrating Trigonometric Functions

Integrating trigonometric functions often comes down to a combination of using the Pythagorean identities, or angle sum identities. Notice very conveniently that the three Pythagorean identities are

- $\sin^2(x) + \cos^2(x) = 1$, and $\frac{\mathrm{d}}{\mathrm{d}x}\sin(x) = \cos(x)$, $\frac{\mathrm{d}}{\mathrm{d}x}\cos(x) = -\sin(x)$,
- $\tan^2(x) + 1 = \sec^2(x)$, and $\frac{\mathrm{d}}{\mathrm{d}x}\tan(x) = \sec^2(x)$, $\frac{\mathrm{d}}{\mathrm{d}x}\sec(x) = \sec(x)\tan(x)$
- $1 + \cot^2(x) = \csc^2(x)$, and $\frac{\mathrm{d}}{\mathrm{d}x}\cot(x) = -\csc^2(x)$, $\frac{\mathrm{d}}{\mathrm{d}x}\csc(x) = -\csc(x)\cot(x)$.

Hence each identity only contains the components of a function and its derivatives. This will prove to be an exceptionally useful tool shortly.

The other trick to remember is the following two identities:

$$\sin^2(x) = \frac{1}{2}(1 - \cos(2x)), \qquad \text{and} \qquad \cos^2(x) = \frac{1}{2}(1 + \cos(2x)).$$

We can immediately use the above to equations to see that

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

Now what happens if our integrand is made up a several different trigonometric functions? For example, how would we deal with something of the form

$$\int \sin^2(x)\cos^3(x)dx?$$

The idea is to use the Pythagorean identities listed above to reduce the integrand to an expression of the form $g(f(x))f'(x)dx$, since from here one can easily apply the substitution method: indeed,

$$\begin{aligned} \int g(f(x))f'(x)dx &= \int g(u)du & \begin{aligned} u &= f(x) \\ du &= f'(x)dx \end{aligned} \\ &= G(f(x)) + C \end{aligned}$$

for some anti-derivative G of g.

Returning to our example, $\int \sin^2(x)\cos^3(x)dx$, we see that by writing

$$\cos^3(x) = \cos^2(x)\cos(x) = (1-\sin^2(x))\cos(x)$$

our integral becomes

$$\int \sin^2(x)\cos^3(x)dx = \int \sin^2(x)(1-\sin^2(x))\cos(x)dx.$$

Our integrand now consists entire of sine functions, and a single cosine function which will serve the role of allowing us to make the appropriate substitution. Hence setting $u = \sin(x)$ so that $du = \cos(x)dx$ we have

$$\begin{aligned} \int \sin^2(x)\cos^3(x)dx &= \int \sin^2(x)[1-\sin^2(x)]\cos(x)dx \\ &= \int u^2(1-u^2)du \\ &= \frac{1}{3}u^3 - \frac{1}{5}u^5 + C \\ &= \frac{1}{3}\sin^3(x) - \frac{1}{5}\sin^5(x) + C. \end{aligned}$$

Let's try another example:

Example 6.7

Determine the integral $\int \sec^4(x)\tan^5(x)dx$.

Solution. One again we conveniently have only secant and tangent functions, which are intimately related to one another through their derivatives. We have to now determine which function we are going to use as a substitution. If we make the substitution $u = \tan(x)$ then $du = \sec^2(x)dx$ and our integral becomes

$$\int \sec^4(x)\tan^5(x)dx = \int \sec^3(x)\tan^4(x)du = \int \sec^3(x)u^4du.$$

But now we are in trouble, since there is no obvious way to turn $\sec^3(x)$ into a function involving only $\tan(x)$.

Let's try the other substitution. If we make the $u = \sec(x)$ substitution, we will have $du = \sec(x)\tan(x)$ and so our integral will become

$$\int \sec^4(x)\tan^5(x)dx = \int \sec^3(x)\tan^4(x)du = \int \tan^4(x)u^3du.$$

This is perfect, since $\tan^4(x) = [\tan^2(x)]^2 = [\sec^2(x) - 1]^2$. Hence we can compute our integral as

$$\begin{aligned}\int \sec^4(x)\tan^5(x)dx &= \int \left[u^2 - 1\right]^2 u^3du \\ &= \frac{1}{8}u^8 - \frac{1}{3}u^6 + \frac{1}{4}u^4 + C \\ &= \frac{1}{8}\sec^8(x) - \frac{1}{3}\sec^6(x) + \frac{1}{4}\sec^4(x) + C.\end{aligned}$$ ■

The previous two example really seem to indicate the importance that the powers of one of the functions have odd degree. In the event that everything is of even degree, we once again have to use our identities on relating $\sin^2(x)$ and $\cos^2(x)$ to $\cos(2x)$.

Example 6.8

Determine the integral $\int \cos^4(x)dx$.

Solution. We know that $\cos^2(x) = \frac{1}{2}(1 + \cos(2x))$ and so our integral becomes

$$\begin{aligned}\int \cos^4(x)dx &= \frac{1}{4}\int [1 + \cos(2x)]\,[1 + \cos(2x)]\,dx \\ &= \frac{1}{4}\int \left[1 + 2\cos(2x) + \cos^2(2x)\right] dx \\ &= \frac{1}{4}[x + \sin(2x)] + \frac{1}{4}\int \cos^2(2x)dx \\ &= \frac{1}{4}x + \frac{1}{4}\sin(2x) + \frac{1}{8}\int [1 + \cos(4x)]\,dx \\ &= \frac{3}{8}x + \frac{1}{4}\sin(2x) + \frac{1}{32}\sin(4x) + C.\end{aligned}$$ ■

One could enumerate a rather large list of all the possible cases and the tricks to use, but this is a rather large waste of time. Instead, the student should just realize that the fundamental idea is to apply trigonometric identities in a way that isolates a single occurrence of a function's derivative and then apply substitution.

6.4 Trigonometric Substitution

The observant student will have recognized something very curious when we were computing the derivatives of the inverse trigonometric functions. In particular, we have

$$\frac{\mathrm{d}}{\mathrm{d}x}\arctan(x) = \frac{1}{1+x^2}, \qquad \frac{\mathrm{d}}{\mathrm{d}x}\arcsin(x) = \frac{1}{\sqrt{1-x^2}}$$

so that the derivative of ostensibly trigonometric functions, yielded quantities which were effectively rational in description. If one recollects how these derivatives were computed, one of the main reasons that rational function appear is the Pythagorean identities:

$$\sin^2(x) + \cos^2(x) = 1, \qquad \tan^2(x) + 1 = \sec^2(x).$$

We will use this, but in reverse, to develop a new integration technique. The idea is that when we see something that looks like $\sqrt{1-x^2}$, we should think of the how this relates to the sine function, and we will make a substitution $x = \sin(\theta)$. When we see something of the form $1+x^2$ we think of the tangent function and will make the substitution $x = \tan(\theta)$.

The way I have written these substitutions is important: up to now many of our substitutions have involved writing something of the form

$$\text{new variable} = \text{ some function of the old variable}$$

and that is effectively what we are doing again, by defining say $\theta = \arctan(x)$. However, this is rather messy to work with as written, so we often write $\tan(\theta) = x$. Computing differentials will give $\sec^2(\theta)d\theta = dx$.

Let's first try this on an integrand for which we know what the result should be:

Example 6.9

Determine the integral $\int \frac{1}{1+x^2}dx$.

Solution. We make the substitution $x = \tan(\theta)$ so that $dx = \sec^2(\theta)d\theta$. Our integral thus becomes

$$\begin{aligned}\int \frac{1}{1+x^2}dx &= \int \frac{\sec^2(\theta)}{1+\tan^2(\theta)}d\theta \\ &= \int \frac{\sec^2(\theta)}{\sec^2(\theta)}d\theta \\ &= \theta + C.\end{aligned}$$

Since $x = \tan(\theta)$ we know that $\theta = \arctan(x)$, so that

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

exactly as we expected. ■

Not all things will be so simple, so let us try a slightly harder question:

Example 6.10

Determine $\int \frac{x^2}{\sqrt{16-x^2}} dx$.

Solution. This time we are tempted to make the substitution $x = \sin(\theta)$ and morally this is correct. However, if we think about it a bit more we can make our lives a lot easier by setting $x = 4\sin(\theta)$. The reason is that

$$16 - x^2 = 16 - (4\sin(x))^2 = 16\left[1 - \sin^2(x)\right] = 16\cos^2(x).$$

This will make our lives much easier, as removing the addition sign will allow us to apply the square root. Now if $x = 4\sin(\theta)$ then $dx = 4\cos(\theta)d\theta$ and we get

$$\begin{aligned}\int \frac{x^2}{\sqrt{16-x^2}} dx &= \int \frac{\left[16\sin^2(\theta)\right]\left[4\cos(\theta)d\theta\right]}{\sqrt{16\cos^2(\theta)}} d\theta \\ &= 16\int \sin^2(\theta) d\theta \\ &= 8\left[\theta - \frac{1}{2}\sin(2\theta)\right] + C\end{aligned}$$

The tricky part comes in changing this back into a function of x. Since $x = 4\sin(\theta)$ we know that $\theta = \arcsin(x/4)$, but the pesky presence of the 2θ means that we cannot just simply plug this back in. Instead, we must make the following substitution:

$$\begin{aligned}\sin(2\theta) &= 2\sin(\theta)\cos(\theta) \\ &= 2\frac{x}{4}\frac{\sqrt{16-x^2}}{4} \\ &= \frac{x\sqrt{16-x^2}}{8}.\end{aligned}$$

We conclude that

$$\int \frac{x^2}{\sqrt{16-x^2}} dx = 8\arcsin\left(\frac{x}{4}\right) - \frac{x\sqrt{16-x^2}}{2} + C.$$ ■

Example 6.11

Compute the integral

$$\int \frac{x}{(x^2+6x+18)^{3/2}} dx.$$

Solution. My hope would be that each student would *try* substitution first, though it is unlikely to succeed given that I am trying to demonstrate the technique of trigonometric substitution. Nonetheless, it should always be your starting point when you recognize a function and its derivative. Unfortunately, we do not have any of the recommended forms above so we must endeavour to manipulate the integrand until it looks amenable to our techniques. Indeed, we may quickly complete the square of the denominator to find that

$$x^2 + 6x + 18 = (x+3)^2 + 9.$$

Referencing our table above, we decide that we should make the substitution $x + 3 = 3\tan\theta$ so that $dx = 3\sec^2(\theta)d\theta$ and

$$(x+3)^2 + 9 = 9(\tan^2(\theta) + 1) = 9\sec^2(\theta).$$

Substituting into our integral we find

$$\begin{aligned}
\int \frac{x}{(x^2+6x+18)^{3/2}}dx &= \int \frac{(3\tan(\theta)-3)}{(9\sec^2(\theta))^{3/2}}(3\sec^2(\theta))d\theta \\
&= \int \frac{9(\tan(\theta)-1)\sec^2(\theta)}{27\sec^3(\theta)}d\theta \\
&= \frac{1}{3}\int \frac{\tan(\theta)-1}{\sec(\theta)}d\theta \\
&= \frac{1}{3}\int (\sin(\theta)-\cos(\theta))d\theta \\
&= -\frac{1}{3}[\cos(\theta)+\sin(\theta)] + C.
\end{aligned}$$

Now $\tan\theta = \frac{x+3}{3}$ which implies (by drawing our triangle) that

$$\cos(\theta) = \frac{3}{\sqrt{x^2+6x+18}}, \qquad \sin(\theta) = \frac{x+3}{\sqrt{x^2+6x+18}}$$

and we conclude that

$$\int \frac{x}{(x^2+6x+18)^{3/2}}dx = -\frac{1}{3}\frac{x+6}{\sqrt{x^2+6x+18}} + C. \qquad \blacksquare$$

6.5 Partial Fractions

The method of partial fractions is a means to turn integrands which are rational functions (that is, the quotient of polynomials) into simpler constitution pieces which may be more simply integrated. Let us quickly recall partial fractions: Let $\frac{p(x)}{q(x)}$ be a rational function, so that p and q are polynomials. If we can write $q(x)$ as in product of linear factors, say $q(x) = (x-r_1)(x-r_2)\cdots(x-r_n)$ then our goal is to write

$$\frac{p(x)}{q(x)} = \frac{C_1}{x-r_1} + \frac{C_2}{x-r_2} + \cdots + \frac{C_n}{x-r_n},$$

for constants $C_i, i = 1, \ldots, n$. If q contains an irreducible quadratic factor, say $q(x) = (x-r_1)(ax^2 + bx + c)$ then we try to write

$$\frac{p(x)}{q(x)} = \frac{A}{x-r_1} + \frac{Bx+C}{ax^2+bx+c}.$$

In general, if $q(x)$ contains an irreducible component of degree k then the numerator of that component can be at most $k-1$. If in the factorization a factor occurs with multiplicity greater than 1, say $q(x) = (x-r_1)(x-r_2)^2$ then we seek an expression of the form

$$\frac{p(x)}{q(x)} = \frac{A}{x-r_1} + \frac{B}{x-r_2} + \frac{C}{(x-r_2)^2}.$$

More generally, if the factor $(x-r)$ occurs with multiplicity m, then we must account for a factor $(x-r)^j$ for all $j = 1, \ldots, m$. All possible combinations of the above rules also hold.

Example 6.12

Find the partial fractions decomposition of the rational function

$$\frac{5x+1}{x^2+x-2}.$$

Solution. We can factor the denominator as $x^2+x-2 = (x+2)(x-1)$ so we look for a decomposition of the form

$$\frac{5x+1}{x^2+x-2} = \frac{A}{x+2} + \frac{B}{x-1}.$$

By cross multiplying, we must thus have $5x+1 = A(x-1) + B(x+2)$. If one so wishes, we could expand this out to find that

$$5x+1 = (A+B)x + (-A+2B), \qquad \begin{aligned} A+B &= 5 \\ -A+2B &= 1 \end{aligned}$$

which is a system of equations that can be solved. On the other hand, realizing that our equation must hold for all values of x, we could substitute convenient values of $x = 1, x = -2$ into our expression to find that

$$6 = 3B, \qquad -9 = -3A.$$

This tells us that $A = 3$ and $B = 2$, so that

$$\frac{5x+1}{x^2+x-2} = \frac{3}{x+2} + \frac{2}{x-1}.$$

■

The process of partial fractions allows us to decompose our integrand into bite size pieces. From here we can exploit the linearity of the integral.

Example 6.13

Determine the integral $\displaystyle\int \frac{5x+1}{x^2+x-2}dx$.

Solution. One might be tempted to try substitution here, but there is no obvious way to make it work. Instead, we exploit the partial fraction decomposition that we computed in Example 6.12 to write

$$\int \frac{5x+1}{x^2+x-2}dx = \int \left[\frac{3}{x+2} + \frac{2}{x-1}\right] dx$$
$$= 3\ln|x+2| + 2\ln|x-1| + C.$$ ■

The problem with partial fractions, as written, is that it is impossible for us to consider the case when the degree of the numerator is greater than or equal to that of the numerator. In such cases, we have to exploit the power polynomial long division:

Example 6.14

Compute $\int \frac{x^3-4x-10}{x^2-x-6}dx$.

Solution. We would like to apply partial fractions, but first need to perform polynomial long division. Some quick calculations show us that

$$\frac{x^3-4x-10}{x^2-x-6} = x+1+\frac{3x-4}{x^2-x-6}.$$

We may now perform partial fractions on the latter part of this expression; namely

$$\frac{3x-4}{x^2-x-6} = \frac{A}{x-3} + \frac{B}{x-2}.$$

One may find that $A = 1$ and $B = 2$. Substituting this into the integral we get

$$\int \frac{x^3-4x-10}{x^2-x-6}dx = \int \left[x+1+\frac{1}{x-3} + \frac{2}{x+2}\right] dx$$
$$= \frac{1}{2}x^2 + x + \ln|x-3| + 2\ln|x-2| + C.$$ ■

As an unnecessarily difficult example, consider the final following problem:

Example 6.15

Compute $\int \frac{2x^3-4x^2+2x-2}{x^4-2x^3+2x^2-2x+1}$.

Solution. Our first goal is to factor the denominator. By the Rational Roots Theorem[25] the only possible rational roots are $x = \pm 1$. Substitution into the denominator reveals that only $x = 1$ is a root, meaning that we can remove a factor of $x-1$ from the denominator. Performing long division, we get that

$$x^4 - 2x^3 + 2x^2 - 2x + 1 = (x-1)(x^3 - x^2 + x + 1).$$

[25] The Rational Root Theorem says that if p/q is a rational number, written in irreducible form, and is a root of $a_n x^n + \cdots + a_0$ where both $a_n, a_0 \neq 0$, then p is a factor of a_0 and q is a factor of a_n.

Once again the only possible rational roots for $x^3 - x^2 + x + 1$ are $x = \pm 1$, and in fact we already know that $x = -1$ cannot be a root. Checking $x = 1$ we again see that 1 is a root, so we may remove another factor of $x - 1$ to get

$$x^4 - 2x^3 + 2x^2 - 2x + 1 = (x-1)^2(x^2+1).$$

As $x^2 + 1$ has no real roots, we cannot factor any further. Our partial fraction decomposition is thus of the form

$$\frac{2x^3 - 4x^2 + 2x - 2}{x^4 - 2x^3 + 2x^2 - 2x + 1} = \frac{A}{x-1} + \frac{B}{(x-1)^2} + \frac{Cx + D}{x^2+1},$$

which if we cross multiply gives us

$$2x^3 - 4x^2 + 2x - 2 = A(x-1)(x^2+1) + B(x^2+1) + (Cx+D)(x-1)^2.$$

Substituting $x = 1$ we get $-2 = 2B$ so that $B = -1$. Now there are no other immediately nice x's to substitution, so let's try $x = 0$:

$$-2 = -A - 1 + D, \qquad \Rightarrow \qquad D - A = -1.$$

Substituting $x = 2$ we get

$$2 = 5A - 5 + 2C + D \qquad \Rightarrow \qquad 5A + 2C + D = 7.$$

Substituting $x = -1$ we get

$$-10 = -4A - 2 + 4(D - C), \qquad \Rightarrow \qquad -4A + 4D - 4C = -8.$$

One can solve this to find $A = 1, D = 0, C = 1$, so that

$$\frac{2x^3 - 4x^2 + 2x - 2}{x^4 - 2x^3 + 2x^2 - 2x + 1} = \frac{1}{x-1} - \frac{1}{(x-1)^2} + \frac{x}{x^2+1}.$$

Integrating we thus have

$$\begin{aligned}\int \frac{2x^3 - 4x^2 + 2x - 2}{x^4 - 2x^3 + 2x^2 - 2x + 1} dx &= \int \left[\frac{1}{x-1} - \frac{1}{(x-1)^2} + \frac{x}{x^2+1}\right] dx \\ &= \log|x-1| + \frac{1}{x-1} + \frac{1}{2}\log|x^2+1| + C.\end{aligned}$$

■

7 Applications of the Integral

In this section we will list a few of the plethora of applications of the integral. The utility of the integral cannot be overstated: it would be impossible to provide a full list of applications, simply because the author cannot begin to know them all. The integral finds its place amongst almost every stem field, from biology (for example, computing glucose tolerance, or other metabolic curves) to engineering (the 'I' in PID-controller stands for integral).

7.1 Volumes

Before the invention of calculus, computing volumes was all but impossible. It was a task relegated to geniuses, and even then the volume of only the most simple objects could be computed, using what was effectively calculus techniques. With the power of calculus, now almost anyone can compute the volume of very complicated shapes!

The textbook presents this subject by enumerating three or four different techniques for computing volumes, depending on how the target shape in generated. I reiterate that our goal is not to teach you algorithms, but to teach you how to problem solve. To this end, I will not present the algorithms for determining these volumes; instead, I will try to show the student how to set up these volume problems so that they can be computed in any situation.

The major theme is to realize that volume can be computed by integrating area. The idea is that integrating 'adds a dimension.' Computing the area under a curve was done by integrating the height of lines $f(x)$ over an interval. Hence integrating length - a one dimensional thing - yielded area. To compute three dimensional volumes, we will thus want to integrate areas. Hence setting up a volume question amounts to determining the *cross-sectional* area, then integrating:

$$V = \int_a^b A(x)dx$$

where V is the volume, and $A(x)$ is the cross-sectional area as a function of x.

To make this somewhat more precise, recall that we did not really 'integrate lines,' but rather, we looked at rectangles with height $f(x)$ and length Δx. In the 'limit,' this Δx comes a dx, giving the infinitesimal length of the interval. Integrating $f(x)dx$ then gave us the infinitesimal area. For volume, we will do the same thing. By partitioning our interval, the elements $A(x)dx$ will represent infinitesimal volume elements, which we will integrate to get volume.

For the sake of simplicity, I will continuously use the same two functions: $f(x) = \sqrt{x}$ and $g(x) = x$. In Example 5.28 we saw how to compute the area between these two curves. In what follows, we will construct a series of three dimensional shapes using the area that they bound, and show the student how to compute the generated volume.

Example 7.1

Consider a space which has as its base the area enclosed by the functions $f(x) = \sqrt{x}$ and $g(x) = x$, and whose cross sections are squares. Determine the volume of this object.

Solution. We are told that the cross sectional area is given by squares; however, the dimension of the squares change depending on the value of x. Choosing a typical point, we see that the base of the square has length $\sqrt{x} - x$ and hence the area of the square is

$$A(x) = (\sqrt{x} - x)^2 = x - 2x^{2/3} + x^2.$$

We integrate for $x \in [0, 1]$ and find the volume to be

$$V = \int_0^1 A(x)dx = \int_0^1 [x - 2x^{3/2} + x^2]dx = \left[\frac{1}{2}x^2 - \frac{4}{5}x^{5/2} + \frac{1}{3}x^3\right]_0^1 = \frac{1}{30}.$$ ■

That was not too bad, but one will rarely be explicitly told the cross sectional area. Often, one has to analyze the object and discover this one his/her own.

Example 7.2

Consider the area enclosed by the functions $f(x) = \sqrt{x}$ and $g(x) = x$. Determine the volume of the object given by rotating that area about the x-axis.

Solution. Once again, we consider a typical cross-section at the point x. The object we get by taking out this slice looks like a washer, whose outer radius is $f(x)$ and whose inner radius is $g(x)$. The cross-sectional area of this element is thus

$$A(x) = \pi f(x)^2 - \pi g(x)^2 = \pi[f(x)^2 - g(x)^2].$$

Integrating we thus get

$$\begin{aligned} V &= \int_0^1 A(x)dx = \int_0^1 \pi \left[\sqrt{x}^2 - x^2\right] dx \\ &= \pi \left[\frac{1}{2}x^2 - \frac{1}{3}x^3\right]_0^1 = \frac{\pi}{6} \end{aligned}$$ ■

Here's an interesting question: What if we have revolved this shape around the y-axis instead? First of all, do we get the same volume? Secondly, how do we handle the integral?

Example 7.3

Determine the object given by rotating the area enclosed by $f(x) = \sqrt{x}$ and $g(x) = x$ about the y-axis.

Solution. This time, taking cross sections about a point x yields really odd shapes whose area is difficult to compute. Instead, let's try taking cross-sections in the y-direction. Notice that the function $y = \sqrt{x}$ can be written as $y^2 = x$, so that the cross section in the y-direction again looks like a washer, but this time with outer radius $(x =)y$ and inner radius $(x =)y^2$. The cross-sectional area is thus $A(y) = \pi[y^2 - y^4]$. Again, our interval is going from 0 to 1, and hence the volume is given by

$$V = \int_0^1 \pi \left[y^2 - y^4\right] dy$$
$$= \left[\frac{1}{3}y^3 - \frac{1}{5}y^5\right]_0^1 = \frac{2\pi}{15}.$$ ■

Let's throw a wrench in the works:

Example 7.4

Determine the volume of the object given by rotating the area enclosed by $f(x) = \sqrt{x}$ and $g(x) = x$ about the line $x = 1$.

Solution. As always, we take a typical cross-section, again in the y-direction this time. We again have a washer whose outer radius is $1 - y^2$ and whose inner radius is $1 - y$. The cross sectional area is thus

$$A(y) = \pi[(1 - y^2)^2 - (1 - y)^2] = \pi(2y - 3y^2 + y^4),$$

and the volume can be computed as

$$V = 2\pi \int_0^1 [2y - 3y^2 + y^4]dy = \pi \left[y^2 - y^3 + \frac{1}{5}y^5\right]_0^1 = \frac{\pi}{5}$$ ■

Alternatives to Sections Another technique for integrating can be to use other shapes instead of just cross-sections. There are times when it may in fact this may be more convenient that using cross-sections, so we introduce it as yet another technique for finding volumes.

As in the previous sections, the trick is to think about an area (representing an infinitesimal piece of volume) and ”sweep out” the shape with which we are concerned. For our first example, instead of using cross-sections, we will use cylinders. Consider a function $f(x)$ defined on $[0, a]$, rotated about the y-axis. A typical slice of the cylinder will have surface area

$$S(x) = 2\pi x f(x).$$

If we were to use our differential framework, we could think about fattening up this cylinder slightly, to gives us a volume $2\pi x f(x)dx$. In that case, the volume of the object can be computed by sweeping these cylinders from 0 to a, to give us the volume

$$V = \int_0^a 2\pi x f(x)dx.$$

Example 7.5

Determine the volume of the object created by rotating the function $\sqrt{x}$ on $[0, 4]$ about the y-axis.

Solution. Using the method of shells, the volume can be computed to be

$$V = \int_0^4 2\pi x\sqrt{x}dx = 2\pi \left[\frac{2}{5}x^{5/2}\right]_0^4 = \frac{128}{5}\pi. \qquad \blacksquare$$

Of course, we can easily generalize this process to the case when we have a difference of functions. For example, consider the area bounded between two function $f(x)$ and $g(x)$, say on $[0, a]$. If we look at the cylinder in this case, the radius of the cylinder remains constant, but the height is given by the difference of the two functions. Hence the surface area is given by

$$S(x) = 2\pi x[f(x) - g(x)].$$

Example 7.6

Determine the volume of the object given by rotating the area enclosed between $\sqrt{x}$ and x^2 about the y-axis.

Solution. We did this exact problem in Example 7.3 where we got a solution of $\frac{2\pi}{15}$. We certainly expect that we should get the same solution this time. Again, using the method of shells we get

$$\begin{aligned} V &= \int_0^1 2\pi x \left[\sqrt{x} - x\right] dx = 2\pi \int_0^1 \left[x^{3/2} - x^2\right] dx \\ &= 2\pi \left[\frac{2}{5}x^{5/2} - \frac{1}{3}x^4\right]_0^1 = \frac{2\pi}{15}. \end{aligned} \qquad \blacksquare$$

As a general rule, I find that using cross-sections more often works out to be the better technique, but the following is a (contrived) example where the shell method proves much simpler:

Example 7.7

Determine the volume when e^{x^2} on $[0, 2]$ is rotated about the y-axis.

Solution. If we were to do this using cross-sections, we would need to find a way of determining the x-coordinate given $y = e^{x^2}$, and while that is doable, the resulting function is not very easy to integrate. Instead, if we use the shell method we have

$$\begin{aligned} V &= \int_0^2 2\pi x e^{x^2} dx & u = x^2, du = 2xdx \\ &= \pi \int_0^4 e^u du = \pi \left. e^u \right|_0^4 = \pi(e^4 - 1). \end{aligned} \qquad \blacksquare$$

Naturally, were we to revolve around the x-axis, the procedure would be the same, though we would have to convert our problem to integrating about the y-direction, and make the appropriate changes to our functions.

What happens if we rotate about a different axis?

Example 7.8

Determine the volume of the object given by rotating the area enclosed by $\sqrt{x}$ and x about the $x = 1$ axis.

Solution. We did this problem in Example 7.4, and found a solution of $\frac{\pi}{5}$. Setting up our cylinder, we see that for a typical slice our radius is now $1 - x$ while our height is $\sqrt{x} - x$; thus our surface area is given by

$$S(x) = 2\pi(1-x)[\sqrt{x} - x].$$

Integrating to find the volume, we have

$$\begin{aligned} V &= \int_0^1 2\pi(1-x)[\sqrt{x}-x]dx = 2\pi \int_0^1 \left[\sqrt{x} - x - x^{3/2} + x^2\right] dx \\ &= 2\pi \left[\frac{2}{3}x^{3/2} - \frac{1}{2}x^2 - \frac{2}{5}x^{5/2} + \frac{1}{3}x^3\right]_0^1 = \frac{\pi}{5}. \end{aligned}$$

■

Example 7.9

Determine the volume of the sphere of radius r.

Solution. Consider the curve $y = \sqrt{r^2 - x^2}$ from $[0, r]$. By rotating this about the y-axis, we will get the upper hemisphere of the sphere, and hence half the area. Using the method of shells, we thus get

$$\begin{aligned} V &= 2\pi \int_0^r x\sqrt{r^2 - x^2}dx & u = r^2 - x^2 \\ &= -\pi \int_{r^2}^0 \sqrt{u}du \\ &= \pi \left[\frac{2}{3}u^{3/2}\right]_0^{r^2} = \frac{2}{3}\pi r^3. \end{aligned}$$

Of course, the total area of the sphere is then twice this amount, giving $\frac{4}{3}\pi r^3$, which is exactly what we expected. ■

8 Improper Integrals

It was essential in defining the integral that we examined *bounded* functions on *bounded* intervals $[a,b]$, so that everything under consideration could be finite. In this section we examine how to now extend the idea of integrals to cover unbounded functions, and functions defined on unbounded intervals.

8.1 First Principles

8.1.1 Infinite Intervals

Our goal is to extend the notion of integration from a finite interval $[a,b]$ to an infinite interval $[a,\infty)$ or $(-\infty,b]$. We will develop the idea for the interval $[a,\infty)$ and leave the details for $(-\infty,b]$ to the student.

Definition 8.1

Let f be a bounded function on the interval $[a,\infty)$ such that f is integrable on $[a,x]$ for every $x > a$. We define the *improper integral* of f on $[a,\infty)$ as

$$\int_a^\infty f(t)dt = \lim_{x\to\infty}\int_a^x f(t)dt.$$

We say that that the improper integral *converges* if this limit is finite, and *diverges* otherwise.

Let us take a moment to think about what this is saying: We are defining a new function

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

, which is precisely the anti-derivative of f. The improper integral then converges if the function $F(x)$ has a horizontal asymptote; that is, the area under the graph of f asymptotically stabilizes to a single, finite number.

Naturally, one then defines the improper integral on $(-\infty,b]$ as

$$\int_{-\infty}^b f(t)dt = \lim_{x\to-\infty}\int_x^b f(t)dt.$$

Example 8.2

Determine $\int_0^\infty e^{-t}dt$, if it exists.

Solution. By definition, we know that $\displaystyle\int_0^\infty e^{-t}dt = \lim_{x\to\infty}\int_0^x e^{-t}dt$. We are very familiar with computing the integral on the right hand side, and know that $\int_0^x e^{-t}dt = -e^{-t}\big|_0^x = 1 - e^{-x}$. Thus

$$\int_0^\infty e^{-t}dt = \lim_{x\to\infty}\int_0^x e^{-t}dt = \lim_{x\to\infty}\left[1-e^{-t}\right] = 1. \qquad \blacksquare$$

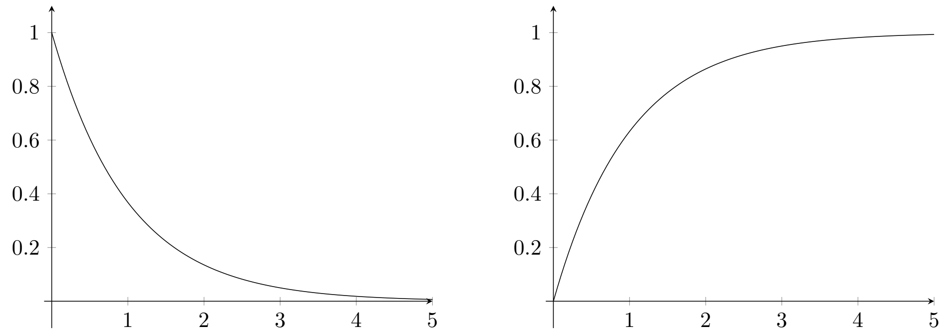


Figure 25: The function $f(x) = e^{-x}$ and an anti-derivative $F(x) = 1 - e^{-x}$. We see that $F(x)$ tends to the number 1 as $x \to \infty$. This occurs because as $x \to \infty$ the graph under the function $f(x)$ becomes very small.

Example 8.3

Determine $\int_0^\infty \sin(x)dx$, if it exists.

Solution. Proceeding by definition, we have that

$$\int_0^\infty \sin(t)dt = \lim_{x\to\infty} \int_0^x \sin(t)dt = \lim_{x\to\infty} [-\cos(t)]$$

which does not exist. If we think about this for a while, we can convince ourselves why it is true. The area under the graph of $\sin(x)$ is constantly oscillating between $+1$ and -1, and this oscillation never stops. Hence even though there are many points where the area is arbitrarily small, it's not enough to ensure that the area ever converges. ■

Intuitively, it seems like functions which tend to zero may have a convergent improper integral. However, this is not always the case. It is necessary that function goes to zero 'fast enough' to ensure that the area is positive.

Proposition 8.4

If $a > 0$ is an arbitrary positive number, then

$$\int_a^\infty \frac{1}{x^p}dx \quad \text{converges if and only if} \quad p > 1.$$

Proof. If $p = 1$ then

$$\int_a^\infty \frac{1}{t}dt = \lim_{x\to\infty} \log(x/a) = \infty$$

so the integral diverges. Thus assume that $p \neq 1$, for which we have

$$\int_a^\infty \frac{1}{t^p}dt = \lim_{x\to\infty} \left[\frac{1}{1-p}\frac{1}{x^{p-1}}\right]_a^x.$$

We know that $1/x^{p-1}$ converges only if and if the power is non-negative; that is, $p-1 \geq 0$. Combining this with $p \neq 1$ tells us that the improper integral converges if and only if $p > 1$. ☐

As was computed explicitly, this implies that $\int_1^\infty \frac{1}{x}$ does not converge, which is a result that often confuses students.

What happens if we want to define the improper integral on $(-\infty, \infty)$?

Definition 8.5

If f is integrable on every interval $[a,b] \subseteq \mathbb{R}$, then we say that $\int_{-\infty}^\infty f(t)dt$ converges if, for any $c \in \mathbb{R}$ we have both

$$\int_{-\infty}^c f(t)dt \text{ converges, and } \int_c^\infty f(t)dt \text{ converges.}$$

In this case, we set[a]

$$\int_{-\infty}^\infty f(t)dt = \int_{-\infty}^c f(t)dt + \int_c^\infty f(t)dt.$$

[a] The student should convince him/herself that the value of the improper integral does not depend on the value of c.

This is very different than simply demanding that

$$\int_{-\infty}^\infty f(t)dt = \lim_{x\to\infty} \int_{-x}^x f(t)dt \text{ exists.}$$

The reason is that the value of the integral should not depend on 'how quickly' we take our limits. For example

$$\lim_{x\to\infty} \int_{-x}^x \sin(t)dt = \lim_{x\to\infty} [\cos(-x) - \cos(x)] = 0$$

since cosine is an even function. On the other hand,

$$\lim_{x\to\infty} \int_{-x}^{2x} \sin(t)dt = \lim_{x\to\infty} [\cos(-x) - \cos(2x)] \text{ does not exist.}$$

Example 8.6

Determine $\int_{-\infty}^\infty e^{-|t|}dt$.

Solution. A natural place to split our interval will be at 0. Now

$$\begin{aligned} \int_0^\infty e^{-|t|}dt &= \int_0^\infty e^{-t}dt && \text{since } t > 0 \\ &= 1 && \text{by Example 8.2.} \end{aligned}$$

Similarly, $\int_{-\infty}^{0} e^{-|t|}dt = 1$, thus

$$\int_{-\infty}^{\infty} e^{-|t|}dt = \int_{-\infty}^{0} e^{-|t|}dt + \int_{0}^{\infty} e^{-|t|}dt = 2.$$

■

8.1.2 Unbounded Functions

The case of unbounded functions often poses even more difficulty, since it is very tempting to just blindly apply the Fundamental Theorem of Calculus without paying attention.

Example 8.7

Compute the integral $\int_{-1}^{1} \frac{1}{x^2}dx$.

Solution. We know that the anti-derivative of $\frac{1}{x^2}$ is $-\frac{1}{x}$, so if we were to just blindly apply the FTC we would get

$$\int_{-1}^{1} \frac{1}{x^2}dx = -\frac{1}{x}\Big|_{-1}^{1} = -2.$$

Unfortunately, **this is completely and totally wrong.** Our first hint at a miscalculation is probably the fact that $\frac{1}{x^2}$ is everywhere positive, yet we somehow ended up with a negative integral. In fact, the integral is infinite. To see this, note that since $\frac{1}{x^2} > 0$ then

$$\begin{aligned} \int_{-1}^{1} \frac{1}{x^2}dx &\geq \int_{\epsilon}^{1} \frac{1}{x^2}dx \\ &= -\frac{1}{x}\Big|_{\epsilon}^{1} \\ &= \frac{1}{\epsilon} - 1, \end{aligned}$$

and that by choosing ϵ to be small enough we can make the integral arbitrarily large. The reason is that the function $\frac{1}{x^2}$ is not integrable on the interval $[-1, 1]$ (all integrable functions are bounded!), and hence we could not apply the FTC. ■

Exercise: Compare the following two expressions:

$$\lim_{\epsilon \to 0^+} \left[\int_{-1}^{-\epsilon} \frac{1}{x}dx + \int_{\epsilon}^{1} \frac{1}{x}dx \right] \qquad \lim_{\epsilon \to 0^+} \left[\int_{-1}^{-\epsilon} \frac{1}{x}dx + \int_{2\epsilon}^{1} \frac{1}{x}dx \right]$$

The way to deal with unbounded functions is precisely the same way that we deal with unbounded intervals: we take a limit.

Definition 8.8

If f is a function on $[a,b]$, unbounded at a, but integrable on $[x,b]$ for every $x > a$ then we define the improper integral

$$\int_a^b f(t)dt = \lim_{x\to a^+} \int_x^b f(t)dt.$$

If this limit is finite we say that the improper integral *converges*; otherwise, we say that the improper integral *diverges*.

Similarly, if f were unbounded at b, we would define the improper integral as

$$\int_a^b f(t)dt = \lim_{x\to b^-} \int_a^x f(t)dt.$$

Example 8.9

Evaluate $\displaystyle\int_3^5 \frac{t}{\sqrt{t^2-9}}dt$, if it exists.

Solution. We immediately recognize that there could be a problem at $t = 3$. Using the substitution $u = t^2 - 9$ we get

$$\begin{aligned}\int_3^5 \frac{t}{\sqrt{t^2-9}}dt &= \lim_{x\to 3^+} \left[\sqrt{t^2-9}\right]_x^5 \\ &= \lim_{x\to 3^+} \left[4 - \sqrt{x^2-9}\right] = 4.\end{aligned}$$ ■

Just as in the case of integrating from $(-\infty, \infty)$ we must also be careful about integrating on both sides of an unbounded function.

Definition 8.10

If f is a function on $[a,b]$ which is unbounded at $c \in [a,b]$ then we say that the improper integral

$$\int_a^b f(t)dt \text{ converges, if and only if } \int_a^c f(t)dt \text{ and } \int_c^b f(t)dt \text{ both exist.}$$

In this case, we set

$$\int_a^b f(t)dt = \int_a^c f(t)dt + \int_c^b f(t)dt.$$

Proposition 8.11

For any $a \neq 0$ we have that

$$\int_0^a \frac{1}{t^p} dt \text{ converges, if and only if } p < 1.$$

Proof. For simplicity, let's assume that $a > 0$. If $p = 1$ then

$$\int_0^a \frac{1}{t} dt = \lim_{x \to 0^+} \int_x^a \frac{1}{t} dt = \lim_{x \to 0^+} \log(t) \Big|_x^a = \infty.$$

If $p \neq 1$ then

$$\int_0^1 \frac{1}{t^p} dt = \lim_{x \to 0^+} \int_x^a \frac{1}{t} dt = \frac{1}{1-p} \lim_{x \to 0^+} \frac{1}{x^{p-1}} \Big|_x^a .$$

The limit converges if and only if $p - 1 \leq 0$ so that $p \leq 1$. Combined with the fact that $p \neq 1$ we get $p < 1$ as required. □

8.2 Comparison Tests

In this section, we develop some techniques to make our lives simpler in terms of dealing improper integrals. The idea is something like the following: Say that you were asked to determine whether the following improper integral converged:

$$\int_0^\infty \left[\frac{2 + \sin(x)}{x} \right] dx,$$

or how about

$$\int_1^\infty \frac{x^2 + 1}{x^4 + 3x^2 - 4x + 1} dx?$$

Neither of these integrals are easy to compute explicitly; in fact, $\frac{\sin(x)}{x}$ cannot be solved in terms of elementary functions, but appears often enough to have a name: The Sine Integral. The goal, as is often the case with mathematics, is to reformulate these problems into ones that are much easier to solve, or that have already been solved. For this, we introduce two tests: The Comparison Test, and the Limit Comparison Test.

8.2.1 The Basic Comparison Test

The Comparison Test is the simplest of the basic tests, and exploits the "monotonicity" of the integral. In particular, the following lemma is rather intuitive:

Lemma 8.12

If F, G are two functions with a domain D, such that $F(x) \leq G(x)$ for all $x \in D$, then

$$\lim_{x \to \infty} F(x) = \infty \quad \Rightarrow \quad \lim_{x \to \infty} G(x) = \infty.$$

Proof. This is a kind of 'Squeeze Theorem' for infinite limits: If the smaller of the two functions goes off to infinity, then certainly the bigger of two must go to infinity as well. The proof is effectively a one-liner. Since $F(x) \xrightarrow{x\to\infty} \infty$ we know that for every $M \in \mathbb{R}$ there exists an $N \in \mathbb{R}$ such that if $x > N$ then $f(x) > M$. We claim that this same N works for G. Indeed, let M be arbitrarily chosen and choose the N guaranteed to exist by above. Then if $x > N$ we have $g(x) \geq f(x) > M$ as required. □

Of use will be this 'spiritual' contrapositive:

Lemma 8.13

If F is a bounded, increasing function on some interval $[a, \infty)$, then $\lim_{x\to\infty} F(x)$ exists.

Proof. Consider the image set of F; namely

$$F[a, \infty) = \{F(x) : x \in [a, \infty)\} = \{y : y = F(x), x \in [a, \infty)\}.$$

Since F is bounded on $[a, \infty)$ we know that $F[a, \infty)$ is bounded, and hence has a supremum L. We claim that L is in fact the limiting point of F as x tends to infinity. Indeed, let $\epsilon > 0$ be arbitrary. By Proposition 5.4, there exists some $x_0 \in [a, \infty)$ such that

$$L - \epsilon < F(x_0) \leq L.$$

Since F is increasing, this in turn implies that for all $x > x_0$ we have

$$L - \epsilon < F(x_0) < F(x) \leq L < L + \epsilon.$$

This is precisely that if $x > x_0$ then $|F(x) - L| < \epsilon$, and so we have shown that F limits to L as required. □

These two lemmas are precisely what we need to state and prove the Basic Comparison Test:

Theorem 8.14: The Basic Comparison Test for Improper Integrals

Let f, g be functions on an interval $[a, \infty)$ such that $0 \leq f(x) \leq g(x)$ for all $x \in [a, \infty)$.

1. If $\int_a^\infty g(t)dt$ converges then $\int_a^\infty f(t)dt$ converges.
2. If $\int_a^\infty f(t)dt$ diverges then $\int_a^\infty g(t)dt$ diverges.

The idea is again a type of 'Squeeze Theorem' argument. If the integral of the bigger function g becomes finite, the monotonicity of the integral cannot allow f to go off to infinity. Similarly, if the integral of the smaller function f goes off to infinity, the larger function's integral must also diverge. A good question at this point is to ask whether any of the integrals could oscillate and hence not converge. The condition that $0 \leq f(x) \leq g(x)$ guarantees that the integrands are positive: this means that the corresponding integrals are increasing functions.

Proof of the Basic Comparison Test. In both cases, define the functions

$$F(x) = \int_a^x f(t)dt, \qquad G(x) = \int_a^x g(t)dt,$$

and note that F, G are increasing functions (since f, g are non-negative) and $F(x) \leq G(x)$ by monotonicity of the integral.

1. Assume that $\int_a^\infty g(t)dt$ converges; that is, the limit

   $$\lim_{x\to\infty} G(x)$$

   exists and is finite. This implies that F is an increasing bounded function. By Lemma 8.13 it follows that

   $$\lim_{x\to\infty} F(x)$$

   also exists and is finite.

2. Assume that $\int_a^\infty f(t)dt$ diverges, in which case

   $$\int_a^\infty f(t)dt = \lim_{x\to\infty} F(x) = \infty$$

   since F is an increasing function. Since $F \leq G$, by Lemma 8.12 this implies that

   $$\int_a^\infty g(t)dt = \lim_{x\to\infty} G(x) = \infty$$

   so that $\int_a^\infty g(t)$ diverges.

□

Example 8.15

Show that $\displaystyle\int_1^\infty \frac{2+\sin(x)}{x}dx$ diverges.

Solution. We want to compare the function $\frac{2+\sin(x)}{x}$ to some function which we know diverges. The idea to keep in mind is that in the limit as $x \to \infty$, the $\sin(x)$ term does not play a significant role; rather, the x-term in the denominator dominates. This suggests that we make the comparison against $\frac{1}{x}$. Indeed, notice that since $-1 \leq \sin(x) \leq 1$ for all x, we have

$$\frac{2+\sin(x)}{x} \geq \frac{1}{x}.$$

By Proposition 8.11, we know that $\int_1^\infty \frac{1}{x}$ diverges, so by the comparison test it follows that

$$\int_1^\infty \frac{2+\sin(x)}{x}dx \text{ diverges.}$$ ■

Example 8.16

Determine whether $\int_0^\infty \frac{x}{\sqrt{x^6+1}}dx$ converges or diverges.

Solution. Once again, we just want to look at which terms in the numerator and denominator dominate in the limit $x \to \infty$. Certainly, we do not expect x^6+1 to be too much different than x^6 for very large x, so we will compare our integrand to the function

$$\frac{x}{\sqrt{x^6}} = \frac{1}{x^2}$$

. Indeed, since $1+x^6 \geq x^6$ we have that $\frac{1}{\sqrt{x^6}} \geq \frac{1}{\sqrt{1+x^6}}$, which in turn implies that

$$\frac{1}{x^2} = \frac{1}{\sqrt{x^6}} \geq \frac{x}{\sqrt{x^6+1}}.$$

Now the integral of the left-hand-side converges by 8.11, so by the Comparison Test we know that

$$\int_0^\infty \frac{x}{\sqrt{x^6+1}}dx \text{ converges.}$$ ■

8.2.2 The Limit Comparison Test

The Basic Comparison Test is just that, basic. Often times the obvious inequality that you want actually ends up going in the wrong direction, yet the integrals are so similar that you feel like you should still be able to compare them. Example 8.18 below will demonstrate precisely this.

The Limit Comparison Test will fix this by asking the question: "Do f and g grow at roughly the same rate?"

Theorem 8.17: The Limit Comparison Test

Let f, g be integrable functions on all subintervals of $[a,\infty)$ and satisfy $0 \leq f(x) \leq g(x)$. If

$$0 < \lim_{x\to\infty} \frac{f(x)}{g(x)} < \infty$$

then $\int_a^\infty f(x)dx$ converges if and only if $\int_a^\infty g(x)dx$ converges.

The statement that $\frac{f(x)}{g(x)}$ converges to some finite, non-zero number, means that f and g grow asymptotically at the same speed, up to some multiplicative constant (which is precisely the value of the limit). Note that if the limit is 0, eventually one must have $g(x) \geq f(x)$ and can use the Basic Comparison Test appropriately. Similarly, if the limit is ∞, then eventually $f(x) \geq g(x)$ and again the Basic Comparison Test can be used.

Also notice that the ratio does not matter, for if $\frac{f(x)}{g(x)} \to L$ which is finite and positive, then $\frac{g(x)}{f(x)} \to \frac{1}{L}$ which is also finite and positive.

Proof. Assume that $\frac{f(x)}{g(x)} \xrightarrow{x\to\infty} L$, so that there exists some $N \in \mathbb{R}$ such that whenever $x > N$ we have

$$L - 1 \leq \frac{f(x)}{g(x)} \leq L + 1.$$

Assume that $\int_a^\infty g(x)dx$ converges, and note that the above equation implies that $f(x) \leq (L+1)g(x)$ for all $x > N$. Hence for $x > N$ we have

$$\int_a^x f(x)dx = \int_a^N f(x) + \int_N^x f(x)dx \leq \int_a^N f(x)dx + (L+1)\int_N^x g(x)dx.$$

In the limit as $x \to \infty$ both terms are the right-hand-side are finite, so by the Basic Comparison Test $\displaystyle\int_a^\infty f(x)dx$ converges.

If one assume that $\displaystyle\int_a^\infty f(x)dx$ converges, then we repeat the argument above but use the fact that $f(x) \geq (L-1)g(x)$ and again apply the Basic Comparison Test. □

Example 8.18

Determine whether $\int_1^\infty \frac{1}{\sqrt{1+x}}dx$ converges or diverges.

Solution. If one were to try to use the Basic Comparison Test, the obvious inequality is that $\frac{1}{\sqrt{1+x}} \leq \frac{1}{\sqrt{x}}$. But this does not tell us anything! The right-hand-side diverges, and so does not impose its will on the left-hand-side. Instead, we recognize that

$$\lim_{x\to\infty} \frac{\sqrt{x}}{\sqrt{1+x}} = 1.$$

Thus by the Limit Comparison Test, since $\int_1^\infty \frac{1}{\sqrt{x}}dx$ diverges, we necessarily have that $\int_1^\infty \frac{1}{\sqrt{1+x}}dx$ diverges as well. ■

Example 8.19

Determine whether $\displaystyle\int_1^\infty \frac{x^2+1}{x^4+3x^2-4x+1}dx$ converges or diverges.

Solution. With a strong enough argument, one might be able to argue this example using the Basic Comparison Test, but the Limit Comparison Test proves much simpler. Again the idea is to look at how the numerator and denominator grow asymptotically. The numerator grows as x^2, while the denominator grows as x^4, meaning that the combined system grows as $\frac{1}{x^2}$. To invoke the Limit comparison Test, we must compute the limit of the ratio of these functions:

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

Since $\int_1^\infty \frac{1}{x^2}dx$ converges (by Proposition 8.11), we conclude by the Limit Comparison Test that $\displaystyle\int_1^\infty \frac{x^2+1}{x^4+3x^2-4x+1}dx$ also converges. ■

9 Sequences and Series

Our goal is to now analyze what happens when we deal with discrete, rather than continuous, information. Our first subject will be to consider sequences and their limits, which will be analogous to the study of differential calculus. Shortly after sequences, we will begin studying series which is the analogous to the integral calculus. At a very rough level, we have seen sequences before. We think of sequences as an ordered collection of elements, such as

$$\left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\right\}. \tag{9.1}$$

Sets are not typically ordered though, so there is a more fruitful way of thinking about sequences that will often allow us to capitalize on our knowledge of calculus.

9.1 The Basics

Definition: The idea is as follows: thus far we have been examining maps $f : \mathbb{R} \to \mathbb{R}$, functions which take a real number as an input and produce a real number as an output. A sequence is a map $a : \mathbb{N} \to \mathbb{R}$ which takes in natural numbers as input and produces real numbers as output. For example, the function $a(n) = \frac{1}{n}$ is such a sequence, and we have

$$a(1) = 1, \quad a(2) = \frac{1}{2}, \quad a(3) = \frac{1}{3}, \quad a(4) = \frac{1}{4}, \ldots.$$

This notation is often clumsy though, so instead we abbreviate $a_1 = a(1), a_2 = a(2)$ and so on, with a general element of the sequence written as $a_n = a(n)$. The *image* of a sequence is the set

$$(a_n)_{n=1}^{\infty} = \{a_n : n \in \mathbb{N}\}$$

Hence the image of the sequence $a_n = 1/n$ is exactly the set given in (9.1). We will often make no distinction between the sequence as a function a_n, and its image set $(a_n)_{n=1}^{\infty}$. Notice that this latter notation is particularly useful if we want to change the index at which the sequence starts. In fact, if the indexing set is implicitly known we may sometimes be lazy and only write (a_n).

One can take infinite subsets of $\mathbb{N}$ and, up to re-ordering, get something which is still labelled by $\mathbb{N}$. Equivalently, it is possible to take infinite subsets of the set $\{a_1, a_2, a_3, \ldots\}$ and hence get another sequence. This leads us to the following definition:

Definition 9.1

If a_n is a sequence, a *subsequence* is an ordered subset[a] (a_{n_k}) of (a_n).

[a] The choice of notation a_{n_k} comes from the fact that if we think of a as a function $a : \mathbb{N} \to \mathbb{R}$ then we can think of n as a function $n : \mathbb{N} \to \mathbb{N}$ for which $n_k = n(k)$. Thus $a_{n_k} = a(n(k))$ is a composition of functions.

For example, the sequence $a_{n_k} = \frac{1}{2^k}$ is a subsequence of the sequence $a_n = \frac{1}{n}$. Indeed, notice that

$$(a_{n_k}) = \left\{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots\right\} \subseteq (a_n) = \left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\right\}.$$

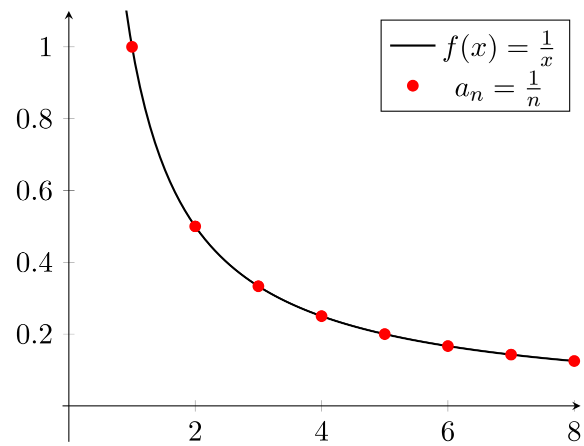


Figure 26: The sequence $a_n = \frac{1}{n}$ can be recognized as the restriction of the function $f(x) = \frac{1}{x}$ to the positive integers. This sequence is decreasing and is bounded.

Relation to Elementary Functions: We know that $\mathbb{N} \subseteq \mathbb{R}$ and so often one can think of sequences as the restriction of a function $f : \mathbb{R} \to \mathbb{R}$ to just the natural numbers. For example,

1. $a_n = \frac{1}{n}$,
2. $a_n = \sin n$,
3. $a_n = \frac{e^n}{n+1}$,
4. $a_n = \sqrt{n}$,
5. $a_n = 2^n$,
6. $a_n = n^n$.

are all sequences which correspond to functions. Consequently, many notions for functions on $\mathbb{R}$ have a corresponding notion for sequences:

Definition 9.2

If a_n is a sequence then we say that

1. a_n is *increasing* if $a_n < a_{n+1}$ and *non-decreasing* if $a_n \leq a_{n+1}$,
2. a_n is *decreasing* if $a_n > a_{n+1}$ and *non-increasing* if $a_n \geq a_{n+1}$,
3. a_n is *bounded above* if there exists $M \in \mathbb{R}$ such that $a_n \leq M$ for all $n \in \mathbb{N}$,
4. a_n is *bounded below* if there exists $m \in \mathbb{R}$ such that $a_n \geq m$ for all $n \in \mathbb{N}$,
5. a_n is *bounded* if it is bounded above and below.

Example 9.3

Show that the sequence $a_n = n + \sin(n)$ is a non-decreasing sequence, bounded from below but not from above.

Solution. Evaluating $\sin(n)$ for an arbitrary integer is not a simple task; however, since we can think of a_n as the restriction of the function $f(x) = x + \sin(x)$, we can use differential calculus to help solve our problem. Notice that $f'(x) = 1+\cos(x) \geq 0$, which means that f is a non-decreasing function. In turn, the sequence a_n must then also be non-decreasing (convince yourself of this).

Since the function is non-decreasing, it is bounded from below by its first term: $a_1 = 1+\sin(1)$. To see that it is not bounded above, we have that for any $M > 0$ the element $a_{M+1} = M+\sin(M) \geq M$, showing that a_n grows unbounded. ■

Exercise: One can convince his/herself that if $f(x)$ is an increasing function, then the sequence defined by $a_n = f(n)$ is also increasing. A similar notion holds for decreasing and bounded. However, the converse is certainly not true. Give an example of a sequence a_n and a real function f such that $a_n = f(n)$, a_n is increasing, but f is not increasing. Give a similar example in the case when a_n is bounded but f is not.

Sequences without Elementary Relations: There are sequences which have (*a priori*) no relationship to functions on $\mathbb{R}$, and hence must really be thought of purely in terms of a function on $\mathbb{N}$. The prototypical example of a sequence with no $\mathbb{R}$-function equivalent[26] is the sequence $a_n = n!$, whose first few terms are given by

$$a_1 = 1,\ a_2 = 2,\ a_3 = 6,\ a_4 = 24,\ a_5 = 120,\ a_6 = 720,\ a_7 = 5040, \ldots.$$

Another class of example are those functions which are defined *recursively*. For example, defining $a_1 = 1$ and setting $a_n = \sqrt{a_{n-1}+1}$ yields the sequence[27]

$$a_1 = 1,\ a_2 = \sqrt{2},\ a_3 = \sqrt{\sqrt{2}+1},\ a_4 = \sqrt{\sqrt{\sqrt{2}+1}+1}, \ldots$$

or equivalently

$$a_1 = \sqrt{1},\ a_2 = \sqrt{1+\sqrt{1}},\ a_3 = \sqrt{1+\sqrt{1+\sqrt{1}}}, a_4 = \sqrt{1+\sqrt{1+\sqrt{1+\sqrt{1}}}}, \ldots$$

A rather famous example is given by the *Fibonacci sequence*. Define $a_0 = a_1 = 1$ and set $a_n = a_{n-1} + a_{n-2}$ for $n \geq 2$. The first few terms of this sequence are given by[28]

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, \ldots.$$

[26] There *is* a function-equivalent of the factorial sequence, though it is not quite amenable to the type of analysis we would like to apply. The function is given by improper integral

$$a_n = \Gamma(n+1) = \int_0^\infty x^n e^{-x} dx.$$

[27] This is a very important example of a recursive sequence. To see why, consider Exercise 9.2

[28] Again, there is a real function which defines this sequence: If $\varphi_\pm = \frac{1}{2}\left(1 \pm \sqrt{5}\right)$ then the function F defined on $[1, \infty)$ given by $F(x) = \frac{1}{\sqrt{5}}\left[\varphi_+^x - \varphi_-^x\right]$ restricts to the Fibonacci sequence.

9.2 Limits of Sequences

Treating sequences as functions, and especially realizing the analogy with real functions, allows us to talk about the notion of convergence of a sequence. Roughly speaking, we will say that a sequence converges if it behaves as though the function $a : \mathbb{N} \to \mathbb{R}$ has a horizontal asymptote. If we have this idea in our mind, we can immediately adapt the notion of a horizontal asymptote for real functions to sequences:

Definition 9.4

A sequence a_n is said to converge to a limit $L \in \mathbb{R}$ if for every $\epsilon > 0$ there exists an $M \in \mathbb{N}$ such that whenever $k \geq M$ we have $|a_k - L| < \epsilon$. In this case we write

$$\lim_{n\to\infty} a_n = L, \qquad \text{or} \qquad (a_n) \xrightarrow{n\to\infty} L.$$

Example 9.5

Show that the sequence $a_n = \frac{1}{n}$ converges to 0.

Solution. We proceed in precisely the same fashion as we would do for a function. Let $\epsilon > 0$ be given, and choose a positive integer M such that $\frac{1}{M} < \epsilon$. Then if $k \geq M$ we have

$$|a_k - 0| = \frac{1}{K} \leq \frac{1}{M} < \epsilon,$$

so $(a_n) \to 0$ as required. ■

Exercise: Show that a sequence (a_n) converges if and only if every subsequence (a_{n_k}) converges.

Many of the Limit Laws we saw in Theorem 2.21 hold for sequences as well:

Theorem 9.6: Limit Laws for Sequences

Let $(a_n) \to L$ and $(b_n) \to M$ be convergent sequences.

1. The sequence $(a_n + b_n)$ is convergent and $(a_n + b_n) \to L + M$,
2. The sequence $(a_n b_n)$ is convergent and $(a_n b_n) \to LM$,
3. For any $\alpha \in \mathbb{R}$ the sequence (αa_n) converges and $(\alpha a_n) \to \alpha L$,
4. If $M \neq 0$ then the sequence (a_n/b_n) converges and $(a_n/b_n) \to L/M$.

Proof. The proof of these are almost identical to those of Theorem 2.21. We will prove (1) and leave the remainder as an exercise.

Assume that $(a_n) \to L$ and $(b_n) \to M$. Let $\epsilon > 0$ be given and choose $M_1, M_2 \in \mathbb{N}$ such that if $k \geq M_1$ then $|a_n - L| < \frac{\epsilon}{2}$ and if $k \geq M_2$ then $|b_n - M| < \frac{\epsilon}{2}$. Let $M = \max M_1, M_2$, so that if $k \geq M$ then

$$|a_n + b_n - (L + M)| \leq |a_n - L| + |b_n - M| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \qquad \square$$

The Squeeze Theorem also carries over directly:

Theorem 9.7: Squeeze Theorem for Sequences

Assume that for sufficiently large k we have $a_n \leq b_n \leq c_n$. If $(a_n) \to L$ and $(c_n) \to L$, then (b_n) is also convergent with limit L.

Once again, the proof of this theorem is almost identical to that of Theorem 2.34.

The two main theorems that we want to prove again following almost identically from those which we have already proven:

Theorem 9.8

Every convergent sequence is bounded.

Proof. The idea is that because the sequence converges, at some point infinitely many terms of the sequence must lie in any ϵ-band of its limit point. Since there are only finitely many other terms, the function cannot grow infinitely large.

More formally, let L be the limit point of (a_n), and choose $N \in \mathbb{N}$ such that for all $k \geq N$ we have $|a_k - L| < 1$. In particular, this means that $L - 1 < a_k < L + 1$, so that the sequence is always bounded after N. Now set

$$M = \max \{|a_1|, \ldots, |a_M|, |L+1|, |L-1|\}.$$

Certainly if $k \leq N$ then $a_k \leq M$ by definition of M, and if $k \geq N$ then $a_k < M$ by our previous line. $\square$

The following theorem is often referred to as the *Monotone Convergence Theorem* and was mentioned in Section 8.2.1. It is important enough to be worth proving again explicitly for sequences, but the diligent student will immediately recognize that the proof is almost identical to that of Theorem 8.13.

Theorem 9.9: Monotone Convergence Theorem

If (a_n) is bounded from above and non-decreasing, then (a_n) is convergent with its limit given by $\sup_{n \in \mathbb{N}} a_n$.

Proof. Let $L = \sup_n a_n$ and let $\epsilon > 0$ be given. By Proposition 5.4 we know that there exists some $M \in \mathbb{N}$ such that

$$L - \epsilon < a_M \leq L.$$

Since (a_n) is non-decreasing, we have that for all $k \geq M$

$$L - \epsilon < a_M < a_k \leq L < L + \epsilon;$$

that is, $|a_n - L| < \epsilon$. Hence $(a_n) \to L$ as required. □

Exercise: Prove the following (equivalent) Corollary of the Monotone Convergence Theorem. If (a_n) is non-increasing and bounded from below, then (a_n) is convergent and its limit is $L = \inf_{n\in\mathbb{N}} a_n$.

Example 9.10

Determine whether the sequence $a_n = \frac{2^n}{n!}$ is convergent.

Solution. We have

$$\frac{a_{n+1}}{a_n} = \frac{2^{n+1}}{(n+1)!}\frac{n!}{2^n} = \frac{2}{n+1},$$

so that if $n \geq 2$ we have $a_{n+1} < a_n$ and the sequence is decreasing. It is easy to see that a_n is always positive, and so by the Monotone Convergence Theorem we know that (a_n) converges. ■

Example 9.11

Consider the sequence defined by $a_1 = 1$ and $a_{n+1} = \frac{n}{2n+1}a_n$. Determine whether this limit converges and if so, find the limit.

Solution. We begin by noticing that this sequence is decreasing, since $\frac{a_{n+1}}{a_n} = \frac{n}{2n+1} < 1$ which implies that $a_{n+1} < a_n$. Furthermore, the sequence is bounded below by 0; that is, we claim that every a_n is positive. This follows quite quickly by induction, for clearly $a_1 = 1 > 0$ and $a_{n+1} = \frac{n}{2n+1}a_n > 0$ by the induction hypothesis. By (Corollary to) the Monotone Convergence Theorem, (a_n) converges, say to L. Now taking the limit as $n \to \infty$ of the equation $a_{n+1} = \frac{n}{2n+1}a_n$ we get

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

for which the only possible real value of L is $L = 0$. ■

Exercise: Consider the sequence $a_1 = 1$ and $a_n = \sqrt{1 + a_{n-1}}$. Show that this sequence is increasing and bounded from above. Conclude that it converges. Show that the limit of this sequence is $L = \frac{1+\sqrt{5}}{2}$: the golden ratio.

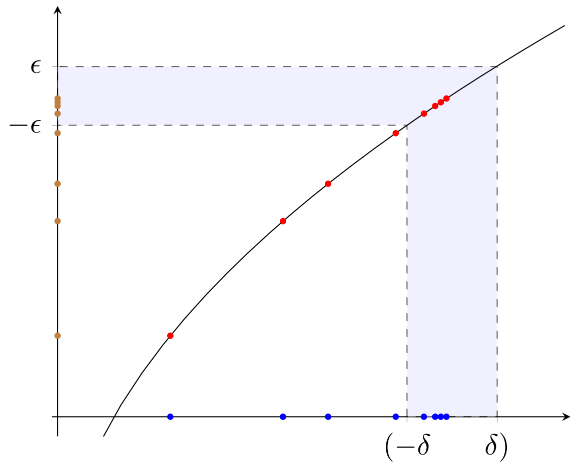


Figure 27: If $(a_n) \to L$, then by going far enough into our sequence (blue) we can guarantee that we will be in δ-neighbourhood of L. The image of these points are the $f(a_n)$ (brown), which live in the desired ϵ-neighbourhood because of the continuity of f.

9.2.1 Continuous Functions

There are a plethora of ways of defining continuous functions, and now we have the tools to introduce yet another.

Theorem 9.12

A function f is continuous if and only if whenever $(a_n) \to L$ then $(f(a_n)) \to f(L)$.

One often says that 'continuous functions preserve convergent sequences' or that 'continuous functions map convergent sequences to convergent sequences.'

Proof. Assume that f is continuous, and let $(a_n) \to L$. We want to show that $(f(a_n)) \to L$. Let $\epsilon > 0$ be given. Since f is continuous, there exists a $\delta > 0$ such that for each x satisfying $|x - L| < \delta$ we have $|f(x) - f(L)| < \epsilon$. Since (a_n) is convergent, there exists an $N \in \mathbb{N}$ such that for all $n \geq N$ we have $|a_n - L| < \delta$. Combining these, we see that whenever $n \geq N$ we have

$$|a_n - L| < \delta, \text{ and so } |f(a_n) - f(L)| < \epsilon.$$

which is exactly what we want to show.

Conversely, assume that f is not continuous, say at c. Hence there exists an $\epsilon > 0$ such that for any $\delta > 0$ there is an x such that $|x - c| < \delta$ and $|f(x) - f(c)| \geq \epsilon$. For each $\delta_n = \frac{1}{n}$, choose an element x_n satisfying $|x_n - c| < \delta_n$ and $|f(x_n) - f(c)| \geq \epsilon$. Then $(x_n) \to c$ but $f(x_n)$ does not converge to $f(c)$. □

9.3 Infinite Series

Section 5.2 introduced sigma notation as a convenient way of writing down finite sums; namely,

$$\sum_{i=m}^{n} a_i = a_m + a_{m+1} + \cdots + a_{n-1} + a_n.$$

The summand here conveniently looks like a sequence, though with finitely many terms. In fact, we will be extending the notion of the finite sum above to an infinite sum, called an *infinite series*.

Definition 9.13

Let $(a_n)_{n=1}^{\infty}$ be a sequence. Define the n^{th} *partial sum* of (a_n) to be

$$S_n = \sum_{k=1}^{n} a_k.$$

We say that the infinite series $\sum_{n=1}^{\infty} a_n$ *converges* if the sequence $(S_n)_{n=1}^{\infty}$ converges as a sequence, and we set

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

We say that the infinite series *diverges* otherwise.

Notice that $S_n = S_{n-1} + a_n$. For example, if we let $a_n = \frac{1}{n}$, then our first few partial sums are given by

$$S_1 = 1, S_2 = S_1 + \frac{1}{2} = 1 + \frac{1}{2} = \frac{3}{2}, S_3 = S_2 + \frac{1}{3} = \frac{11}{6}, S_4 = S_3 + \frac{1}{4} = \frac{25}{12}, \ldots.$$

It is not at all clear from looking at the partial sums whether or not this sequence converges. Shortly, we will analyze techniques that will allow us to determine whether series converge (though we will rarely be able to determine the limit to which the series converges).

Here is an important distinction to make, as students often confuse sequences and series:

- A *sequence* is a collection of elements $(a_n)_{n=1}^{\infty}$. Note that a_n need not be in any way related to a_{n-1}. The analogy to keep in mind that is sequences behave like functions.
- A *series* is a sum of elements of a sequence, with the partial sums being related via $S_n = S_{n-1} + a_n$. The analogy to keep in mind is that a sum is like an improper integral of a function.

This analogy between sequences/series and functions/integrals is an important one to keep in mind.

Since the convergence of a series is defined in terms of the limits of its partial sums (a sequence), the limit laws for sequences immediately give us the following result:

Theorem 9.14

Let $\sum_{k=1}^{\infty} a_n$ and $\sum_{k=1}^{\infty} b_n$ be convergent series.

1. The sum of the series is convergent and
$$\sum_{k=1}^{\infty}(a_n + b_n) = \sum_{k=1}^{\infty} a_n + \sum_{k=1}^{\infty} b_n.$$
2. For any $\alpha > 0$, we have
$$\sum_{k=1}^{\infty}(\alpha a_n) = \alpha \sum_{k=0}^{\infty} a_n.$$

Proof. Let $(s_n) \to L$ be the partial sums of $\sum_{k=1}^{\infty} a_n$ and $(t_n) \to M$ be the partial sums for $\sum_{k=1}^{\infty} b_n$. The partial sum of the sum is given by
$$u_n = \sum_{k=1}^{\infty}(a_n + b_n) = \sum_{k=1} a_n + \sum_{k=1} b_n = s_n + t_n.$$
By Theorem 9.6, we have $(u_n) = (s_n + t_n) \to L + M$, and so
$$\sum_{k=1}^{\infty}(a_n + b_n) = \sum_{k=1}^{\infty} a_n + \sum_{k=1}^{\infty} b_n.$$
The proof for (2) is similar. □

Intuitively, one should expect that we require $(a_n) \to 0$ to even have a chance for the infinite series to converge. Otherwise, if $(a_n) \to L \neq 0$ then the infinite series would effectively be 'adding L to itself infinitely many times,' which would certainly yield a non-finite number. The following is the more precise way of saying this:

Theorem 9.15

If $\sum_{k=1}^{\infty} a_k$ converges, then $(a_k) \to 0$.

Proof. Let (S_n) be the partial sums, so that $(S_n) \to L$ for some limit L. Since $S_n = S_{n-1} + a_n$ we have $a_n = S_n - S_{n-1}$, so that $(a_n) \to L - L = 0$. □

The contrapositive of this is exactly what we mentioned above: "If $(a_k) \not\to 0$ then $\sum_{k=1}^{\infty} a_k$ does not converge." However, the converse of Theorem 9.15 is not true. We do not yet have the techniques to show that this is the case, but the series
$$\sum_{k=1}^{\infty} \frac{1}{k} \quad \text{does not converge,}$$

despite the fact that $\frac{1}{k} \xrightarrow{k\to\infty} 0$.

9.3.1 Some Special Series

Geometric Series: A geometric series is a series defined by a sequence (a_n) where $a_n = ra_{n-1}$ for some $r \in \mathbb{R}$. For example,

$$a_0 = 1,\ a_1 = \frac{1}{2},\ a_2 = \frac{1}{4},\ a_3 = \frac{1}{8},\ a_4 = \frac{1}{16}, \ldots$$

satisfies the relation $a_n = \frac{1}{2}a_{n-1}$. We can write such series as

$$\sum_{k=0}^{\infty} r^k.$$

Theorem 9.16

For any $|r| < 1$ we have $\displaystyle\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$, and the series diverges otherwise.

Proof. The partial sums are given by $S_n = 1 + r + r^2 + \cdots + r^n$ and so

$$\begin{aligned}(1-r)s_n &= 1 + r + r^2 + r^3 + r^4 + \cdots + r^n \\ &\qquad - r - r^2 - r^3 - r^4 + \cdots - r^n - r^{n+1} \\ &= 1 - r^{n+1}\end{aligned}$$

so that $s_n = \dfrac{1-r^{n+1}}{1-r}$. Now in the limit as $n \to \infty$ we have that $r^{n+1} \xrightarrow{n\to\infty} 0$, so

$$\sum_{k=0}^{\infty} r^k = \lim_{n\to\infty} \frac{1-r^{n-1}}{1-r} = \frac{1}{1-r}.$$ ☐

An interesting if not sophistic argument that proves the same result is to see that

$$\begin{aligned}(1-r)(1 + r + r^2 + r^3 + \cdots) &= 1 + r + r^2 + r^3 + \cdots \\ &\qquad - r - r^2 - r^3 - \cdots \\ &= 1.\end{aligned}$$

In an entirely formal sense (that is, treating r purely as a symbol without assigning it any value) we see that $(1-r)$ is the multiplicative inverse of $(1 + r + r^2 + \cdots)$, giving the desired results as well. Determining from this which values of r actually make sense is something that we will see in a later chapter.

Telescoping Series: Telescoping series are ones in which many of the internal summands cancel one another out. For example, an example of a telescoping sum arises when we consider the Riemann sum of a constant function. If $f(x) = c$ and $P = \{a = x_0, x_1, \ldots, x_{n-1}, x_n = b\}$ is a partition, then the Riemann sum is

$$U_f(P) = \sum_{i=1}^{n} f(x_i^*)(x_i - x_{i-1}) = c \sum_{i=1}^{n}(x_i - x_{i-1}).$$

Notice what happens if we expand out a few terms of the sum: we get

$$(x_1 - x_0) + (x_2 - x_1) + (x_3 - x_2) + \cdots + (x_{n-1} - x_{n-2}) + (x_n - x_{n-1}).$$

For that majority of terms, both x_i and $-x_i$ appear in the summation and hence cancel one another out. The only terms which survive are the first and last, so that the sum reduces to $x_n - x_0$.

$$U_f(P) = c(x_n - x_0) = c(b - a)$$

exactly as expected.

Example 9.17

Determine the value of the series $\displaystyle\sum_{k=1}^{\infty} \frac{1}{k(k+1)}$ if it converges.

Solution. The trick is to realize our sum as a telescoping series. Using partial fractions, we get that

$$\frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1}$$

so that the n^{th} partial fraction is given by

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

We now take the limit as $n \to \infty$ and we find that

$$\sum_{k=1}^{\infty} \frac{1}{k(k+1)} = \lim_{n\to\infty} \sum_{k=1}^{n} \frac{1}{k(k+1)} = \lim_{k\to\infty} \frac{k}{k+1} = 1. \qquad \blacksquare$$

9.4 Convergence Tests

The analogy between infinite series and improper integrals continues in this section with the introduction of the basic and limit comparison test. Later, we will develop several more tests which allow us to infer convergence of infinite series.

9.4.1 Comparison Tests

In section 8.2 we saw the Basic Comparison Test (Theorem 8.14) and Limit Comparison Test (Theorem 8.17) for improper integrals. The statement and proof of these tests for series are almost identical. We will give the statement, but leave the proofs as an exercise:

Theorem 9.18: Basic Comparison Test for Series

Suppose that (a_n) and (b_n) are sequences such that $0 \leq a_k \leq b_k$ for sufficiently large k.

1. If $\sum_k b_k$ converges, then $\sum_k a_k$ converges.
2. If $\sum_k a_k$ diverges, then $\sum_k b_k$ diverges.

The idea here is precisely the same as that of Theorem 8.14: If the larger of the two series converges, it forces the smaller sequence to also converge. In the opposite direction, if the smaller sequence diverges, then the larger sequence must also diverge.

Example 9.19

Determine whether the sequence $\displaystyle\sum_{k=1}^{\infty} \frac{7^n}{8^n+2}$ converges.

Solution. By considering the terms which grow fastest in this expression, we suspect that the 2 in the denominator should not affect the long term behaviour of the function. We thus have the comparison

$$\frac{7^n}{8^n+2} \leq \frac{7^n}{8^n} = \left(\frac{7}{8}\right)^n.$$

Now the series $\sum_k \left(\frac{7}{8}\right)^n$ is a geometric series with ratio less than 1, and hence converges. We conclude by the Basic Comparison Test the $\sum_k \frac{7^n}{8^n+2}$ converges as well. ■

Theorem 9.20: Limit Comparison Test for Series

If (a_n) and (b_n) are sequences with positive terms, and $\displaystyle\lim_{n\to\infty} \frac{a_k}{b_k} = L$ for some $0 < L < \infty$ then

$$\sum_k a_k \text{ converges} \quad \Leftrightarrow \quad \sum_k b_k \text{ converges.}$$

The fact that a_k/b_k converges to some positive, non-zero number means that a_k and b_k grow at 'approximately the same speed.' Hence convergence or divergence of one will immediately imply convergence/divergence of the other.

Example 9.21

Determine whether the sequence $\sum \frac{k^{10}+25k^7+1}{k^{12}-20}$ converges or diverges.

Solution. As mentioned above, we only care about the terms which grow the fastest, which means if we want to guess we just consider the biggest factors in each of the numerator and the denominator. Those terms are $\frac{k^{10}}{k^{12}} = \frac{1}{k^2}$ and since we know that the series $\sum \frac{1}{k^2}$ converges, it seems likely that so too will our series above. Indeed, this actually gives us the other series we should use in our Limit Comparison Test. Notice that

$$\begin{aligned} \lim_{k\to\infty} \frac{\frac{1}{k^2}}{\frac{k^{10}+25k^7+1}{k^{12}-20}} &= \lim_{k\to\infty} \frac{k^{12}-20}{k^{12}+25k^9+k^2} \frac{1/k^{12}}{1/k^{12}} \\ &= \lim_{k\to\infty} \frac{1-20/k^{12}}{1+25 1/k^2 + 1/k^{10}} \\ &= 1. \end{aligned}$$

Thus the Limit Comparison Test tells us that the series converges. ■

Example 9.22

If a_n are such that $\sum \frac{1}{a_n}$ converges, show that $\sum \frac{1}{a_n+M}$ converges for any $M > 0$.

Solution. Since constants neither shrink nor grow, we do not expect them to affect convergence of the limit, and we can use both the Basic Comparison Test.

We realize that $a_n < a_n + M$ so that $\frac{1}{a_n} > \frac{1}{a_n+M}$. Thus $\sum \frac{1}{a_n+M} < \sum \frac{1}{a_n}$ and $\sum \frac{1}{a_n}$ converges, so $\sum \frac{1}{a_n+M}$ will also converge.

On the other hand, by using the Limit Comparison Test we can actually extend this result to work for any $M \in \mathbb{R}$. Recall that since $\sum \frac{1}{a_n}$ converges, it must be the case that $\frac{1}{a_n} \to 0$ as $n \to \infty$. The Limit Comparison Test then tells us that

$$\begin{aligned} \lim_{n\to\infty} \frac{1/a_n}{1/a_n+M} &= \lim_{n\to\infty} \frac{a_n+M}{a_n} \\ &= \lim_{n\to\infty} 1 + \frac{M}{a_n} = 1 \end{aligned}$$

and so again we conclude that $\sum \frac{1}{a_n+M}$ converges. ■

9.4.2 Non-Comparison Tests

The following tests are, in a sense, more appealing than the above test, since they do not require us to finagle a sequence against which to compare.

The Integral Test: This tests takes our series/integral analogy to the next level, but allowing us to directly compare them.

Theorem 9.23: The Integral Test

If f is a continuous, non-negative, decreasing function on $[1, \infty)$ then

$$\sum_{k=1}^{\infty} f(k) \text{ converges} \quad \Leftrightarrow \quad \int_1^{\infty} f(x)dx \text{ converges.}$$

Proof. Since f is continuous, it is integrable on every subinterval $[1, b]$ of $[1, \infty)$; that is, for any partition P of $[1, b]$ we have

$$L_f(P) \leq \int_1^n f(x)dx \leq U_f(P)$$

Consider the interval $[1, n]$ and the partition $P = \{1, 2, 3, \ldots, n\}$ consisting n equal subintervals, each of length 1. Since our function is decreasing, we know that

$$\min_{x \in [i,i+1]} f(x) = f(i+1), \qquad \max_{x \in [i,i+1]} f(x) = f(i), \quad i = 1, \ldots, n-1$$

so that $f(i+1) \leq f(x) \leq f(i)$ for all $x \in [i, i+1]$. In terms of upper and lower Riemann sums, we get

$$\sum_{k=2}^{n} f(k) = L_f(P) < \int_1^n f(x)dx < U_f(P) = \sum_{k=1}^{n-1} f(k).$$

From the second inequality we get

$$\begin{aligned} \int_1^{\infty} f(x)dx &= \lim_{n\to\infty} \int_1^n f(x)dx \\ &\leq \lim_{n\to\infty} \sum_{k=1}^{n-1} f(k) \\ &= \sum_{k=1}^{\infty} f(k) \end{aligned}$$

which tells us that if $\sum_k f(k)$ converges, then so too does $\int_1^\infty f(x)dx$. Conversely, the other inequality shows us that

$$\begin{aligned} \sum_{k=1}^{\infty} f(k) &= f(1) + \lim_{n\to\infty} \sum_{k=2}^{n} f(k) \\ &< f(1) + \lim_{n\to\infty} \int_1^n f(x)dx \\ &= f(1) + \int_1^{\infty} f(x)dx. \end{aligned}$$

Thus if $\int_1^\infty f(x)dx$ converges, so too must the infinite series. □

Example 9.24

Consider the series whose terms are given by $a_n = \frac{1}{n}$. Determine the convergence/divergence of this series.

Solution. By now, it is my hope that you have seen this example an enumerable number of times. This is what is called the *Harmonic Series* and is the classic example of how to trick first year students into making mistakes. It is clear that $a_n \to 0$ as $n \to \infty$, but $\sum a_n$ diverges. Indeed, the integral test tells us that $\sum \frac{1}{n}$ converges if and only if $\int_1^n \frac{1}{x} dx$ converges, but

$$\int_1^n \frac{1}{x} dx = \log(x) \Big|_1^\infty = \infty.$$

However, regardless of how many times I emphasize the point that "Just because the sequence converges to zero, does not mean the series converges" there are countless students who get this wrong! Do not make the same mistake, and learn this classic example immediately. ∎

Example 9.25

Determine whether the following series converges: $\sum e^{-n}$.

Solution. Here, the terms of the series are given by the function $f(x) = e^{-x}$. Applying the integral test, we have

$$\int_1^\infty e^{-x} dx = -e^{-x} \Big|_1^\infty = \frac{1}{e}$$

so $\sum e^{-n}$ converges. Note that since the index occurs as a power, we can also apply the root test. Indeed, setting $a_n = e^{-n}$ we have that $\sqrt[n]{e^{-n}} = e^{-1}$. Since $e^{-1} < 1$ it then follows that the series converges. ∎

Ratio Test: One of the most powerful test, the ratio test is somewhat similar to the Limit Comparison Test at first glance, but only uses sequential terms of the sequence itself.

Theorem 9.26: Ratio Test

Let (a_n) be a sequence such that $\left|\frac{a_{k+1}}{a_k}\right| \xrightarrow{n\to\infty} L$.

1. If $L < 1$ then $\displaystyle\sum_{k=1}^\infty a_k$ converges,
2. If $L > 1$ then $\displaystyle\sum_{k=1}^\infty a_k$ diverges,
3. If $L = 1$ then the test is inconclusive.

Proof. The idea of the proof is as follows: If the ratio converges to $L \neq 1$, then for sufficiently large N we can treat $a_{k+1} = La_k$ for all $k \geq N$. At this point the series effectively behaves like a geometric series $a_N \sum_i L^i$, which will converge if $L < 1$ and diverge if $L > 1$.

Assume that $L < 1$ and choose some ℓ such that $L < \ell < 1$. Since the ratio converges to L, we know there exists some $N \in \mathbb{N}$ such that if $k \geq N$ then $|a_{k+1}/a_k| < \ell$ (convince yourself of this, using ϵ-N if necessary). In particular, this means that $|a_{k+1}| < \ell |a_k|$. In particular,

$$|a_{N+1}| < \ell |a_N|, \quad |a_{N+2}| < \ell^2 |a_N|, \quad |a_{N+3}| < \ell^3 |a_N|, \ldots, |a_{N+k}| < \ell^k |a_n|.$$

Using the Basic Comparison Test, we have $|a_k| \leq \ell^k |a_N|$, with the right-hand-side being a geometric series with common ratio $\ell < 1$ and hence converging. We conclude that $\sum_k |a_k|$ converges as well, so $\sum_k a_k$ converges.

Precisely the same reasoning holds for (2), though this train of argument cannot be used when $L = 1$. □

Example 9.27

Determine whether the series $\sum \frac{2^k k!}{k^k}$ converges or diverges.

Solution. We might be tempted to use the root test here but the presence of the factorial term suggests that we should use the ratio test instead. Let $a_k = \frac{2^k k!}{k^k}$ so that

$$\begin{aligned}
\lim_{k\to\infty} \frac{a_{k+1}}{a_k} &= \lim_{k\to\infty} \frac{2^{k+1}(k+1)!}{(k+1)^{k+1}} \frac{k^k}{2^k k!} \\
&= \lim_{k\to\infty} \frac{2^{k+1}}{2^k} \frac{(k+1)!}{k!} \frac{k^k}{(k+1)^{k+1}} \\
&= 2 \lim_{k\to\infty} \frac{k^k (k+1)}{(k+1)^{k+1}} \\
&= 2 \lim_{k\to\infty} \left(\frac{k}{k+1} \right)^k \\
&= \frac{2}{e}
\end{aligned}$$

where the last line is an exercise in L'Hôpital's rule. By the Ratio Test, we then know that the series converges. ■

Root Test: The root test is similar in both statement and proof to the ratio test, but is far less useful. I will thus present the statement and an example, but leave the proof as an exercise.

Theorem 9.28: The Root Test

Let (a_n) be a sequence with non-negative terms, such that $(a_k^{1/k}) \xrightarrow{n\to\infty} L$.

1. If $L < 1$ then $\sum a_k$ converges,
2. If $L > 1$ then $\sum a_k$ diverges,
3. If $L = 1$ then the test is inconclusive.

Example 9.29

Determine whether the series $\sum \frac{k^k}{3^{k^2}}$ converges or diverges.

Solution. Since everything involves the index in the power, we should probably apply the root test. Indeed, set $a_k = \frac{k^k}{3^{k^2}}$ so that $\sqrt[k]{a_k} = \frac{k}{3^k}$. It is then clear (using L'Hôpital's rule) that $\sqrt[k]{a_k} \to 0$ as $k \to \infty$. By the root test, we then have that the series converges. ■

Example 9.30

Determine whether the series $\sum \frac{3^k+k^9}{k^k}$ converges or diverges.

Solution. This one is a bit tricky, as it will require us to use two tests. First of all, we notice that the terms which grow that fastest are given by $\frac{3^k}{k^k}$, and so we use the Limit Comparison Test to see that

$$\lim_{k\to\infty} \frac{3^k + k^9}{k^k}\frac{k^k}{3^k} = \lim_{k\to\infty} 1 + \frac{k^9}{3^k} = 1$$

where the last equality is an exercise in L'Hôpital's rule. Thus $\sum \frac{3^k+k^9}{k^k}$ converges if and only if $\sum \frac{3^k}{k^k}$ converges. Setting $a_k = \frac{3^k}{k^k}$ the root test tells us that

$$\begin{aligned}\lim_{k\to\infty} \sqrt[k]{\frac{3^k}{k^k}} &= \lim_{k\to\infty} \sqrt[k]{\left(\frac{3}{k}\right)^k} \\ &= \lim_{k\to\infty} \frac{3}{k} = 0\end{aligned}$$

and since this is certainly less than 1, we know that $\sum \frac{3^k}{k^k}$ converges. We conclude that the original series converges as required. ■

9.5 Kinds of Convergence

Alternating Series: If (a_n) is a sequence of positive numbers, an *alternating sequence* is any sequence of the form $\sum_k (-1)^k a_k$. Luckily, there is a very simple test to determine whether an alternating series converges:

Theorem 9.31: Alternating Series Test

If (a_k) is a positive decreasing sequence, the alternating series $\sum_k (-1)^k a_k$ converges if and only if $\lim_{k\to\infty} a_k = 0$.

The proof of this theorem is not terribly enlightening, so we omit it and refer the reader to the textbook.

Example 9.32

Determine whether $\displaystyle\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}$ converges.

Proof. By taking a negative sign out of the summation, our series is equivalent to $-\sum_k \frac{(-1)^k}{k}$, and this is an alternating series with $a_k = \frac{1}{k}$. These a_k are clearly limiting to zero, and hence by the alternating series test, the limit converges. □

Notice that the alternating series given in Example 9.32 is closely related to the Harmonic series which we actually know does not converge, and is often called the *alternating harmonic series*. In fact, we will (very luckily) be able to determine the exact value of the alternating harmonic series after we have seen Taylor series.

Absolute Convergence: If (a_n) is a sequence, the series $\sum a_k$ is said to *converge absolutely* if $\sum |a_k|$ converges. There is an important relationship between series which converge, and which converge absolutely.

Theorem 9.33: Absolute Convergence

If $\sum_k |a_k|$ converges then $\sum a_k$ converges; that is, all absolutely convergent series are convergent.

Proof. The crux of this proof follows from the fact that taking absolute values means that the terms of the series get larger, and hence are less likely converge. In fact, we shall use the Basic Comparison Test to show exactly this. Notice that since

$$-|a_k| \leq a_k \leq |a_k| \quad \text{then} \quad 0 \leq a_k + |a_k| \leq 2|a_k|.$$

The series formed by the terms on the right-hand-side is just $2\sum_k |a_k|$ and hence converges by assumption. By the Basic Comparison Test, it follows that the series $\sum_k (a_k + |a_k|)$ converges. Since $a_k = (a_k + |a_k|) - |a_k|$ the linearity of the series implies that

$$\sum a_k = \underbrace{\sum_k (a_k + |a_k|)}_{\text{finite}} - \underbrace{\sum |a_k|}_{\text{finite}} < \infty.$$ □

A natural question to ask at this point is whether the converse is true: Are all convergent series also absolutely convergent. The answer in this case is no: we saw in Example 9.32 that the alternating harmonic series is convergent, while taking the absolute value gives the harmonic series which is not convergent. Hence the alternating harmonic series is a convergent series which is not absolutely convergent. We often call such series *conditionally convergent*.

Example 9.34

Determine whether the series $\sum_k \frac{\cos(\pi k)}{k\sqrt{k}}$ converges or diverges.

Solution. Notice that $\cos(\pi k) = (-1)^{k+1}$, so that the terms of our series are actually given by $a_k = \frac{(-1)^{k+1}}{k\sqrt{k}}$. In absolute value, we have

$$|a_k| = \frac{1}{k\sqrt{k}}.$$

By the integral test, we know that $\sum_k |a_k| = \sum_k \frac{1}{k^{3/2}}$ converges. Since our series is absolutely convergent, by our theorem we know that it is actually convergent.

Note that we could also just use the alternating series test. ■

10 Power Series

10.1 Taylor Series

We now begin our study of using infinite series to write down functions, and is both a topic of great interest and great application. The idea is to use an 'infinite polynomial,' for which evaluating at a value $x = a$ will give the value of the function $f(a)$.

10.1.1 Taylor Polynomials and Taylor Remainder

As a prelude to discussing an infinite series which depends on a variable x, we should start by discussing the finite dimensional case: a polynomial. Our goal then is as follows: given a function f, find a polynomial p_n of degree n such that p_n is a good approximation to f. One way of doing this is via *Lagrange interpolation*, where we choose a random smattering of points and construct an n^{th} order polynomial which must pass through those points. If we are intelligent about which points to choose, our corresponding Lagrange polynomial may be a fair facsimile.

The problem with Lagrange interpolation is it is not canonical: there seems to be no determinate way of choosing the interpolation points. Instead, we will look towards using the derivatives of a function as an approximation. For example, we have seen that we can reconstruct curves, often just by knowing something about its critical points and inflection points. Taking this logic further, maybe we should try to specify our polynomial p_n to have its first n derivatives agree with f.

We'll have to be a bit more specific than that; namely, we should specify that the derivatives of f *at a point* a agree with those of p_n at a. To see how to formulate the theory, let us start by setting $a = 0$, after which we will appropriately generalize. Write

$$p_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0 = \sum_{k=0}^{n} a_k x^k.$$

Our goal is to choose the a_k such that $p_n^{(k)}(0) = f^{(k)}(0)$. Starting this process, notice that

$$\begin{aligned} p_n(0) &= a_0 \\ p_n'(0) &= a_1 \\ p_n''(0) &= 2a_2 \\ p_n^{(3)}(0) &= 3!a_3 \\ &\vdots \\ p_n^{(k)}(0) &= k!a_k. \end{aligned}$$

Hence if we want $f^{(k)}(0) = p_n^{(k)}(0) = k!a_k$ we should set $a_k = \dfrac{f^{(k)}(0)}{k!}$, and our polynomial is thus

$$p_n(x) = f(0) + f'(0)x + \frac{f''(0)}{2}x^2 + \cdots + \frac{f^{(n)}(0)}{n!}x^n = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!}x^k.$$

Now what happens if we want to perform this expansion around a more general point a instead of just 0? Instead of doing more work, we just translate our p_n above by the term $x - a$. More precisely, consider the polynomial

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

for which we have $p_n^{(k)}(a) = k!a_k$. We would like to set this to $f^{(k)}(a)$, meaning that we should take $a_k = \frac{f^{(k)}(a)}{k!}$, and our polynomial becomes

$$p_{n,a}(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k.$$

This is called the n^{th} *order Taylor polynomial at* a.

Example 10.1

Determine the n^{th} order Taylor polynomial of $f(x) = e^x$ at $x = 0$.

Solution. We are well familiar with the fact that $f^{(k)}(x) = e^x$ and so $f^{(k)}(0) = e^0 = 1$. Thus the Taylor polynomial is

$$p_{n,0}(x) = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!}x^k = \sum_{k=0}^{n} \frac{x^k}{k!}.$$ ■

Example 10.2

Determine the n^{th} order Taylor polynomial of $f(x) = \log(1+x)$ at $x = 0$.

Solution. This requires that we determine a general form for the derivative $f^{(k)}(x)$. Checking our first few derivatives, we find that

$$\begin{aligned}
f(x) &= \log(1+x) & f(0) &= 0 \\
f'(x) &= \frac{1}{1+x} & f'(0) &= 1 \\
f''(x) &= -\frac{1}{(1+x)^2} & f''(0) &= -1 \\
f^{(3)}(x) &= \frac{2!}{(1+x)^3} & f^{(3)}(0) &= 2! \\
&\vdots \\
f^{(k)}(x) &= \frac{(-1)^{k-1}(k-1)!}{(1+x)^k} & f^{(k)}(0) &= (-1)^{k-1}(k-1)!
\end{aligned}$$

The veracity of this most general form can be checked by induction, and is left as an exercise for the student. Hence the n^{th}-order Taylor polynomial is

$$\begin{aligned}
p_{n,0}(x) &= \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!} x^k = \sum_{k=0}^{n} \frac{(-1)^{k-1}(k-1)!}{k!} x^k \\
&= \sum_{k=0}^{n} \frac{(-1)^{k-1}}{k} x^k.
\end{aligned}$$

■

Sums and Products of Taylor Polynomials: The great thing about Taylor polynomials is that they behave exactly as we expect they should under addition and multiplication.

Proposition 10.3

Let f, g be functions with n^{th} order Taylor polynomials $p_{n,a}(x)$ and $q_{n,a}(x)$ respectively.

1. The Taylor polynomial of $f+g$ is $p_{n,a}(x) + q_{n,a}(x)$.
2. The Taylor polynomial of fg is $p_{n,a}(x)q_{n,a}(x)$ (and is of order $2n$).

Proof. 1. The sum is easiest to show, since we know that the derivative is linear. Indeed, $\frac{\mathrm{d}}{\mathrm{d}x^n}\Big|_{x=a}[f+g] = f^{(n)}(a) + g^{(n)}(a)$ and hence the n-th order Taylor polynomial is

$$\begin{aligned}
\sum_{k=0}^{n} \frac{(f+g)^{(k)}(a)}{k!}(x-a)^k &= \sum_{k=0}^{n} \left[\frac{f^{(k)}(a)}{k!}(x-a)^k + \frac{g^{(k)}}{k!}(x-a)^k \right] \\
&= \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + \sum_{k=0}^{n} \frac{g^{(k)}(a)}{k!}(x-a)^k \\
&= p_{n,a}(x) + q_{n,a}(x).
\end{aligned}$$

2. The product is somewhat more difficult to discern. One can check via induction that the n-th

derivative of fg is

$$(fg)^{(n)}(a) = \sum_{k=0}^{n} \binom{n}{k} f^{(k)}(a) g^{(n-k)}(a), \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Hence the Taylor polynomial of the product is given by

$$\sum_{k=0}^{2n} \frac{(fg)^{(k)}(a)}{k!}(x-a)^k = \sum_{k=0}^{2n}\sum_{\ell=0}^{k} \binom{n}{k} \frac{f^{(\ell)}(a) g^{(n-\ell)}(a)}{k!}(x-a)^k \tag{10.1}$$

On the other hand, we have

$$\begin{aligned}
p_{n,a}(x)q_{n,a}(x) &= \left[\sum_{r=0}^{n} \frac{f^{(r)}(a)}{r!}(x-a)^r\right]\left[\sum_{s=0}^{n} \frac{g^{(s)}(a)}{s!}(x-a)^s\right] \\
&= \sum_{r=0}^{n}\sum_{s=0}^{n} \frac{f^{(r)}(a) g^{(s)}(a)}{s!r!}(x-a)^{r+s} \\
&= \sum_{k=0}^{2n}\sum_{r=0}^{n} \frac{f^{(r)}(a) g^{(k-r)}(a)}{r!(k-r)!}(x-a)^{k} && k = r+s \\
&= \sum_{k=0}^{2n}\sum_{r=0}^{n} \frac{k!}{r!(k-r)!}\frac{f^{(r)}(a) g^{(k-r)}(a)}{k!}(x-a)^{k} \\
&= \sum_{k=0}^{2n}\sum_{r=0}^{k} \binom{n}{k} \frac{f^{(r)}(a) g^{(n-r)}(a)}{k!}(x-a)^{k}
\end{aligned}$$

Comparing this to (10.1) we see that the equation are equal, giving the desired result. ☐

Example 10.4

Compute the Taylor polynomial of $x\sin(x)$ about 0.

Solution. It will quickly become laborious for us to differentiate the function $f(x) = x\sin(x)$, and in fact it will be effectively impossible to derive a closed form expression for a generic n-th derivative. Instead, the Taylor polynomial of x is just x itself, so we need only compute the Taylor polynomial of $\sin(x)$.

We know that

$$\begin{aligned}
&\sin(0) = 0 && \left.\frac{d^2}{dx^2}\right|_{x=0} \sin(x) = -\sin(x)|_{x=0} = 0 \\
&\left.\frac{d}{dx}\sin(x)\right|_{x=0} = \cos(x)|_{x=0} = 1 && \left.\frac{d^3}{dx^3}\right|_{x=0} = -\cos(x)|_{x=0} = -1
\end{aligned}.$$

Since the derivatives of sine repeat with periodicity 4, this is all we need to compute. In particular though, we notice that only the odd terms will survive, and they will switch sign. By writing our odd numbers as $2n+1$, the sign of the $2n+1$-st derivative is $(-1)^n$ and we get the Taylor polynomial (or order $2n+1$)

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

The Taylor polynomial of $x\sin(x)$ is thus

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

Approximation Error: How well does this polynomial do at actually approximating $f(x)$? The following theorem gives us the ability to talk about the 'Taylor remainder.'

Theorem 10.5: Taylor's Theorem

If f is $n+1$ times differentiable in a neighbour U of a and $p_{n,a}(x)$ is the Taylor polynomial of f of order n, then there exists a function $r_{n,a}(x)$ called the *Taylor remainder of order* n such that for every $x \in U$ we have

$$f(x) - p_{n,a}(x) = r_{n,a}(x).$$

Moreover, $r_{n,a}(x) \xrightarrow{x\to\infty} 0$. More explicitly, one can write

1. $r_{n,a}(x) = \dfrac{1}{n!}\displaystyle\int_a^x f^{(n+1)}(s)(x-s)^n ds$ called the *integral from* of the remainder,
2. $r_{n,a}(x) = \dfrac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1}$ for some $c \in U$, called the *Lagrange form* of the remainder.

The Lagrange form of the remainder in particular allows us to determine upper bounds on the error of using Taylor series. Indeed, assume that $f^{(k+1)}$ is continuous. If we wish to approximate $f(x)$ on an interval $[a-c, a+c]$ for some $c > 0$, then by the Extreme Value Theorem we know that $|f^{(k+1)}|$ attains its max M on $[a-c, a+c]$, and hence

$$|r_k(x)| = \left|\frac{f^{(k+1)}(c)}{(k+1)!}\right| |x-a|^{k+1} \leq \frac{M}{(k+1)!}|x-a|^{k+1}$$

which can be computed explicitly.

Example 10.6

Let $\epsilon = 0.001$ and $f(x) = e^x$. Find the order of Taylor polynomial (about 0) necessary to ensure that one can approximate e^x to within ϵ on $[-1, 1]$.

Solution. Recall that the Taylor polynomial for e^x about 0 is given by

$$p_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!},$$

and the Taylor remainder on the interval I can be bounded by

$$|r_n(x)| \leq \left[\max_{x\in I}\left|f^{(n+1)}(x)\right|\right] \frac{|x|^{n+1}}{(n+1)!}.$$

The function $f(x) = e^x$ is its own derivative and is increasing, hence

$$\max_{x\in[-1,1]} |f^{(n+1)}(x)| = \max_{x\in[-1,1]} e^x = e,$$

and $r_n(x) \leq e\frac{x^{n+1}}{(n+1)!} \leq \frac{e}{(n+1)!}$. The first n for which this is true is $n = 6$, hence the 6-th order Taylor polynomial is required. ■

10.2 Power Series

The idea of a power series is an exceptionally fundamental thing within many fields of mathematics, such as statistics, combinatorics, algebra, analysis, etc. The idea is to look at infinite series of the form

$$\sum_{k=0}^{\infty} a_k x^k$$

where the $a_k \in \mathbb{R}$ are just real numbers. The role of the x^k is much more subtle. Sometimes the x^k just serve as containers, indicating that we want 'degree k' information to be stored in its coefficients. This is especially useful in for things such as *generating functions*, which are found in the description of *moments* (via the moment generating function) in statistics, and in the field of *combinatorial enumeration*. Other times they are designed to be purely abstract containers, such as in algebra.

Where we are going to be interested in power series was alluded to earlier: We know that the Taylor polynomial $p_n(x)$ corresponding to the function f serves as an approximation for f. Taylor's Theorem then says that as we let n get larger, the approximation becomes better and better. Does it make sense to take a limit as $n \to \infty$, in which case our infinite series would then just 'be' the function?

We have seen that infinite series can be present some tricky situations, in which we must be particularly careful with notions such as convergence. Let us start with some general discussion of power series, before moving on to Taylor series in particular.

Convergence of power series: Thinking of $\sum_{k=0}^{\infty} a_k x^k$ as a function in x, we need to ask whether it makes sense for the series to converge. [29] This will naturally depend on the value of x that we substitute. For example, the power series

$$f(x) = \sum_{k=0}^{\infty} x^k \tag{10.2}$$

represents the geometric series with common ratio x. We already know that the geometric series will converge if and only if our ratio x satisfies $|x| < 1$. This means that the function defined by (10.2) only converges if we input a number $x \in (-1, 1)$.

[29] If we do not care about convergence of the power series, as is the case with generating functions, we say we are dealing with *formal power series*.

Theorem 10.7

If $f(x) = \sum_{k=0}^{\infty} a_k x^k$ is a power series function, there exists a unique $r \geq 0$ (possibly $r = \infty$) such that $f(x)$ converges for all $|x| < r$ and diverges for all $|x| > r$. [a]

[a] Note that nothing has been said about what happens when $|x| = r$, since this actually differs dramatically depending on the function at hand.

This effectively gives us three possible scenarios:

1. $r = 0$ means that our power series only converges at a single point. This is not an enlightening case,
2. $0 < r < \infty$ means that our power series converges, but only within some specific range for the value of x,
3. $r = \infty$ means that our power series always converges, regardless of what value of x we use.

We can use the convergence tests that we learned about in Section 9.4.2 to determine radii of convergence, just like in the following example:

Example 10.8

Find the radius of convergence for the power series $\sum_{k=0}^{\infty} \frac{x^k}{k^2 2^k}$.

Solution. We would like to determine the values of x such that the series $\sum b_k$ defined by setting $b_k = \frac{|x|^k}{k^2 2^k}$ converges. Using the ratio test, we have

$$\begin{aligned}
\lim_{k\to\infty} \frac{b_{k+1}}{b_k} &= \lim_{k\to\infty} \frac{x^{k+1}}{(k+1)^2 2^{k+1}} \frac{2^k k^2}{x^k} \\
&= \frac{|x|}{2} \lim_{k\to\infty} \left(\frac{k}{k+1} \right)^2 \\
&= \frac{|x|}{2}.
\end{aligned}$$

We know that we get convergence when $\frac{|x|}{2} < 1$ and so $|x| < 2$. ■

The footnote in Theorem 10.7 emphasized the fact that we do not know how to stipulate endpoint conditions for convergence on an interval. More precisely, we know that there exists an $r \geq 0$ so that a power series will converge on $(-r, r)$, but what about that the points $-r$ and r themselves? A brute force solution is to plug $\pm r$ into the power series and just check convergence directly. The *Interval of Convergence* is the whole interval (including the endpoints) on which the power series converges.

Example 10.9

Determine the interval of convergence for the power series

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

Solution. Set $a_k = \dfrac{3^k}{k}|2x-4|^k$ and apply the ratio test to get

$$\begin{aligned}
\lim_{k\to\infty} \frac{a_{k+1}}{a_k} &= \lim_{k\to\infty} \frac{3^{k+1}|2x-4|^{k+1}}{k+1} \frac{k}{3^k|2x-4|^k} \\
&= 3|2x-4| \lim_{k\to\infty} \frac{k}{k+1} \\
&= 3|2x-4|.
\end{aligned}$$

Setting $3|2x-4| < 1$ we equivalently get $|x-2| < \frac{1}{6}$, so our radius of convergence if $\frac{1}{6}$, and our interval so far is $\left(\frac{11}{6}, \frac{13}{6}\right)$.

To determine endpoint convergence, we plug these number directly into our series. At $x = \frac{11}{6}$ we get

$$\sum_{k=0}^{\infty} \frac{3^k}{k}\left(2\frac{11}{6} - 4\right)^k = \sum_{k=0}^{\infty} \frac{(-1)^k 3^{k-1}}{k!}.$$

Since $\frac{3^{k-1}}{k!} \xrightarrow{x\to\infty} 0$, this series converges by the Alternating Series Test. On the other hand, at $x = \frac{13}{6}$ we have

$$\sum_{k=0}^{\infty} \frac{3^k}{k}\left(2\frac{13}{6} - 4\right)^k = \sum_{k=0}^{\infty} \frac{3^{k-1}}{k!}$$

which converges by the Ratio Test. Hence the interval of convergence included both endpoints and is $\left[\frac{11}{6}, \frac{13}{6}\right]$. ■

Taylor Series: By extending the idea of Taylor polynomials to power series, we get the definition of the *Taylor series* of the function f about the point $x = a$. In particular, consider the following result:

Theorem 10.10

Let f be an infinitely differentiable function and consider the power series defined by

$$g(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k.$$

If I is the interval of convergence of this power series, then on I the Taylor series of g is precisely its power series representation.

Note however that even if g converges, it need not agree with f. We reserve a special name for such functions. We say that the function f is *analytic* on the open interval I if for each $a \in I$ there exists a $\delta > 0$ such that f agrees with its Taylor series on $(a-\delta, a+\delta) \subseteq I$.

One must be careful to realize that infinitely differentiable is not the same thing as analytic [30]. Define the function

$$f(x) = \begin{cases} e^{-\frac{1}{x}} & x > 0 \\ 0 & x \leq 0 \end{cases}.$$

This function is infinitely differentiable. However, it is not analytic at $x = 0$ as its Taylor series is identically 0, and the function is not constantly zero in any neighbourhood of $x = 0$. With a bit of work, one can show that the Taylor series of f at 0 has radius of convergence 0.

1. $e^x = \displaystyle\sum_{k=0}^{\infty} \frac{x^k}{k!}$, interval of convergence $\mathbb{R}$.

2. $\sin(x) = \displaystyle\sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!}, \quad \cos(x) = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!}$, interval of convergence $\mathbb{R}$.

3. $\displaystyle\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k$, interval of convergence $(-1, 1)$.

4. $\log(1+x) = \displaystyle\sum_{k=1}^{\infty} \frac{(-1)^{k+1} x^k}{k}$, interval of convergence $(-1, 1]$.

5. $\arctan(x) = \displaystyle\sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{2k+1}$, interval of convergence $(-1, 1]$.

An easy way to remember property (2): we know $\sin(x)$ is odd as a function, and indeed its Taylor series only has powers in odd order. Similarly, $\cos(x)$ is even as a function and its Taylor series only has powers of even order. We are not yet in a position to compute (5), but we provide it here as an interesting and motivating example which we will use (and prove) later.

It is worth noting that I am cheating here. Theorem 10.10 only tells us that the function is equal to its Taylor series on the interior of the interval of convergence and not at the endpoints. It is another theorem to show that if convergence holds at the endpoints then the function still converges there, but let's sweep that under the rug for now.

[30] Unless you're doing calculus on $\mathbb{C}$, in which case these are the same thing.

Exercise: Notice how the Taylor series for e^x looks very similar to the Taylor series for $\sin(x)$ and $\cos(x)$. Indeed, if we ignore the $(-1)^k$ terms we see that $\sin(x)$ contains the odd terms of e^x while $\cos(x)$ contains the even terms. If we can account for these minus signs, maybe we can get a relationship between *a priori* unrelated exponential function and the trigonometric functions.

1. Recall that in the complex numbers, we can define a number i such that $i^2 = -1$. Using their Taylor series expansions, show that $e^{ix} = \cos(x) + i\sin(x)$.
2. We know that every point on the unit circle can be written as $(\cos(\theta), \sin(\theta))$ where θ is the angle between the point and the positive x-axis. If we think about $\cos(x) + i\sin(x)$ as $(\cos(x), \sin(x))$ convince yourself that e^{ix} sweeps out the unit circle as x traverses through $[0, 2\pi)$.

10.3 Differentiation and Integration of Taylor Series

A natural question to ask (in a calculus course at least) is whether one can differentiate and integrate the terms of the Taylor series, and if so

- how does the radius of convergence change?
- does the new power series represent the Taylor series of the derivative/integral?

The answer to both of these questions lies in the following theorem, which we are not going to prove:

Theorem 10.11

Let $f(x) = \sum_{k=0}^{\infty} a_k x^k$ be a power series with radius of convergence r.

1. If f is differentiable then $\sum_{k=1}^{\infty} k a_k x^{k-1}$ has radius of convergence r, and moreover
$$f'(x) = \sum_{k=1}^{\infty} k a_k x^{k-1}.$$
2. If f is integrable then $F(x) = \sum_{k=1}^{\infty} \frac{a_k}{k+1} x^{k+1}$ has radius of convergence r, and moreover
$$\int f(x)dx = \int F(x)dx + C.$$

Effectively, this theorem says that it is fine to integrate and differentiate the terms of a Taylor series. This is the same as saying that for a Taylor series, we can interchange the sum with the derivative/integral. Indeed,

$$f'(x) = \frac{\mathrm{d}}{\mathrm{d}x} f(x) = \frac{\mathrm{d}}{\mathrm{d}x} \sum_{k=0}^{\infty} a_k x^k = \sum_{k=0}^{\infty} a_k \left[\frac{\mathrm{d}}{\mathrm{d}x} x^k \right] = \sum_{k=1}^{\infty} k a_k x^{k-1},$$

and similar for integration. This is not the same thing as linearity, and we should not assume that this result follows immediately. We encourage the student to accept the theorem as stated, but not to underestimate the subtlety of the situation.

Note however that while the radius of convergence does not change, *the interval of convergence might be different.* For example, we know that

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k \text{ on } (-1,1), \qquad \log(1+x) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1} x^k}{k} \text{ on } (-1,1].$$

These functions are not immediately related, but they can be via some manipulation. Notice that

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

and the interval of convergence is still $(-1,1)$. Now differentiating $\log(1+x)$ term by term we get

$$\begin{aligned} \frac{1}{1+x} = \frac{\mathrm{d}}{\mathrm{d}x} \log(1+x) &= \sum_{k=1}^{\infty} \left[\frac{\mathrm{d}}{\mathrm{d}x} \frac{(-1)^{k+1} x^k}{k} \right] \\ &= \sum_{k=1}^{\infty} (-1)^{k+1} x^{k-1} \\ &= \sum_{k=0}^{\infty} (-1)^k x^k. \end{aligned}$$

We knew that this would be true from our theorem, but notice that the interval of convergence of $\log(1+x)$ is $(-1,1]$ while its derivative has interval of convergence $(-1,1)$. So the radius is unchanged, but the interval is not.

Example 10.12

Confirm that the Taylor series of $\arctan(x)$ is given by $\displaystyle\sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{2k+1}$.

Solution. We know that $\arctan(x) = \displaystyle\int_0^x \frac{1}{1+t^2} dx$, and we know the Taylor series for $\frac{1}{1+t^2}$ since it is given by

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

Integrating term by term, we get

$$\arctan(x) = \sum_{k=0}^{\infty} \int_0^x (-1)^k t^{2k} dt = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{2k+1}$$

as required. ■

10.4 Applications of Taylor Series

Calculators and calculating: One of the historic applications of power series is the ability a function to arbitrary precision in terms of polynomials. In the case of transcendental functions especially (such as $\log, \sin, \cos, \exp$), this is how modern day computers and calculators often compute values.

Computing Limits and Asymptotics: Mathematicians often cheat when looking at limits, in the sense that we can often guess that value of the limit if we know a little something about the Taylor series of the terms involved.

Before proceeding, let us introduce a little notation. Recall in Section 4.5.2 that we defined 'little-o' notation, and said that $f = o(g)$ if f and g grow at the same rate; that is, f and g shared the same asymptotics in the limit as f and g go to infinity. In a similar vein, we may define 'big-O' notation as describing the asymptotics of f and g when they go to zero [31]. In effect, we will write $f = O(g)$ if f is at worst g in the limit as $x \to 0$. For example, the Taylor series of e^x is

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

If we are playing with numbers very near zero, and are only interested in what is happening up to the quadratic term, we might write this as

$$e^x = 1 + x + \frac{x^2}{2} + O(x^3)$$

where here $O(x^3)$ indicates that the dominant factor in the error is of order x^3 (for x near 0, the term x^4 will be much smaller than x^3, and so on). When taking limits as $x \to 0$, anything with $O(x^k)$ $k \geq 1$ will vanish, meaning that we are only interested in the constant terms.

Now we often used Taylor series to create L'Hôpital problems, such as the following:

Example 10.13

Consider the limit $\lim_{x\to 0} \frac{e^x - 1 - x}{x^2}$.

Solution. When we made this example, we used the fact that we knew what the Taylor series for e^x. Indeed, the numerator becomes

$$e^x - 1 - x = \left(1 + x + \frac{x^2}{2} + O(x^3)\right) - 1 - x = \frac{x^2}{2!} + O(x^3).$$

[31] There is also a version of big-O for infinite asymptotics.

Hence applying this to the limit, we have

$$\begin{aligned}\lim_{x\to 0}\frac{e^x-1-x}{x^2} &= \lim_{x\to 0}\frac{\frac{x^2}{2!}+O(x^3)}{x^2}\\ &= \lim_{x\to 0}\left[\frac{1}{2!}+O(x)\right]\\ &= \frac{1}{2}.\end{aligned}$$

One can check this is the same answer we get by applying L'Hôpital. ■

Example 10.14

(Example 4.28) Determine the limit $\lim_{x\to 0}\frac{xe^{nx}-x}{1-\cos(nx)}$ for $n\neq 0$.

Solution. Using Taylor series, we have

$$\begin{aligned}xe^{nx}-x &= x\left[1+(nx)+\frac{(nx)^2}{2}+O(x^3)\right]-x\\ &= nx^2+\frac{n^2x^3}{2}+O(x^4).\end{aligned}$$

On the other hand, the denominator is

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

Taking the limit, we have

$$\begin{aligned}\lim_{x\to 0}\frac{xe^{nx}-x}{1-\cos(nx)} &= \lim_{x\to 0}\frac{nx^2+O(x^3)}{\frac{n^2x^2}{2}+O(x^4)}\\ &= \lim_{x\to 0}\frac{n+O(x)}{\frac{n^2}{2}+O(x^2)}\\ &= \frac{2}{n},\end{aligned}$$

and this is exactly what we found in Example 4.28. ■

Difficult Integrals: We have seen a few examples of functions that cannot be integrated, and with some more advanced theory (called differential Galois theory), one can actually prove that some of these integrals do not have anti-derivatives that can be expressed with elementary functions. However, the simplicity of integrating polynomials means that, so long as we are content with a power series representation, one can still get a closed form expression for these integrals.

Example 10.15

Using power series, integrate $\int \frac{\sin(x)}{x}dx$.

Solution. The power series for $\frac{\sin(x)}{x}$ is given by

$$\begin{aligned}\frac{\sin(x)}{x} &= \frac{1}{x}\sum_{k=0}^{\infty}\frac{(-1)^k x^{2k+1}}{(2k+1)!}\\ &= \sum_{k=0}^{\infty}\frac{(-1)^k x^{2k}}{(2k+1)!}.\end{aligned}$$

Since we can integrate term-by-term, we get

$$\begin{aligned}\int \frac{\sin(x)}{x}dx &= \sum_{k=0}^{\infty}\int\frac{(-1)^k x^{2k}}{(2k+1)!}dx\\ &= \sum_{k=0}^{\infty}\frac{(-1)^k x^{2k+1}}{(2k+1)^2(2k)!}.\end{aligned}$$ ■

Example 10.16

Using power series, integrate $\int_0^x e^{-x^2}dx$.

Solution. The power series for e^{-x^2} is given by

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

Integrating term-by-term, we get

$$\begin{aligned}\int_0^x e^{-t^2}dt &= \sum_{k=0}^{\infty}\frac{(-1)^k}{k!}\int_0^x t^{2k}dt\\ &= \sum_{k=0}^{\infty}\frac{(-1)^k x^{2k+1}}{(2k+1)k!}.\end{aligned}$$ ■

Differential Equations: In a similar vein to integrals, one can solve differential equations using Taylor series. One begins by assuming that the solution can be written as a power series, then uses the differential equation to determine a relationship amongst the coefficients.

Example 10.17

Use a power series to find a solution to $y'' + y = 0$.

Solution. We have already seen that the most general solution to this equation is $y(x) = A\cos(x) + B\sin(x)$ for variables A and B. We expect that we should arrive at the same solution using power series. Indeed, suppose that we can write $y = \sum_k a_k x^k$, so that

$$y = \sum_{k=0}^{\infty} a_k x^k \qquad y' = \sum_{k=1}^{\infty} k a_k x^{k-1}$$
$$y'' = \sum_{k=2}^{\infty} k(k-1) a_k x^{k-2}$$

Plugging these into our differential equation, we get

$$\begin{aligned} 0 = y'' + y &= \left[\sum_{k=2}^{\infty} k(k-1)a_k x^{k-2}\right] + \left[\sum_{k=0}^{\infty} a_k x^k\right] \\ &= \left[\sum_{k=0}^{\infty}(k+2)(k+1)a_{k+2}x^k\right] + \left[\sum_{k=0}^{\infty} a_k x^k\right] && \text{re-labelling the first sum to start at 0} \\ &= \sum_{k=0}^{\infty}\left[(k+2)(k+1)a_{k+2} + a_k\right] x^k \end{aligned}$$

Now the power series is zero if and only if each coefficient is identically zero. This means that we can solve for a_{k+2} in terms of a_k as

$$a_{k+2} = -\frac{a_k}{(k+2)(k+1)}.$$

Hence once we have determine a_0 and a_1, we can find all the remaining coefficients. Set $a_0 = A$ and $a_1 = B$ and notice that

$$\begin{array}{ll} a_2 = -\frac{A}{2\cdot 1} & a_3 = -\frac{B}{3\cdot 2} \\ a_4 = \frac{-a_2}{4\cdot 3} = \frac{A}{4!} & a_5 = -\frac{a_3}{5\cdot 4} = \frac{B}{5!} \\ \vdots & \vdots \\ a_{2n} = (-1)^n \frac{A}{(2n)!} & a_{2n+1} = (-1)^n \frac{B}{(2n+1)!} \end{array}$$

By splitting the even and odd powers of the power series, our solution thus looks like

$$y = A\sum_{k=0}^{\infty}\frac{(-1)^k x^{2k}}{(2k)!} + B\sum_{k=0}^{\infty}\frac{(-1)^k x^{2k+1}}{(2k+1)!} = A\cos(x) + B\sin(x),$$

exactly as we expected. ■