Real Analysis (MAST20026) University of Melbourne
ver. November 2021

Contents

Weekly Coverage

Week	**Topics**	**Coverage**
1	Proof and Formal Logic	1.1-1.2
2	Proof Techniques	1.3
3	Set Theory	2.1 - 2.3
4	The Real Numbers	3.1 - 3.2
5	Sequences and Convergence Part I	4.1 - 4.2.1
6	Sequences and Convergence Part II	4.2.2-5.1
7	Limits and Continuity Part I	5.1
8	Limits and Continuity Part II	5.2
9	Differentiation and Integration Part I	6.1-6.2
10	Differentiation and Integration Part II	6.2-6.3
11	Series Part I	7.1-7.2
12	Series Part II	7.2

A Note on the Text

These notes were created during the Winter 2021 semester at the University of Melbourne. They were written out of necessity; the ongoing effects of the COVID-19 pandemic lead to a full cessation of on-campus activities. Each week, students were assigned to read roughly twenty pages of notes. These readings were paired with short videos (5-7 minutes) designed to introduce the broad approach of the subject material at hand. Additionally, there were weekly drop-in office hours with the instructor to answer questions about subject material, as well as twice weekly tutorial sessions. For these reasons, there are few deep numerical examples in these notes. Such examples were presented in other formats during the semester.

At the University of Melbourne, *Introduction to Real Analysis* plays the dual role of introducing real analysis and also introducing students to proof and mathematical convention. Students traditionally taking this subject have previously seen material in both calculus and linear algebra, but have seen little in the way of formal mathematics. As such, the implied reader of these notes has no experience with mathematical proof, writing or formalism.

Each section of the notes ends with a part titled *Test Your Understanding.* In general, these questions are not designed to challenge a reader. Instead they are meant to be a check of broad understanding of the key ideas in the section. Challenging the learner to engage more fully with the material is a task left to assessed work, tutorial sessions and problem sheets.

Throughout the text the reader will find short diversions under the heading **Aside**. These fragments of text often present ideas beyond the learning outcomes and sometimes require students to have a more mature mathematical background. These parts of the text can be fully ignored without detriment.

The approach and ordering of the topics in this text is adapted from existing lecture slides for the subject. The presentation of some materials herein is influenced by *Transition to Higher Mathematics: Structure and Proof, 2Ed.* by Dumas and McCarthy and *Real Analysis: A Long-Form Mathematics Textbook* by Cummings. Notably, the introductory text given in Section 1 is adapted from text in the former reference.

Errors are the sole responsibility of the original author. In some places, however, this text tells stretches the truth in pursuit of broader learning objectives.

-cd

1 Proof and Mathematical Logic

A primary activity in mathematics is proving mathematical claims. This may seem surprising for you, given that most of the mathematics you have studied has been the application of techniques to deriving solutions of relatively concrete problems. In actuality, the *mathematics* is more in proving that the techniques work as opposed to applying them to specific problems. And so it makes sense to begin our discussion in this course by thinking about mathematical proofs.

What is a mathematical proof? The nature of a mathematical proof depends on the context. There is a formal notion of a mathematical proof:
A finite sequence of formal mathematical statements such that each statement either

- *is an axiom or assumption, or*
- *follows by formal rules of logical deduction from previous statements in the sequence.*

We will return to formal mathematical proofs in Section 1.3 of the course. However, most mathematicians do not think of mathematical proofs as formal mathematical proofs, and few mathematicians write formal mathematical proofs. This is because, as we will see, a formal proof is often a tremendously cumbersome thing. Rather, mathematicians write proofs that are sequences of statements in a combination of natural language and formal mathematical symbols (interspersed with diagrams, questions, references and other devices that are intended to assist the reader in understanding the proof) that can be thought of as representing a purely formal argument.

A good practical definition of a mathematical proof is: *an argument in favour of a mathematical statement that will convince the preponderance of knowledgeable peers of the truth of the mathematical statement*

This definition is imprecise and mathematicians can disagree on whether an argument is a proof, particularly for extremely difficult or deep arguments. However, for virtually all mathematical arguments, after some time for careful consideration, the mathematical community reaches a unanimous consensus on whether it is a proof.

Inherent within our definition of is an idea of shared vocabulary and notation. When we communicate mathematical ideas to one another it is important that we agree on what the words and the symbols all mean. Throughout the semester we will define a number of terms and symbols. Usually, new terms and symbols are shorthand for more complex mathematical ideas. When we use these words and these symbols we are referring back to the agreed upon definition from the notes/lecture. The symbols themselves are not the mathematics; they represent mathematical ideas.

Our goal in this first part of the course is to develop some ideas and some strategies that will help us in constructing proofs. As our practical definition of a mathematical proof is not wholly unrelated to formal mathematical proofs, we will begin our study with the study of formal mathematical proofs. This study will help give us the framework we need when we are writing our own proofs.

Throughout this course you will be learning to use the conventions of mathematical grammar and argument. Like most conventions, these are often determined by tradition or precedent. It can be quite difficult, initially, to determine whether your mathematical exposition meets the standards for this course. This is why graders provide feedback on assignments and opportunities for revision are available.

Why do mathematicians care about proofs? Mathematicians depend on proofs for certainty and explanation. Once a proof is accepted by the mathematical community, it is a rare occasion that the result is subsequently refuted. This was not always the case: in the 19th century there were serious disputes as to whether results had really been proved or not. This led to our modern notion of a "rigorous" mathematical argument.

For very complicated results, writing a detailed proof helps us convince ourselves of the truth of the claim. After a we have hit upon the key idea behind an argument, there is a lot of hard work left developing the details of the argument. Many promising ideas fail as the author attempts to write a detailed argument based on the idea. Finally, proofs often provide a deeper insight into the result and the mathematical objects that are the subject of the proof.

Mathematical proofs are strongly related to formal proofs in a purely logical sense; the existence of an informal mathematical proof is overwhelming evidence for the existence of a formal mathematical proof. If it is not clear that the informal proof could conceivably be interpreted into a formal argument, it is doubtful that the informal argument will be accepted by the mathematical community. Consequently, mathematical arguments have a transparent underlying logical structure. For this reason we shall begin our discussion of mathematical proofs with an introduction to formal logic.

Aside. *The text above is adapted from Chapter 3 of* Transition to Higher Mathematics: Structure and Proof *(2nd Ed.) by Dumas and McCarthy.*

A Note on Reading Mathematics: Reading mathematics is difficult. One cannot read mathematics the way one does a novel. In general sentences in mathematics texts are carefully constructed to provide precise meaning. Though text below attempts to bridge the gap between terse mathematics textbook and readable prose, one should read slowly and carefully and be sure they understand the meanings of new words and notation before continuing on. The mathematics in this course (and in general) is a scaffold – it difficult to truly understand the top layers until one grapples with the lower levels.

Try not to be discouraged when things don't make sense, it is possible they are poorly written or there is an error. Good luck!

1.1 Propositional Logic

We begin with a definition of the underlying object of study in propositional logic: statements.

Definition 1.1 (Statement). *A statement is a sentence or expression that is either true or false.*

Aside. *Statements are sometimes called propositions; hence propositional logic.*

We use T and F to refer, respectively, to *true* and *false*. A statement takes on the role of a logical variable (the variables are *true* and *false*). We generally use lower case letters $p, q, r, \dots$ to represent statements. For example, we may say:
Let p be the statement

$$1 + 1 = 3.$$

The statement p is false.

Statements are sentences that, unambiguously, can be categorised as being either true or false. The statement above, $1 + 1 = 3$, is a sentence in disguise:

One plus one equals three.

This is sentence is unambiguously false, and so we are certain that it is a statement. On the other hand, consider the sentence:

$f(x)$ is continuous.

This is not a statement as it is neither true nor false without some further information (i.e., knowing which function $f(x)$ we are considering). Take a moment to determine which of the following are statements. Don't forget that false statements are still statements.

- 6 is the largest integer.
- $det(A) = 0$.
- If it is Saturday, then there is no MAST20026 lecture.
- $x > 2$
- For every $x \in \mathbb{Z}$, $x^2 \geq 0$.
- Every even number greater than 2 is the sum of two primes.

- $\begin{bmatrix} 1 & 1 \\ 3 & 4 \end{bmatrix}$ is invertible and $1 + 1 = 3$

The last example on this list is slightly interesting; it is made up from two statements:

1. $\begin{bmatrix} 1 & 1 \\ 3 & 4 \end{bmatrix}$ is invertible
2. $1 + 1 = 3$

We can combine statements to form new statements. Such statements are called compound statements (or logical formulae). They are constructed from simpler statements, connectives (such as *and*, *or*, etc...) and parentheses (to remove ambiguity). Some of our connectives we will study correspond to how we combine statements in natural English writing and speech:

- not (negation)
- and (conjuction)
- or (disjunction)

So that we all agree what we mean when we use these connectives[1], let us take a moment to define their meanings and assign to them symbols we can use to build sentences in the language of formal logic.

Definition 1.2. [2] *Let p be a statement. The negation of p is the statement "not p" and is denoted by $\sim p$. The truth value of $\sim p$ is the opposite of p (ie., if p is true then $\sim p$ is false).*

p	$\sim p$
T	F
F	T

- Let p be the statement "$5 > 0$". The statement $\sim p$ is the statement "5 is not > 0" (i.e "$5 \leq 0$".)
- Let r be the statement "these notes are boring". The statement $\sim r$ is the statement "these notes are not boring"

In defining negation we used a table to outline all of the possibilities for the truth value of $\sim p$ given a particular truth value of p. The column on the left listed out the possible values for p (T or F) and the column on the right listed out the corresponding values for $\sim p$. These truth tables are remarkably useful in defining our other connectives.

Definition 1.3. *Let p and q be statements The conjunction of p and q is the statement "p and q" and is denoted by $p \wedge q$. The truth value of $p \wedge q$ depends on the values of p and q and is given by the following truth table*

[1] The word *connective* is a bit of a misnomer. When we negate a statement we are not connecting two statements together

[2] A note about definitions: definitions are not universal. The terms/notation we will define are done so in the context this course. When you consult outside resources you may see differences in notation and terminology.

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

Let p be the statement "$\begin{bmatrix} 1 & 1 \\ 3 & 4 \end{bmatrix}$ is invertible". Let q be the statement $1 + 1 = 3$. The statement $p \wedge q$ is the statement "$\begin{bmatrix} 1 & 1 \\ 3 & 4 \end{bmatrix}$ is invertible *and* $1 + 1 = 3$" Since p is true (T) and q is false (F), looking at the second row of the truth table tells is that the statement $p \wedge q$ is false. The truth table for conjunction is consistent with how we use the word "and" as part of English speaking and writing. Similarly, we define a connective that, mostly, corresponds to how we use the word "or" as part of English speaking and writing.

Definition 1.4. *Let p and q be statements. The disjunction of p and q is the statement "p or q" and is denoted by $p \vee q$. The truth value of $p \vee q$ depends on the values of p and q and is given by the following truth table*

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

The first row of the truth table for disjunction is particularly interesting. When p and q are both true, their disjunction, $p \vee q$, is also true. In mathematics disjunction is *inclusive.* The statement $p \vee q$ is true when at least one of p and q is true. This is slightly different from how we use the world "or" when speaking and writing in standard English. In our day-to-day use of "or" we permit only one part to be true: *the weather will be warm or it will be cold.* In mathematics, this is not the case. The statement

$$\text{“}\begin{bmatrix} 1 & 1 \\ 3 & 4 \end{bmatrix} \text{ is invertible or } 1 + 1 = 2\text{”}$$

is true.

Using our connectives we can consider much more complicated logical formulae. We can truth tables to determine the value of a compound statement for particular values of the constituent statements.

Example 1.5. *For which values of p and q are the following compound statements true:*

1. $(\sim p \wedge q) \wedge (p \vee q)$
2. $(p \vee \sim p) \wedge \ (q \vee \sim q)$

Solution*: We proceed by constructing and filling a truth table for these compound statements. On the far left we have columns for p and q. We need to consider all possible combinations of p and q being true and false; and so we will have four rows.*

p	q	
T	T	
T	F	
F	T	
F	F	

To determine the truth value of $(\sim p \wedge q) \wedge (p \vee q)$ *for particular truth values of* p *and* q *we first need to know the truth values for* $\sim p \wedge q$ *and* $p \vee q$*. We compute these using the definitions above for* $\sim, \wedge$ *and* $\vee$

p	q	$\sim p$	$\sim p \wedge q$	$p \vee q$
T	T	F	F	T
T	F	F	F	T
F	T	T	T	T
F	F	T	F	F

We now have sufficient information to fill a column for $(\sim p \wedge q) \wedge (p \vee q)$*.*

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

Therefore $(\sim p \wedge q) \wedge (p \vee q)$ *is true only when* p *is false and* q *is true.*

Exercise 1.6. *Complete part (b) of the example above.*

Our result in part (b) is seems to be a particularly special occurrence! (*... you did do the exercise, right?*). This outcome, as well as the opposite outcome, is interesting enough to be given its own special name.

Definition 1.7. *Let* A *be a compound statement. We say* <u>A *is a tautology*</u> *when* A *is always true regardless of the truth values of the simpler statements used to build* A*.*

Definition 1.8. *Let* A *be a compound statement. We say* <u>A *is a contradiction*</u> *when* A *is always false regardless of the truth values of the simpler statements used to build* A*.*

Taking a moment to look at our work so far, perhaps we have enough expressive power with our three connectives to be able to encode any sort of statement. However, two quick examples perhaps convince us otherwise:

- If it is Saturday, then there is no MAST20026 lecture.
- For every $x \in \mathbb{Z}$, $x^2 \leq 0$.

Both of these are indeed statements; the first is true and the second is false. They seem to be slightly more complicated than just a simple statement, but they do not seem to use any of the connectives we have learned about so far. The first of these examples is made up from two statements:

$$\begin{aligned} p &: \text{It is Saturday.} \\ q &: \text{There is no MAST20026 lecture.} \end{aligned}$$

but none of our connectives, negation, conjunction or disjunction, seem to let us model this statement using propositional logic. For this we need[3] a further connective.

Definition 1.9. *Let p and q be statements. The implication "p implies q" is denoted by $p \Rightarrow q$. The truth value of $p \Rightarrow q$ depends on the truth values of p and q and is given by the following truth table*

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

In the implication $p \Rightarrow q$ we refer to p as the hypothesis and q as the conclusion.

Notice, if p is true and $p \Rightarrow q$ is true, then q must be true. Consider the following example.

Let p be the statement:

"The temperature outside is colder than $0°C$."

And let q be the statement

"Chris wears a beanie when he is outside"

The compound statement $p \Rightarrow q$ is true. And so when the hypothesis is true (i.e., temperature $\leq 0°C$) necessarily the conclusion (Chris will wear a beanie when he is outside) is true.

Consider now the implication $q \Rightarrow p$. Just because Chris wears a beanie when the temperature is $\leq 0°C$, does not tell us anything about the temperature when Chris is seen wearing a beanie. It might be that Chris wears his beanie as an ironic fashion statement in the summer, or that Chris always wears a beanie regardless of the temperature.

Definition 1.10 (converse)**.** *The converse the implication $p \Rightarrow q$ is the implication $q \Rightarrow p$.*

In general, knowing $p \Rightarrow q$ is true tells us nothing about the truth value of the converse.

To finish up our work on connectives, we return to a fact from our study of Linear Algebra. Let $A = \begin{bmatrix} 1 & 1 \\ 3 & 4 \end{bmatrix}$. And consider the statement

$$A \text{ is invertible if and only if } det(A) \neq 0$$

[3] this isn't strictly true, but it is much easier to introduce a connective for implication than to try to model implication using only negation, conjunction and disjunction

This statement seems to be made of two constituent statements:

$$\begin{aligned} &p : A \text{ is inverible} \\ &q : det(A) \neq 0 \end{aligned}$$

From our work in Linear Algebra we recognize this "if and only if" statement to be a true statement. The following connective helps us model "if and only if" statements in propositional logic.

Definition 1.11. *Let p and q be statements. The biconditional "p if and only if q" is denoted by $p \Leftrightarrow q$. The truth value of $p \Leftrightarrow q$ depends on the values of p and q and is given by the following truth table*

p	q	$p \Leftrightarrow q$
T	T	T
T	F	F
F	T	F
F	F	T

Notice $p \Leftrightarrow q$ is true exactly when p and q have the same truth value. From our study of linear algebra, we recall that the condition of a square matrix being invertible is equivalent to the condition that the matrix has a non-zero determinant: either the matrix is invertible and it has non-zero determinant, or the matrix is not invertible and its determinant is 0.

Just as with our previous connectives, we can use truth tables to study more compound statements formed using implications and biconditionals. Take a moment to fill in the following truth table:

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

There is something curious to notice here. The column for $(\sim q) \Rightarrow (\sim p)$ is identical to the column for $p \Rightarrow q$. These two compound statements always have the same truth value. Let us go back to a motivating example from above to make sense of this.

Let

$$\begin{aligned} &p : \text{It is Saturday} \\ &q : \text{There is no MAST20026 lecture.} \end{aligned}$$

Negating these statements we have:

$$\begin{aligned} &\sim p : \text{It is not Saturday} \\ &\sim q : \text{There is a MAST20026 lecture.} \end{aligned}$$

Consider now the statement $(\sim q) \Rightarrow (\sim p)$:

If there is a MAST20026 lecture, then it is not Saturday.

If $p \Rightarrow q$ is true, then necessarily so must $(\sim q) \Rightarrow (\sim p)$ be true. Similarly, if $(\sim q) \Rightarrow (\sim p)$ is true then so must $p \Rightarrow q$ be true.

Definition 1.12 (contrapositive)**.** *Let p and q be statements. The contrapositive of the implication $p \Rightarrow q$ is the implication $(\sim q) \Rightarrow (\sim p)$.*

As we saw above, an implication and its contrapositive always have the same truth value.

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

A pair of compound statements have the same column in a truth table reveals some sense of equivalence between the statements. You may have noticed that so far we have not used the equals sign (=) when referring to this sense of equivalence compound statements. As we proceed through the course, we will continue as we have in this section to have agreed upon definitions for terms and notation. And so if we want to write $p \Rightarrow q = (\sim q) \Rightarrow (\sim p)$ we must first agree what it means for a pair of compound statements to be equal.

Mathematically there is a choice to be made here, but, unfortunately our preparation for this subject has not prepared us to contextualise why we will never write $p \Rightarrow q = (\sim q) \Rightarrow (\sim p)$. Rather than using the verb "equals" to relate two compound statements with the same truth table, we instead will use the word "equivalent".

We frame our definition of equivalence of compound statements using biconditionals and tautologies. Notice that if a pair of compound statements have identical columns of a truth table, then the the biconditional of these two compound statements will be a tautology. Similarly, if the biconditional of two compound statements is a tautology, then the two compound statements have identical columns in the truth table.

Take a moment to convince yourself: construct the truth table for

$$(p \Rightarrow q) \Leftrightarrow [(\sim q) \Rightarrow (\sim p)]$$

Definition 1.13. *Let r and s be statements. We say r and s are logically equivalent when $r \Leftrightarrow s$ is a tautology. When r and s are logically equivalent we write $r \equiv s$.*

At first this is an awkward definition to think about, however it does offer an advantage. Working with this definition does not require us to invoke the concept of a truth table. By defining logical equivalence using tautologies, we need only invoke concepts we have carefully defined as part of our study of propositional logic. As long as we all agree on the meaning of the symbol $r \Leftrightarrow s$ and the word tautology, then we agree on the meaning of $r \equiv s$.

Exercise 1.14. *Show $\sim (p \vee q) \equiv (\sim p) \wedge (\sim q)$.*
Strategy: To show $\sim (p \vee q)$ and $(\sim p) \wedge (\sim q)$ are logically equivalent we must convince our

reader that these two compound statements satisfy the definition of logically equivalent. Re-reading the definition of logically equivalent we do this by demonstrating that $\sim (p \vee q) \Leftrightarrow [(\sim p) \wedge (\sim q)]$ *is a tautology. Looking back at our definition of tautology, we do this by ensuring that this statement is always true, regardless of the truth values of* p *and* q*.*

Solution: We proceed by constructing a truth table and verifying that the statement $\sim (p \vee q) \Leftrightarrow (\sim p) \wedge (\sim q)$ is a tautology.

p	q	$p \vee q$	$\sim (p \vee q)$	$\sim p$	$\sim q$	$(\sim p) \wedge (\sim q)$	$\sim (p \vee q) \Leftrightarrow [(\sim p) \wedge (\sim q)]$

Test Your Understanding

Translate the following into the language of formal logic.

1. (a) Six is not prime or eleven is not prime

   (b) The square of 10 is 50 and the cube of 5 is 12.

   (c) If 7 is an integer then 6 is not an integer.

   (d) If both 2 and 5 are prime then 2×5 is not prime.

   Construct truth tables for the following statements.

2. (a) $(p \wedge q) \vee (\sim p \wedge \sim q)$

   (b) $[\sim q \wedge (p \Rightarrow q)] \Rightarrow \sim p$

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

Test Your Understanding Answers

1. Let $\mathcal{P}$ = the set of prime numbers.

   (a) $(\sim (6 \in \mathcal{P})) \vee (\sim (11 \in \mathcal{P}))$

   (b) $(10^2 = 50) \wedge (5^3 = 12)$

   (c) $(7 \in \mathbb{Z}) \Rightarrow (\sim (6 \in \mathbb{Z}))$

   (d) $((2 \in \mathcal{P}) \wedge (5 \in \mathcal{P})) \Rightarrow (\sim (2 \times 5 \in \mathcal{P}))$

   Construct truth tables for the following statements.

   (a) The truth table for $(p \wedge q) \vee (\sim p \wedge \sim q)$ is:

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

   (b) The truth table for $[\sim q \wedge (p \Rightarrow q)] \Rightarrow \sim p$ is:

p	q	$[\sim q \wedge (p \Rightarrow q)] \Rightarrow \sim p$
T	T	T
T	F	T
F	T	T
F	F	T

   (c) The truth table for $[(p \vee q) \wedge r] \Rightarrow (p \wedge r)$ is:

p	q	r	$[(p \vee q) \wedge r] \Rightarrow (p \wedge r)$
T	T	T	T
T	T	F	T
T	F	T	T
T	F	F	T
F	T	T	F
F	T	F	T
F	F	T	T
F	F	F	T

1.2 First-Order Logic

In previous section we spent our time developing terminology and notation to enable us to model mathematical statements using propositional logic. Using statements and connectives we can build increasingly complex formal mathematical statements. However, the tools of propositional logic do present some shortcomings. Consider the following examples:

- $f(x)$ is continuous and differentiable.
- $x^2 \geq 9$.
- The set A has ten elements.

These sentences look a lot like statements, expect they are not: these sentences are not unambiguously true or false. The truth value of these statements depends on some parameter. For example, when $x = 4$ the second sentence is true, but when $x = 2$ is not. These parameters can be variables (as we used to seeing them in Calculus), or other types of mathematical objects: sets, functions, etc.. To simplify matters, we use term variable to refer to unknown mathematical objects of any kind (not just numbers.)

Mathematical sentences whose truth depend on variables (mathematical objects) are called conditions. Our three examples from above are conditions.

- Let $p(f)$ be the condition "$f(x)$ is continuous and differentiable".
- Let $q(x)$ be the condition "$x^2 \geq 9$".
- Let $r(A)$ be the condition "The set A has ten elements".

When we consider a particular value for the variable in the condition the result is a statement. For example, let $p(f)$ be the condition:

$$f(x) \text{ is continuous on } \mathbb{R}$$

The statement denoted by $p(2x+1)$ is the statement

$$2x+1 \text{ is continuous on } \mathbb{R}$$

From our experience in studying calculus, we know the statement $p(2x+1)$ to be true. On the other hand, the statement $p\left(\frac{1}{x+2}\right)$ is false.

In all likelihood, you have seen conditions before, though perhaps not explicitly. Conditions are hiding in the background when we use *set-builder notation*. For example, consider the set

$$\{x \mid x^2 \geq 9\}$$

Using our notation from above, we can denote this set as

$$\{x \mid q(x)\}$$

In the example above, I expect that many of us implicitly thought about x as a real number. However, when instead we think of x as an integer, we arrive at a different set of values for which $q(x)$ is true. That is

$$\{x \in \mathbb{R} \mid x^2 \geq 9\} \neq \{x \in \mathbb{Q} \mid x^2 \geq 9\}$$

And so when we consider conditions we must also keep in mind the domain of the variable. Sometimes our conditions are explicit and other times we know them from context alone. Consider the following examples. Do you have enough information to know the domain of the variable for each condition?

1. x is even.
2. $x^2 \geq 9$.
3. $n \geq 0$.
4. A is invertible.

As we saw above, when we consider a condition for a particular value in the domain, the result is a statement. And so we can use our tools from Propositional Logic to create compound conditions.

$$\sim p(x)$$

$$p(x) \vee q(x)$$

$$p(x) \wedge q(x)$$

$$p(x) \Rightarrow q(x)$$

$$p(x) \Leftrightarrow q(x)$$

All of these conditions mean what you would expect. For example, let $p(A)$ be the following condition over the set of square matrices:

$$A \text{ is invertible.}$$

And let $q(A)$ be the following condition over the set of square matrices:

$$det(A) \neq 0$$

The condition

$$A \text{ is invertible if and only if } det(A) \neq 0$$

is denoted as $p(A) \Leftrightarrow q(A)$.

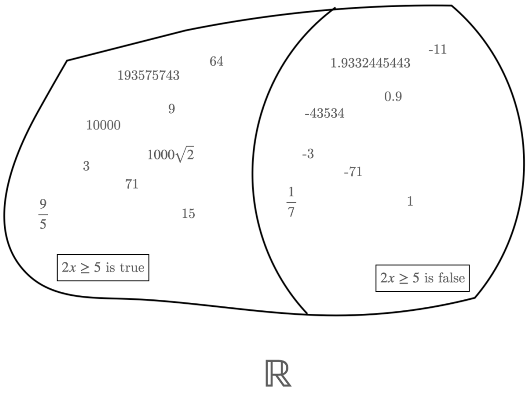
Over a domain, is possible that some variables lead to true statements, but others lead to false statements. The condition

$$A \text{ is invertible if and only if } det(A) \neq 0$$

is true for every matrix in the domain.

However, for some conditions there may be some elements of the domain that makes the condition false. For example, consider the condition $r : 2x \geq 5$ over the real numbers. For some choices of $x \in \mathbb{R}$ the statement $2x \geq 5$ is true (e.g. $x = 10$), whereas for other choices the statement is false (e.g. $x = -5$).

Truth values partition[4] the domain into those for which corresponding statement is true and those for which the corresponding statement is false.



There are four (overlapping) possibilities for this partition.

Let q be a condition.

1. $q(x)$ is true for every x in the domain.
2. $q(x)$ is true for at least one x in the domain.
3. $q(x)$ is false for every x in the domain.
4. $q(x)$ is false for at least one x in the domain.

To better understand these possibilities, we proceed with an example. Let $\mathcal{P}_1$ denote the set of polynomials of degree 1

$$\mathcal{P}_1 = \{f \mid f(x) = ax + b, a \neq 0\}$$

For $f \in \mathcal{P}_1$, let $p(f)$ be the condition

$$f(x) \text{ crosses the } x\text{-axis}$$

From our experience in studying calculus, we know the condition $p(f)$ satisfies possibilities (1) and (2) above, but does not satisfy possibilities (3) and (4).

On the other hand, for $f \in \mathcal{P}_1$, the condition

$$q : f'(x) > 0$$

is true only for linear functions with positive slope. In this case, this condition satisfies possibilities (2) and (4).

Consider the following statement:

$$\text{for every } f \in \mathcal{P}_1, \text{ the statement } p(f) \text{ is true.}$$

[4] When we carefully define partition later in the course, we will see that this statement is slightly incorrect.

As discussed above, this statement is true. On the other hand, the statement:

$$\text{for every } f \in \mathcal{P}_1\text{, the statement } q(f) \text{ is true.}$$

is false.

Looking back at the tools we developed in our work on Propositional Logic in 1.1, it does not seem as if our connectives can be used to model these statements. These two statements do not use negations, conjuction, disjunctions, implications or bi-conditionals. To express these statements similar to (1)-(4) above, we need some further tools.

Definition 1.15 (existential quantifier, universal quantifier)**.** *Let $p(x)$ be a condition over a domain D.*

- *The statement "$p(x)$ is true for every x is the domain" is denoted as $\underline{(\forall x \in D)\ p(x)}$. We refer to the symbol $\forall$ as the universal quantifier and we say $(\forall x \in D)\ p(x)$ is a universally quantified statement.*
- *The statement "$p(x)$ is true for at least one x is the domain" is denoted as $\underline{(\exists x \in D)\ p(x)}$. We refer to the symbol $\exists$ as the existential quantifier and we say $(\exists x \in D)\ p(x)$ is a existentially quantified statement.*

For example, consider $x \in \mathbb{R}$ and let $p(x) : x^2 \geq 3$. Certainly there exists at least one value of x in the domain such that $p(x)$ is true (e.g. $x = 2$). Therefore the statement $(\exists x \in \mathbb{R})\ p(x)$ is true. On the hand, there are values in the domain so that $p(x)$ is false (e.g $x = 0$). Therefore the statement $(\forall x \in \mathbb{R})\ p(x)$ is false.

As universally quantified statements are statements, their use with connectives has meaning. Let $q(x)$ be a condition over a domain D. Consider the following statement and its truth table

$$\sim(\exists x \in D\ p(x))$$

$(\exists x \in D)\ p(x)$	$\sim[(\exists x \in D)\ p(x)]$
T	F
F	T

When the statement $(\exists x \in D)\ p(x)$ is false it means that for every $x \in D$ it must be that $p(x)$ is false. In other words, for every $x \in D$ it must be that $\sim p(x)$ is true. In other words, the statement $(\forall x \in D)\ \ \sim p(x)$ is true.

Similarly, when $(\exists x \in D)\ p(x)$ is true it must mean that $(\forall x \in D)\ \ \sim p(x)$ is false. Adding $(\forall x \in D)\ \ \sim p(x)$ to the truth table we have:

$(\exists x \in D)\ p(x)$	$\sim[(\exists x \in D)\ p(x)]$	$(\forall x \in D)\ \ \sim p(x)$
T	F	F
F	T	T

Thus $\sim[(\exists x \in D)\ p(x)] \Leftrightarrow (\forall x \in D)\ \ \sim p(x))$ is a tautology and so $\sim[(\exists x \in D)\ p(x)] \equiv (\forall x \in D)\ \ \sim p(x))$.

Using similar reasoning we can show $\sim[(\forall x \in D)\ p(x)] \equiv (\exists x \in D)\ \ \sim p(x)$.

Returning to our line of reasoning above, we now have four statements that correspond to our four possibilities of how truth values partition the domain of a condition:

English Sentence	**Formal Logic Statement**
$q(x)$ is true for every x in the domain.	$(\forall x \in D)\ q(x)$
$q(x)$ is true for at least one x in the domain.	$(\exists x \in D)\ q(x)$
$q(x)$ is false for every x in the domain.	$(\forall x \in D)\ \sim q(x)$
$q(x)$ is false for at least one x in the domain.	$(\exists x \in D)\ \sim q(x)$

Let us return back to the familiar mathematical territory to see some more examples of quantified statements.

Example 1.16. *Express the following statement in the language of formal logic.*

There exists a natural number n so that $2n$ is odd

Thought: Depending on the value of n, the truth vale of the statement "$2n$ is odd" will change. Thus "$2n$ is odd" is a condition

For $n \in \mathbb{N}$ let $p(n)$ be the condition

$2n$ is odd

Expressed in the language of formal logic our statement above becomes:

$$(\exists n \in \mathbb{N})\ p(n)$$

Example 1.17. *Let $\mathcal{V}_3$ be the set of all vector spaces of dimension 3. We know from our work in linear algebra that every vector space of dimension 3 has a basis. For $V \in \mathcal{V}_3$ let $p(V)$ be the statement*

V has a basis.

The statement

Every vector space of dimension 3 has a basis

can be expressed in formal logic as

$$(\forall V \in \mathcal{V}_3)\ p(V)$$

Consider now the statement

There exists a vector space of dimension 3 with no basis.

Expressed in formal logic, this statement becomes

$$(\exists V \in \mathcal{V}_3)\ \sim p(V)$$

Example 1.18. *Consider the statement:*

For every $x \neq 0$ and every $y \neq 0$, the product of x and y is not zero.

For $x, y \in \mathbb{R}$ let $p($... wait... something is different here. In all of our examples so far there has only been one variable. But here, the quantified condition, $xy \neq 0$, is a function of two variables!

Fortunately our notation is flexible enough to naturally handle this case. For $x, y \in \mathbb{R}$ let $p(x, y)$ be the condition $xy \neq 0$. And so we have:

$$(\forall x, y \in \mathbb{R})\ p(x, y)$$

As our condition here is already, in some sense, mathematical, we need not introduce the label p. That is, we may write:

$$(\forall x, y \in \mathbb{R})\ xy \neq 0$$

Multivariable conditions given us even more expressive power. From our work in differential calculus, we know that the derivative of a polynomial is another polynomial. Let $\mathcal{P}$ be the set of polynomials. For $f, g \in \mathcal{P}$ let $p(f, g)$ be the condition $g(x)$ is the first derivative of $f(x)$. For example, the statement denoted by $p(x^2, 2x)$ is the statement

$$g(x) = 2x \text{ is the first derivative of } f(x) = x^2$$

This is a true statement and so the statement $p(x^2, 2x)$ is true. On the other hand, the statement $p(x^3 + 1, 2x)$ is false.

To make our transition into doubly quantified statements a little more smooth, let us fix $f(x) = x^2 + 2$ and consider the condition $p(x^2 + 2, g)$. This condition has one variable, g. For some values of g the corresponding statement is false. For example $p(x^2 + 1, 3)$ is false: $g(x) = 3$ is not the first-derivative of $f(x) = x^2 + 1$. Therefore the statement

$$(\forall g \in \mathcal{P})\ p(x^2 + 1, g)$$

is false. On the other hand, since $p(x^2 + 2, 2x)$ is true, the statement

$$(\exists g \in \mathcal{P})\ p(x^2 + 1, g)$$

is true.

We know from our work in calculus that no matter which polynomial $f(x)$ we consider its derivative will always be another polynomial. In our work above, if we swap $f(x) = x^2+1$ for any other element of $\mathcal{P}$, then the statement

$$(\exists g \in \mathcal{P})\ p(f, g)$$

remains true. That is to say, for every $f \in \mathcal{P}$ the statement

$$(\exists g \in \mathcal{P})\ p(f, g)$$

is true. In other words, the statement

$$(\forall f \in \mathcal{P})[(\exists g \in \mathcal{P})\ p(f, g)]$$

is true. Reading this statement of formal logic as a sentence in English we have:

For every polynomial f we can find a polynomial g so that $g(x)$ is the first derivative of $f(x)$.

We know from our work in differential calculus that this is a true statement.

Doubly quantified statements can be tricky to work with (we'll get more practise with this in Tutorial 1). For now, let us consider one final example.

For example, consider the condition $q : x + 1 > y$ and the statement

$$(\forall x \in \mathbb{R})[(\exists y \in \mathbb{R})\ q(x, y)]$$

To determine if this statement is true, we must determine if the condition $[(\exists y \in \mathbb{R})\ q(x, y)]$ is true for every $x \in \mathbb{R}$. Let us try some sample values of x to try understand better what is happening here.

If we choose $x = 0$, then we can find $y \in \mathbb{R}$ (e.g. $y = -1$) so that $q(0, y)$ is true. If we choose $x = -50$, then we can find $y \in R$ (e.g. $y = -60$) so that $q(-50, y)$ is true. We suspect that no matter which $x \in \mathbb{R}$ we choose, we can find $y \in \mathbb{R}$ so that $q(x, y)$ is true. Therefore we suspect $(\forall x \in D)[(\exists y \in)q(x, y)]$ is true.

Consider now the statement

$$(\exists x \in \mathbb{R})[(\forall y \in \mathbb{R})\ q(x, y)].$$

As an English sentence this says:

There exists $x \in \mathbb{R}$ so that for every $y \in \mathbb{R}$ we have $x + 1 > y$

For any chosen $x \in \mathbb{R}$ the piece of notation $[(\forall y \in \mathbb{R})\ q(x, y)]$ is a statement. For example, when $x = 10$ the statement $[(\forall y \in \mathbb{R})\ 10 + 1 > y]$ is false. It is not true that $11 > y$ for every $y \in \mathbb{R}$.

When $x = 100$ the statement $[(\forall y \in \mathbb{R})\ 100 + 1 > y]$ is false. It is not true that $101 > y$ for every $y \in \mathbb{R}$.

If we are convinced that the statement $[(\forall y \in \mathbb{R})\ q(x, y)]$ is false no matter which $x \in \mathbb{R}$ we consider, then we conclude that the statement $(\exists \in \mathbb{R})[(\forall y \in \mathbb{R})q(x, y)]$ is false.

Aside. *Which of the following is easier for you to understand?*

- *Since* $(\exists A^{-1} \Leftrightarrow det(A) \neq 0) \wedge det(A) = 0, \therefore \sim \exists A^{-1}$.
- *Recall* A^{-1} *exists if and only if* $det(A) \neq 0$. *Since* $det(A) = 0$, *it then follows that* A^{-1} *does not exist.*

It is the second one, right?

The goal of mathematical communication is to communicate mathematical ideas. When we doctor up our writing with more mathematical symbols than necessary we are not being more mathematical. Instead we are doing a bad job at communicating.

At no time have these notes for this section used any of our new notation: $\vee, \wedge, \neg, \Rightarrow, \Leftrightarrow, \forall$ *and* $\exists$ *as part of English sentences to replace the words* or, and, not, implies, if and only if, for each *and* there exists. *These are symbols in formal logic and should avoided when* *we are trying to communicate mathematics to another human (such as to our tutor in our assignment solutions). Using these symbols instead of communicating in plain language the words we want to express puts a significant barrier between us and our readers.*

(Instructors sometimes use these symbols on slides to save time/space and because they are verbally explaining the concepts in parallel with using the logical symbols.)

Test Your Understanding

1. Translate each of the following to the language of formal logic
   (a) All rational numbers are larger than 6.
   (b) There is a real-number solution to $x^2 + 3x - 7 = 0$.
   (c) There is a natural number whose cube is 8.
   (d) The set of all numbers that aren't multiples of 7.
2. Rewrite each of the following with an English language sentence.
   (a) $(\forall a \in \mathbb{Q})\ a + 0 = a$
   (b) $(\forall x \in \mathbb{R})\ x^2 > 1$
   (c) $(\exists y \in \mathbb{R})[(\forall x \in \mathbb{R})\ x^2 > y]$
   (d) $(\forall x \in \mathbb{R})[(\exists x \in \mathbb{R})\ x^2 > y]$

Test Your Understanding Answers

1. (a) $(\forall r \in \mathbb{Q})\ \ r > 6$

   (b) $(\exists x \in \mathbb{R});\ \ x^2 + 3x - 7 = 0$

   (c) $(\exists n \in \mathbb{N})\quad n^3 = 8$

   (d) $\{n \in \mathbb{N} \mid \forall p \in \mathbb{N}\ \ n \neq 7p\}$

2. (a) Adding zero to any rational number doesn't change the number.

   (b) The square of any real number is greater than one.

   (c) There is at least one value $y \in \mathbb{R}$ so that $x^2 > y$ for every $x \in \mathbb{R}$

   (d) For each $x \in \mathbb{R}$ we can find $y \in \mathbb{R}$ so that $x^2 > y$

1.3 Mathematical Proofs

Before we delve into our first look at mathematical proofs, let us briefly return to thinking about the motivation for writing proofs. In addition to the aim of convincing ourselves that a statement is true, we should also remember that many of the proofs that mathematicians write down are written with the goal of convincing others that a statement is true.

Much of mathematical communication is written. And so when we write mathematics, we should have a clear idea in our heads of our ideal reader[5]. That is, a hypothetical reader who has the background and context to understand our writing.

Just as these notes are written for you, a student in MAST20026, so too should your written work in this course be written for this audience. The best way for you to display your understanding of the course material is to communicate it at a level that can be understood by your peers in this course.

With these thoughts in mind, we can think of a proof as a conversation. The writer is telling the reader how to convince themselves that a statement is true. And so not only does the proof need to be logically correct, but it also need to be written so as to be easily understood by the intended audience[6]. We'll talk more about achieving this aim in future sections, but for now let us dive in to thinking about to create logically correct proofs.

Most mathematical statements (i.e., theorems) that we encounter appear to be of the form

$$\text{If } p, \text{ then } q.$$

For example:

Theorem. *Let $f : \mathbb{R} \to \mathbb{R}$. If $f$is differentiable, then f is continuous.*

Theorem. *Let $f : \mathbb{R} \to \mathbb{R}$. If f is a quadratic function, then f crosses the x-axis at most twice.*

Theorem. *Let V be a vector space. If V is finite-dimensional, then V has a basis.*

Let us look more closely at first theorem. The implication has two parts:

$$\text{(Hypothesis) } p(f) : f \text{ is differentiable.}$$

$$\text{(Conclusion) } q(f) : f \text{ is continuous}$$

The truth of these statements depend on the particular choice of f. In fact, these are conditions. This theorem tells us that no matter which f we choose, the implication $p(f) \Rightarrow q(f)$ is true. And so, as a statement of formal logic we can encode the first theorem as

$$(\forall f \in \mathcal{F}) \;\; p(f) \Rightarrow q(f)$$

In thinking about how to prove an implication, let us recall the truth table for $\Rightarrow$.

[5] https://en.wikipedia.org/wiki/Reader-response criticism

[6] many textbooks supposedly written for undergraduates seem to forget this fact.

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

Looking at the rows, we see that when p is false, the implication $p \Rightarrow q$ is necessarily true. The only way an implication can be false is if p is true and q is false. Thus to prove $p \Rightarrow q$ it suffices to assume p is true and deduce that q is necessarily true. This strategy (assume p is true and deduce q is true) is the basis of our first proof technique: direct proof.

In a direct proof one assumes that the hypothesis is true and deduces the conclusion is true. This technique, of course, presents us with a question: how do we communicate our deductions? In the previous section we briefly discussed the difference between formal and informal proof. Let us consider a toy example to see what these two *modes* of proof are.

Theorem 1.19. *Let x be an integer. If x is even, then x^2 is even.*

We begin with an *informal proof*

Proof. (informal) Let x be an even integer. Since x is even there exists $k \in \mathbb{Z}$ so that $x = 2k$. Computing we find $x^2 = 2(2k^2)$. Thus there exists $\ell \in \mathbb{Z}$ so that $x^2 = 2\ell$. In particular, $\ell = 2k^2$. Therefore x^2 is even. Therefore if x is even, then x^2 is also even. □

How, then, can turn our informal proof into a formal proof? This is where our work in formal logic comes in handy. Let $p(x)$ be the condition

$$(\exists k \in \mathbb{Z})\, x = 2k$$

And let $q(x)$ be the condition

$$(\exists k \in \mathbb{Z})\, x^2 = 2k$$

Looking back at our strategy for direct proof, we want to deduce $q(x)$ is true, given that $p(x)$ is true.

Consider the following sequence of true statements and their justifications.

Proof. (formal)

1. $p(x)$ (premise)
2. $(\exists k \in \mathbb{Z})\quad x = 2k$ (1.)
3. $x^2 = 2(2k^2)$. (algebra)
4. $(\exists \ell \in \mathbb{Z})\quad x^2 = 2\ell$ (3.)
5. $q(x)$ (4.)
6. $p(x) \Rightarrow q(x)$. (1.,5.)

□

Translating these back to English prose we have:

Proof. (semi-formal)

1. x is an even integer. (premise)
2. There exists $k \in \mathbb{Z}$ so that $x = 2k$ (1.)
3. $x^2 = 2(2k^2)$. (algebra)
4. There exists $\ell \in \mathbb{Z}$ so that $x^2 = 2\ell$. (3.)
5. x^2 is an even integer. (4.)
6. If x is an even integer, then x^2 is an even integer. (1.,5.)

□

The sequence of statements in our formal proof match almost exactly we wrote as part of our informal proof above. The only difference is that in our informal proof we incorporated the justifications as part of the sentences, whereas here we have stated them in parentheses.

Broadly, we can define a mathematical proof as follows:

Definition 1.20. *A mathematical proof, is a sequence of statements where each statement is*

- *a declaration of notation,*
- *a premise, or*
- *a true statement that follows from*
  - *a definition,*
  - *algebra, or*
  - *previous true statements*

When we consider formal proofs, we can make this definition even more precise.

Definition 1.21. *a formal mathematical proof is a finite sequence of statements*

$$A_1, A_2, \ldots\ldots, A_n$$

such that each A_i is either

- *known or assumed true or*
- *can be inferred from a known or assumed true statement A_j, with $j < i$ (ie., a previous statement).*

Our definition for mathematical proof states that each statement that isn't a premise or a declaration of notation must follow from a definition, algebra or a previous true statement. Let us deal with these possibilities one at a time.

As was mentioned in the previous section, part of participating in the mathematical community is to become versed in mathematical nomenclature. Give a common framework for everyone working in and learning about a subject. Definitions (and notation) are a mathematical shorthand – they allow us to communicate precise mathematical ideas using only a word or two. Understanding the meaning of the definitions in this course is a key step to understanding the proofs that we will write.

Quite often, students who are struggling to understand course material are in fact struggling to understand the definitions of new terms and notation. Each time we introduce new terms in this course, the newly defined word or phrase will be underlined. From then forward, that word of phrase should be taken to mean precisely what the definition states.

Though much of your day-to-day mathematics through school has been education in techniques for algebraic manipulation. Unless particularly insightful, steps such as factoring, grouping and expanding will often be skipped over during lecture/in these notes.

Finally, let us consider what is meant by *previous true statement* in our definition of proofs. Consider the following example.

Theorem 1.22. *Let x and y be integers. If x and y are even, then $(x+y)^2$ is even.*

Proof. Let x and y be even integers. We compute $(x+y)^2 = x^2 + 2xy + y^2$. By Theorem 1.19 each of x^2 and y^2 is even. By the definition of even integer, $2xy$ is even. The sum of three even integers is necessarily even. Therefore $(x+y)^2$ is even. Therefore if x and y are both even, then $(x+y)^2$ is even. □

Re-written as a formal proof we have

Proof. $p(x) : (\exists k \in \mathbb{Z})\, x = 2k$

1. $p(x) \wedge p(y)$ (premise)
2. $p(x)$ (1., defn of $\wedge$)
3. $p(y)$ (1., defn of $\wedge$)
4. $(x+y)^2 = x^2 + 2xy + y^2$. (algebra)
5. $p(z) \Rightarrow p(z^2)$ (Thm 1.19)
6. $p(x^2)$ (2.,5., defn of $\Rightarrow$) [7]
7. $p(y^2)$ (3.,5.)
8. $(\exists k \in \mathbb{Z})\, 2xy = 2k$ (algebra)

[7] Here we conclude $p(x^2)$ is true because we know that $p(x)$ is true and $p(x) \Rightarrow p(x^2)$ is true. This logical deduction is called *Modus Ponens*. Translated from the Latin, this phrase means: "The method that affirms the conclusion by affirming the hypothesis".

9. $p(2xy)$ (8.)

10. $p(x^2) \wedge p(y^2) \wedge p(2xy) \Rightarrow p(x^2 + y^2 + 2xy)$ (???)

11. $p(x^2 + y^2 + 2xy)$ (6.,7.,9., defn of $\wedge$, 10, defn of $\Rightarrow$)

12. $p((x + y)^2)$. (4.)

13. $p(x) \wedge p(y) \Rightarrow p((x + y)^2)$ (1.,12.)

□

In the proof above, each of the lines is labeled with its justification. For example, 6. is true because $p(x)$ is true (this is line 2.) and the statement $p(z) \Rightarrow p(z^2)$ is true (this is line 5.). And so when $z = x$ we can conclude $p(x^2)$ is true. However there is one justification missing

$$10.\ \ p(x^2) \wedge p(y^2) \wedge p(2xy) \Rightarrow p(x^2 + y^2 + 2xy) \quad (???)$$

Written in prose, this says

If x^2, y^2 and $2xy$ are all even, then their sum is even.

If we wanted to fully adhere to the definition of formal proof, then this statement should have a justification. And so, we would need to prove a proof of the following theorem.

Theorem 1.23. *Let a, b and c be integers. If a, b and c are even, then $a + b + c$ is even.*

Proof. proof omitted. □

With a proof of Theorem 1.23 in hand, we could then write:

$$10.\ \ p(x^2) \wedge p(y^2) \wedge p(2xy) \Rightarrow p(x^2 + y^2 + 2xy) \quad \text{(Thm 1.23)}$$

With this example, perhaps we start to understand why it may be impractical to give full formal proofs for every mathematical statement. And so in this course our proofs will generally be similar to the informal proof given above for Theorem 1.19.

As we are learning for the first time to write proofs, it is difficult for us to write clear and cogent prose. To bridge the gap between formal proofs (a sequence of mathematical statements written in first-order logic) and informal proofs (paragraphs with full sentences), we can use *a semi-formal* proof like the one given in. In *semi-formal* proofs, we give our mathematical statements in plain English, but we number our lines and parenthesise our justifications. This technique helps to remind us that every true statement must be justified. In this course when you are asked to *prove* or *show*, feel welcome to use a *semi-formal* proof.

Test Your Understanding

1. Prove Theorem 1.23 with a direct proof

2. Prove the following theorem with a direct proof.

Theorem. *Let x be an even integer and let y be an odd integer. The product, xy, is an even integer*

Write you proof as both a formal proof and as an informal proof.

Test Your Understanding Solution

1. Informal: Let a, b, c be even integers. Therefore there exists integers k_a, k_b, k_c so that

$$\begin{aligned} a &= 2k_a \\ b &= 2k_b \\ c &= 2k_c \end{aligned}$$

   Therefore $a + b + c = 2(k_a + k_b + k_c)$. Therefore $a + b + c$ is even.

2. Let x be an even integer and let y be an odd integer. Therefore there exits integers k_x and k_y so that

$$\begin{aligned} x &= 2k_x \\ y &= 2k_y + 1 \end{aligned}$$

   We compute $xy = (2k_x)(2k_y + 1) = 4k_xk_y + 2k_x = 2(2k_xk_y + k_x)$. Therefore xy is even.

   Formal:

   - $p(z) : x$ is even
   - $q(z) : z$ is odd

   (a) $p(x)$ (premise)

   (b) $q(y)$ (premise)

   (c) $(\exists k_x \in \mathbb{Z})\ x = 2k_x$ (a)

   (d) $(\exists k_y \in \mathbb{Z})\ x = 2k_y + 1$ (b)

   (e) $xy = (2k_x)(2k_y + 1) = 4k_xk_y + 2k_x = 2(2k_xk_y + k_x)$. (a,b, algebra)

   (f) $(\exists \ell \in \mathbb{Z})\ xy = 2\ell$ (e)

   (g) $p(xy)$. (f)

   (h) $p(x) \wedge q(y) \Rightarrow p(xy)$ (a,b,g)

1.3.1 Indirect Methods

In the previous section we took our first look at constructing mathematical proofs. We introduced the notion of direct proof – a proof technique where we assume the hypothesis of an implication and then use a sequence of logical deductions to confirm the conclusion to be true. Here we introduce some other techniques for proving implications.

Contrapositive

Recall the following fact from our work in formal logic.

$$p \Rightarrow q \equiv \sim q \Rightarrow \sim p$$

This logical equivalence gives us an alternate approach when trying to prove an implication.

Consider the following implication

Theorem 1.24. *Let $x \in \mathbb{Z}$. If $x^2 - 6x + 5$ is even, then x is odd.*

Let us try to prove this statement with a direct proof. We begin with our premises and the definition of even.

1. Let x be an integer so that $x^2 - 6x + 5$ is even.
2. There exists $k \in \mathbb{Z}$ so that $x^2 - 6x + 5 = 2k$.

To show x is odd, ideally we would find an integer ℓ so that $x = 2\ell + 1$. But, it is difficult to algebraically manipulate the expression $x^2 - 6x + 5 = 2k$ to isolate x. Let us try then a different approach.

Rather than prove the statement

$$\text{If } x^2 - 6x + 5 \text{ is even, then } x \text{ is odd}$$

instead let us consider the contrapositive:

$$\text{If } x \text{ is even, then } x^2 - 6x + 5 \text{ is odd.}$$

Let us try to prove this with a direct proof. We assume our hypothesis is true and try to conclude the conclusion is true.

Proof.

1. x is an even integer (premise)
2. there exists $k \in \mathbb{Z}$ so that $x = 2k$ (1, defn of even)
3. $x^2 - 6x + 5 = (2k)^2 - 6(2k) + 5 = 4k^2 - 12k + 5$ (2, algebra)
4. $4k^2 - 12k + 5 = 4k^2 - 12k + 4 + 1 = 2(2k^2 + 6k + 2) + 1.$ (3, algebra)
5. there exists $\ell \in \mathbb{Z}$ so that $x^2 - 6x + 5 = 2\ell + 1$ (4)
6. $x^2 - 6x + 5$ is odd (5, defn of odd)

7. if x is an even integer, then $x^2 - 6x + 5$ is odd (1,6)

□

Since the statement

$$\text{If } x \text{ is even, then } x^2 - 6x + 5 \text{ is odd.}$$

is true. And an implication is logically equivalent to its contrapositive, then the statement

$$\text{If } x^2 - 6x + 5 \text{ is even, then } x \text{ is odd}$$

is true.

This technique is called proof by contrapositive. In a proof by contrapositive one first re-writes the implication as its contrapositive and then uses a direct proof to prove the contrapositive.

Contradiction

Our definition for mathematical proof states that each statement that isn't a premise or a declaration of notation must follow from a definition, algebra or a previous true statement. When the premise indeed is true, then as a consequence all of the subsequent statements are true. However, what would happen if the premise was false?

Consider the following sequence of statements.

1. $\frac{1}{2}$ is an integer. (premise)
2. $1 + \frac{1}{2}$ is an integer (1, integers are closed under addition)
3. $\frac{3}{2}$ is an integer. (2)
4. There is an integer between 1 and 2. (3)

If (1) is true, then necessarily (4) is true. Since (4) is false, necessarily (1) is false.

Since we cannot deduce a false statement from a true premise, if we manage to deduce a false statement it must be that our premise is false.

This technique is called proof by contradiction. Let us see it in action.

Theorem 1.25. *Let a and b be real numbers. If a is a rational number and b is irrational number, then the sum $a + b$ is irrational.*

We prove this theorem by contradiction. To do so we assume that the hypothesis is true and the conclusion is false. That is, we assume:

$$\text{(hypothesis) } a \text{ is a rational number and } b \text{ is an irrational number}$$

is true. And

$$\text{(conclusion) } a + b \text{ is an irrational number}$$

is false. We then derive a contradiction. That is, we deduce something we know to be false.

In this proof our contradiction will be producing a real number that is both rational and irrational. Since this is impossible, one of our assumptions must be false. If we maintain that the hypothesis is true, then it must be the conclusion must also be true.

Proof. We proceed by contradiction. Let a and b be real numbers. Assume a is a rational. Therefore there exists $u, v \in \mathbb{Z}$ so that $a = u/v$. Assume b irrational. Therefore for every $p, q \in \mathbb{Z}$ we have $b \neq p/q$.

Assume $a + b$ is rational. Therefore there exists $x, y \in \mathbb{Z}$ so that $a + b = x/y$. Therefore

$$b = (a + b) - a = \frac{vx - uy}{vy}.$$

Since $vx - uz \in \mathbb{Z}$ and $vy \in \mathbb{Z}$, it follows that b is rational. This contradicts that b is irrational.

Therefore if a is a rational number and b is irrational number, then the sum $a + b$ is irrational. □

Here our proof has three premises:

- a is rational
- b is irrational
- $a + b$ is rational

Assuming these three premises lead us to the following contradiction:

$$b \text{ is rational and } b \text{ is irrational}$$

Thus, if our first two premises are true, then the third must be false. In other words, if a is rational and b is irrational, then $a + b$ must be irrational.

To use a proof by contradiction to prove an implication, one assumes that the hypothesis is true and the conclusion is false. From these two premises, one then attempts to deduce a false statement.

Test Your Understanding

1. Prove the following statement using a proof by contrapositive:

   Let $n \in \mathbb{Z}$. If n^4 is even, then n is even.

2. Prove the following statement using a proof by contradiction:

   Let $p, q \in \mathbb{N}$ with $p, q > 0$. If $pq = 1$, then $p = q = 1$.

Test Your Understanding Solution

1. *Proof.* Assume n is odd. Since n is odd there exists $m \in \mathbb{Z}$ so that $n = 2m+1$. We compute

$$\begin{aligned} n^4 &= (2m+1)^4 \\ &= 32m^4 + 32m^3 + 24m^2 + 8m + 1 \\ &= 2(16m^4 + 16m^3 + 12m^2 + 4m) + 1 \end{aligned}$$

   Therefore there exists an integer ℓ so that $n^4 = 2\ell + 1$. Hence n^4 is odd.

   By the contrapositive if n^4 is even, then n is even. □

2. *Proof.* We proceed by contradiction. That is, we assume $pq = 1$ but $p \neq 1$ or $q \neq 1$.

   As $pq = 1$ it must be $p, q \neq 0$.

   Assume $p \neq 1$. Since $p \neq 0$ necessarily $p \geq 2$. Therefore $pq > 1$. Therefore $pq \neq 1$. This contradicts that $pq = 1$.

   Assume $q \neq 1$. Since $q \neq 0$ necessarily $q \geq 2$. Therefore $pq > 1$. Therefore $pq \neq 1$. This contradicts that $pq = 1$.

   □

1.3.2 Quantified Statements

So far we have talked about how to prove an implication. But not every mathematical theorem is an implication. For example, consider the following statement

Theorem 1.26. *There are infinitely many prime numbers*

This certainly is a mathematical statement; it is unambiguously true or false. Looking at our work from above, however, this statement is markedly different than the ones we saw above; it not an implication. In fact, this is a quantified statement in disguise. Another way to convey that there are infinitely many prime numbers is with the following quantified

For every prime number p, there exists a prime number q so that $q > p$

Let $\mathcal{P}$ be the set of prime numbers. Written in formal logic, we can express this statement as follows

$$(\forall p \in \mathcal{P})[(\exists q \in P)\ q > p]$$

Recall that quantified statements come in two flavours: existentially quantified statements and universally quantified statements.

Proving Existentially Quantified Statements
We proceed with an example. Let $\mathcal{P}_1$ denote the set of linear functions. And let $q(f)$ be the following condition over $\mathcal{P}_1$.

$f(x)$ *passes through the origin*

Consider the following statement.

$$(\exists f \in \mathcal{P}_1)\ \ q(f)$$

This statement is certainly true. We know this because can produce an example of an element of $\mathcal{P}_1$ for which $q(f)$ is true. For example, $f(x) = 3x$. Since $q(3x)$ is true, it follows that $(\exists f \in \mathcal{P}_1)q(f)$ is true.

In summary, if our statement asserts that an object with a particular property exists, then producing an example of such an object is all we need to prove that our statement is true.

Producing an example also serves as a strategy to disprove a some quantified statements. Let $\mathcal{P}$ be the set of prime numbers and consider the statement

Every prime number is odd.

This statement is false. We know this statement is false because we can produce a counterexample. In other words, we can produce a prime number that isn't odd. Namely, 2.

In summary, if our statement asserts that every object of a particular type has a particular property, then producing an example of an object that does not have the particular property serves to disprove the statement. We refer to such examples as counter examples

We can use examples to prove an existential statement and disprove universal statements. But what about the opposite? What strategies do we have to disprove existential statements and prove universal statements?

Proving Universally Quantified Statements

Let us restrict our attention for the moment to conditions over the natural numbers. As opposed to statements of the form $(\exists n \in \mathbb{N})\, p(n)$, justifying the truth of a universally quantified statement, $(\forall n \in \mathbb{N})\, p(n)$, seems like a lot more work – we cannot just provide an example of particular values of n for which $p(n)$ is true. Instead we must argue that $p(n)$ is true for each and every possible value of n.

One common tool for proving universally quantified statements over the natural numbers is *mathematical induction*. Proof by mathematical induction is a technique that lets us justify the truth of a universally quantified statement.

Remembering one of our reasons for writing proofs, a proof by induction is a roadmap for your reader. It tells your reader how to prove $p(n)$ is true for any particular value of n they may care about. If, from what you have written, your reader can verify $p(n)$ is true for any particular value of n that they may care about, then necessarily $p(n)$ is true for every possible value of n.

Imagine your reader wanted to verify $p(7)$ was true for some formula $p(n)$ over $\mathbb{N}$. Instead of telling them directly how to prove $p(7)$ is true, you told them the following two facts:

(1) $p(0)$ is true; and

(2) for every $k \in \mathbb{N}$, if $p(k)$ is true, then $p(k+1)$ is true.

When $k = 6$, (2) says:

if $p(6)$ *is true, then* $p(7)$ *is true.*

Thus, for your reader to verify $p(7)$ is true is suffices for them to verify $p(6)$ is true and then apply (2) with $k = 6$.

Similarly, when $k = 5$, (2) says:

if $p(5)$ *is true, then* $p(6)$ *is true.*

Thus, for your reader to verify $p(6)$ is true it suffices for them to verify $p(5)$ is true and then apply (2) with $k = 5$. And so, for your reader to verify $p(7)$ is true, it suffices for them to verify $p(5)$ is true.

Continuing in this fashion, we can conclude for your reader to verify $p(7)$ is true, it suffices for them to verify $p(0)$ is true. Looking at statement (1), we have told our reader that $p(0)$ is true. And so by knowing (1) and (2) are true, the reader is certain $p(7)$ is true.

Statement (1) tells us that $p(0)$ is true. Statement (2) with $k = 0$ then tells us that $p(1)$ is true. Statement (2) with $k = 1$ then tells us that $p(2)$ is true. Continuing in this fashion, we should be convinced that when (1) and (2) hold, then $p(n)$ is true for any $n \in \mathbb{N}$.

Fluent speakers of western English may recognize the adage:

If you give a someone a fish they will be hungry again tomorrow. If you teach them to catch a fish you feed them for a lifetime

Proof by induction is a *teach someone to fish* proof technique. Instead of telling your reader (or yourself!) how to verify $p(n)$ is true for every $n \in \mathbb{N}$. You are showing them how to verify $p(n)$ is true for any $n \in \mathbb{N}$ they may care about.

Let us make these ideas more concrete with an example. Let n be a non-negative integer. Consider the following sum:

$$s(n) = \sum_{i=0}^{n} 2^i = 2^0 + 2^1 + 2^2 + \cdots + 2^n$$

When n is small it is easy to compute values of $s(n)$:

$$\begin{aligned} s(0) &= 2^0 = 1 \\ s(1) &= 2^0 + 2^1 = 3 \\ s(2) &= 2^0 + 2^1 + 2^2 = 7 \\ s(3) &= 2^0 + 2^1 + 2^2 + 2^3 = 15 \\ s(4) &= 2^0 + 2^1 + 2^2 + 2^3 + 2^4 = 31 \end{aligned}$$

You may notice each of these values is one fewer than a power of 2.

$$\begin{aligned} s(0) &= 2^1 - 1 \\ s(1) &= 2^2 - 1 \\ s(2) &= 2^3 - 1 \\ s(3) &= 2^4 - 1 \\ s(4) &= 2^5 - 1 \end{aligned}$$

We wonder does this pattern hold in general. That is, does $s(n) = 2^{n+1} - 1$ for each non-negative integer n? Let us consider the case $n = 5$, but rather than compute directly, let us try a different approach:

$$s(5) = 2^0 + 2^1 + 2^2 + 2^3 + 2^4 + 2^5$$

Notice $2^0 + 2^1 + 2^2 + 2^3 + 2^4 = s(4)$. We have already confirmed $s(4) = 2^5 - 1$. Thus

$$s(5) = \left(2^0 + 2^1 + 2^2 + 2^3 + 2^4\right) + 2^5 = s(4) + 2^5 = \left(2^5 - 1\right) + 2^5$$

Simplifying, we notice $2^5 + 2^5 = 2(2^5) = 2^6$. Therefore $s(5) = 2^6 - 1$.

Using this same technique, we can verify $s(6) = 2^7 - 1$ as

$$s(6) = 2^0 + 2^1 + 2^2 + 2^3 + 2^4 + 2^5 + 2^6 = s(5) + 2^6$$

Using $s(6) = 2^7 - 1$ we can use the same technique to verify $s(7) = 2^8 - 1$. We could continue on indefinitely. For any particular value of k, once we have verified that our formula holds for $s(k)$, we can then verify that our formula holds for $s(k + 1)$.

Let us think now about how we can express these ideas using our language of formal logic. Let $p(n)$ be the following condition over $\mathbb{N}$

$$\sum_{i=0}^{n} 2^i = 2^{n+1} - 1$$

For any particular value of $n \in \mathbb{N}$, this is a mathematical statement – it is either true for false. We have confirmed that each of $p(0), p(1), p(2), p(3), p(4)$ and $p(5)$ is true. In asking if the statement

$$\sum_{i=0}^{n} 2^i = 2^{n+1} - 1$$

holds for each integer $n \geq 0$, we are asking if $p(n)$ is true for each $n \geq 0$.

Think back a moment to the argument we made to verify that $p(5)$ is true. Consider the sum

$$2^0 + 2^1 + 2^2 + 2^3 + 2^4 + 2^5$$

Since $p(4)$ is true, we have $2^0 + 2^1 + 2^2 + 2^3 + 2^4 = 2^5 - 1$. Therefore

$$2^0 + 2^1 + 2^2 + 2^3 + 2^4 + 2^5 = \big(2^0 + 2^1 + 2^2 + 2^3 + 2^4\big) + 2^5 = (2^5 - 1) + 2^5 = 2^6 - 1.$$

In this argument there is nothing particular special about $n = 4$ and $n = 5$. Using this same argument, we can verify that if $p(k)$ is true for some $k \in \mathbb{N}$, then necessarily $p(k+1)$ is true:

For some $k \in \mathbb{N}$, if we knew for certain that

$$\sum_{i=0}^{k} 2^i = 2^{k+1} - 1$$

then we could conclude:

$$\sum_{i=0}^{k+1} 2^i = \left(\sum_{i=0}^{k} 2^i\right) + 2^{k+1} = \big(2^{k+1} - 1\big) + 2^{k+1} = 2^{k+2} - 1$$

Thus if $p(k)$ is true for some $k \in \mathbb{N}$, then necessarily $p(k + 1)$ is true. Since $p(0)$ is true, this then should convince us that $p(n)$ is true for every $n \in \mathbb{N}$. That is, the statement $(\forall n \in \mathbb{N})\, p(n)$ is true.

Let us consider another example and employ a similar technique. For a set X, let $\mathbf{2}^X$ denote the set of all subsets of X.

For example, when $X = \{x_1, x_2, x_3\}$ we have

$$\mathbf{2}^X = \{\{\}, \{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\}, \{x_2, x_3\}, \{x_1, x_2, x_3\}\}$$

We call $\mathbf{2}^X$ the power set of X.

Notice here that X has 3 elements and $\mathbf{2}^X$ has $2^3 = 8$ elements.

Consider now the empty set. We have $\mathbf{2}^{\{\}} = \{\{\}\}$. This set has a single element: the empty set [8]. And so we notice the empty set has 0 elements and $\mathbf{2}^{\{\}}$ has $2^0 = 1$ elements.

Let $p(n)$ be the formula

There are 2^n subsets of a set n elements.

Above we have verified that $p(0)$ and $p(3)$ are true. Take a moment to convince yourself that $p(1)$ and $p(2)$ are both true.

Consider now the case $n = 4$. The two columns below list out all of the subsets of the set $X = \{x_1, x_2, x_3, x_4\}$.

$$\begin{array}{cc}
\{\} & \{x_4\} \\
\{x_1\} & \{x_1, x_4\} \\
\{x_2\} & \{x_2, x_4\} \\
\{x_3\} & \{x_3, x_4\} \\
\{x_1, x_2\} & \{x_1, x_2, x_4\} \\
\{x_1, x_3\} & \{x_1, x_3, x_4\} \\
\{x_2, x_3\} & \{x_2, x_3, x_4\} \\
\{x_1, x_2, x_3\} & \{x_1, x_2, x_3, x_4\}
\end{array}$$

Notice that the eight sets on the left are exactly the subsets of the set $\{x_1, x_2, x_3\}$. The sets on the right are those same subsets but with the addition of the element x_4.

We see that the subsets of X can be partitioned into those that do not contain x_4 and those that do contain x_4. The subsets that do not contain x_4 are the subsets of $\{x_1, x_2, x_3\}$. Since $p(3)$ is true there are exactly 8 of these.

For every subset of X that does not contain x_4, there is a corresponding subset of X that does contain x_4. Therefore the number of subsets of $\{x_1, x_2, x_3, x_4\}$ is exactly two times the number of subsets of $\{x_1, x_2, x_3\}$. Since there are 2^3 subsets of a set with three elements (we know this because $p(3)$ is true), there must be $2 \times 2^3 = 2^4$ subsets of a set with four elements. Therefore $p(4)$ is true.

This same argument can be used to show that there are 2^5 subsets of a set with five elements. We can write down all of the subsets of the set $\{x_1, x_2, x_3, x_4, x_5\}$ in two columns. The first of these columns are all of the subsets that do not contain x_5. The second of these columns are the subsets we get by adding x_5 to each subset in the first column. The first column is the set of subsets of the set $\{x_1, x_2, x_3, x_4\}$. Since $p(4)$ is true, this column has 2^4 subsets. Therefore there are $2 \times 2^4 = 2^5$ subsets of $\{x_1, x_2, x_3, x_4, x_5\}$. And so we see that $p(5)$ is true.

[8] We'll return to tread sets more carefully in Section 3. For now, just accept that the empty set is a set and that it is a subset of every set.

Using this same argument, we can verify that a set with $k+1$ elements has twice as many subsets as a set with k elements. In other words, if a set with k elements has 2^k subsets, then necessarily a set with $k+1$ elements has 2^{k+1} subsets.

We can write down all of the subsets of the set $\{x_1, x_2, \ldots, x_{k+1}\}$ in two columns. The first of these columns are all of the subsets that do not contain x_{k+1}. The second of these columns are the subsets we get by adding x_{k+1} to the subset to each subset in the first column. The first column is the set of subsets of the set $\{x_1, x_2, \ldots, x_k\}$. If $p(k)$ is true, this column has 2^k subsets. Therefore there are $2 \times 2^k = 2^{k+1}$ subsets of $\{x_1, x_2, \ldots, x_{k+1}\}$. And so we see that if $p(k)$ is true, then $p(k+1)$ is true.

That is to say, if $p(k)$ is true for some $k \in \mathbb{N}$, then necessarily $p(k+1)$ is true. Since $p(0)$ is true, this then should convince us that $p(n)$ is true for every $n \in \mathbb{N}$. That is, the statement $(\forall n \in \mathbb{N})\, p(n)$ is true.

In both of our examples ($s(n)$ and counting subsets) we justified the truth the universally quantified statement: $(\forall n \in \mathbb{N})\, p(n)$ by showing to be true the following two statements:

1. $p(0)$
2. for every $k \in \mathbb{N}$, if $p(k)$ is true, then $p(k+1)$ is true.

This technique is so ubiquitous that this proof technique is its own theorem:

Theorem (The Principle of Mathematical Induction)**.** *Let $p(n)$ be a condition over $\mathbb{N}$. If the following two statements are true, then $p(n)$ is true for each integer $n \geq 0$.*

(1) $p(0)$; and

(2) for each $k \geq 0$, if $p(k)$ is true, then $p(k+1)$ is true.

Though stated as a theorem, the Principle of Mathematical Induction is usually thought of as a proof technique. If we can show that the hypotheses of this theorem hold for some particular formula $p(n)$ over $\mathbb{N}$, then applying the theorem above tells us the statement $(\forall n \in \mathbb{N})\, p(n)$ is true.

Let us return to our first-example and use the Principle of Mathematical Induction to prove $\sum_{i=0}^{n} 2^i = 2^{n+1} - 1$ for each $n \in \mathbb{N}$.

Theorem 1.27. *For every $n \in \mathbb{N}$ we have $\sum_{i=0}^{n} 2^i = 2^{n+1} - 1$*

Proof. We proceed by induction on n. Let $p(n)$ the following formula over $\mathbb{N}$:

$$\sum_{i=0}^{n} 2^i = 2^{n+1} - 1$$

We show that both (1) and (2) hold in the hypothesis of the Principle of Mathematical Induction for $p(n)$.

$\underline{\text{(1) } p(0) \text{ is true:}}$
$p(0)$ is the statement

$$2^0 = 2^1 - 1$$

Since $2^0 = 1$ and $2^1 - 1 = 1$, then $p(0)$ is true.

(2) for each $k \geq 0$, if $p(k)$ is true, then $p(k+1)$ is true
Consider some integer $k \geq 0$ so that $p(k)$ is true. Since $p(k)$ is true, we know

$$\sum_{i=0}^{k} 2^i = 2^{k+1} - 1.$$

Consider the statement $p(k+1)$:

$$\sum_{i=0}^{k+1} 2^i = 2^{k+2} - 1$$

Notice

$$\sum_{i=0}^{k+1} 2^i = \left(\sum_{i=0}^{k} 2^i\right) + 2^{k+1}$$

Since $p(k)$ is true, we have

$$\sum_{i=0}^{k+1} 2^i = \left(2^{k+1} - 1\right) + 2^{k+1}$$

Therefore

$$\sum_{i=0}^{k+1} 2^i = 2^{k+2} - 1$$

And so it follows that $p(k+1)$ is true.

Since both of the hypotheses in the Principle of Mathematical Induction hold, necessarily the conclusion holds. That is, $p(n)$ is true for each integer $n \geq 0$. Therefore for every $n \in \mathbb{N}$ we have

$$\sum_{i=0}^{n} 2^i = 2^{n+1} - 1.$$

□

The statement of (2) is essential a mini-theorem in and of itself. In showing that (2) holds, we assume the hypothesis is true (i.e., we assume $p(k)$ is true for some $k \in \mathbb{N}$) and we then show that the conclusion is true (i.e., $p(k+1)$ is true). We usually accomplish this by way of direct proof.

Aside. *If you have seen* `proof by induction` *before you may wonder why the phrases* `base case`, `induction premise` and `induction step` *have not yet appeared. In standard use, the term* `base case` *refers to (1) in the statement of the Principle of Mathematical Induction. The term* `induction premise` *refers to the hypothesis of (2) in the statement of the Principle of Mathematical Induction. The term* `induction step` *refers to proving the conclusion of (2) in the statement of the Principle of Mathematical Induction.*

Using these terms is not wrong, not at all! If you are comfortable with these terms, then please continue to use them. I have avoided these terms because for students seeing these ideas for the first time, hiding (1) and (2) behind these terms sometimes makes proof by induction feel like magic. As we get more comfortable with induction we will come to realize that we need not even invoke the ideas of induction directly in our proofs. Just mentioning to our reader that we are proceeding by induction will be enough to communicate our general proof strategy.

In our use of the Principle of Mathematical Induction we are proving that some formula over $\mathbb{N}$ holds for every value of n. To verify $p(k+1)$ is true for some particular value of k, we use the hypothesis that $p(k)$ is true.

When $p(n)$ was the formula

$$\sum_{i=0}^{n} 2^i = 2^{n+1} - 1,$$

we used the truth of $p(4)$ to verify that $p(5)$ was true. Even though $p(3), p(2), p(1)$ and $p(0)$ were also true, we did not directly invoke these facts in verifying that $p(5)$ was true. However the technique of proof by induction is flexible enough for conditions where we need to invoke the truth of many previous cases in order to determine the truth of the case we are considering. We turn now to consider this possibility.

The Principle of Strong Mathematical Induction

Mathematical induction comes in many forms, all based on the techniques above. To see another use of induction, let us proceed with another example.

Our modern base-ten place-value system of notation is incredibly useful. It is so ingrained in us, that most of the time we don't even notice. Recall that the notation

$$1342$$

Refers to the number equal to

$$1 \times 10^4 + 3 \times 10^2 + 4 \times 10^1 + 2 \times 10^0$$

This system is based upon powers of 10 and uses the digits $0-9$. Such a system is likely based upon us humans having ten fingers.

Imagine what our number system would be like if we only had thumbs. We may have developed a place-value system based upon powers of 2 using digits 0 and 1. For example we would write:

$$1101$$

to refer to the number equal to

$$1 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0.$$

We would express this number as 13 in our base-ten notation.

For such a system to be useful as our base ten system, we would need to ensure that every natural number is either power of 2 or can be expressed as a sum of powers of 2.

Numbers on the left are in base ten. The equivalent representation in base two is on the right. The equals sign means that the two pieces of notation represent the same integer.

$$
\begin{aligned}
0 &= 0\\
1 &= 1\\
2 &= 10\\
3 &= 11\\
4 &= 100\\
5 &= 101\\
6 &= 110
\end{aligned}
$$

So far so good, now what about 7? Rather than continue with the pattern, let us try and be slightly clever. We notice that $7 = 4 + 3$. The integer 4 can be expressed as 100 in base two. The integer 3 can be expressed as 011 in base two. Thus

$$
\begin{aligned}
4 &= 1 \times 2^2 + 0 \times 2^1 + 0 \times 2^0\\
3 &= 1 \times 2^1 + 1 \times 2^0
\end{aligned}
$$

Adding these together yields

$$7 = 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0$$

Thus 7 can be expressed as 111 in base-two.

Marching on, let us think about 8. Since $8 = 2^3$, this isn't terribly interesting– 8 is a power of two.

$$
\begin{aligned}
7 &= 111\\
8 &= 1000
\end{aligned}
$$

For 9 we can use the same strategy as for 7. We notice $9 = 8 + 1 = 2^3 + 2^0$ and so 9 can be expressed as 1001 in base-two.

We could continue on with this, but this may take a while. Let us assume we have been successful in expressing each integer up to, say, 24, as a power of 2 or as a sum of powers of 2. To express 25 as a sum of distinct powers of 2, we notice $25 = 16 + 9$. As we have assumed we have been successful in expressing each integer up to 24 as a power or 2 or sum of powers of 2, then we know that each of 16 and 9 can be expressed in this way: $16 = 2^4$ and $9 = 2^3 + 2^0$. Therefore 25 can be expressed as a sum of powers of 2.

Let us assume we have been successful in expressing each integer up to, say, k, as a power of 2 or as a sum of powers of 2. If the integer $k + 1$ is a power of 2, then it can be expressed as a power of 2. Otherwise there is some largest integer ℓ so that ℓ is a power of 2 and $\ell < k + 1$. Since each of ℓ and $k + 1 - \ell$ are less than $k + 1$, as we worked our way up to $k + 1$ we must have been able to express each of ℓ and $k + 1 - \ell$ as a power of 2 or as a sum of powers of 2. Adding together these sums gives us $k + 1$ as a sum of powers of 2.

Let $p(n)$ be the following formula over $\mathbb{N}$

n is a power of 2 or can be expressed as a sum of powers of 2.

In our work above we have verified the following information

1. $p(0)$ is true
2. For any $k \in \mathbb{N}$, if each of $p(0), p(1), \ldots, p(k)$ is true, then $p(k + 1)$ is true.

When $k = 0$, (2) tells us that $p(1)$ is true. When $k = 1$, (2) tells us that $p(2)$ is true. Continuing in this manner, we should be convinced that $p(n)$ is true for each $n \in \mathbb{N}$.

To hone our intuition further, let us consider another example.

Recall that a prime number is an integer $n \geq 2$ so that the only positive divisors of n are 1 and itself. The study of prime numbers has some surprisingly important applications in the cryptography – the art of communicating in secret code. For example, most security systems for online commerce depend on some fundamental facts about prime numbers.

Any integer $n \geq 2$ that is not prime is called composite. Necessarily, for any composite integer n there exists a pair of integers a and b in the range $[2, n - 1]$ so that $n = ab$. If each of a and b are composite, then again they can be expressed as the product of two integers. This process will stop when all of the factors we have found are prime.

For example, consider the integer 30. Since 30 even we have

$$30 = 2 \times 15$$

Since $15 = 3 \times 5$ we have

$$30 = 2 \times 3 \times 5$$

As each of 2, 3 and 5 is prime, we can continue this process no further. Notice that we have expressed the integer 30 a product of prime numbers.

For another example consider the integer 6136. This integer is divisible by 4 (this is not obvious) and so we find

$$6136 = 4 \times 1534$$

The integer 4 is composite. As is 1534. And so

$$6136 = 2 \times 2 \times 2 \times 767$$

Continuing this process we can find $767 = 13 \times 59$. And so

$$6136 = 2 \times 2 \times 2 \times 13 \times 59$$

Each of these integers $2, 13$ and 59 is prime. And so we have written 6136 as a product of prime numbers.

It seems reasonable that we could repeat this sort of process with any composite integer $n \geq 2$. In doing so we may be able to save ourselves some time by using factorisations we have already found.

Consider the integer $n = 184080$. With some calculator work we can find

$$184080 = 30 \times 6136$$

We have already expressed each of 30 and 6136 as a product of prime numbers. And so it follows that 184080 can be expressed as a product of prime numbers. Therefore

$$\begin{aligned} 184080 &= (2 \times 3 \times 5) \times (2 \times 2 \times 2 \times 13 \times 59) \\ &= 2^4 \times 3 \times 5 \times 13 \times 59 \end{aligned}$$

Let $p(n)$ be the following formula over $\mathbb{N}$

$n + 2$ *is prime or* $n + 2$ *be expressed as a product of prime numbers*

From our work above it seems reasonable that the statement $(\forall n \in \mathbb{N})\, p(n)$ is true. Certainly $p(0)$ is true; 2 is prime. Let k be an integer. Imagine we have verified that each of $p(0), p(1), \ldots, p(k)$ is true. If the integer $k+3$ is prime, then $p(k+1)$ is true. Otherwise, $k+3$ is composite and so there exists integers a and b in the range $[2, k+2]$ so that $k + 3 = ab$.

Since we have verified that each of $p(0), p(1), \ldots, p(k)$ is true, necessarily $p(a-2)$ and $p(b-2)$ are both true. Therefore each of a and b are either prime or can be expressed as a product of prime numbers. Therefore $k+3$ can be expressed as a product of prime numbers. And so we see $p(k+1)$ is true.

Since $p(0)$ is true, and whenever each of $p(1), p(2), \ldots p(k)$ is true it follows that $p(k+1)$ is true, then we should be convinced that $p(n)$ is true for each $n \in \mathbb{N}$.

Just as we did for the Principle of Mathematical Induction, we state this proof technique as a theorem.

Theorem (The Principle of Strong Mathematical Induction)**.** *Let $p(n)$ be a formula over $\mathbb{N}$. If the following two statements are true, then $p(n)$ is true for each integer $n \geq 0$.*

1. *$p(0)$; and*
2. *for each $k \geq 0$, if each of $p(0), p(1), \ldots, p(k)$ is true, then $p(k+1)$ is true.*

The Principle of Strong Mathematical Induction is a proof technique. If we can show that the hypotheses hold for some particular formula $p(n)$ over $\mathbb{N}$, then applying the theorem above tells us the statement $(\forall n \in \mathbb{N})\, p(n)$ is true.

Let us return to our second example and use the Principle of Strong Mathematical Induction to prove that every integer that is at least two is prime or can be expressed as a product of prime numbers.

Theorem 1.28. *Every integer $n \geq 2$ is prime or can be expressed as a product of prime numbers.*

Proof. We proceed using the Principle of Strong Induction. Let $p(n)$ be the following formula over $\mathbb{N}$

$$n + 2 \textit{ is prime or } n + 2 \textit{ be expressed as a product of prime numbers}$$

1. $p(0)$ is true

The statement $p(0)$ is

$$2 \textit{ is prime or } n + 2 \textit{ be expressed as a product of prime numbers}$$

which is true.

2. if each of $p(0), p(1), \ldots, p(k)$ is true, then $p(k+1)$ is true.

Let $k \geq 0$ be an integer. Assume each of $p(0), p(1), \ldots, p(k)$ is true. That is, assume that for each $k' \in \{0, 1, \ldots k\}$ the following statement is true:

$$k' + 2 \textit{ is prime or } k' + 2. \textit{ be expressed as a product of prime numbers}$$

Consider now $n = k + 1$. If $(k + 1) + 2$ is prime, then the statement

$$(k + 1) + 2 \textit{ is prime or } (k + 1) + 2 \textit{ can be expressed as a product of prime numbers}$$

is true.

Otherwise, if $(k+1)+2$ is not prime, then there exists $a, b \in [2, k]$ such that $ab = (k+1)+2$. Since $a, b \in [2, k + 2]$, each of $p(a - 2)$ and $p(b - 2)$ is true. In other words, there exists not necessarily distinct primes $p_1, p_2, \ldots, p_t$ and $q_1, q_2, \ldots q_\ell$ so that

$$\begin{aligned} a &= p_1, p_2, \cdots p_t \\ b &= q_1, q_2, \cdots q_\ell \end{aligned}$$

Therefore $(k + 1) - 2 = (p_1, p_2 \cdots, p_t)(q_1, q_2 \cdots q_\ell)$ and so $p(k + 1)$ is true.

Since both hypotheses of the Principle of Strong Mathematical Induction hold, necessarily so does the conclusion. Therefore every integer $n \geq 2$ is prime or can be expressed as a product of prime numbers. □

Theorem 1.28 is actually a weaker version of a much stronger (and more interesting) result.

Theorem 1.29 (The Fundamental Theorem of Arithmetic)**.** *Every positive integer $n \geq 2$ is prime or can be uniquely expressed as a product of prime numbers.*

Proving the Principle of Mathematical Induction

Recall the statement of the Principle of Mathematical Induction

Theorem (The Principle of Mathematical Induction)**.** *Let $p(n)$ be a formula over $\mathbb{N}$. If the following two statements are true, then $p(n)$ is true for each integer $n \geq 0$.*

(1) $p(0)$; and

(2) for each $k \geq 0$, if $p(k)$ is true, then $p(k+1)$ is true

As this is a theorem there ought to be a proof that convinces us that it is true. We proceed by contradiction. That is, we will assume that the hypothesis of the theorem is true and the conclusion of the theorem is false. From this we will derive an absurdity (i.e., a contradiction.) The contradiction we will find is a natural number ℓ so that $p(\ell)$ is both true and false. This then convinces us that if the hypothesis is true, then the conclusion must also be true.

In this theorem the hypothesis is:

The following two statements hold

(1) $p(0)$ is true; and

(2) for each $k \geq 0$, if $p(k)$ is true, then $p(k+1)$ is true

and the conclusion is:

$p(n)$ is true for each integer $n \geq 0$.

If the conclusion is false, then there must be at least one value of n for which $p(n)$ is false. Let F be the set of such values. That is,

$$F = \{t \mid p(t) \text{ is false } \}$$

Since the conclusion is false, necessarily $F \neq \{\}$. Of all of the elements of F, let us use the label ℓ to refer to the smallest. Is it possible that $\ell = 0$? (Take a moment to think about this before you go on).

.

.

.

.

.

.

As we assumed our hypothesis to be true, it must be that $p(0)$ is true. Therefore $0 \notin F$ and so $\ell \neq 0$.

To find our contradiction let us consider the number $\ell - 1$. Since $\ell \geq 1$, necessarily $\ell - 1 \geq 0$. What can we say about the truth value of $p(\ell - 1)$? (Take a moment to think about this before you go on. How did we choose ℓ?).

.

.

.

.
.
.

Recall that ℓ is the smallest element of F. That is, it is the smallest natural number so that $p(\ell)$ is false. Since $\ell - 1 < \ell$, necessarily $p(\ell - 1)$ is true.

Let us return now to thinking about the hypothesis. We have assumed true the statement

for each $k \geq 0$, if $p(k)$ is true, then $p(k+1)$ is true.

When $k = \ell - 1$ this statement is:

if $p(\ell - 1)$ is true, then $p(\ell)$ is true.

We have just shown that $p(\ell-1)$ is true. Therefore $p(\ell)$ is true. This statement contradicts the fact that $p(\ell)$ is false. Since $p(\ell)$ cannot be both true and false, then ℓ does not exist! That is to say, there is no smallest natural number in the set F. Therefore $F = \{\}$. Since F is empty, it must be that $p(n)$ is true for every $n \in \mathbb{N}$. This last statement

$p(n)$ is true for every $n \in \mathbb{N}$

is exactly the conclusion of our theorem!

Let us write this all down succinctly as a proof of the Principle of Mathematical Induction.

Proof. We proceed by contradiction. Let $p(n)$ be a formula over $\mathbb{N}$ so that the statement $(\forall n \in \mathbb{N})\, p(n)$ is false, but the following two statements are true:

(1) $p(0)$

(2) for each $k \geq 0$, if $p(k)$ is true, then $p(k+1)$ is true

Let $F = \{t \mid p(t) \text{ is false }\}$. By (1), we have $0 \notin F$. Let ℓ be the smallest element of F. By our choice of ℓ we have $\ell - 1 \notin F$. Therefore $p(\ell - 1)$ is true. By (2) it follows that $p(\ell)$ is true. This is a contradiction as ℓ was chosen so that $p(\ell)$ is false. □

Aside. *Our proof of the Principle of Mathematical Induction turns on the following sentence:*

Let ℓ be the smallest element of F

In our proof, F is a non-empty subset of $\mathbb{N}$. With this sentence we are asserting the existence of some smallest element. How do we know such a smallest element exists?

At the lowest level, abstract mathematics depends on a list of axioms, statements that we assume to be true. Most of the time we don't ever think about the axioms that underpin our work as they are so common-sense that they seem to not require proof.

The `well-ordering principle` *is the axiom that let's us assert the existence of the smallest element of F. Assuming that the well-ordering principle holds allows us to prove the statement of the Principle of Mathematical Induction.*

The respective wikipedia pages on the `Peano Axioms` *and the* `well-ordering principle` *provide more information on these sorts of things should be you interested.*

Test Your Understanding

1. Use a counterexample to show that the following statements is false.

   (a) If the product of two integers is even then both of those integers are even.

   (b) For all real numbers, if $x^2 = y^2$ then $x = y$.

2. Using a proof by induction, prove the following condition is true for all $n \in \mathbb{N}$

$$1 \cdot 2^0 + 2 \cdot 2^1 + 3 \cdot 2^2 + 4 \cdot 2^3 + \cdots + (n+1) \cdot 2^n = 1 + n2^{n+1}$$

3. In talking about expressing each integer $n \geq 2$ as a product of prime numbers, why did we define $p(n)$ to be the formula

   $n + 2$ *is prime or can be expressed as a product of prime numbers*

   instead of

   n *is prime or can be expressed as a product of prime numbers*

Test Your Understanding Solutions

1. Consider $3 \times 4 = 12$ which is even, but 3 is odd, so the statement is false.
2. Put $x = 1$ and $y = -1$. Then $x^2 = y^2$, but $x \neq y$.
3. (a) We proceed by induction. Let $p(n)$ be the condition

   $$\sum_{j=0}^{n}(j+1)2^j = 1 + n2^{n+1}$$

   $p(0)$ is the statement

   $$1 \cdot 2^0 = 1 + (0)2^1$$

   which is true.

   Let $k \in \mathbb{N}$ and assume $p(k)$ holds. Therefore

   $$\sum_{j=0}^{k}(j+1)2^j = 1 + k2^{k+1}$$

   Consider now the expression $\sum_{j=0}^{k+1}(j+1)2^j$. We have

   $$\sum_{j=0}^{k+1}(j+1)2^j = (\sum_{j=0}^{k}(j+1)2^j) + (k+1)(2^k).$$

   Since $p(k)$ is true, we have $\sum_{j=0}^{k}(j+1)2^j = 1 + k2^{k+1}$ Therefore

   $$\sum_{j=0}^{k+1}(j{+}1)2^j = \left(\sum_{j=0}^{k}(j+1)2^j\right){+}(k{+}2)(2^{k+1}) = 1{+}k2^{k+1}{+}(k{+}2)(2^{k+1}) = 1{+}(k{+}1)2^{k+2}$$

   Therefore $p(k)$ is true.

   The result now follows from the Principle of Mathematical Induction.
4. If we choose the latter condition then the domain of the condition is not all of $\mathbb{N}$, a requirement in the hypothesis of the Principle of Strong Mathematical Induction. (We notice, however, that the statement of the Principle of (strong)Mathematical induction can be modified to be of use for conditions whose domain is $n \geq k$ for any $k \in \mathbb{Z}$)

2 Set Theory

In just about every mathematics class you have taken, sets were lurking beneath the surface. For example, to be able to describe a function, we need a way to communicate the collection of values for which the function is defined (i.e., the domain). To talk about a vector space, we need a way to communicate what the collection of objects is. And so, in a conversation about justifying mathematical truths, it makes sense to spend some time thinking about the humble set.

We begin with our common understanding of the meaning of the word set:

Definition 2.1. *A set is a collection of unique objects. Objects contained in a set are called elements. When A is a set and x is an element of A we write $x \in A$.*

Let $A = \{\mathbb{Z}, \mathbb{N}, 0, \mathbb{Q}, \mathbb{R}\}$. Since $\mathbb{Z}$ is an element of A we can write $\mathbb{Z} \in A$.

By definition sets are unordered. What this means is that the order in which we list the elements of a set does not change the set. For example

$$\{\mathbb{Z}, \mathbb{N}, \mathbb{R}, \mathbb{Q}, 0\} = \{\mathbb{Z}, \mathbb{N}, 0, \mathbb{Q}, \mathbb{R}\}$$

Looking again at our definition for set, there is nothing in our definition of set that disallows a set from being an element of another set. For example, consider again the set

$$\{\mathbb{Z}, \mathbb{N}, \mathbb{R}, \mathbb{Q}, 0\}$$

This set has five elements. Four of those elements are sets. The fifth element is the number 0.

Unfortunately this observation, sets can be elements of sets, leads us to discover a shortcoming in our naive definition above. Consider the possibility of a set being an element of itself.

For example, let S be the set of things that are not elephants. The set S has many elements. For example, the following objects are each elements of S: *cat*, *dog*, $\mathbb{N}$, 1, S.

The last object on this list, S itself, perhaps gives us a moment of pause. Certainly the set of things that are is not elephants is itself not an elephant. And so we can conclude S is a member of itself. That is, $S \in S$.

Some sets are elements of themselves. (For example, S above.) Whereas other sets are not elements of themselves. (For example, A above.)

Let P be the set of sets that are not elements of themselves. From our work above, we have $S \notin P$ and $A \in P$

Since P is a set, it seems reasonable to ask if it is an element of itself. Exactly one of the following statements is true:

- $P \in P$
- $P \notin P$

The statement $P \in P$ means that P is an element of P. However, if P is an element of P then P is an element of itself. And thus by the definition of P, we conclude $P \notin P$.

The statement $P \notin P$ means that P is not an element of P. However, if P is not an element of P, then P is not an element of itself. And thus by the definition of P we conclude $P \in P$.

The premise $P \in P$ leads to the conclusion $P \notin P$. However, the premise $P \notin P$ leads to the conclusion $P \in P$. So which is it? $P \in P$ or $P \notin P$? Each option leads to a contradiction.

Yuck!

This seemingly inescapable contradiction is called Russell's Paradox. Russell's Paradox arises when we think of sets as collections of unique objects, without restriction.

As sets underpin just about every mathematical idea, we take some time in Section 2 to think more carefully about our intuition surrounding mathematical sets and common associated ideas.

Before we do so, let us take a moment to think about the integers. To describe the integers as a set we usually write

$$\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$$

However, the integers are much more than just a set. We think of the integers as being in an order. For example, 1 comes before 2 because 1 is smaller than 2. However, sets are, by definition, unordered collections of objects. And so when we think about the integers, we think of them as having more structure than just set; they are ordered.

The integers come with another structure with which we are intimately familiar. In our primary school education we learned how to add integers. However there is nothing in our definition of set that states there must be a way to add elements of a set. And so as well as having structure arising from their order, the integers also have structure arising from addition.

In 2.2 and 2.3 we look at these types of structures for sets and formalize some of our intuition around orders and operations.

Aside. *Russell's Paradox is named for Bertrand Russell, noted 20th century philosopher and mathematician. Though mathematics has been applied to solve myriad problems for thousands of years, it is only the relatively recent past that researchers have considered the problem of ensuring that our mathematical foundations are solid. Russell's Paradox was also contemporaneously discovered by Ernst Zemelo. Both mathematicians suggested solutions to this paradox. Zermelo's contribution, as we will see in the coming section, eventually was adopted.*

2.1 Preliminaries

Recall from Linear Algebra the following definition:

Definition. *A vector space is a set V together with an addition and a scalar multiplication, so that the following properties hold:*

1. *$x + y \in V$ for all $x, y \in V$ (addition is closed)*
2. *$ax \in V$ for all $x \in V$ and $a \in \mathbb{R}$. (scalar multiplication is closed)*
3. *$x + y = y + x$ for all $x, y \in V$ (addition is commutative)*
4. *$(x + y) + z = x + (y + z)$ for all $x, y, z \in V$ (addition is associative)*
5. *there exists an element $0 \in V$ so that $x + 0 = x$ for all $x \in V$ (additive identity)*
6. *for all $x \in V$ there exists x' such that $x + x' = 0$. We denote x' as $-x$. (additive inverses)*
7. *$1x = x$ for all $x \in V$ (multiplicative identity)*
8. *$a(x + y) = ax + by$ and $(a + b)x = ax + bx$ for all $x, y \in V$ and all $a, b \in \mathbb{R}$ (multiplication distributes over addition)*

We call the elements of a vector space vectors or points.

The Vector Space Axioms (i.e., items 1-8) give the framework for working with vector spaces. A collection of objects that satisfies the Vector Space Axioms qualifies to be called a vector space. These axioms tells us which properties one can assume every vector space to have.

Just as a list of axioms specify the properties of a vector space, so too does a list of axioms specify the properties of sets. As sets underpin just about every mathematical concept one can describe, let us take a moment to look at these axioms.

Unfortunately, the Set Axioms are quite technical; much more than we will need to consider in this course. And so we present a simplified version of these axioms. They are enumerated here so that we may talk about a few of them in more detail to see how the facts about sets we already know fit in to their axiomatic construction.

Definition 2.2 (Set Axioms[1])**.**

- *Equality: Two sets are equal when they have the same elements*
- *Specification: If U is a set and $p(x)$ is a condition on U expressed in first-order logic, then the collection of all elements of U for which $p(x)$ is true is a set.*
- *Pairing: If A and B are sets, then there is a set that contains A and B.*
- *Union: If A and B are sets, then there is a set that contains all of the elements of A together with all of the elements of B.*

[1] These axioms are usually referred to as ZF+C. The first eight axioms are usually collectively referred to as the Zermelo–Fraenkel axioms. The first paragraph of the Wikipedia article *Zermelo–Fraenkel set theory* is an excellent place to start if you are looking for more information.

- *Power Set: If A is a set then there is a set that contains all of the subsets of A.*
- *Replacement: The range of a function is a set.*
- *Inductive Set: There exists a set A so that if $x \in A$, then $\{A, \{x\}\}$ is a set.*
- *Foundation: Every non-empty set A contains an element y so that A and y have no common elements.*
- *Axiom of Choice. Given a collection of sets, there is a set that has exactly one element from each set.*

With these informal descriptions of the axioms, it is not clear that we have resolved the problem solved by Russell's Paradox. For reasons that go beyond the scope of this course, one can use the Axiom of Foundation to prevent Russell's Paradox.

Many of axioms here refer to fine-grained details about sets that are not generally important to our regular use of sets as a tool to represent collections of mathematical objects. However, we will examine a few of them more closely and see how they give rise to concepts with which we are likely already familiar.

Axiom of Equality
The axiom of equality gives meaning to the equals sign for sets. If A and B are sets, we should probably all agree on the meaning of the notation $A = B$ before we start using it. In this case, equality means exactly what we expect it to mean.

Definition 2.3. *Let A and B be sets. We say A and B are equal when they have the same elements. In other words, we say A and B are equal when the following statement is true*

$$(\forall x)\ x \in A \Leftrightarrow x \in B$$

When A and B are equal we write $A = B$.

Axiom of Union
The Axiom of Union tells us that it is permitted in the world of set theory to combine the elements of two sets to create a new set. This new set contains no duplicates, even if the two sets have some common elements. We know this operation as *set union.*

Definition 2.4. *Let A and B be sets. The union of A and B is the set containing all elements of A and all elements of B. We denote the union of A and B as $A \cup B$. That is,*

$$A \cup B = \{x \mid (x \in A) \vee (x \in B)\}$$

For example

$$\{\mathbb{R}, \mathbb{Z}, \mathbb{N}\} \cup \{\mathbb{R}, 0, 1\} = \{0, 1, \mathbb{R}, \mathbb{N}, \mathbb{Z}\}$$

Axiom of Specification
The Axiom of Specification tells is that we may use conditions to specify membership

in a set. That is, if U is a set and $p(x)$ is a condition on U, then the collection of all elements of U for which $p(x)$ is true forms a set. We denote such a set as follows:

$$\{x \in U \mid p(x)\}$$

For example, if $U = \{1, 2, 3, 4\}$ is a set, and $p(x) : x < 4$, then

$$\{x \in U \mid p(x)\} = \{1, 2, 3\}$$

This method of specifying a set is usually referred to as *set-builder notation.*

As $p(x)$ is a statement for any particular value of x, we may use connectives when we specify conditions. Continuing our example above, consider the set

$$\{x \in U \mid \sim p(x)\} = \{4\}$$

The set $\{x \in U \mid \sim p(x)\}$ is what remains in U when we remove all elements of the set $\{x \in U \mid p(x)\}$. In other words, the set $\{x \in U \mid \sim p(x)\}$ is the complement of the set $\{x \in U \mid p(x)\}$.

Definition 2.5. *Let U be a set, let $p(x)$ be a condition on U and let $A = \{x \in U \mid p(x)\}$. The complement of A is the set $\{x \in U \mid \sim p(x)\}$. We denote the complement of A as $U \setminus A$.*

In our example above we have $U = \{1, 2, 3, 4\}, A = \{1, 2, 3\}$ and $U \setminus A = \{4\}$.

Continuing with our example, consider now the condition $q(x) : x > 4$ over the domain $U = \{1, 2, 3, 4\}$. In this case there are no elements of U for which $q(x)$ is true. Thus the set $\{x \in U \mid q(x)\}$ contains no elements. That is

$$\{x \in U \mid q(x)\} = \{\}$$

The Axiom of Specification implies the existence of a set that contains no elements.

Definition 2.6. *Let A be a set. When A contains no elements we say A is the empty set. When A is the empty set we write $A = \{\}$.*

Aside. *Some authors use the notation $\emptyset$ to denote the empty set. Feel welcome to use $\emptyset$ in your submitted work. In these notes we will generally avoid its use as when we only the notation $\emptyset$ we may forget we are referring to a set and not some other type of mathematical object.*

Looking at our Axioms of Set Theory, there is no axiom that tells is that the empty set necessarily exists. The existence of the empty set follows as a consequence of the Axiom of Specification.

When we use the Axiom of Specification (i.e, set builder notation) to construct a new set from an existing set, necessarily all of the elements of the new set are also contained in the existing set. In other words, the new set is a *subset* of the existing set. And so we define the following notation and terminology.

Definition 2.7. *Let A and B be sets. We say B is a subset of A when every element of B is also an element of A. In other words, we say B is a subset of A when the following statement is true*

$$(\forall x)\ x \in B \Rightarrow x \in A$$

When B is a subset of A we write $B \subseteq A$.

When B is a subset of A and B and A are not equal, we say B is a proper subset of A and we write $B \subsetneq A$.

Aside. *Some authors use the notation $\subset$ in place of $\subsetneq$. Feel welcome to use either in your submitted work.*

The Axiom of Separation has one last gift for us. Looking at the Axioms of Set Theory, there is no axiom that directly permits us to consider the intersection of a pair of sets. However, as the Axiom of Separation permits us to build new sets using conditions, it is possible to use to Axiom of Separation to define set intersection.

For example, let $U = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Let $A = \{1, 2, 4, 6\}$ and let $B = \{1, 2, 3, 4, 5\}$. Let $p(x) : (x \in A) \wedge (x \in B)$ be a condition over U. Consider the set

$$\{x \in U \mid p(x)\} = \{1, 2, 4\}$$

From our past experience in mathematics, I suspect we agree $A \cap B = \{1, 2, 4\}$ is a true statement. And so let us define set intersection to mean what we expect it to mean:

Definition 2.8. *Let U be a set and let A and B be subsets of U. The intersection of A and B is the subset of U that contains all elements common to A and B. We denote the intersection of A and B as $A \cap B$. In other words*

$$A \cap B = \{x \in U \mid (x \in A) \wedge (x \in B)\}$$

Looking at our work above we defined have three ways to build new sets from old sets: set union, set intersection and set complement. We have also defined two ways to compare sets: set equality and subset.

Let A and B be sets. The notation $A \cup B$ refers to a set. It is the set that contains all of the elements of A together with all of the elements of B. The notation $A \cup B$ is not a statement; it is neither true nor false. On the other hand, the notation $A \subseteq B$ is a statement; it is either true or false.

Using our new notation and our techniques from Section 1.3, we can prove those facts about sets with which we are likely already familiar. For example, you may be familiar with the following statement

$$A = B \text{ if and only if } A \subseteq B \text{ and } B \subseteq A.$$

We state this as a theorem and consider how we may prove this theorem:

Theorem 2.9. *Let A and B be sets. We have $A = B$ if and only if $A \subseteq B$ and $B \subseteq A$.*

Combining our work examining connectives from Section 1, and our work definiing notions of sets from above, we can now make complete meaning of this statement. Expressed in first-order logic we can re-write this statement as:

$$A = B \Leftrightarrow A \subseteq B \wedge B \subseteq A$$

Let $p : A = B$ and $q : A \subseteq B \wedge B \subseteq A$. We want to prove $p \Leftrightarrow q$ is true. Recall the following logical equivalence

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

And so to prove $p \Leftrightarrow q$ is true, it suffices to prove $p \Rightarrow q$ is true and $q \Rightarrow p$ is true. Translating back to English prose, we want to prove the following two statements

If $A = B$, then $A \subseteq B$ and $B \subseteq A$.

If $A \subseteq B$ and $B \subseteq A$, then $A = B$.

We provide both a formal and an informal proof for the first of these two statements by appealing to the definitions of $=$ and $\subseteq$. We leave the second half as an exercise.

Proof. Let A and B be sets so that $A = B$. Since $A = B$, for all elements x, we have $x \in A$ if and only if $x \in B$.

Let x be an element of A. Since $x \in A$, and $x \in A$ if and only if $x \in B$, it then follows that $x \in B$. Therefore every element of A is an element of B. In other words $A \subseteq B$.

Let x be an element of B. Since $x \in B$ and $x \in B$ if and only if $x \in A$, it then follows that $x \in A$. Therefore every element of B is an element of A. In other words, $B \subseteq A$.

By the previous two arguments, it follows that if $A = B$, then $A \subseteq B$ and $B \subseteq A$. □

Proof.

1. $A = B$ (premise)
2. $(\forall x)\ x \in A \Leftrightarrow x \in B$ (defn of $=$)
3. $x \in A \Rightarrow x \in B$ (2)
4. $A \subseteq B$ (3, defn of $\subseteq$)
5. $x \in B \Rightarrow x \in A$ (2)
6. $B \subseteq A$ (5, defn of $\subseteq$)
7. $(A \subseteq B) \wedge (B \subseteq A)$ (4,6,defn of $\wedge$)
8. $A = B \Rightarrow [(A \subseteq B) \wedge (B \subseteq A)]$ (1,7)

□

Exercise 2.10. *Give an informal proof of the following statement:*

$$[(A \subseteq B) \wedge (B \subseteq A)] \Rightarrow A = B$$

Before we leave our work on the Axioms of Set Theory, let us take a moment to consider one last method of constructing a new set from an old set. Quite likely, we are all familiar with the following notation

$$\mathbb{R}^2 = \{(x, y) \mid x \in \mathbb{R} \text{ and } y \in \mathbb{R}\}$$

And asked to write down an analogous definition for the notation $\mathbb{R}^3$ we would write:

$$\mathbb{R}^3 = \{(x, y, z) \mid x \in \mathbb{R}, y \in \mathbb{R} \text{ and } z \in \mathbb{R}\}$$

These sets, $\mathbb{R}^2$ and $\mathbb{R}^3$, are constructed from $\mathbb{R}$ but don't seem to arise by the use of set union, intersection or complement. Certainly if we want to be able to think about functions $f : \mathbb{R} \to \mathbb{R}$ we require the notion of *ordered pair*. But, as yet, our Axioms of Set Theory have no given us a method to construct ordered pairs.

It turns out that ordered pairs are actually sets in disguise. The following definition is provided for those curious about how to construct an ordered pair using only sets. (i.e., you are not expected to spend anytime making sense of this definition)

Definition. *Let A and B be sets. For $a \in A$ and $b \in B$ we define the notation (a, b) to be the set:*

$$(a, b) = \{a, \{a, b\}\}$$

We call (a, b) an ordered pair.

For example the ordered pair $(1, 2)$ refers to the set $\{1, \{1, 2\}\}$.

With this definition in place (or by just accepting that the notation (a, b) means what we expect it to) we can define the meaning of the notation $\mathbb{R}^2$.

Definition 2.11. *Let A and B be sets. The Cartesian product of A and B is the set of all ordered pairs whose first entry is in A and whose second entry is in B. We denote the Cartesian product of A and B as $A \times B$. That is*

$$A \times B = \{(a, b) \mid a \in A \text{ and } b \in B\}$$

For example, we have

$$\{1, 6, dog\} \times \{6, cat\} = \{(1, 6), (1, cat), (6, 6), (6, cat), (dog, 6), (dog, cat)\}$$

Aside. *The Cartesian Product is named for René Descartes, who is considered to have been the first to connect algebraic techniques with geometric concepts. It is a wonderful coincidence that the drawing we make of $\mathbb{R}^2$ (i.e., a pair of perpendicular lines) is named the Cartesian plane and that* Cartography *is the study of map making.*

The notation $\mathbb{R}^2$ is shorthand for the notation $\mathbb{R} \times \mathbb{R}$. Which, in turn, is shorthand for the set

$$\{(x, y) \mid x, y \in \mathbb{R}\}$$

One can analogously define the Cartesian product of three sets as:

$$A \times B \times C = \{(a,b,c) \mid a \in A, b \in B \text{ and } c \in C\}$$

With this in mind, the notation $\mathbb{R}^3$ is shorthand for the notation $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$. Which, in turn, is shorthand for the set

$$\{(x,y,z) \mid x,y,z \in \mathbb{R}\}$$

Similarly, the notation $\mathbb{R}^n$ is shorthand for the notation $\underbrace{\mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}}_{n \text{ times}}$. Which, in turn, is shorthand for the set

$$\{(x_1, x_2, \ldots, x_n) \mid x_1, x_2, \ldots, x_n \in \mathbb{R}\}$$

Test Your Understanding

1. Let $A = \{1,2,3,4\}$, $B = \{1,3,5,7\}$ and $C = \{2,3\}$.

   (a) Find the following sets:

   i. $A \cup B$

   ii. $A \cap B$

   iii. $A \cup C$

   iv. $A \cap C$

   v. $B \times C$

   (b) For each piece of notation below, determine if it is true or false or neither/ambiguous

   i. $3 \subseteq B$

   ii. $\{\} \subseteq A$

   iii. $7 \in B \cup C$

   iv. $C \subseteq B$

   v. $C = B \cup C$

   vi. $C \in A$

   vii. $(\forall a \in A)\ a \leq 4$

   viii. $(\forall b \in B)[(\exists c \in C)\ b - c \geq 0]$

2. Let A, B and C be sets. Using the definition of subset, prove that if $A \subseteq B$ and $B \subseteq C$, then $A \subseteq C$.

Test Your Understanding - Answers

1. Let $A = \{1, 2, 3, 4\}$, $B = \{1, 3, 5, 7\}$ and $C = \{2, 3\}$.

   (a) Find the following sets:

      i. $A \cup B = \{1, 2, 3, 4, 5, 7\}$

      ii. $A \cap B = \{1, 3\}$

      iii. $A \cup C = A$

      iv. $A \cap C = C$

      v. $B \times C = \{(1, 2), (1, 3), (3, 2), (3, 3), (5, 2), (5, 3), (7, 2), (7, 3)\}$

   (b) i. False: Here, 3 is an element of B, not a subset of B.

      ii. True: the empty set is a subset of all sets. $x \in X \Rightarrow x \in Y$ is true when $x \notin X$.

      iii. True

      iv. False: 2 is in C, but not in B.

      v. False: B contains elements that aren't in C.

      vi. False

      vii. True

      viii. False: The number 1 is in B, and there is no number in C you can subtract from it that gives a result that is non-negative.

2. *Proof.* Let A, B and C be sets so that $A \subseteq B$ and $B \subseteq C$. To show $A \subseteq C$ we must show that every element of A is an element of C. Let x be an element of A. Since $A \subseteq B$, x is an element of B. Since $B \subseteq C$ and x is an element of B, then x is an element of C. Therefore every element of A is an element of C. In other words, $A \subseteq C$. □

2.2 Ordered and Bounded Sets

Take a moment to draw a picture of the set integers.

Quite likely (though not certainly) you drew for yourself a picture of the number line marked with the integers. You put 0 in the middle. To the right you put the positive integers, proceeding as $1, 2, 3, \ldots$, and to the left you put the negative integers, proceeding as $-1, -2, -3, \ldots$.

Alternatively you may have written the following set

$$\mathbb{Z} = \{\ldots, 3, -2, -1, 0, 1, 2, 3, \ldots\}$$

In either case you thought about the integers as being in some sort of order. However, looking above at our previous material on sets, there is no mention of order in sight. In fact, quite explicitly, a set is unordered.

$$\{1, 2, 7\} = \{7, 1, 2\}$$

And so to be able to mathematically describe the integers as being a set with a particular ordering, we must agree on what it means for a set to have an order.

Certainly if we have placed a collection of objects in some sort of order, then for any pair of objects we can decide which object comes first. Moreover, an ordering of objects must be consistent: if a comes before b and b comes before c, then necessarily a comes before c. And so let us define the meaning of ordering to be exactly these two properties.

Definition 2.12. *Let S be a set. An order on S is a relation, denoted $<$, so that the following two statements are true:*

(O1) for $x, y \in S$ exactly one of the following is true:

- $x < y$
- $y < x$
- $x = y$

(O2) for $x, y, z \in S$, if $x < y$ and $y < z$, then $x < z$.

Additionally, let $x > y$ denote the statement $y < x$ and let $x \leq y$ denote the statement $(x < y) \wedge (x = y)$.

Though we may not realize it, we are familiar with orders in more contexts than just numbers. For example, consider the ordering of words in a dictionary. It is not unreasonable to write

$$\textit{catastrophic} < \textit{mouse}$$

Aside. *It is reasonable to find Definition 2.12 a bit dissatisfying. This definition states that an order is a* relation. *However, we have not agreed upon (i.e., defined) the meaning of the word relation. Let us momentarily ignore this sense of dissatisfaction. In Section 4 of the course we will return to this thought and define carefully the meaning of the word relation.*

The integers (as well as the rational numbers and the real numbers) have an ordering we are well-familiar with.

Of course not every set we are familiar with comes with a natural sense of ordering. For example, there is no natural ordering in $\mathbb{C}$. Nor is there one in $\mathbb{R}^3$. (What should it mean for one vector to be less than another? I have no idea!)

From our sense of order on $\mathbb{R}$, lots of mathematics follows. An ordering on a set permits us to consistently compare things to one another. For example, one may write:

$$\text{The function } f(x) = \tfrac{1}{x} \text{ is bounded below by } 0 \text{ on the domain } x > 0$$

From our work in calculus we take this sentence to mean

$$\text{For every } x > 0, \text{ we have } \tfrac{1}{x} > 0$$

Similarly, one may write

$$\text{The function } f(x) = \tfrac{1}{x} \text{ is not bounded above on the domain } x > 0$$

We take this sentence to mean that no matter what upper bound we might consider, the function $f(x) = \frac{1}{x}$ exceeds that bound for some value of x. This is a doubly quantified statement in disguise:

- no matter what upper bound we might consider: $(\forall k \in \mathbb{R})$
- the function $f(x) = \frac{1}{x}$ exceeds that bound for some value of x: $(\exists x > 0)\ \ 1/x > k$

As our work in Real Analysis will (eventually) focus on the behaviour of functions, let us define these terms a little more carefully.

Definition 2.13. *Let S be an ordered set and let A be a subset of S. We say A is bounded below in S when there exists $\beta \in S$ so that the following statement is true*

$$(\forall x \in A)\ \ \beta \leq x$$

We say β is a lower bound for A in S.

Definition 2.14. *Let S be an ordered set and let A be a subset of S. We say A is bounded above in S when there exists $\beta \in S$ so that the following statement is true*

$$(\forall x \in A)\ \ \beta \geq x$$

We say β is a upper bound for A in S

Returning to our example above, let $A = \left\{ \frac{1}{r} \mid r \in \mathbb{R}_{>0} \right\}$ and let $S = \mathbb{R}$.

The value $\beta = 0$ is a lower bound for the set A in $\mathbb{R}$. The following statement is true:

$$(\forall x \in A)\ \ 0 \leq x$$

However, it is not the only lower bound. For example, the value $\beta = -10$ is a lower bound for the set A in $\mathbb{R}$. The following statement is true:

$$(\forall x \in A)\ \ -10 \leq x$$

In parlance from calculus, we use the words "tight" and "loose" to refer to how good a bound is. In Real Analysis, we use the terms *infimum* and *supremum* to respectively refer to the tightest possible lower and upper bounds.

Definition 2.15. *Let S be an ordered set, let A be a subset of S, let U_A be the set of upper bounds of A in S, and let $\alpha \in U_A$. We say α is a least upper bound of A in S when for all upper bounds $\beta \in U_A$ we have $\alpha \leq \beta$. In other words, α is a least upper bound of A in S when the following statement is true*

$$(\forall \beta \in U_A)\ \ \alpha \leq \beta$$

When α is the least upper bound of A in S we say α is the supremum of A in S and we write $\sup A = \alpha$.

Definition 2.16. *Let S be an ordered set, let A be a subset of S, let L_A be the set of lower bounds of A in S, and let $\alpha \in L_A$. We say α is a greatest lower bound of A in S when for all lower bounds $\beta \in L_A$ we have $\alpha \geq \beta$. In other words, α is a greatest lower bound of A in S when the following statement is true*

$$(\forall \beta \in L_A)\ \ \alpha \geq \beta$$

When α is the greatest lower bound of A in S we say α is the infimum of A in S and we write $\inf A = \alpha$.

Continuing with our example above, the value $\alpha = 0$ is the infimum for the set $\left\{ \frac{1}{r} \mid r \in \mathbb{R}_{>0} \right\}$ in $\mathbb{R}$.

On the other hand, the set $A = \left\{ \frac{1}{r} \mid r \in \mathbb{R}_{>0} \right\}$ does not have a supremum in $\mathbb{R}$. Let us take a moment to convince ourselves of this fact.

Let us proceed by contradiction. That is, assume that $\left\{ \frac{1}{r} \mid r \in \mathbb{R}_{>0} \right\}$ has a supremum in $\mathbb{R}$. Let k be this supremum. Since k is the supremum of $\left\{ \frac{1}{r} \mid r \in \mathbb{R}_{>0} \right\}$ in $\mathbb{R}$, necessarily k is an upper bound of $\left\{ \frac{1}{r} \mid r \in \mathbb{R}_{>0} \right\}$ in $\mathbb{R}$. (This is part of the definition of supremum)

However for $x > 0$, when $x < \frac{1}{k}$, necessarily $\frac{1}{x} > k$. Therefore there exists a value in $\left\{ \frac{1}{r} \mid r \in \mathbb{R}_{>0} \right\}$ that is greater than k. In other words, k is not an upper bound for $\left\{ \frac{1}{r} \mid r \in \mathbb{R}_{>0} \right\}$ in $\mathbb{R}$. Since k is not an upper bound for $\left\{ \frac{1}{r} \mid r \in \mathbb{R}_{>0} \right\}$ in $\mathbb{R}$ it cannot be that k is the supremum of $\left\{ \frac{1}{r} \mid r \in \mathbb{R}_{>0} \right\}$ in $\mathbb{R}$.

Let us hone our comfort with these definitions with a few more examples.

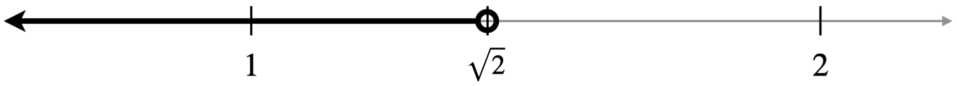
Example 2.17. *Consider* $\{1, 2, 3, 4, 10\} \subseteq \mathbb{N}$. *In our definitions above we have* $S = \mathbb{N}$ *and* $A = \{1, 2, 3, 4, 10\}$.

$\beta = 11$ *is an upper bound for* A. *We know this because* $\beta = 11$ *satisfies all parts of the definition for upper bound:*

- $11 \in \mathbb{N}$ *(i.e,* $\beta \in S$*) is true*
- $(\forall x \in \{1, 2, 3, 4, 10\})\ \ x \leq 11$ *is true.*

However, $\beta = 11$ *is not the least upper bound of* A. *Though* $\beta = 11$ *is an upper bound, it is not the smallest of all upper bounds. The value* $\alpha = 10$ *is the least upper bound. It is an upper bound for* A *in* $\mathbb{N}$ *and no upper bound of* A *in* $\mathbb{N}$ *is less than* $\alpha = 10$. *Therefore* $\sup\{1, 2, 3, 4, 10\} = 10$.

Example 2.18. *Let* $A = \{x \in \mathbb{R} \mid x^2 < 2\} \cup \{x \mid x < 0\} \subseteq \mathbb{R}$.



$\beta = 2$ *is an upper bound for* A *in* $\mathbb{R}$. *We know this because* $\beta = 2$ *satisfies all parts of the definition for upper bound:*

- $2 \in \mathbb{R}$ *(i.e,* $\beta \in S$*) is true*
- $(\forall x \in \{x \in \mathbb{R} \mid x^2 < 2\} \cup \{x \mid x < 0\})\ \ x \leq 2$ *is true.*

However, $\beta = 2$ *is not the least upper bound of* A *in* $\mathbb{R}$. *We can observe that* $\alpha = \sqrt{2}$ *is the least upper bound of* A *in* $\mathbb{R}$. *If there is a smaller upper bound, say* γ*, then necessarily* $\gamma < \sqrt{2}$*, which then implies* $\gamma \in A$. *However,*

$$\gamma < \frac{\gamma + \sqrt{2}}{2} < \sqrt{2}$$

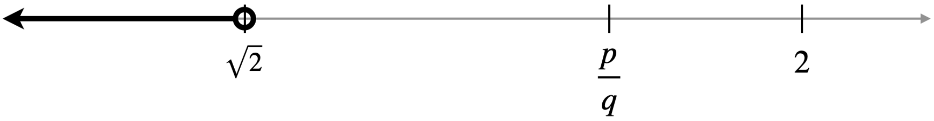
Therefore γ *is not an upper bound of* A *in* $\mathbb{R}$. *Therefore* $\sqrt{2}$ *is the least upper bound of* A *in* $\mathbb{R}$. *In other words,* $\sup A = \sqrt{2}$.

Example 2.19. *Let* $A = \{x \in \mathbb{Q} \mid x^2 < 2\} \cup \{x \mid x < 0\} \subseteq \mathbb{Q}$.

This seems to be the same as the previous example, however here we have swapped $\mathbb{R}$ *for* $\mathbb{Q}$. *As before,* $\beta = 2$ *is an upper bound for* A *in* $\mathbb{Q}$. *However, proceeding as before, we notice something different.*

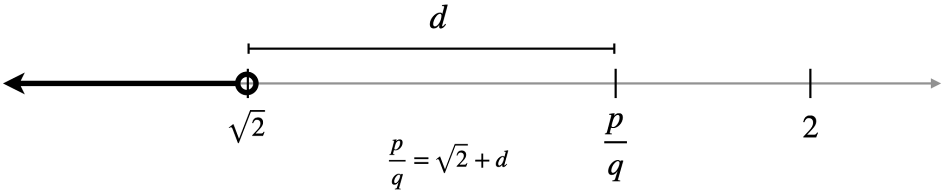
Considering A *as a subset of* $\mathbb{Q}$*, we can be no longer consider* $\sqrt{2}$ *as a possible upper bound for* A *in* $\mathbb{Q}$. *This is because we have* $\sqrt{2} \notin \mathbb{Q}$. *And so we wonder, does* A *have a least upper bound in* $\mathbb{Q}$*?*

As we did in our study of $\frac{1}{x}$ *above, let us consider a potential supremum for* A *in* $\mathbb{Q}$*, say* $\frac{p}{q}$



To show $\frac{p}{q}$ is not the supremum of A in $\mathbb{Q}$ it suffices to find an element of $\mathbb{Q}$ in the region between $\frac{p}{q}$ and $\sqrt{2}$.

Let $d = \frac{p}{q} - \sqrt{2}$. Since $\frac{p}{q} > \sqrt{2}$, necessarily $d > 0$. By Theorem 1.25[2]*, necessarily d is irrational.*



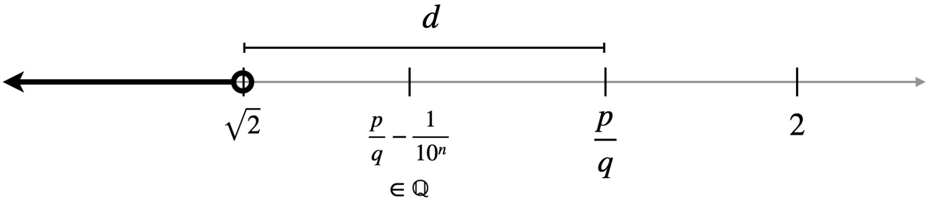
Certainly[3] *since $d > 0$, there exists n so that $d > \frac{1}{10^n}$.*

0
$\frac{1}{10^n}$
d

And so, the following inequalities hold:

$$\sqrt{2} < \frac{p}{q} - \frac{1}{10^n} < \frac{p}{q}$$

Since $\frac{p}{q} - \frac{1}{10^n} \in \mathbb{Q}$, it follows that $\frac{p}{q}$ is not the least upper bound of A in $\mathbb{Q}$.



From this argument, we conclude that $\frac{p}{q}$ is not the least upper bound of A in $\mathbb{Q}$. And so we conclude that A has no supremum in $\mathbb{Q}$.

[2] psst... you did copy down Theorem 1.25 in your own notes, right? Go find it.

[3] Are we certain of this? We will confirm this fact in Section 3.

In this latter example, our intuition is pushing towards the following statement

There is at least one rational number between every rational number and every irrational number.

We return to this statement in Section 3.

The $\sqrt{2}$ examples above are strange. The set $\{x \in \mathbb{Q} \mid x^2 < 2\} \cup \{x \mid x < 0\}$ does not have a least upper bound in $\mathbb{Q}$, but it has one in $\mathbb{R}$. Surprisingly it turns out that every non-empty subset of $\mathbb{R}$ that has an upper bound has a least upper bound.

Theorem. *Every non-empty subset of* $\mathbb{R}$ *that is bounded above has a least upper bound in* $\mathbb{R}$.

From our work above, we see that $\mathbb{Q}$ does not have this property. In Section 3 we examine this least upper bound property of $\mathbb{R}$ and return to the proof of this theorem.

Our work in this section has been mostly scattershot. Rather than consider a systematic way to establish suprema and infima, we have mostly proceeded with our intuition. Looking back at our definitions, we can perhaps see how to turn our scattershot approach into a formal method.

Let S be an ordered set, let $A \subseteq S$, and let $k \in S$. Let U_A be the set of all upper bounds of A in S. To prove $\sup A = k$, we must prove two things:

1. $(\forall x \in A)\ \ x \leq k$ (k is a upper bound for A)
2. $(\forall \beta \in U_A)\ \ k \leq \beta$ (k is not bigger than any upper bound of A.)

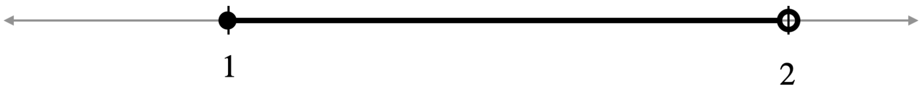
Let L_A be the set of all lower bounds of A in S. To prove $\inf A = k$, we must prove two things:

1. $(\forall x \in A)\ \ k \leq x$ (k is a lower bound for A)
2. $(\forall \beta \in L_A)\ \ \beta \leq k$ (k is not smaller than any lower bound of A)

Aside. *As you take your own notes from these notes, these criteria are an excellent example of the kind of thing you should have written in your own notes*

Let us use this method to consider the infimum and supremum of the following subset of $\mathbb{R}$

$$A = \{x \in \mathbb{R} \mid 1 \leq x < 2\}$$



We begin with the infimum. Looking at our picture, it seems as if $k = 1$ is a good candidate for the infimum of A in $\mathbb{R}$.

(1) $(\forall x \in A)\ \ 1 \leq x$:

i. Consider $x \in A$ (premise)

ii. $1 \leq x$. (defn of A.)

(2) $(\forall \beta \in L_A)\ \beta \leq 1$:

i. Consider $\beta \in L_A$. (premise)

ii. $1 \in A$ (defn of A.)

iii. $\beta \leq 1$ (defn of lower bound, ii)

By (1) and (2) and the definition of infimum, it follows that $\inf A = 1$.

Consider now the supremum. Looking at our picture, it seems as if $k = 2$ is a good candidate for the supremum of A in $\mathbb{R}$. We proceed in a similar manner as we did when we established the infimum.

(1) $(\forall x \in A)\ x \leq 2$

i. Consider $x \in A$ (premise)

ii. $x < 2$. (defn of A.)

iii. $x \leq 2$ (ii)

(2) $(\forall \beta \in U_A)\ 2 \leq \beta$

i. Consider $\beta \in U_A$. (premise)

ii. $2 \in A$

iii. ... wait. No. That's not right. In this interval the end point is not included... We can't just proceed as we did in establishing the infimum.

Let's try again, this time proceeding by contradiction.

If 2 is not the supremum, then there exists a smaller upper bound β. Therefore $\beta < 2$. But if $\beta < 2$, then perhaps we can find an element of A that is greater than β, contradicting that β is an upper bound. For instance, halfway between β and 2 we can find the real number $\beta + \frac{2-\beta}{2}$

$$\beta < \beta + \frac{2-\beta}{2} < 2$$

(2) $(\forall \beta \in U_A)\ 2 \leq \beta$

i. Consider $\beta \in U_A$. (premise)

ii. $\beta < 2$. (premise)

iii. $2 - \beta > 0$. (ii, algebra)

iv. $0 < \frac{2-\beta}{2} < 2 - \beta$ (iii, algebra)

v. $\beta < \beta + \frac{2-\beta}{2} < 2$ (iv, algebra)

vi. $\beta + \frac{2-\beta}{2} \in A$ (v, defn of A)

vii. $\beta \notin U_A$ (vi, defn of upper bound)

viii. Contradiction. (i, vii)

By (1) and (2) and the definition of supremum, it follows that $\sup A = 2$.

In (2) we are being slightly clever in considering the interval $\beta < \beta + \frac{2-\beta}{2} < 2$. We are showing that β appears before 2 in our ordering by finding an element of our ordering that is greater than β but less than 2. This element, $\beta + \frac{2-\beta}{2}$, together with property (O2) of our definition of order, ensures $\beta < 2$.

Just as there is no one-size-fits-all method of *doing proofs*, there is no one-size-fits-all method of establishing suprema and infima. Certainly there are frameworks we can work within, and patterns we will begin to notice. But in both cases, we require intuition, formal understanding, and, at times, a dash of creative insight.

In the upcoming section we will spend some time looking carefully at the properties of the real numbers. One of the properties we will study carefully is the existence of a supremum for any non-empty subset of the real numbers that is bounded above. Such a supremum is guaranteed by the *Completeness Axiom*:

every subset of $\mathbb{R}$ that is bounded above in $\mathbb{R}$ has a least upper bound in $\mathbb{R}$.

In establishing 2 as the supremum of the set $A = \{x \in \mathbb{R} \mid 1 \leq x < 2\}$ we proceeded by contradiction and imagined some smaller element of $\mathbb{R}$ was the supremum. We can proceed this way in general to establish a value as the supremum of a set of real numbers.

Let $B \subset \mathbb{R}$ be bounded above and non-empty. Let γ be the supremum of B. Since γ is the supremum of B, then for every $\epsilon > 0$, $B - \epsilon$ is not an upper bound for the set. Therefore there must be some $b \in B$ so that $b > B - \epsilon$.

Surprisingly the converse of this argument holds, and we arrive at the following theorem.

Theorem 2.20. *Let $B \subseteq \mathbb{R}$ be bounded above and non-empty. Let $\gamma \in \mathbb{R}$ be an upper bound of B in $\mathbb{R}$. We have $\sup B = \gamma$ if and only if for every $\epsilon \in \mathbb{R}$ with $\epsilon > 0$ there is an element of A greater than $\gamma - \epsilon$.*

For now, this theorem is not much use or interest to us. However this theorem will play an unexpectedly important role in our future work in the subject. We leave the proof of this theorem as an exercise in Test Your Understanding.

Test Your Understanding

1. Find the supremum and infinum of the set A (where $A \subseteq \mathbb{R}$), if they exist, and if they do, state whether the supremum or infimum is an element of A. You do not need to prove your answer is correct. To get yourself started, it may help to draw a picture of the set on the real line.

   (a) $A = \{x \in \mathbb{R} \mid x^2 \leq 9\}$

   (b) $A = \{x \in \mathbb{R} \mid \ |x - 2| < 3\}$

   (c) $A = \{x \in \mathbb{R} \mid \ |2x + 1| < 5\}$

   (d) $A = \{x \in \mathbb{Q} \mid x^2 \leq 7\}$

2. Let $A \subseteq \mathbb{R}$ so that A is bounded above. Let $x \in \mathbb{R}$. And let $c + A$ denote the following set
$$c + A = \{c + x \mid x \in A\}$$
Prove $c + A$ is bounded above.

3. Fill in the blanks in the proof of Theorem 2.20.

   Proof of Theorem 2.20. Let $A \subseteq \mathbb{R}$ be bounded above and let $\gamma \in \mathbb{R}$ be an upper bound of A in $\mathbb{R}$. By the Completeness Axiom, A has a supremum.

   Assume $\sup A = \gamma$. Since γ is the supremum of A, we have $\gamma \geq$ ________ for every $\beta \in A$. We proceed by contradiction. Assume there exists $\epsilon > 0$ so that for every $\beta \in A$ we have $\beta \leq \gamma - \epsilon$. Therefore $\gamma - \epsilon$ is an ____________ of A in $\mathbb{R}$. Notice $\gamma - \epsilon <$ _______. This contradicts that γ is the ____________________ of A in $\mathbb{R}$. Therefore if $\sup A = \gamma$, then no element of A is greater than γ and for every $\epsilon > 0$ there is an element of A greater than $\gamma - \epsilon$.

   Assume for every $\epsilon > 0$ there is an element of A greater than $\gamma - \epsilon$. We proceed by contradiction. Assume $\sup A \neq \gamma$. Since γ is an upper bound of A in $\mathbb{R}$ and $\sup A$ is the least such upper bound, we have $\sup A < \gamma$. Let $\delta = \gamma - \sup A$ and $\epsilon = \delta/2$. By hypothesis, there exists $r \in A$ so that $r > \gamma -$ __________. Recall $\delta = \gamma - \sup A$. Therefore
$$r > \gamma - \delta/2 = \sup A + \underline{\qquad\quad} > \sup A.$$
   Since $r \in A$ and $r > \sup A$, then $\sup A$ is not ____, a contradiction. Therefore if for every $\epsilon > 0$ there is an element of A greater than $\gamma - \epsilon$, then $\sup A = \gamma$. □

4. Using Theorem 2.20, prove 1 is the supremum of the set $\{x \in \mathbb{Q} \,|x < 1\}$.

5. Let $q \in \mathbb{Q}$. Prove q is the supremum of the set $\{x \in \mathbb{Q} \,|x < q\}$.

Test Your Understanding - Answers

1. (a) $\sup A = 3$ and $\inf A = -3$. Both are in A.

   (b) $\sup A = 5$ and $\inf A = -1$. Neither is in A.

   (c) $\sup A = 2$ and $\inf A = -3$. Neither is in A.

   (d) $\sup S = \sqrt{7}$ and $\inf S = -\sqrt{7}$. Neither is in A.

2. *Proof.* Assume A is bounded above. Therefore there exists $\beta \in \mathbb{R}$ so that for every $x \in A$ we have $x \leq \beta$. Therefore $x + c \leq \beta + c$ for every $x \in A$. Therefore $\beta + c$ is an upper bound for $c + A$. Therefore $c + A$ is bounded above. □

2.3 Sets and Operations

Much like the sets we are familiar with sometimes come with an ordering, so too do some sets have some extra structure: operations.

Rather than reach back to numbers to contextualise let us instead consider matrices.

Let M be the set of 2×2 matrices with the following form:

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$$

In other words, let

$$M = \left\{ \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \mid a, b \in \mathbb{R} \right\}$$

Our experience in linear algebra tells us how to multiply and add elements of M.

$$\begin{bmatrix} a_1 & -b_1 \\ b_1 & a_1 \end{bmatrix} + \begin{bmatrix} a_2 & -b_2 \\ b_2 & a_2 \end{bmatrix} = \begin{bmatrix} a_1 + a_2 & -(b_1 + b_2) \\ b_1 + b_2 & a_1 + a_2 \end{bmatrix}$$

$$\begin{bmatrix} a_1 & -b_1 \\ b_1 & a_1 \end{bmatrix} \begin{bmatrix} a_2 & -b_2 \\ b_2 & a_2 \end{bmatrix} = \begin{bmatrix} a_1a_2 - b_1b_2 & -(b_1a_2 + b_1b_2) \\ b_1a_2 + b_1b_2 & a_1a_2 - b_1b_2 \end{bmatrix}$$

We notice that when we perform such operations, the resulting matrix is again an element of M. That is to say, the following statements are true.

- $(\forall x, y \in M)\ \ x + y \in M$
- $(\forall x, y \in M)\ \ xy \in M$

Aside. *It feels uncomfortable for us to use lower case letters to refer to a matrix. But, as we will see, it is for a worthy cause.*

Both addition and multiplication are commutative on M. That is to say, the following statements are true.

- $(\forall x, y \in M)\ \ x + y = y + x$
- $(\forall x, y \in M)\ \ xy = yx$

Addition and multiplication on M also has some other familiar properties: distributivity and associativity. That is to say, the following statements are true.

- $(\forall x, y, z \in M)\ \ x(y + z) = xy + xz$
- $(\forall x, y, z \in M)\ \ (x + y) + z = (x + y) + x$
- $(\forall x, y, z \in M)\ \ (xy)z = x(yz)$

Looking back at these seven statements, we can notice that these sorts of statements are true for other sets of objects that we can add and multiply. For example, if we replace M with $\mathbb{Q}$ in each of these seven statements, then the statements remain true. The same holds if we replace M with $\mathbb{R}$ or $\mathbb{C}$.

The set $\mathbb{R}$ (as well as $\mathbb{Q}$ and $\mathbb{C}$) also have some features with respect to addition and multiplication for which we can find analogous notions in M. For examples, if we let 0 denote the all zeroes matrix and we let 1 denote the identity matrix, then the following statements are true:

- $(\forall x \in M)\ \ x + 0 = x$
- $(\forall x \in M)\ \ 1x = x$

If we think about manipulating expressions of elements of M, the properties listed above make the process feel very similar to working with real numbers. Consider the following algebraic manipulations.

$$\begin{aligned} x + xy - 1 &= 0 \\ x(1+y) &= 1 \\ x &= 1(1+y)^{-1} \end{aligned}$$

Without context, one likely assumes that these manipulations are of real-valued variables. However, if we let 0 denote the 3×3 all zero matrix, 1 denote the 3×3 identity matrix, then all of a sudden these can be considered as manipulations of variables taking values in M.

For every matrix $A \in M$, there is a matrix $B \in M$ so that $A + B = 0$. We generally denote this matrix B as $-A$. Similarly for every matrix $A \in M$ with $A \neq 0$, there is a matrix $B \in M$ so that $AB = 1$. We generally denote the matrix B as A^{-1}. In other words, the following statements are true:

- $(\forall x \in M)[(\exists y \in M)\ \ x + y = 0]$
- $(\forall x \in M \setminus \{0\})[(\exists y \in M \setminus \{0\})\ \ xy = 1]$

As in our statements above, replacing M with $\mathbb{R}$ or $\mathbb{Q}$ or $\mathbb{C}$ again yields a true statement.

In the latter half of our Linear Algebra course we took as a pre-requisite for MAST20026 we studied *Vector Spaces*. We noticed that there are other collections of mathematical objects that behave the same as vectors with respect to addition and scalar multiplication. So as to save ourselves time and work, instead of proving facts about $\mathbb{R}^n$ we could instead prove facts about general vector spaces. Any statement that is true for an arbitrary vector space is true for any particular example of a vector space. For example, by proving that every finite-dimensional vector space has a basis we are saved from having to prove that any particular example of a finite dimensional vector space has a basis.

We consider the same goal with sets of mathematical objects for which addition and multiplication is defined. For example, in $\mathbb{R}$ it is true that *if $x + y = x + z$, then $y = z$*. This statement is is also true when we replace real numbers with M or $\mathbb{Q}$. In our efforts to provide mathematical justifications, it would be good to not have to prove this fact both for real numbers and again for M. And so for this end we introduce the following definition.

Definition 2.21. *Let F be a set equipped with two operations, addition (+) and multiplication (×). We say F is a field when the following statements are true for F.*

(A1) $(\forall x, y \in F)\ \ x + y \in F$ *(addition is closed)*

(A2) $(\forall x, y \in F)\ \ x + y = y + x$ *(addition is commutative)*

(A3) $(\forall x, y, z \in F)\ \ (x + y) + z = (x + y) + x$ *(addition is associative)*

(A4) $(\exists 0 \in F)[(\forall x \in F)\ \ x + 0 = x]$ *(additive identity)*

(A5) $(\forall x \in F)[(\exists y \in F)\ \ x + y = 0]$ *(additive inverses)*

(M1) $(\forall x, y \in F)\ \ xy \in F$ *(multiplication is closed)*

(M2) $(\forall x, y \in F)\ \ xy = yx$ *(multiplication is commutative)*

(M3) $(\forall x, y, z \in F)\ \ (xy)z = x(yz)$ *(multiplication is associative)*

(M4) $(\exists 1 \in F \setminus \{0\})[(\forall x \in F)\ \ 1x = x]$ *(multiplicative identity)*

(M5) $(\forall x \in F \setminus \{0\})[(\exists y \in F \setminus \{0\})\ \ xy = 1]$ *(multiplicative inverses)*

(D) $(\forall x, y, z \in F)\ \ x(y + z) = xy + xz$ *(multiplication distributes over addition)*

Informally, these statements say the following:

(A1) if we add together two things in F the result is in F.

(A2) the order in which we add elements doesn't matter

(A3) the order in which perform additions doesn't matter.

(A4) there is an element of F so that adding with than element does nothing. (We usually denote this element as 0.)

(A5) Every element has a companion so that their sum equals 0.

(M1) if we multiply together two things in F the result is in F.

(M2) the order in which we multiply elements doesn't matter

(M3) the order in which perform multiplication doesn't matter.

(M4) there is an element of F so that multiplying with that element does nothing. (We usually call this element 1.)

(M5) Every element has a companion so that their product equals 1.

(D) multiplication distributes over addition

The statements listed above are true when $F = \mathbb{R}$. And so we say that $\mathbb{R}$ is a field. They are also true when $F = \mathbb{Q}$ and when $F = M$. Therefore $\mathbb{Q}$ is a field and M is a field.

Consider now the $\mathbb{Z}$. Certainly (A1)-(A5) are true for the integers. As are (M1)-(M4). However (M5) is not true for the integers. There is no integer that can be multiplied by 2 so that the result is 1. Therefore $\mathbb{Z}$ is not a field.

In Linear Algebra our definition of vector space had a very similar format as above. A vector space is a collection of mathematical objects for which we can add and scalar multiply. Just as a vector space can be thought of as *a collection of mathematical objects* *that behave like vectors*, a field can be thought of as *a collection of mathematical objects that behave like the real numbers.*

Aside. *A first attempt at definition of fields was due to Dedekind in the late 19th century. However, Dedekind hadn't made the connection between the operations as he saw them in $\mathbb{R}$ and $\mathbb{C}$, and the corresponding operations for other mathematical objects. The first definition of field that is similar to the one is use today arrived nearly twenty years after Dedekind's first attempt to generalize the operations in $\mathbb{R}$ and $\mathbb{C}$. Dedekind's work in writing down a careful definition for the real numbers lead him on the path to define fields for the first time. For our work in this class, we are taking the opposite approach. In Section 3 we will use the definition of field to study Dedekind's definition of the real numbers.*

Let us take a moment to consider more closely items (A4) and (A5). Statement (A4) is telling us that there is an element of F so that adding with than element does nothing. In $\mathbb{R}$, this element is the real number 0. In M this element is the all zeros matrix.

Recalling our study of vector spaces in Linear Algebra, we saw a similar type of property for vectors. Every vector space has a zero vector. A vector for which adding to that vector does nothing: $\overrightarrow{x} + \overrightarrow{0} = \overrightarrow{x}$. In $\mathbb{R}^3$ the zero vector is $(0, 0, 0)$. In a vector space of polynomials, the zero vector is the 0-polynomial.

And so perhaps we recall the following definitions from our time in Linear Algebra:

Definition 2.22. *Let F be a field and let $x, y \in F$ We say 0 is the <u>additive identity</u>. When $x + y = 0$ we say <u>y is the additive inverse of x</u>. When y is the additive inverse of x we denote y as $-x$.*

Items (M4) and (M5) are similar to (A4) and (A5), but replace addition with multiplication. In $\mathbb{R}$, the element for which multiplication does nothing is the real number 1. In M the corresponding element is I_2, the 2×2 identity matrix. And so we define analogous terms corresponding to (M4) and (M5).

Definition 2.23. *Let F be a field and let $x, y \in F$ We say 1 is the <u>multiplicative identity</u>. When $xy = 1$ we say <u>y is the multiplicative inverse of x</u>. When y is the multiplicative inverse of x we denote y as x^{-1}.*

Aside. *Consider the set of all invertible functions from $\mathbb{R} \to \mathbb{R}$. The set of invertible functions has a nicely defined notion of addition. Define $f + g$ so that $(f + g)(x) = f(x) + g(x)$*

If we define multiplication to mean function composition, then the set of all invertible functions from $\mathbb{R} \to \mathbb{R}$ forms a field. The multiplicative identity of this field is the function $I : \mathbb{R} \to \mathbb{R}$ so that $I(x) = x$ for all $x \in \mathbb{R}$.

When f is an invertible function we denote its inverse as f^{-1}. An invertible function and its inverse satisfy the following property

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

In other words, $f \circ f^{-1} = I$.

Statements (A1)-(A5), (M1-M5) and (D) are called the field axioms. In linear algebra we use the vector axioms to prove statements about all vectors spaces. Similarly can use the field axioms to prove facts about fields.

For example, using the definition of field, we can prove additive cancellation works as we expect.

Theorem 2.24. *Let F be a field and let $x, y, z \in F$. If $x + y = x + z$, then $y = z$.*

Proof.

1. Let F be a field and let $x, y, z \in F$ so that $x + y = x + z$. (premise)
2. There exists $x' \in F$ so that $x + x' = 0$. (A5)
3. $x' + (x + y) = x' + (x + z)$ (algebra)
4. $(x' + x) + y = (x' + x) + z$ (3,A3)
5. $0 + y = 0 + z$ (2,4)
6. $y + 0 = z + 0$ (5,A2)
7. $y = z$ (6,A4)
8. if $x + y = x + z$, then $y = z$ (1,7)

□

Since $\mathbb{R}$ is a field, replacing F with $\mathbb{R}$ gives a proof that additive cancellation is valid in $\mathbb{R}$. Similarly, since $\mathbb{Q}$ is a field, this proof tells us that additive cancellation is valid in $\mathbb{Q}$. Similarly, since $\mathcal{D}_3$ is a field, this proof tells us that additive cancellation is valid in $\mathcal{D}_3$.

With a similar technique we can[4] verify the following:

Theorem 2.25. *Let F be a field and let $x, y, z \in F$ with $x \neq 0$. If $xy = xz$, then $y = z$.*

Aside. *Just about every algebraic technique we learned about in school is true in any field. For the remainder of this course we will only be considering $\mathbb{R}$, the real field. If you find the idea of abstracting away extraneous details and only focusing on underlying algebraic structure (i.e., you find vector spaces more interesting than $\mathbb{R}^n$) you may want to consider taking MAST30005 (Abstract Algebra). In this subject we study the algebraic laws satisfied by familiar objects such as integers, polynomials and matrices. This abstraction simplifies and unifies our understanding of these structures and enables us to apply our results to interesting new cases.*

If you find the work we do in this course particularly interesting, you may want to consider taking MAST30021 (Complex Analysis), where the topics we consider are again considered after replacing $\mathbb{R}$ with $\mathbb{C}$.

Consider the following sequence of subsets

$$\mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}$$

[4] this is a great exercise for students to demonstrate their understanding on, say, a final exam.

The set of integers is ordered, but it is not a field. It satisfies all of the field axioms except (M5). Though $\mathbb{Z}$ has a multiplicative identity, its elements do not have a multiplicative inverse.

If we wanted to add things to $\mathbb{Z}$ to make it into a field we would have to add in all of the multiplicative inverses of all of the integers. Namely, we would have to add in everything of the form $\frac{1}{n}$ for all $n \neq 0$. However, in order to continue to satisfy properties (A2) and (M2) we would then need to add in all elements of form $\frac{m}{n}$, for every $m, n \in \mathbb{Z}$ and $n \neq 0$. In other words, if we want to add things to $\mathbb{Z}$ to turn it in to a field, then the set of numbers we end up with is $\mathbb{Q}$.

Compare now $\mathbb{Q}$ and $\mathbb{R}$. Both are ordered fields, but only $\mathbb{R}$ has the least upper bound property. If we consider $\mathbb{Q}$ as arising from $\mathbb{Z}$ by turning $\mathbb{Z}$ into a field, then perhaps $\mathbb{R}$ arises from $\mathbb{Q}$ by adding elements so that the resulting ordered field has the least upper bound property. This fact turns out to be true.

Theorem 2.26. *The real numbers*

- *are an ordered field,*
- *contain $\mathbb{Q}$ as a subfield*[5]*.*
- *have the property that every subset that is bounded above in $\mathbb{R}$ has a supremum in $\mathbb{R}$.*

To prove this theorem, however, we would need a better description of $\mathbb{R}$ than *"... umm.. everything on the number line"*. We consider this problem, how can we define $\mathbb{R}$, in Section 3 of the course.

Aside. *In fact, the real numbers are the only set of numbers that satisfy the three properties in Theorem 2.26. That is to say, if a set A satisfies the three properties in Theorem 2.26, then necessary A is the real numbers. Unfortunately, for one to confirm this fact, one needs to first grapple with what the definition of "is" is, which is beyond the scope of this course.*

---

[5] we have not defined the meaning of the word subfield. However, maybe we understand intuitively what it means. If we know what a subspace of a vector space is then we consider a subfield of a field to be defined analogously; it a subfield is a field that is wholly contained inside another field

Test Your Understanding

1. Using your knowledge of matrix computations, explain how you know M satisfies axiom (A2).
2. Using the Field Axioms, prove the following theorem:

   Theorem. *Let F be a field and let $x, y, z \in F$ with $x \neq 0$. If $xy = xz$, then $y = z$.*

Test Your Understanding - Answers

1. Let $A_1 = \begin{bmatrix} a_1 & -b_1 \\ b_1 & a_1 \end{bmatrix}$ and $A_2 = \begin{bmatrix} a_2 & -b_2 \\ b_2 & a_2 \end{bmatrix}$ We have

$$A_1 + A_2 = \begin{bmatrix} a_1 & -b_1 \\ b_1 & a_1 \end{bmatrix} + \begin{bmatrix} a_2 & -b_2 \\ b_2 & a_2 \end{bmatrix} = \begin{bmatrix} a_1 + a_2 & -(b_1 + b_2) \\ b_1 + b_2 & a_1 + a_2 \end{bmatrix}$$

$$A_2 + A_1 = \begin{bmatrix} a_2 & -b_2 \\ b_2 & a_2 \end{bmatrix} + \begin{bmatrix} a_1 & -b_1 \\ b_1 & a_1 \end{bmatrix} = \begin{bmatrix} a_2 + a_1 & -(b_2 + b_1) \\ b_1 2 b_1 & a_2 + a_1 \end{bmatrix}$$

Since

$$\begin{bmatrix} a_1 + a_2 & -(b_1 + b_2) \\ b_1 + b_2 & a_1 + a_2 \end{bmatrix} = \begin{bmatrix} a_2 + a_1 & -(b_2 + b_1) \\ b_2 + b_1 & a_2 + a_1 \end{bmatrix}$$

we have $A_1 + A_2 = A_2 + A_1$. Therefore M satisfies axiom (A2)

2. *Proof.* Let F be a field and let $x, y, z \in F$ so that $xy = xz$ and $x \neq 0$ By axiom (M5), there exists $x' \in F$ so that $xx' = 1$. Consider the expression

$$x'(xy) = x'(xz)$$

Since multiplication in a field is associative, we have

$$(x'x)y = (x'x)z$$

Since multiplication in a field is commutative, we have

$$(xx')y = (xx')z$$

Since $xx' = 1$ we have

$$1y = 1z$$

And so by axiom (M4) we have

$$y = z$$

Therefore if $xy = xz$, then $y = z$. □

3 The Real Numbers

A Caveat About the Section 3 *This section is unavoidably notationally heavy. We will be defining lots of new pieces of notation and then using that notation to prove various mathematical facts. It is easy to get lost in the mess that will follow. This .pdf has an index and a glossary of notation. If you find yourself not understanding a argument, take the time to review the definitions of the notation and the terms.*

Good luck!

Section 2 ended with the statement of the following theorem:

Theorem. *The real numbers*

- *are an ordered field*
- *contain* $\mathbb{Q}$ *as a subfield*
- *have the property that every subset that is bounded above in* $\mathbb{R}$ *has a supremum in* $\mathbb{R}$.

As we spending time in this course *justifying mathematical truths*, let us spend a moment and think about how we might go about proving this statement. Unlike many of the theorems we have encountered so far in this class, this theorem does not appear to be an implication. Nor does it appear to be a universally quantified statement. Instead, it is an assertion that some particular mathematical object satisfies some properties. In particular, this theorem asserts the following:

- $\mathbb{R}$, together with its ordering and its operations, satisfies the definition of an order and all of the field axioms.
- In the same way that we can find subspace of vector spaces, we can find the field $\mathbb{Q}$ contained wholly within the field $\mathbb{R}$.
- Every subset of $\mathbb{R}$ that is bounded above in $\mathbb{R}$ has a supremum in $\mathbb{R}$. In other words, for every $A \subsetneq \mathbb{R}$ that is bounded above, there is a least upper bound of A in $\mathbb{R}$.

In the last section we considered the field M, where

$$M = \left\{ \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \mid a, b \in \mathbb{R} \right\}$$

Since we have an explicit description of the elements of M and knowledge of how addition and multiplication behave in M, we could take the time to prove that M satisfies all of the field axioms. However, as yet, we do not have such a description of $\mathbb{R}$. At present, our definition of $\mathbb{R}$ is little more than a vague intuition about points on the *real line*.

Our goal in Section 3 of the course is to discover a description of $\mathbb{R}$ that lets us prove that our theorem above is true.

At the end of 2.3, we considered the sequence of subsets:

$$\mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}$$

We saw that $\mathbb{Q}$ is constructed by using $\mathbb{Z}$ as a starting point.

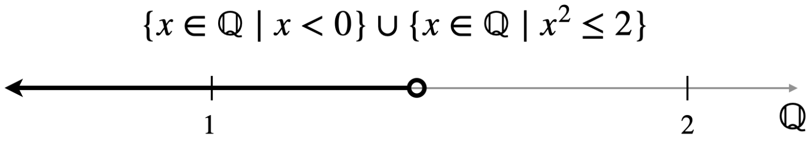
$$\mathbb{Q} = \left\{ \frac{p}{q} \mid p, q \in \mathbb{Z}, q \neq 0 \right\}$$

It turns out that we can *construct* $\mathbb{R}$ by using $\mathbb{Q}$ as a starting point.

Consider the following subset of $\mathbb{Q}$:

$$A = \{x \in \mathbb{Q} \mid x^2 \leq 2\} \cup \{x \in \mathbb{Q} \mid x < 0\} \subseteq \mathbb{Q}.$$

As a subset of $\mathbb{Q}$ this set is bounded above, but does not have a supremum.

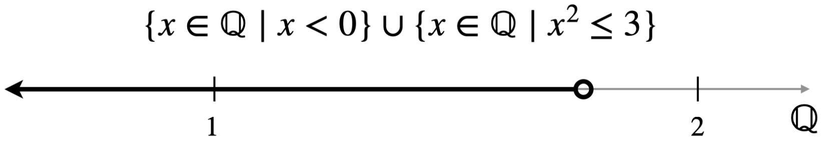

$\{x \in \mathbb{Q} \mid x < 0\} \cup \{x \in \mathbb{Q} \mid x^2 \leq 2\}$
1
2
$\mathbb{Q}$

Let us consider a seemingly absurd sentence:

Let $\sqrt{2}$ *denote the set* $\{x \in \mathbb{Q} \mid x^2 \leq 2\} \cup \{x \in \mathbb{Q} \mid x < 0\}$

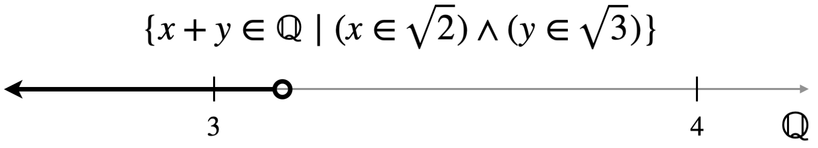
And in the same vein let us consider a similar seemingly absurd sentence

Let $\sqrt{3}$ *denote the set* $\{x \in \mathbb{Q} \mid x^2 \leq 3\} \cup \{x \in \mathbb{Q} \mid x < 0\}$


$\{x \in \mathbb{Q} \mid x < 0\} \cup \{x \in \mathbb{Q} \mid x^2 \leq 3\}$
1
2
$\mathbb{Q}$

And finally, let us consider a third seemingly absurd sentence

Let $\sqrt{2} + \sqrt{3}$ *denote the set* $\{x + y \in \mathbb{Q} | (x \in \sqrt{2}) \wedge (y \in \sqrt{3})\}$


$\{x + y \in \mathbb{Q} \mid (x \in \sqrt{2}) \wedge (y \in \sqrt{3})\}$
3
4
$\mathbb{Q}$

This last one is even stranger than the first two!

With these absurdities in mind, what should we make of the condition

$$(x \in \sqrt{2}) \wedge (y \in \sqrt{3})$$

Well, if we are using $\sqrt{2}$ to denote a set, then writing $x \in \sqrt{2}$ has meaning for us. The set $\sqrt{2} + \sqrt{3}$ is the set of all rational numbers we can get by adding together an element of $\sqrt{2}$ with an element of $\sqrt{3}$

Thinking of the notations $\sqrt{2}, \sqrt{3}$ and $\sqrt{2}+\sqrt{3}$ as we traditionally understand them, we have the following chain of inequalities:

$$\sqrt{2} < \sqrt{3} < \sqrt{2}+\sqrt{3}$$

Using our absurd new meanings of the notation $\sqrt{2}$, $\sqrt{3}$ and $\sqrt{2}+\sqrt{3}$ we have the following chain of subsets

$$\sqrt{2} \subsetneq \sqrt{3} \subsetneq \sqrt{2}+\sqrt{3}$$

We could prove this using our definition of $\subsetneq$ and our absurd definitions of $\sqrt{2}$, $\sqrt{3}$ and $\sqrt{2}+\sqrt{3}$ as subsets of $\mathbb{Q}$.

If our goal is to find a description of elements of $\mathbb{R}$ so that we can prove facts about addition, multiplication, ordering and bounded subsets, then this seems like a good start. We have given a description of three elements of $\mathbb{R}$ (namely $\sqrt{2}$, $\sqrt{3}$ and $\sqrt{2}+\sqrt{3}$) by using only $\mathbb{Q}$ together with set operations. In doing so, our notion of ordering of $\mathbb{R}$ has been replicated by subsets in $\mathbb{Q}$.

This, then, will be our approach: We will define each irrational number k as a subset $A \subsetneq \mathbb{Q}$ so that A contains all elements of $\mathbb{Q}$ that we want to be smaller than k.

$$\sqrt{2} = \{x \in \mathbb{Q} \mid x^2 \leq 2\} \cup \{x \in \mathbb{Q} \mid x < 0\} \subseteq \mathbb{Q}$$

Once we have done this, we will then define the meaning of $<$, $+$ and $\times$ in terms of these sets. In doing so, we will ensure that the ordering axioms [(O1), (O2)] and the field axioms are satisfied.

Just like Vector Spaces, there is a set of axioms for the real numbers:

Definition 3.1. *Let R be an ordered set equipped with two operations: addition $(+)$ and multiplication $(\times)$ We say* <u>*R is the real numbers*</u>*, when the following statements are true.*

- *(O1) for $x, y \in R$ exactly one of the following is true:*

  $x < y$

  $y < x$

  $x = y$

- *(O2) for $x, y, z \in R$, if $x < y$ and $y < z$, then $x < z$.*
- *(A1)* $(\forall x, y \in F)\ \ x + y \in F$ *(addition is closed)*
- *(A2)* $(\forall x, y \in F)\ \ x + y = y + x$ *(addition is commutative)*
- *(A3)* $(\forall x, y, z \in F)\ \ (x + y) + z = x + (y + z)$ *(addition is associative)*
- *(A4)* $(\exists 0 \in F)[(\forall x \in F)\ \ x + 0 = x]$ *(additive identity)*
- *(A5)* $(\forall x \in F)[(\exists y \in F)\ \ x + y = 0]$ *(additive inverses)*
- *(M1)* $(\forall x, y \in F)\ \ xy \in F$ *(multiplication is closed)*

- *(M2)* $(\forall x, y \in F)\ \ xy = yx$ *(multiplication is commutative)*
- *(M3)* $(\forall x, y, z \in F)\ \ (xy)z = x(yz)$ *(multiplication is associative)*
- *(M4)* $(\exists 1 \in F \setminus \{0\})[(\forall x \in F)\ \ 1x = x]$ *(multiplicative identity)*
- *(M5)* $(\forall x \in F \setminus \{0\})[(\exists y \in F \setminus \{0\})\ \ xy = 1]$ *(multiplicative inverses)*
- *(D)* $(\forall x, y, z \in F)\ \ x(y+z) = xy+xz$ *(multiplication distributes over addition)*
- *(OA)* $(\forall x, y, c \in F)\ \ (x < y) \Rightarrow (x + c < y + c)$ *(addition preserves order)*
- *(OM)* $(\forall x, y \in F)\ \ [(0 < x) \wedge (0 < y)] \Rightarrow (0 < xy)$ *(multiplication preserves order)*
- *(C) every subset of R that is bounded above in R has a least upper bound in R.*

When R is the real numbers we write $R = \mathbb{R}$.

Let us take a moment to sit in the absurdity of this definition.

Just as a vector space is any collection of objects that satisfy the Vector Space Axioms, We call a set the real numbers when it satisfies all of the Real Number Axioms. This, then, perhaps implies there can be more set we can refer to as the real numbers. This turns out not to be the case. The real numbers as we know them are the only set that satisfies the properties above.

Comparing this list to the definition of ordering and the Field Axioms we note three additional items: (OA), (OM) and (C). The first two of these new axioms ensure that ordering is consistent with addition and multiplication in the way we expect. This third axiom (C) is the third statement in our theorem above. We refer to (C) as the Completeness Axiom

We proceed as follows. We define each irrational number k as a subset A of rational numbers so that A contains all elements of $\mathbb{Q}$ that are smaller than k. For example, we define $\sqrt{2}$ to refer to the following set:

$$\sqrt{2} = \{x \in \mathbb{Q} \mid x^2 \leq 2\} \cup \{x \in \mathbb{Q} \mid x < 0\} \subseteq \mathbb{Q}$$

When then equip the set of all such possible sets with an ordering, a meaning for addition and a meaning for multiplication. When then show that our set, together with $<, +$ and $\times$ satisfies the Axioms for the Real Numbers.

3.1 Dedekind Cuts

Recall our approach for defining irrational numbers: We will define each irrational number k as a subset $A \subsetneq \mathbb{Q}$ so that A contains all elements of $\mathbb{Q}$ we want to smaller than k. And so, perhaps, this is a reasonable definition for an irrational number, k.

Definition. *Let k be an irrational number. Let k denote the following subset of $\mathbb{Q}$*

$$k = \{x \in \mathbb{Q} \mid x < k\}$$

Except, hmm... This definition is inherently circular. With this definition has to already know what k is before one can define k. This doesn't seem like a good approach.

Rather than start with our definition, let us imagine for a moment we already know how to construct the corresponding set for each irrational number k. In other words, let $p(x)$ be a condition on $\mathbb{Q}$ so that the following set defines an irrational number k.

$$k = \{x \in \mathbb{Q} \mid p(x)\}$$

When $k = \sqrt{2}$ we had $p(x) : (x^2 \leq 2) \vee (x < 0)$.

If we have succeeded as defining k as a subset $A \subsetneq \mathbb{Q}$ so that A contains all elements of $\mathbb{Q}$ we want to smaller than k, then the following must be true about the set k:

1. For $x, y \in \mathbb{Q}$, if $x \in k$ and $y < x$, then $y \in k$.

If the set k is to contain all rational numbers smaller than the irrational number k and $x \in k$, then any rational number smaller than x must also be an element of k.

Back in the previous section we spent some time thinking about intervals between rational numbers and irrational numbers. We came upon the following thought:

There is at least one rational number between every rational number and every irrational number.

And so, looking back at our set k, we expect the following must be true:

2. For $x \in \mathbb{Q}$, if $x \in k$, there exists $z \in k$ such that $z > x$.

In other words, if x is smaller than the irrational number k (i.e., $x \in k$), then we expect to be able to find another rational number between x and k. And so we must be able to find an element of the set k that is larger than x.

Unfortunately, there are some examples of sets that satisfy these two properties that aren't helpful in our goal to define irrational numbers using a set of rational numbers.

Consider the set $\mathbb{Q}$. The set $\mathbb{Q}$ satisfies both properties, but doesn't seem to correspond to an irrational number in the same way that the set $\sqrt{2}$ did. Similarly, the empty set, $\{\}$, satisfies both properties. But again, the empty set doesn't seem to nicely correspond to an irrational number in the same way that the set $\sqrt{2}$ did.

With these thoughts in mind, let us give a name of proper and non-empty subsets of $\mathbb{Q}$ that have these two properties:

Definition 3.2. *Let A be a subset of $\mathbb{Q}$. We say A is a cut when the following three statements are true*

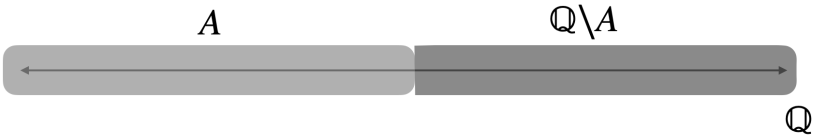
1. *For $x, y \in \mathbb{Q}$, if $x \in A$ and $y < x$, then $y \in A$.*
2. *For $x \in \mathbb{Q}$, if $x \in A$, there exists $z \in A$ such that $z > x$.*
3. *$A \neq \{\}$ and $A \neq \mathbb{Q}$.*

For example, $\sqrt{2}$ (as defined in the above) is a cut:

$$\sqrt{2} = \{x \in \mathbb{Q} \mid x^2 \leq 2\} \cup \{x \in \mathbb{Q} \mid x < 0\}$$

We can verify that this set satisfies all parts of the definition above.

In general, a cut splits $\mathbb{Q}$ into two parts.



The threshold of this split will be the real number corresponding to the cut.

Let $\mathbf{R}^\star$ be the set of all cuts. In other words, define $\mathbf{R}^\star$ so that

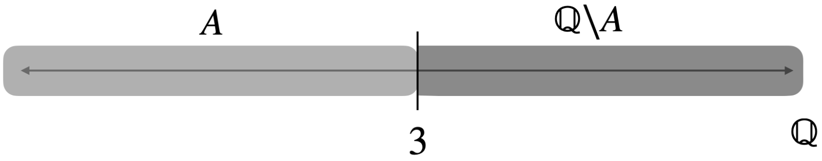
$$\mathbf{R}^\star = \{A \mid A \text{ is a cut}\}$$

From our work, so far we have that each of the sets $\sqrt{2}$, $\sqrt{3}$ and $\sqrt{2} + \sqrt{3}$ are cuts. And so, $\sqrt{2}, \sqrt{3}, \sqrt{2} + \sqrt{3} \in \mathbf{R}^\star$

Consider now the following subset of $\mathbb{Q}$

$$A = \{x \in \mathbb{Q} \mid x < 3\}$$

Take a few moments to convince to yourself that this is a cut. (Does it satisfy the definition of cut above?)

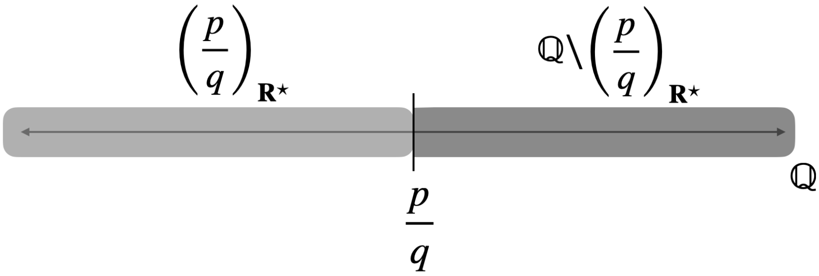


It seems that our set of all cuts, $\mathbf{R}^\star$, also includes sets that correspond to rational numbers! For example, the cut corresponding to $0 \in \mathbb{Q}$ is the cut

$$\{x \in \mathbb{Q} \mid x < 0\}$$

For $\frac{p}{q} \in \mathbb{Q}$, let $\left(\frac{p}{q}\right)_{\mathbf{R}^\star}$ denote the following cut:

$$\left(\frac{p}{q}\right)_{\mathbf{R}^\star} = \left\{x \in \mathbb{Q} \mid x < \frac{p}{q}\right\}$$

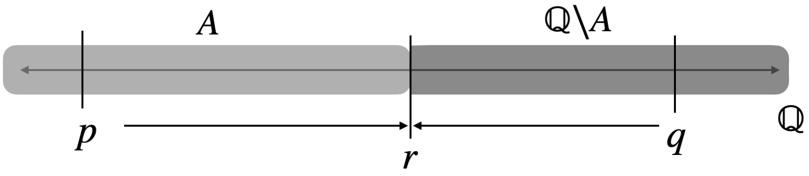

$\left(\frac{p}{q}\right)_{\mathbf{R}^\star}$
$\mathbb{Q}\setminus\left(\frac{p}{q}\right)_{\mathbf{R}^\star}$
$\frac{p}{q}$
$\mathbb{Q}$

Not only does the set $\mathbf{R}^\star$ seem to contain elements (i.e., sets) that correspond to irrational numbers, but also elements (i.e., sets) that correspond to rational numbers.

It turns out that $\mathbf{R}^\star$ is in fact $\mathbb{R}$ in disguise! Every element of $\mathbb{R}$ corresponds to exactly one element of $\mathbf{R}^\star$, and vice versa. To see what is meant by this, consider A, an arbitrary element of $\mathbf{R}^\star$.

The set A is a subset of $\mathbb{Q}$. Since it is a cut, it is neither the empty set, nor all of $\mathbb{Q}$. In other words, there is at least one element, $p \in \mathbb{Q}$ that is contained in A and one element of $q \in \mathbb{Q}$ that is not contained in A.

Consider moving to the right and left respectively from p and q. By part 2 in the definition of a cut, we can move right from p in $\mathbb{Q}$ and continue to encounter elements of A. (We may have to move very slowly!) As we move left from p we may or may not encounter elements of A. However, as soon as we encounter an element of A then everything we encounter will continue to be in A. And so there is a threshold where the elements of A meet the non-elements of A.



This threshold is exactly the element of $\mathbb{R}$ that the cut A refers to.

Now that we can see how to associate elements of $\mathbf{R}^\star$ with elements of $\mathbb{R}$, we will define meanings for $<$ $+$ and $\times$ for elements of $\mathbf{R}^\star$ so that the outcomes are exactly the expected outcomes in $\mathbb{R}$. In other words, we define the meaning of $<$ $+$ and $\times$ for elements of $\mathbf{R}^\star$ so that the Real Number axioms are satisfied.

Aside. *These objects are traditionally called Dedekind Cuts. Dedekind was a late 19 century German mathematician who worked in a number of different foundational areas of mathematics.*

The Ordering Axioms Let us begin with axioms (O1) and (O2). To be able to satisfy these axioms we need a meaning for $<$ on the set of all cuts. As we saw in the introduction, we can accomplish this using subsets.

Definition 3.3. *Let A and B be cuts. We write $\underline{A < B}$ when $A \subsetneq B$.*

For example when we write $(0)_{\mathbf{R}^\star} < (1)_{\mathbf{R}^\star}$ we mean:

$$\{x \in \mathbb{Q} \mid x < 0\} \subsetneq \{x \in \mathbb{Q} \mid x < 1\}$$

Using this definition we can prove axioms (O1) and (O2) hold for $\mathbf{R}^\star$:

- (O1) for $x, y \in \mathbf{R}^\star$ exactly one of the following is true:

  $x < y$

  $y < x$

  $x = y$

- (O2) for $x, y, z \in \mathbf{R}^\star$, if $x < y$ and $y < z$, then $x < z$.

In these statements x, y and z are cuts. Therefore they are sets. We discuss (O2) and leave (O1) as an exercise.

Theorem 3.4. *Let $x, y, z \in \mathbf{R}^\star$. If $x < y$ and $y < z$, then $x < z$.*

Before we dive in and prove this, let us take a moment to think about what this theorem is saying:

$$\text{Hypothesis: } x < y \text{ and } y < z$$

$$\text{Conclusion: } x < z$$

In the hypothesis each of x, y and z are elements of $\mathbf{R}^\star$. Therefore they are cuts. From our definition above, the notation $x < y$ is taken to mean $x \subsetneq y$. Re-writing our hypothesis and our conclusion with this observation in mind, we have:

$$\text{Hypothesis: } x \subsetneq y \text{ and } y \subsetneq z$$

$$\text{Conclusion } x \subsetneq z$$

In Tutorial 3 we prove the following statement

Theorem. *Let A, B and C be sets. If $A \subseteq B$ and $B \subseteq C$, then $A \subseteq C$.*

Taking $A = x$, $B = y$ and $C = z$, this theorem gives directly that Theorem 3.4 is true. In other words, $\mathbf{R}^\star$ satisfies Axiom (O2).

The Addition Axioms We turn now to the Addition Axioms: (A1)-(A5) and (OA). To do so, we must define was addition means for elements of $\mathbf{R}^\star$. Looking back at our work in introduction, we made the following statement:

$$\sqrt{2} + \sqrt{3} = \{x + y \in \mathbb{Q} \mid (x \in \sqrt{2}) \wedge (y \in \sqrt{3})\}$$

And so, let us take this as our definition of addition on $\mathbf{R}^\star$

Definition 3.5. *Let A and B be cuts. We define the set $A + B$ as follows*

$$A + B = \{x + y \in \mathbb{Q} \mid (x \in A) \wedge (y \in B)\}$$

The set $A+B$ is the set of all possible sums we can take from elements of A and elements of B.

To check that this decision is reasonable, let us consider the sum

$$(0)_{\mathbf{R}^\star} + (1)_{\mathbf{R}^\star}$$

By definition, we have

$$\begin{aligned}(0)_{\mathbf{R}^\star} &= \{x \in \mathbb{Q} \mid x < 0\}\\ (1)_{\mathbf{R}^\star} &= \{x \in \mathbb{Q} \mid x < 1\}\end{aligned}$$

Consider the set

$$\{x + y \in \mathbb{Q} \mid (x < 0) \wedge (x < 1)\}$$

If our addition works as we expect, we should hope for the following result

$$(0)_{\mathbf{R}^\star}+(1)_{\mathbf{R}^\star} = \{x \in \mathbb{Q} \mid x < 0\}+\{x \in \mathbb{Q} \mid x < 1\} = \{x+y \in \mathbb{Q} \mid (x < 0)\wedge(y < 1)\} = (1)_{\mathbf{R}^\star}$$

Here the first equals sign follows from the definition of our notation. The second equals sign follows from our definition of addition in $\mathbf{R}^\star$. It is the third equals sign we need to confirm.

Recall that for sets A and B we have $A = B \Leftrightarrow (A \subseteq B) \wedge (B \subseteq A)$. And so to show third equality holds, it suffices to show the following statement[1] is true:

$$(\{x + y \in \mathbb{Q} \mid (x < 0) \wedge (y < 1)\} \subseteq (1)_{\mathbf{R}^\star}) \wedge [(1)_{\mathbf{R}^\star} \subseteq \{x + y \in \mathbb{Q} \mid (x < 0) \wedge (y < 1)\}]$$

We first show

$$\{x + y \in \mathbb{Q} \mid (x < 0) \wedge (x < 1)\} \subseteq (1)_{\mathbf{R}^\star}$$

Let $z \in \{x + y \in \mathbb{Q} \mid (x < 0) \wedge (x < 1)\}$. Since $z \in \{x + y \in \mathbb{Q} \mid (x < 0) \wedge (x < 1)\}$ there exists $x, y \in \mathbb{Q}$ so that $x + y = z$, $x < 0$, and $y < 1$. Since $x < 0$ and $y < 1$, it follows that $x + y < 1$. Therefore $x + y \in (1)_{\mathbf{R}^\star}$. And so $z \in (1)_{\mathbf{R}^\star}$.

We now show

$$(1)_{\mathbf{R}^\star} \subseteq \{x + y \in \mathbb{Q} \mid (x < 0) \wedge (y < 1)\}$$

Let $z \in (1)_{\mathbf{R}^\star}$. We proceed based on the interval in which we can find z. In each case we find $x \in (0)_{\mathbf{R}^\star}$, $y \in (1)_{\mathbf{R}^\star}$ so that $x + y = z$.

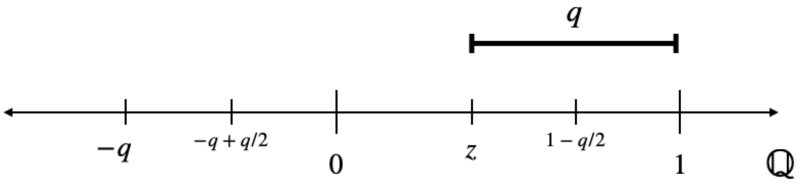
If $z < 0$, then let $x = z$ and $y = 0$. Notice $x + y = z$, $x \in (0)_{\mathbf{R}^\star}$ and $y \in (1)_{\mathbf{R}^\star}$.

If $z = 0$, then let $x = -0.1$ and $y = 0.1$. Notice $x + y = z$, $x \in (0)_{\mathbf{R}^\star}$ and $y \in (1)_{\mathbf{R}^\star}$.

Otherwise, assume $0 < z < 1$. Since $0 < z < 1$ there exists $q \in \mathbb{Q}$ so that $z + q = 1$ and $0 < q < 1$.

[1] If you want to impress your non-math friends, show them this expression and tell them that this statement proves $0 + 1 = 1$. When you are done with that, get down from you academic high horse and be sure that you value the contributions of social scientists. Everything about this expression depends on the mathematical culture you are learning about this semester.

Let[2] $x = -q + q/2$ and $y = 1 - q/2$. Notice $x \in (0)_{\mathbf{R}^\star}$, $y \in (1)_{\mathbf{R}^\star}$ and $x + y = z$.



In other words $z \in \{x + y \in \mathbb{Q} \mid (x < 0) \wedge (y < 1)\}$.

From these two arguments it follows that

$$\{x + y \in \mathbb{Q} \mid (x < 0) \wedge (y < 1)\} = (1)_{\mathbf{R}^\star}$$

And so we conclude

$$(0)_{\mathbf{R}^\star} + (1)_{\mathbf{R}^\star} = (1)_{\mathbf{R}^\star}$$

Aside. *Is our work above that verifies $(0)_{\mathbf{R}^\star} + (1)_{\mathbf{R}^\star} = (1)_{\mathbf{R}^\star}$ a proof? That depends on what we mean by proof. If our reader accepts that we can algebraically manipulate elements of $\mathbb{Q}$ in the way we expect, then the above argument would suffice as an informal proof. To turn this informal proof into a formal proof, we would have to first accept (or prove) that $\mathbb{Q}$ is an ordered field. We would then have to use the various axioms to prove that we can algebraically manipulate expression in $\mathbb{Q}$ in the way we expect.*

Exercise 3.6. *Give an informal proof that $(1)_{\mathbf{R}^\star} + (1)_{\mathbf{R}^\star} = (2)_{\mathbf{R}^\star}$*

One of the conclusions to draw here is that working with elements of $\mathbf{R}^\star$ directly is tedious. Look at all of the work we had do to verify $0 + 1 = 1$.

Noting this tedium, let us opt not to spend our time painstakingly verifying axioms (A1)-(A5) and (OA). Instead, let us trust that verifying these axioms is not overly difficult, but takes clear organization and understanding of the meaning of the notations.

The Multiplication Axioms Unfortunately, the work to verify the multiplication axioms for $\mathbf{R}^\star$ is significantly less straightforward than verifying the addition axioms. In the interest of brevity, we present only the definition of multiplication for elements of $\mathbf{R}^\star$ and be confident that one could use this definition to prove the multiplication axioms hold for $\mathbf{R}^\star$. (Though this is not a straightforward task!)

Definition 3.7. *Let A and B be cuts. Define the product, $A \times B$ as follows:*

$$A \times B = \begin{cases} \{xy \in \mathbb{Q} \ \mid \ (x \in A) \wedge (y \in B)\} & (0_{\mathbf{R}^\star}) \leq A, B \\ -(A \times (-B)) & 0_{\mathbf{R}^\star} \leq A, \ \textit{and } B < (0_{\mathbf{R}^\star}) \\ (-A) \times (-B) & A, B < (0_{\mathbf{R}^\star}) \end{cases}$$

[2] How did we know how to choose these particular values for x and y? We didn't. We knew our goal was to find x and y so that $x + y = z$. To do this ourselves we would have to spend a few minutes trying to be clever. Unfortunately, being able to do this quickly comes only with experience.

Before we move on, however, let us take a moment to compare axioms (AO) and (MO):

- (OA) $(\forall x, y, c \in F)\ \ x < y \Rightarrow (x + c < y + c)$ (addition preserves order)
- (OM) $(\forall x, y \in F)\ \ [(0 < x) \wedge (0 < y)] \Rightarrow 0 < xy$ (multiplication preserves order)

Axiom (OA) tells us that adding a constant to both sides of an inequality preserves the inequality. However Axiom (OM) is not the corresponding statement for multiplication by a positive constant. Using Axiom (OM) we prove that multiplication by a positive constant preserves inequalities.

Theorem 3.8. *Let $x, y, z \in \mathbb{R}$ with $z > 0$. If $x < y$, then $xz < yz$.*

Proof. Let $x, y, z \in \mathbb{R}$ so that $x < y$ and $z > 0$. By (AO), we have $x + (-x) < y - x$. By (A5), we have $0 < y - x$. By (MO), we have $(y - x)z > 0$. By (D) we have $yz - xz > 0$. By (AO) and (A5) we have $yz > xz$. □

Aside. *This proof isn't quite complete. For example, we use (D) to assert $(y - x)z = yz - xz$. However, Axiom (D) tells us that multiplication distributes over addition, not subtraction. Actually, now that we mention it, have we defined subtraction anywhere...?*

The Completeness Axiom Before we dive in to the completeness axiom, let us take a moment to think about boundedness in $\mathbf{R}^\star$. Elements of $\mathbf{R}^\star$ are sets. And so when we talk about boundedness of a subset of $\mathbf{R}^\star$ we are talking about boundedness of a set of sets where $\leq$ is taken to mean $\subseteq$

When we write $A \leq B$ for elements of $\mathbf{R}^\star$ we mean $A \subseteq B$. Let E be a set of cuts. For E to be bounded above in $\mathbf{R}^\star$ it means that there exists a cut $B \in \mathbf{R}^\star$ so that $A \leq B$ for every $A \in E$. In other words, $A \subseteq B$ for every $A \in E$. Upper bounds in $\mathbf{R}^\star$ are sets that contain as a subset every element (i.e. set) of E.

Recall the Completeness Axiom:

(C) every subset of R that is bounded above in R has a least upper bound in R.

To prove the Completeness Axiom is true for $\mathbf{R}^\star$ we will use set unions to construct upper bounds for subsets of $\mathbf{R}^\star$. For example, consider following subset of $\mathbf{R}^\star$:

$$E = \{(0)_{\mathbf{R}^\star}, (1)_{\mathbf{R}^\star}, (7)_{\mathbf{R}^\star}\}$$

The set $\{0, 1, 7\}$ is bounded above in $\mathbb{R}$ and its supremum is 7. Coincidentally,

$$(0)_{\mathbf{R}^\star} \cup (1)_{\mathbf{R}^\star} \cup (7)_{\mathbf{R}^\star} = (7)_{\mathbf{R}^\star}$$

In other words, the set E is bounded above in $\mathbf{R}^\star$ and the supremum of E in $\mathbf{R}^\star$ is given by the union of the elements of E. This fact turns out to be true every time a subset of $\mathbf{R}^\star$ is bounded above in $\mathbf{R}^\star$.

To ease our work, let us introduce a piece of notation. Recall from our time in Calculus, $\sum$ notation:

$$\sum_{n=1}^{5} n^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2$$

As we know it, $\sum$ notation is good, but it is not terribly flexible. What if, say, we wanted to only add the squares of the odd terms? We can use a condition to communicate which terms we want to take the sum of.

$$\sum_{n\in\{1,3,5\}} n^2 = 1^2 + 3^2 + 5^2$$

We can be even more succinct by naming the set $B = \{1, 3, 5\}$

$$\sum_{n\in B} n^2 = 1^2 + 3^2 + 5^2$$

In our work to verify the Completeness Axiom, we will need to consider taking unions of many sets. Let us define the notation $\bigcup$ analogously to $\sum$. For example if A_1, A_2 and A_3 are sets, and $E = \{A_1, A_2, A_3\}$ then

$$\bigcup_{A\in E} A = A_1 \cup A_2 \cup A_3$$

Let E be a set of cuts. The notation

$$\bigcup_{A\in E} A$$

is the union of all of the sets contained in E. For example, when $E = \{(0)_{\mathbf{R}^\star}, (1)_{\mathbf{R}^\star}, (7)_{\mathbf{R}^\star}\}$ we have

$$\bigcup_{A\in E} A = (0)_{\mathbf{R}^\star} \cup (1)_{\mathbf{R}^\star} \cup (7)_{\mathbf{R}^\star}$$

In our quest to prove the Completeness Axiom, we first prove that if a subset $E \subseteq \mathbf{R}^\star$ is bounded, then the union of all of the elements that can be found within an element of E is a cut. That is, we prove

$$\bigcup_{A\in E} A \in \mathbf{R}^\star$$

We do this by showing that the set $\bigcup_{A\in E} A$ satisfies all parts of the definition of a cut.

This proof depends on a key insight[3]

$$z \in \bigcup_{A\in E} A \text{ if and only if there exists } A \in E \text{ so that } z \in A$$

Lemma 3.9. *Let E be a set of cuts that is bounded in $\mathbf{R}^\star$. The union of all elements of E is a cut.*

[3] this insight can be taken to be the definition of the notation $\bigcup_{A\in E} A$

Proof. Let E be a set of cuts that is bounded above in $\mathbf{R}^\star$. Let B be an upper bound of E in $\mathbf{R}^\star$. We show the set

$$\alpha = \bigcup_{A \in E} A$$

satisfies the definition of cut.

1. For $x, y \in \mathbb{Q}$, if $x \in \alpha$ and $y < x$, then $y \in \alpha$
Let $x, y \in \mathbb{Q}$ so that $x \in \alpha$ and $y < x$. Since $x \in \alpha$ there exists $A \in E$ so that $x \in A$. Since A is a cut and $y < x$, we have $y \in A$. Therefore $y \in \alpha$.

2. For $x \in \mathbb{Q}$, if $x \in \alpha$, there exists $z \in \alpha$ such that $z > x$
Let $x \in \mathbb{Q}$ so that $x \in \alpha$. Since $x \in \alpha$ there exists $A \in E$ so that $x \in A$. Since A is a cut, there exists $z \in A$ so that $z > x$. Since $z \in A$ it follows that $z \in \alpha$. Therefore there exists $z \in \alpha$ such that $z > x$.

3. $\alpha \neq \{\}$ and $\alpha \neq \mathbb{Q}$
Since cuts are necessarily not empty and α is a union of cuts, necessarily α is non-empty. Since E is bounded above by B, we have $A \leq B$ for all $A \in E$. Since B is a cut, there is an element of $\mathbb{Q}$ that is not an element of B. Let p be such an element. In other words, $p \notin B$. Since $A \leq B$ for all $A \in E$, we have $A \subseteq B$ for all $A \in E$. Since $p \notin B$, we have $p \notin A$ for every $A \in E$. Therefore $p \notin \bigcup_{A \in E} A$. In other words, $p \notin \alpha$. And so we conclude $\alpha \neq \mathbb{Q}$.

Since α satisfies all three parts of the definition of cut, it follows that α is a cut. In other words,

$$\bigcup_{A \in E} A \in \mathbf{R}^\star.$$

□

Lemma 3.9 implies that when E is bounded above, the union of all elements of E is a cut. We now prove, in fact, that this cut is the supremum of E. We accomplish this by proving $\bigcup_{A \in E} A$ satisfies all parts of the definition of supremum.

Lemma 3.10. *If E is a set of cuts that is bounded above in $\mathbf{R}^\star$, then*

$$\sup E = \bigcup_{A \in E} A$$

Proof. Let $E \subseteq \mathbf{R}^\star$ be bounded above in $\mathbf{R}^\star$. Therefore there exists $B \in \mathbf{R}^\star$ so that $A \leq B$ for all $A \in E$. Let

$$\alpha = \bigcup_{A \in E} A$$

By Lemma 3.9, α is a cut. We claim α is the least upper bound of E in $\mathbf{R}^\star$.

We first show α is an upper bound of E in $\mathbf{R}^\star$. Notice that for all $A \in E$ we have $A \subseteq \bigcup_{A \in E} A$. Therefore $A \subseteq \alpha$. Therefore $A \leq \alpha$ for all $A \in E$. In other words, α is an upper bound of E in $\mathbf{R}^\star$.

We now show α is the least upper bound of E in $\mathbf{R}^\star$. Recall that since $E \subseteq \mathbf{R}^\star$ is bounded above in $\mathbf{R}^\star$, there exists $B \in \mathbf{R}^\star$ so that $A \leq B$ for all $A \in E$. We show $\alpha \leq B$.

Let $x \in \alpha$. Since $x \in \alpha$ there exists $A \in E$ so that $x \in A$. Since B is an upper bound for E, we have $A \leq B$. Therefore $A \subseteq B$. Since $x \in A$ and $A \subseteq B$, it then follows that $x \in B$. Therefore $\alpha \subseteq B$. And so $\alpha \leq B$. □

Using Lemma 3.10 we show that $\mathbf{R}^\star$ satisfies the Completeness Axiom.

Theorem 3.11. *The ordered set* $\mathbf{R}^\star$ *satisfies the Completeness Axiom.*

Proof. Let $E \subseteq \mathbf{R}^\star$ so that E is bounded above. By Lemma 3.10 we have

$$\sup E = \bigcup_{A \in E} A$$

Therefore every subset of $\mathbf{R}^\star$ that is bounded above in $\mathbf{R}^\star$ has a least upper bound in $\mathbf{R}^\star$. □

Before we end this section, let us take a moment to look back and marvel at what we have accomplished. At the start of this section, real numbers were something whose definition we likely hadn't paid much attention to in the past. We took their existence of granted.

Now that we have completed our work, we have a new understanding (and perhaps appreciation) for the real numbers. Each real number can be considered as a set of rational numbers that satisfy some particular properties (i.e., a cut). We can use subsets to define an ordering for our real numbers, and use addition and multiplication $\mathbb{Q}$ to define these operations in $\mathbb{R}$.

Just as $\mathbb{Z}$ arises from $\mathbb{N}$ by adding in additive inverses, and $\mathbb{Q}$ arises from $\mathbb{Z}$ by turning $\mathbb{Z}$ into a field, so too does $\mathbb{R}$ arise from $\mathbb{Q}$ by mandating that every subset that is bounded above has a supremum.

Test Your Understanding

1. Let $A_1 = \{\}, A_2 = \{1, 2, 3\}, A_3 = \{3, 5, 7\}$. Let $E = \{A_1, A_2, A_3\}$. Determine

$$\bigcup_{A \in E} A$$

2. What is the additive inverse of $(2)_{\mathbf{R}^\star}$ in the field $\mathbf{R}^\star$?

3. What is the supremum of the following subset of $\mathbf{R}^\star$?

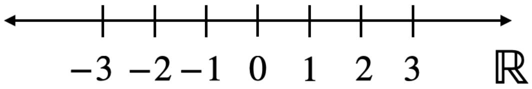
$$E = \{\{x \in \mathbb{Q} \mid x < -1\}, \{x \in \mathbb{Q} \mid x < 1\}, \{x \in \mathbb{Q} \mid x < 0\}\}$$

Test Your Understanding - Answers

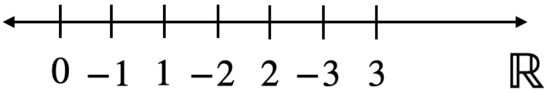
1. $\bigcup_{A \in E} A = \{1, 2, 3, 5, 7\}$
2. $-(2)_{\mathbf{R}^\star} = \{x \in \mathbb{Q} \mid x < -2\} = (-2)_{\mathbf{R}^\star}$
3. $\sup E = (1)_{\mathbf{R}^\star}$

3.2 Properties of (Subsets of) the Real Numbers

With our new found confidence in the existence of the real numbers, let us look at a familiar object with new eyes. Consider the real line.



Let us attempt to forge a connection between the definition of the real numbers as a set of axioms and our mental picture of the real numbers as occupying their expected position on the real line. For example, how do we know that all of the additive inverses of the natural numbers are on the opposite side of the natural numbers? Is there anything in our definition of the real numbers that tells us that the following picture is impossible?



Indeed if the real numbers look the way we expect, then we ought to be able to use the Real Number Axioms to prove the following *obvious* facts.

Theorem 3.12. *For all $x, y \in \mathbb{R}$, if $x < y$, then $-y < -x$.*

Proof. Let $x, y \in \mathbb{R}$ so that $x < y$.

1. By Axiom (AO) we have $x + (-x) < y + (-x)$.
2. By Axiom (A5) we have $0 < y + (-x)$.
3. By Axiom (AO) we have $0 + (-y) < (y + (-x)) + (-y)$.
4. By Axioms (A2), (A3), and (A5) we have $-y < (y + (-y)) + (-x)$.
5. By Axiom (A5) we have $-y < 0 + (-x)$.
6. By Axiom (A4) we have $-y < -x$.

□

Theorem 3.13. *For all $x \in \mathbb{R}$, if $x > 0$, then $-x < 0$*

Proof. Let $x \in \mathbb{R}$ and so $x > 0$. By Theorem 3.12 we have $-x < -0$. By definition of multiplicative inverse, we have $0 + (-0) = 0$. Since addition is commutative we have $-0 + 0 = 0$. Therefore -0 satisfies the property of the additive identity. In other words $-0 = 0$. Therefore $-x < 0$. □

These two theorems confirm to us that the real line looks the way we expect: once we lay out the positive real numbers as proceeding in order off to the right, then necessarily the negative real numbers proceed in the expected order off to the left. By making a connection between the Real Number Axioms and the geometry of the real line, we can use the Real Number Axioms to prove statements about the real numbers that are *obvious* when we look at the real line. In particular, in this section we look at how we can use the Real Number Axioms to prove some *obvious* facts about inequalities, intervals and absolute values.

Aside. *The two proofs above are written in two very different styles. Which one is easier for you to read?*

Intervals and Inequalities on the Real Line Throughout lots of the work we have done in Calculus, we have considered sequences of inequalities:

$$a < b < c$$

This notation is shorthand for the following statement

$$a < b \text{ and } b < c$$

Our experiences with the real line tell us that we can't find two points that are *next to* each other. In other words, given a pair of points on the real line, we can always find another point in between them. But, of course, there is nothing in our axioms that asserts this is true. And so we prove that for any $x, y \in \mathbb{R}$ with $x < y$ the following inequalities hold:

$$x < \frac{x+y}{2} < y$$

Theorem 3.14. *Let $x, y \in \mathbb{R}$. If $x < y$, then $x < \frac{x+y}{2}$ and $x < \frac{x+y}{2} < y$*

Proof. Let $x, y \in \mathbb{R}$ so that $x < y$. Since addition of a constant preserves inequality, we have $x+y < y+y$. Since multiplication distributes over addition we have $x+y < (1+1)y$. Since $1+1 = 2$, we have $x+y < 2y$. By Theorem 3.8, we have $\frac{x+y}{2} < y$.

A similar argument shows $x < \frac{x+y}{2}$ □

Theorem 3.14 tells us that between any two distinct real numbers we can find another real number. Notice that this property is not true when we swap $\mathbb{R}$ for $\mathbb{Z}$.

This property is commonly referred to as *density*. We say that real numbers are dense because between any two elements one can find a third element. Informally, density tells is that intervals have no gaps: we cannot find a *hole* between a pair of real numbers. And so continuous intervals of real numbers have no gaps. Let us take a moment to remind ourselves about the notation we use for intervals.

Definition 3.15. *Let $a, b \in \mathbb{R}$ with $a < b$. The closed interval from a to b is the set $\{x \in \mathbb{R} \mid a \leq x \leq b\}$. The closed interval from a to b is denoted as $[a, b]$. That is*

$$[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$$

Definition 3.16. *Let $a, b \in \mathbb{R}$ with $a < b$. The open interval from a to b is the set $\{x \in \mathbb{R} \mid a < x < b\}$. The open interval from a to b is denoted as (a, b). That is*

$$(a, b) = \{x \in \mathbb{R} \mid a < x < b\}$$

Recalling our terminology from Section 2.2 perhaps we are convinced of the following statements:

1. $\inf(a, b) = \inf[a, b] = a$
2. $\sup(a, b) = \sup[a, b] = b$

The flexibility of our notation for open and closed intervals allows us to make sense of notation that refers to intervals that are neither open or closed as defined above.

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

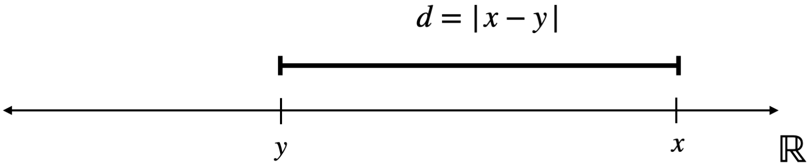
Absolute Value Recall from somewhere[4] the following notation

Definition 3.17. *Let $x \in \mathbb{R}$. We define $|x|$ so that*

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

This definition is perfectly reasonable, but doesn't really get at the usefulness of absolute value notation. We can relate absolute value and the geometry of the real line by thinking about lengths. And so, consider the following alternate definition of absolute value notation.

Definition. *Let $x, y \in \mathbb{R}$. The notation $|x - y|$ refers to the distance between x and y on the real line. When $y = 0$ we write $|x|$ in place of $|x - 0|$.*



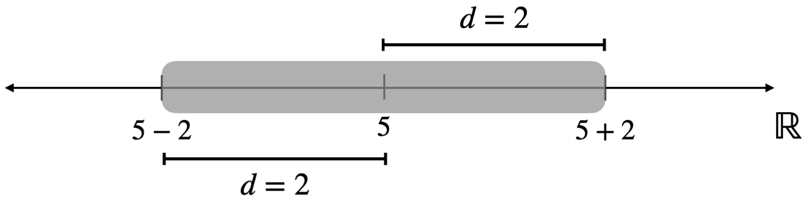
[4] I have no idea at what age students learn about absolute value.

Interpreting value as distances helps when we consider inequalities with absolute values. Consider the inequality:

$$|x-5|<2$$

Interpreting this absolute value as a distance, in *solving* this inequality we are asking:

For which values of x is the distance between x and 5 less than 2?



Consider now the inequality

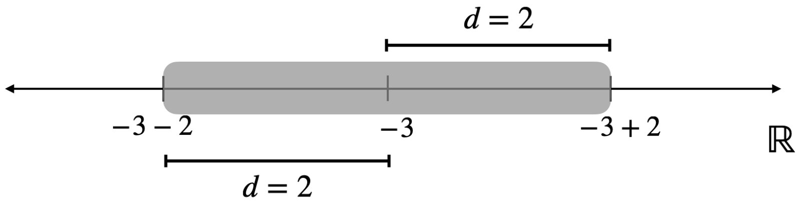
$$|x+3|<2$$

We can re-write this inequality as

$$|x-(-3)|<2$$

And so in solving this inequality we are asking

For which values of x is the distance between x and -3 less than 2?

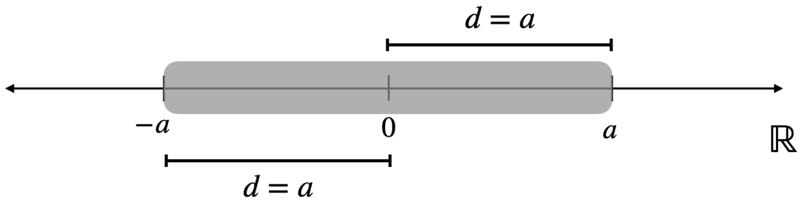


Consider now the inequality

$$|x|\leq a$$

With our definition of absolute value as distance, in solving this inequality we are asking:

For which values of x is the distance between x and 0 no more than a.



This solution will be useful to us in future discussions on *distance*. And so we immortalize this solution with a theorem:

Theorem 3.18. *Let $x \in \mathbb{R}$ and $a \in \mathbb{R}^+$. We have $|x| \leq a$ if and only if $-a \leq x$ and $x \leq a$*

Unfortunately, our definition of absolute value based on distance isn't much help in proving theorems. As we have not formally defined the meaning of the word *distance* we cannot use this interpretation of absolute value in our proofs. In proving statements about absolute value we must appeal to the definition given in Definition 3.17.

We see this in action by using the Real Number Axioms to prove the Triangle Inequality:

Theorem 3.19 (The Triangle Inequality)**.** *For every $x, y \in \mathbb{R}$ we have*

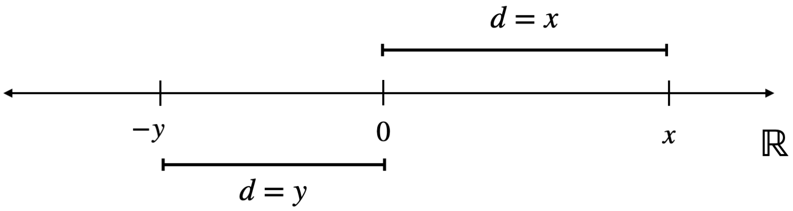
$$|x+y| \leq |x| + |y|$$

Before we dive in to our proof, let us take a moment to conceptualize the Triangle Inequality using our interpretation of absolute value as distance. Notice $|x+y| = |x-(-y)|$. The statement

$$|x-(-y)| \leq |x| + |y|$$

asserts the following:

The sum of the distance from 0 to x and 0 to y is no more than the distance between x and $-y$



(Caution: This picture is misleading! Certainly it is possible to have $x < 0$ or $y < 0$)

When we consider arbitrary values of x and y in the statement of the Triangle Inequality either $x+y \geq 0$ or $x+y < 0$. We proceed in our proof by showing that Triangle Inequality holds in each of these two cases. Since the Triangle Inequality holds in each of these cases, the Triangle Inequality must hold in every case.

Proof of Triangle Inequality. Let $x, y \in \mathbb{R}$. We proceed in cases based on whether $x+y$ is negative.

Case 1: $x+y \geq 0$:
If $x+y \geq 0$ then $|x+y| = x+y$. By definition of absolute value, we have $y \leq |y|$ and $x \leq |x|$ And so by Axiom (AO) we have $x+y \leq x+|y|$. Similarly, since $x \leq |x|$, we have $x+|y| \leq |x|+|y|$. Therefore

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

By Axiom (O2) we have $|x+y| \leq |x|+|y|$.

Case 2: $\underline{x + y < 0}$
If $x + y < 0$, then $|x + y| = -(x + y)$. By Axiom (D) we have $-(x + y) = (-x) + (-y)$. By the definition of absolute value, we have $-x \leq |x|$ and $-y \leq |y|$. Proceeding as in Cast 1, we conclude

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

By Axiom (O2) we have $|x + y| \leq |x| + |y|$. □

Before we leave our work on the Real Numbers, let us take a moment to recall an observation from Section 2:

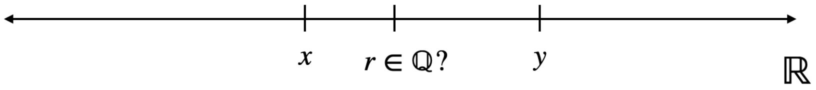
There is at least one rational number between every rational number and every irrational number.

Theorem 3.14 almost confirms this, but not quite. If x is rational and y is irrational, then $\frac{x+y}{2}$ is irrational. However, with our work above, we are now are now able to prove this observation.

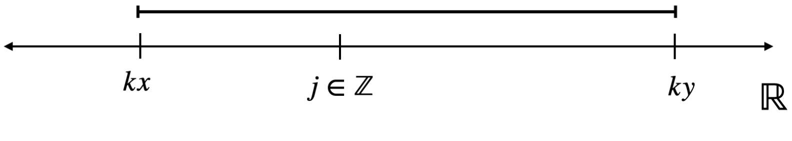
The Archimedian Property Though our goal is to prove that there exists a rational number between every real number and ever rational number, we take a slightly circuitous route. Rather than prove this statement directly, instead we prove that one may find a rational number between any two real numbers.

Theorem 3.20. *For every $x, y \in \mathbb{R}$ with $x < y$ there exist $r \in \mathbb{Q}$ so that $x < r < y$.*

To simplify matters, let us restrict ourselves to the case $x, y > 0$.



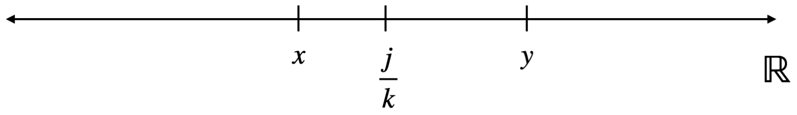
Let k be a large positive integer. Intuitively, when we multiply y and x each by k, the resulting values kx and ky get far apart on the number line. In particular, given y and x, perhaps we can find a value of k so that kx and ky are more than one unit apart.



If one can find an integer somewhere in this gap, then dividing through by k gives the following sequence of inequalities:

$$x < \frac{j}{k} < y$$

We have found a rational value between two elements of $\mathbb{R}$.



Let us try to formalize this argument.

1. Let $x, y \in \mathbb{R}$ so that $0 < x < y$ (premise)
2. There exists $k \in \mathbb{N}$ so that $ky - kx > 1$. (1,???)
3. There exists $j \in \mathbb{N}$ so that $kx < j < ky$. (2, ???)
4. $x < \frac{j}{k} < y$ (3,Thm 3.8)

To complete our argument we require two facts:

i. If $x, y > 0$ and $x < y$, then there exists $k \in \mathbb{R}$ so that $k(y - x) > 0$

ii. For $x, y \in \mathbb{R}$, if $ky - kx > 1$ then there exists $j \in \mathbb{R}$ so that $kx < j < ky$

It turn out that these two facts follow from a seemingly unrelated statement about the natural numbers.

Theorem 3.21 (Archimedian Principle v1)**.** *The set of Natural Numbers is unbounded above in* $\mathbb{R}$

We proceed in this proof by contradiction. We assume $\mathbb{N} \subsetneq \mathbb{R}$ is bounded above in $\mathbb{R}$.

If $\mathbb{N}$ is bounded above, then the Completeness Axiom tells us that $\mathbb{N}$ has a supremum, α. We find a contradiction by exhibiting a natural number that is larger than α.

Proof. We proceed by contradiction. Suppose $\mathbb{N}$ is bounded above. By the Completeness Axiom, $\mathbb{N}$ has a least upper bound in $\mathbb{R}$. Let $\alpha = \sup \mathbb{N}$. Since $1 \in \mathbb{N}$ we have $\alpha \geq 1$.

Since $\alpha = \sup \mathbb{N}$, and $\alpha/2 < \alpha$, it follows that $\alpha/2$ is not an upper bound for $\mathbb{N}$. Therefore there exists $n \in \mathbb{N}$ such that $\alpha/2 < n$. Thus $\alpha < 2n$. Since $2n \in \mathbb{N}$, it follows that α is not an upper bound for $\mathbb{N}$. This contradicts $\alpha = \sup \mathbb{N}$. □

The two facts that we want to prove follow from statements that are equivalent to the Archimedian Principle v1.

Theorem 3.22 (Archimedian Principle v2)**.** *For every* $x \in \mathbb{R}^+$*, there exists* $n \in \mathbb{N}$ *so that* $n - 1 \leq x \leq n$

Theorem 3.23 (Archimedian Principle v3)**.** *If* $x, y \in \mathbb{R}^+$*, then there exists* $n \in \mathbb{N}$ *so that* $y < nx$.

In other words, the following is true

$$\begin{tikzcd}
& \text{Archimedian Principle v1} \arrow[dl, Leftrightarrow] \arrow[dr, Leftrightarrow] & \\
\text{Archimedian Principle v2} \arrow[rr, Leftrightarrow] & & \text{Archimedian Principle v3}
\end{tikzcd}$$

Before we leave this section, let us take a moment to see how the v2 and v3 of the Archimedian Principle relate to v1.

Consider first v2. Since $\mathbb{N}$ is unbounded, then x is not an upper bound of $\mathbb{N}$. In other words, the following set is not empty

$$\{m \in \mathbb{N} \mid x \leq m\}$$

Choosing the smallest element of this set gives n so that $n - 1 \leq x \leq n$.

Consider now v3. Dividing through by x yields

$$\frac{y}{x} < n$$

Let $t = y/x$.

$$t < n$$

To verify v3 we must show that for every $t \in \mathbb{R}$ there exists $n \in \mathbb{N}$ so that $n > t$. In other words, t is not an upper bound for $\mathbb{N}$. Such a statement follows directly from v1.

Using the three versions of the Archimedean Principle, one can prove i. and ii. above. However, these notes are already long enough and we are all tired. So let us omit these proofs and conclude our work in this section.

Test Your Understanding

1. Give the solutions to the following inequalities in terms of intervals
   (a) $|1 + 2x| \leq 4$
   (b) $|x + 2| \geq 5$
   (c) $|x - 5| < |x + 1|$
2. Use the Archimedian Principle to prove the following statement.

$$(\forall \epsilon > 0) \ \ [(\exists n \in \mathbb{N}) \ 1/n < \epsilon]$$

Test Your Understanding - Answers

1. (a) $x \in [-\frac{5}{2}, \frac{3}{2}]$
   (b) $x \in (-\infty, -7] \cup [3, \infty)$
   (c) $x \in (2, \infty)$
2. Proceed using v3 with $y = 1$ and $x = \epsilon$.

4 Functions and Sequences

Consider the Pingala Sequence[1]

$$0,\ 1,\ 1,\ 2,\ 3,\ 5,\ 8,\ \dots$$

Following the first two terms in the sequence, subsequent terms come from taking the sum of the previous two terms.

$$f_{k+1} = f_k + f_{k-1}$$

The terms of this sequence seem to increase without bound. For example, $f_{50} = 12586269025$. Perhaps then it is reasonable to write

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

On the other hand, consider the sequence

$$(g_n) = \left(1,\ \frac{1}{2},\ \frac{1}{4},\ \frac{1}{8},\ \frac{1}{16},\ \dots\right)$$

In this case, perhaps it is reasonable to write:

$$\lim_{n\to\infty} g_n = 0$$

However, our learning from the first part of this course should make us pause before writing such things. If we want to communicate mathematical ideas with one another, then we must first agree on the meaning of our mathematical notation. Further, our definitions must provide us with a method to prove that mathematical objects in fact satisfy the definition. We can only prove that (g_n) converges to 0 once we have a definition for the phrase converges to 0.

Our main goal in Section 4 of the course is to develop tools to let us study the long-term behaviour of sequences. To do so we must first agree on what we mean when we use the word *sequence*. To come to such an agreement on terminology, surprisingly we go back to our discussion about ordered sets.

[1] more commonly referred to as the Fibonacci Sequence

4.1 Functions and Relations

Back in our study of orderings on sets we considered the following definition for order:

Definition. *Let S be a set. An* <u>*order*</u> *on S is a relation on S, denoted $<$, so that the following two statements are true:*

(O1) for $x, y \in S$ exactly one of the following is true:

- $x < y$
- $y < x$
- $x = y$

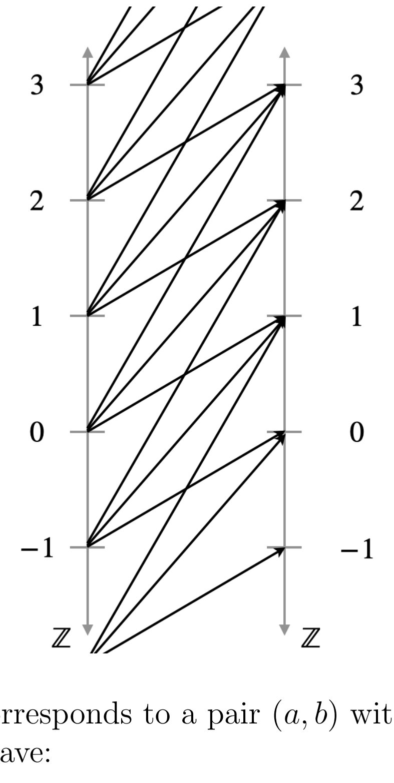
(O2) for $x, y, z \in S$, if $x < y$ and $y < z$, then $x < z$.

The first part of this definition may strike us as unsatisfying:

An <u>*order*</u> *on S is a relation...*

Throughout the course we have leaned on definitions as a strategy to structure our proofs. For example, to prove β is a supremum we must prove that β satisfies all parts of the definition of supremum. So, then, how can we prove something is an order? For us to agree on the meaning of *order* we must first agree on the definition of *relation*.

To gain some intuition here, let us think about our usual ordering of elements of $\mathbb{Z}$. Though rather than our usual picture, let us instead draw a slightly different picture:



In this picture every arrow corresponds to a pair (a, b) with $a < b$. If we imagine the set of all such arrows/pairs, we have:

$$\{(a, b) \in \mathbb{Z}^2 \mid a < b\}$$

If we could find a way to express $a < b$ without appealing to our prior knowledge of the ordering on $\mathbb{Z}$, we could use this set to define the ordering on $\mathbb{Z}$. Notice we have $a < b$ if and only if $b - a$ is a positive integer. Consider the following set

$$O_{\mathbb{Z}} = \{(a, b) \in \mathbb{Z}^2 \mid b - a \in \mathbb{N}^+\}$$

Every element of this set is a pair of elements (a, b) so that $a < b$.

It turns out that a set of pairs let us encode relational data in other circumstances.

Think for a moment about Twitter. As a company, Twitter is an amoral behemoth, but as a social network, Twitter must exist in some form in some database. One way to encode the Twitter network is as a set of pairs

$$T = \{(u, v) \mid \text{ user } u \text{ follows user } v\}$$

Let U be the set of all Twitter users. Notice

$$T \subseteq U \times U$$

Looking back at our ordering example we have

$$O_{\mathbb{Z}} \subseteq \mathbb{Z} \times \mathbb{Z}$$

Relational data can be encoded using a set by taking a subset of a Cartesian product.

One last example before we proceed with our formal definitions. Let S be the set of subjects offered at the University of Melbourne and let P be the set of students enrolled at the University of Melbourne. Deep inside a university database lives the set

$$E = \{(p, s) \in P \times S \mid \text{student } p \text{ is enrolled in class } s\} \subseteq P \times S$$

Each of $O_{\mathbb{Z}}$, T and E are sets that encode structure on a collection of objects. The set $O_{\mathbb{Z}}$ encodes the structure of ordering on the integers. The set T encodes the network structure of Twitter. The set E encodes the enrollment for students at the University of Melbourne.

Definition 4.1. *Let A, B and R be sets. We say R is a relation on A and B when R is a subset of $A \times B$.*

Each of $O_{\mathbb{Z}}$, T and E are relations.

It turns out that we can use the idea of relations to give a useful definition for the word function. Our intuition around function is likely some sense of "assigning". That is, a function assigns a value (an output) to each input. A function, then, in some sense, can be considered to be a set of pairs: the input and the assigned output.

Consider the following set

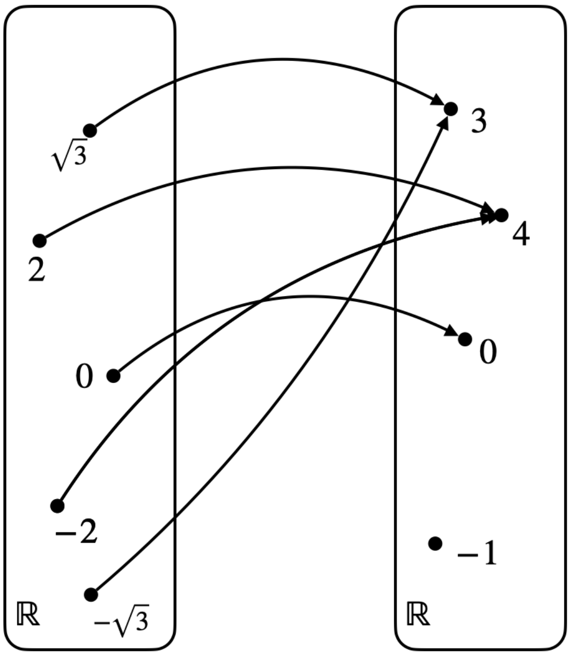
$$A = \{(z, z^2) \mid z \in \mathbb{R}\}$$

Using our notation from our discussions on set theory we notice $A \subseteq \mathbb{R} \times \mathbb{R}$. Recalling our definition of the notation $\subseteq$ and the notation $\times$ we should be convinced that every element of A is an ordered pair whose first entry is from $\mathbb{R}$ and whose second entry is from $\mathbb{R}$.

The set A has many elements, let us examine some of them.

- $(0, 0) \in A$ as $0 \in \mathbb{R}$ and $0 = 0^2$
- $(-3, 9) \in A$ as $-3 \in \mathbb{R}$ and $9 = (-3)^2$
- $\left(\sqrt{2.1}, 2.1\right) \in A$ as $\sqrt{2.1} \in \mathbb{R}$ and $2.1 = \left(\sqrt{2.1}\right)^2$

Just as we did when we drew a picture of $O_{\mathbb{Z}}$, we can draw a picture of A to build some intuition.



With not too much thought, we should be able to convince ourselves that the elements of A are exactly the points of the parabola $f(x) = x^2$. Similarly, we can construct such a set for any function. For example consider the function $g(x) = \sin(x)$. The set of points of this curve are exactly the points in the set

$$\{(x, \sin(x)) \mid x \in \mathbb{R}\}$$

Let us consider another example. Let S be the set of students registered in MAST20026. Let $G = \{H1, H2A, H2B, H3, P, N\}$. Consider the set

$$R = \{(s, g) \in S \times G \mid \text{ student } s \text{ gets grade } g \text{ in MAST20026}\}$$

Aside. *... this set doesn't exist yet. Your grade in the course is far from determined! If you are worried this course is not going well for us, there are lots of resources available* *to you! Feel welcome, at any time, to make an appointment to chat with your instructor about how the course is going and what you can do to be sure you end up meeting your goals in the course.*

Much like our set representation for $f(x) = x^2$ we have a set of ordered pairs in this set. In this case the ordered pair has a student as its first entry and that student's grade as the second entry. Much like our function $f(x) = x^2$ we can consider the the second entry in the ordered pair to be "assigned" to the first entry.

With these examples in mind, we define a function as follows.

Definition 4.2. *Let A, B be sets and let f be a relation on A and B. We say $\underline{f \text{ is a function from } A \text{ to } B}$ when each element of A appears as the first entry of an ordered pair exactly once. When f is a function from A to B we write $f : A \to B$. We say that A is the $\underline{\text{domain of } f}$ and that B is the $\underline{\text{codomain of } f}$*

The ordered pairs in f tell us which element of the codomain is assigned to each element of domain. To ensure that each element of A is assigned exactly one element of the codomain, we require that every element of A appears as the first entry of an ordered pair exactly once. Notice that we have no restriction on how many times an element of the codomain can appear? For $f(x) = x^2$ we are okay with having $(2, 4) \in f$ and $(-2, 4) \in f$.

Quite likely we understand the notation $f(2) = f(-2) = 4$. Put in terms of our new definition of a function, this is equivalent to saying $(2, 4) \in f$ and $(-2, 4) \in f$. As every function can be represented as set, we can extend our use of this piece of notation for any function.

Our experience with functions have equipped us with a wide vocabulary of words/notation we can use to describe various related concepts. With our definition of function above, we can give formal definitions for many concepts with which we are already familiar.

Definition 4.3. *Let A and B be sets and let f be a function from A to B. For $(a, b) \in f$ we say $\underline{b \text{ is the image of } a}$ and $\underline{a \text{ is a pre-image of } b}$. When b is the image of a we write $f(a) = b$.*

Definition 4.4. *Let A and B be sets, let f be a function from A to B and let b be an element of B. The $\underline{\text{pre-image of } b}$ is the set $\{a \in A | f(a) = b\}$. We denote this set as $f^{-1}(b)$.*

For example, if $(\textit{Emily Reeve}, \textit{H2A}) \in R$, then we would write $R(\textit{Emily Reeve}) = \textit{H2A}$. The set $R^{-1}(\textit{H2A})$ is the set of all students whose grade is $\textit{H2A}$.

Aside. *Look at all of the things that we need to understand before we can fully understand the meaning of the notation $f^{-1}(b)$. We need to know what a set is. We need to know the definition of a function. We need to know about set builder notation. We need to know the meaning of the symbol $f(a)$.*

We think of functions and related things to be relatively straight-forward ideas and they are. But trying to precisely define these concepts takes work! However, there is a payoff: we are developing a shared vocabulary. What is more, you are sharing this vocabulary with *fellow students and mathematicians throughout the (western) world! The standardization of these notations is a relatively recent occurrence in the history of mathematics. It is only in the last hundred years or so that these notations have become standard. In fact, it isn't really true to say that all of this notation is standard. For example, just as some people exclude* 0 *from* $\mathbb{N}$*, so too do some use the symbol* $\subset$ *to mean proper subset. There is nothing inherently correct or incorrect about these choices. They are just standards that make it easier for us to communicate.*

When we read a definition that we don't understand, a usual reason is that we don't yet understand all of the meanings of the words in the definition. As we grapple with new definitions we continuously need to backtrack at times to remember the prescribed meanings of all of the words in our new definition.

Definition 4.5. *Let A and B be sets and let f be a function from A to B. The range of f, denoted $range(f)$, is the set of elements of B that are the image of some element of A. That is, we have*

$$range(f) = \{f(a) \in B | a \in A\}$$

Back to our familiar example of $f = \{(z, z^2) \mid |z \in \mathbb{R}\}$. We claim $range(f) = \mathbb{R}_{\geq 0}$.

Consider the statement

$$range(f) = \mathbb{R}_{\geq 0}.$$

This is a statement about set equality[2]. This statement means that these two sets have the elements. To prove this we would have to prove two things:

$$range(f) \subseteq \mathbb{R}_{\geq 0} \text{ and } \mathbb{R}_{\geq 0} \subseteq range(f)$$

Let $x \in range(f)$. Therefore there exists $z \in \mathbb{R}$ so that $f(z) = x$. Therefore $z^2 = x$. Therefore $x \geq 0$. And so $x \in \mathbb{R}_{\geq 0}$. Since every element of $range(f)$ is an element of $\mathbb{R}_{\geq 0}$, we have $range(f) \subseteq \mathbb{R}_{\geq 0}$.

Let $x \in \mathbb{R}_{\geq 0}$. Therefore $x \geq 0$. Therefore $f(\sqrt{x}) = x$. Since $\sqrt{x} \in \mathbb{R}$, we have $x \in range(f)$. Since every element of $\mathbb{R}_{\geq 0}$ is an element of $range(f)$, we have $\mathbb{R}_{\geq 0} \subseteq range(f)$.

Therefore $range(f) = \mathbb{R}_{\geq 0}$.

Aside. *Are you starting to notice a pattern for set equality proofs? There are always two parts. The premise of the two parts are always the same. The conclusions of the two parts are always the same.*

[2] Phew! It is a good thing we have the Set Axioms to tell us about set equality!

4.2 Sequences and Convergence

Continuing our discussion about functions, let us restrict our attention to functions $f : \mathbb{N}^+ \to \mathbb{R}$. As we have a natural ordering for $\mathbb{N}^+$ we can enumerate the images of the elements of $\mathbb{N}^+$ in order:

$$f(1), f(2), f(3), f(4), \dots$$

For example, let $f(n) = n^2 + 1$. Our enumeration becomes

$$2, 5, 10, 17, \dots$$

A function whose domain in $\mathbb{N}^+$ is a sequence!

Definition 4.6. *Let A and B be sets and let f be a function from A to B. We say $\underline{f \text{ is a sequence}}$ when $A = \mathbb{N}^+$ and $B = \mathbb{R}$.*

One can denote a sequence in a number of ways. One traditionally denotes the elements of a sequence as

$$(f_n) = (f_1, f_2, f_3, \dots)$$

rather than

$$f(1), f(2), f(3), \dots$$

In returning to f from above we have $f_1 = 2, f_2 = 5, f_3 = 10$, etc...

One of the main focuses of Real Analysis is examining sequences that approximate some real number L. For example, consider the following sequence (f_n)

$$(f_n) = (3,\ 3.1,\ 3.14,\ 3.142,\ 3.1416,\ 3.14159,\ \dots)$$

Each of entries in the sequence is an approximation of π. And subsequent entries in the sequence are better approximations than previous ones. Let us take a moment to think about why such a sequence may be useful.

Mathematical tools permit us to build mathematical models of real-world physics. However, in the real world our precision for measurement and construction is limited by tools at hand. And so in construction projects, one cannot expect complete precision. Even ignoring the physical realities of engineering projects, as π is an irrational number, it cannot be easily be fully stored in computer memory with complete precision. Sequences like the one above allow us to estimate π to within whatever precision we would like. For example, if our computations need to be accurate to the third decimal place[3], then we can choose 3.1415 as our representation for π.

Our intuition for the decimal digits of π together with our experiences in Calculus perhaps permits us to make meaning of the following piece of notation

$$(f_n) \to \pi$$

This piece of notation is meant to describe the *long term* behaviour of the sequence (f_n).

[3] Somewhere a physicist is crying. This explanation of *significant figures* leaves a lot to be desired.

But of course, this sort of behaviour (i.e., *approaching* a value) is not the only possibility for the long-term behaviour of a sequence. Consider the sequence

$$(-1, 1, -1, 1, -1, 1, \dots)$$

This sequence does not *approach* any particular value. Nor does the sequence

$$(2, 4, 6, 8, 10, \dots)$$

For each of these sequences it is easy to tell at a glance the long-term behaviour of the sequence. However, for most sequences, this is not the case. For example, consider the sequence $f_n = 4 + (-1)^n \frac{2}{\sqrt{n}}$. At this point we cannot *prove* anything about this sequence because as of yet we do not have any definitions in place for words like *converge.*

Aside. *There is a notational subtlety to notice here.* $f_n = 4 + (-1)^n \frac{2}{\sqrt{n}}$ *is not a sequence in and of itself. The statement* $f_n = 4 + (-1)^n \frac{2}{\sqrt{n}}$ *is a formula that tells us how to compute* $f(n)$ *for any* $n \in \mathbb{N}^+$*. We use the notation* (f_n) *to refer to the sequence in its entirety.*

To build some intuition about what *convergence* ought to mean, let us take a closer look at the terms of this sequence

$$(f_n) = \left(4 - \frac{2}{\sqrt{1}},\ \ 4 + \frac{2}{\sqrt{2}},\ \ 4 - \frac{2}{\sqrt{3}},\ \ 4 + \frac{2}{\sqrt{4}},\ \ 4 - \frac{2}{\sqrt{5}},\ \ \dots\right)$$

Converting these square roots terms to rounded decimals we have

$$(2,\ \ 5.414,\ \ 2.845,\ \ 5,\ \ 3.106,\ \ 4,816,\ \ 3.244,\ \ 4.707, \dots)$$

The terms seem to be oscillating on either side of 4, getting closer with each subsequent term. Let us compute some larger values to see how close this sequence gets to 4.

$$\begin{aligned}
f_{100} &= 4 + \frac{2}{\sqrt{10}} \approx 4.189 \\
f_{1000} &= 4 + \frac{2}{\sqrt{1000}} \approx 4.0632 \\
f_{10000} &= 4 + \frac{2}{\sqrt{1000}} \approx 4.02 \\
f_{100000} &= 4 + \frac{2}{\sqrt{100000}} \approx 4.006 \\
f_{1000000} &= 4 + \frac{2}{\sqrt{1000000}} \approx 4.002
\end{aligned}$$

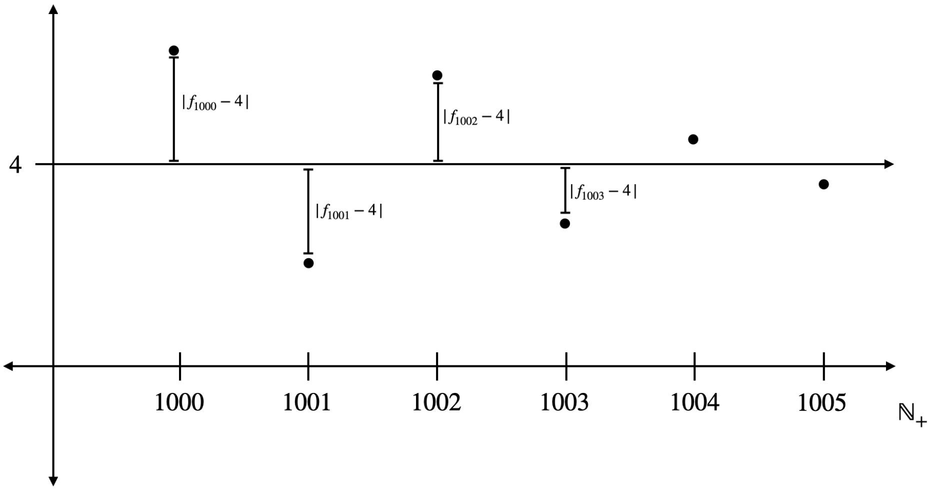
On Assignment 1 we looked at the limit of $f(x)$ as x goes to infinity. We had the following definition:

Definition. *Let $f : \mathbb{R} \to \mathbb{R}$ be a function. We say* <u>*the limit of $f(x)$ as x goes to ∞ is ∞*</u> *when for every $r \in \mathbb{R}$ there exists $k \in \mathbb{R}$ so that $f(x) > r$ whenever $x > k$.*

Our intuition for this definition was the following idea

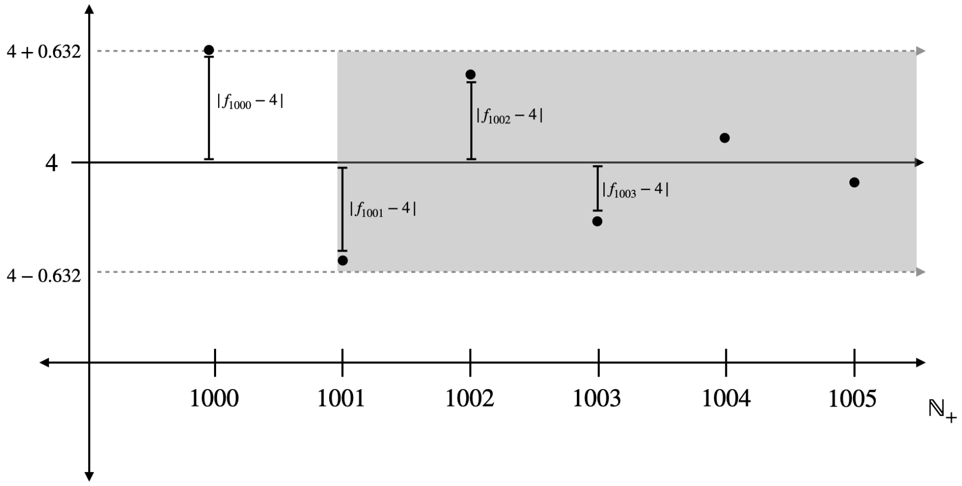
the limit of $f(x)$ as x goes to ∞ is ∞ when $f(x)$ is eventually bigger and stays bigger than any value $r \in \mathbb{R}$

Looking at our sequence above, we see similar behaviour. Except instead of f_n being bigger (and staying bigger) than any value $r \in \mathbb{R}$, here we notice that the distance between f_n and 4 gets smaller (and stays smaller) than any $r \in \mathbb{R}$.



For example,

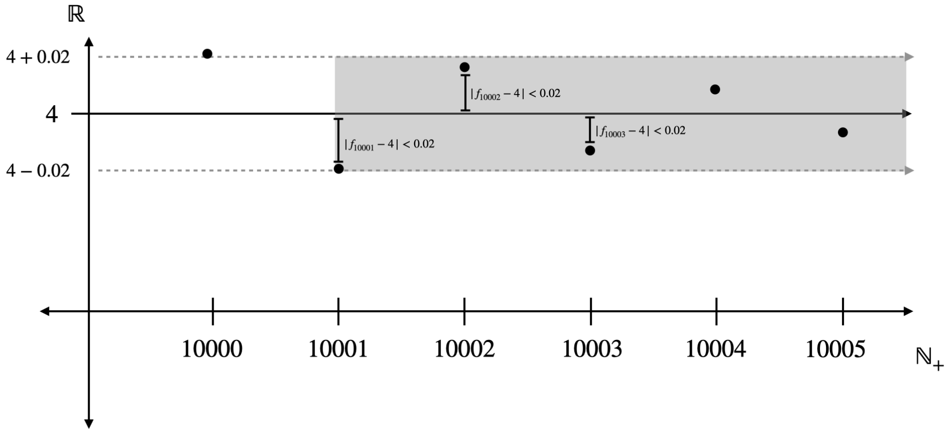
- for all $n > 1000$, we have $|f_n - 4| < 0.0632$.
- for all $n > 10000$, we have $|f_n - 4| < 0.02$
- for all $n > 100000$, we have $|f_n - 4| < 0.006$
- etc...



Consider the following condition $p(n, \epsilon)$

$$|f_n - 4| < \epsilon$$

When $\epsilon = 0.0623$, this condition is true for every $n > 1000$. When $\epsilon = 0.02$, this condition is true for every $n > 10000$.



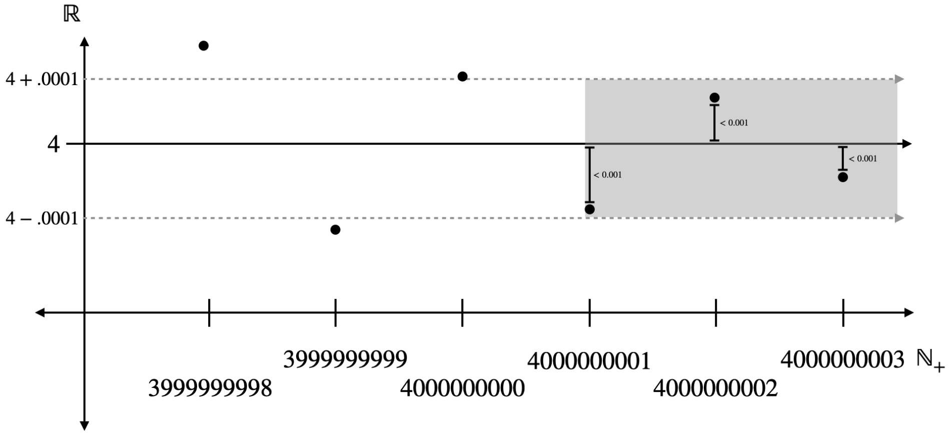
Aside. *The symbol ϵ is the greek letter* epsilon*. There is no reason other than tradition to use ϵ rather than any other variable name. Most of the time when you encounter ϵ in mathematics, limit-type concepts are lurking somewhere nearby.*

As we change the value of ϵ, the bounds on n for which $p(n, \epsilon)$ is true also changes. We saw a similar phenomena for a condition on Assignment 1 that depended on two variables.

For example, consider $\epsilon = 0.0001$. Let us see if we can find a value for M so that

$P(n, .0001)$ is true whenever $n > M$

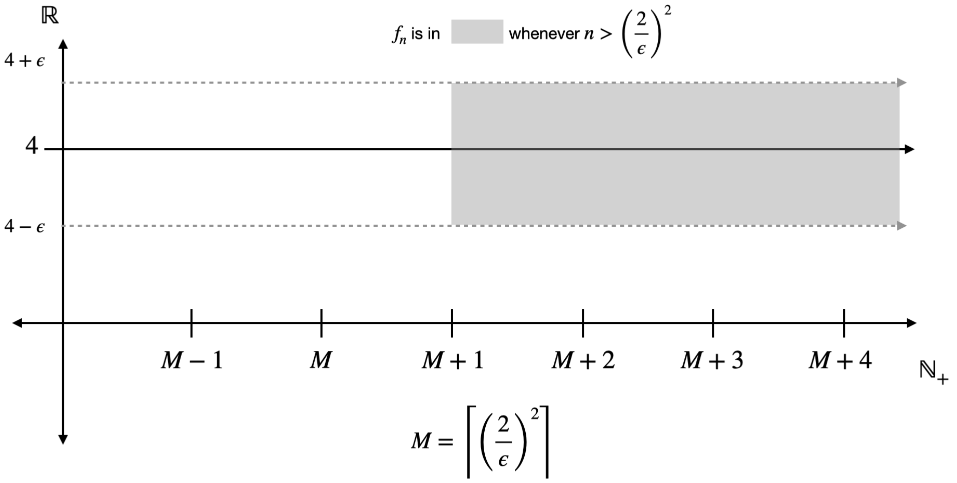
$$\begin{aligned}
|f_n - 4| &< 0.0001 \\
|4 + (-1)^n \frac{2}{\sqrt{n}} - 4| &< 0.0001 \\
\frac{2}{\sqrt{n}} &< 0.0001 \\
\left(\frac{2}{.0001}\right)^2 &< n \\
4000000000 &< n
\end{aligned}$$



When n is greater than $M = 400000000$, we have $|f_n - 4| < 0.0001$. Notice, however, that in our algebra nothing is special about $\epsilon = 0.0001$. Performing this same computation for arbitrary ϵ yields

$$\begin{aligned}
|f_n - 4| &< \epsilon \\
|4 + (-1)^n \frac{2}{\sqrt{n}} - 4| &< \epsilon \\
\frac{2}{\sqrt{n}} &< \epsilon \\
\left(\frac{2}{\epsilon}\right)^2 &< n
\end{aligned}$$

And so given any $\epsilon > 0$, one can find a value M so that the difference between f_n and 4 is less than ϵ whenever $n > M$.


$\mathbb{R}$
f_n is in ▨ whenever $n > \left(\frac{2}{\epsilon}\right)^2$
$4+\epsilon$
4
$4-\epsilon$
$M-1$
M
$M+1$
$M+2$
$M+3$
$M+4$
$\mathbb{N}_+$
$M = \left\lceil \left(\frac{2}{\epsilon}\right)^2 \right\rceil$

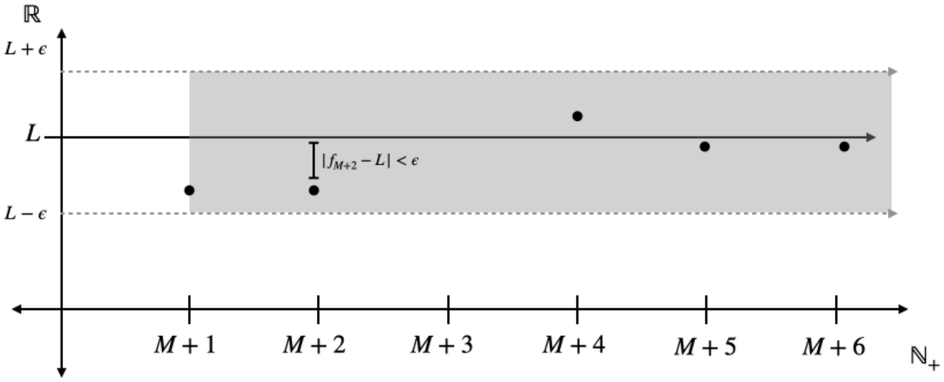
With these ideas in mind, let us define the meaning of *convergence* for sequences.

Let us begin with a slightly squishy intuitive version of our definition.

Definition. *Let f_n be a sequence and let $L \in \mathbb{R}$. We say $\underline{f_n \text{ converges to } L}$ when the distance between f_n and L eventually gets smaller (and stays smaller) than any value $\epsilon > 0$.*

If the difference between f_n and L is to get smaller (and stay smaller) than ϵ then there must exist some $M \in \mathbb{N}^+$ so that the distance between f_n and L is less than ϵ whenever $n > M$. In other words, $|f_n - L|$ is less than ϵ whenever $n > M$.

Definition 4.7. *Let f_n be a sequence and let $L \in \mathbb{R}$. We say $\underline{(f_n) \text{ converges to } L}$ when for every $\epsilon > 0$ there exists $M \in \mathbb{N}^+$ so that $|f_n - L| < \epsilon$ whenever $n > M$. When (f_n) converges to L we write $f_n \to L$ or $\lim\limits_{n\to\infty} f_n = L$.*



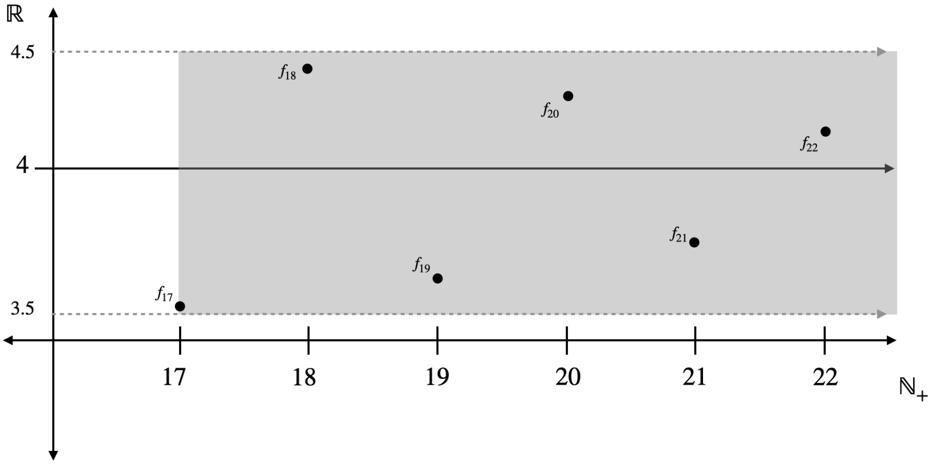
To make a little more sense of this definition, let us return to our example above. Back to our example: $f_n = 4 + (-1)^n \frac{2}{\sqrt{n}}$.

As we saw above, we have $|f_n - 4| < 0.0001$ whenever $n > 4000000000$. Moreover, we can change 0.0001 to any value of $\epsilon > 0$ we wish and produce a value M so that $|f_n - 4| < \epsilon$

whenever $n > M$. For example, for $\epsilon = 0.5$ we can compute

$$\left(\frac{2}{\epsilon}\right)^2 = \left(\frac{2}{0.5}\right)^2 = \frac{4}{.25} = 16$$

Therefore $|f_n - 4| < 0.5$ whenever $n > 16$.



Since we can find an appropriate M for any $\epsilon > 0$, we may write

$$\lim_{n\to\infty} 4 + (-1)^n \frac{2}{\sqrt{n}} = 4$$

Let us consider another example of a sequence that *obviously* converges: $f_n = 1 - 1/n$. As n gets very large, $1/n$ gets very small and so we expect $f_n \to 1$.

Let us see how we can verify this using the definition of the notation $f_n \to 1$.

To be able to write $f_n \to 1$ we must verify that for any $\epsilon > 0$ we can find M so that $|(1 - 1/n) - 1| < \epsilon$ whenever $n > M$. Manipulating this inequality we find

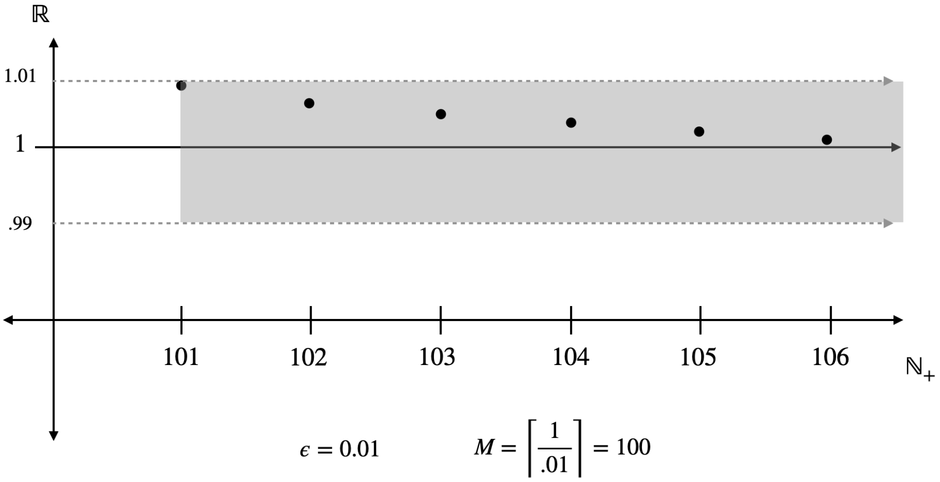
$$\begin{aligned} |(1 - 1/n) - 1| &< \epsilon \\ |-1/n| &< \epsilon \\ 1/n &< \epsilon \\ 1/\epsilon &< n \end{aligned}$$

For example, when $\epsilon = 0.01$, this computation tells us

$$|f_n - 1| < 0.01$$

whenever

$$n > 100 \quad (1/\epsilon = 1/0.1 = 100)$$



Given any particular value of ϵ, we have $|f_n - 1| < \epsilon$ whenever $n > \lceil 1/\epsilon \rceil$. Therefore, for every $\epsilon > 0$ there exists $M \in \mathbb{N}^+$ so that $|f_n - 1| < \epsilon$ whenever $n > M$. And so we conclude $f_n = 1 - 1/n$ converges to 1.

Aside. *Why are we taking $M = \lceil 1/\epsilon \rceil$ rather than $M = 1/\epsilon$? In our pictures we are looking at f_{M+1}, f_{M+2}, etc... Since the the domain of f is $\mathbb{N}^+$ we need each of $M+1, M+2, \ldots$ to be natural numbers. For almost every value of $\epsilon > 0$, the number $1/\epsilon$ is not a natural number. However $\lceil 1/\epsilon \rceil \geq 1/\epsilon$ and $\lceil 1/\epsilon \rceil$ is a natural number. Taking $n > \lceil 1/\epsilon \rceil$ guarantees $n > 1/\epsilon$.*

Using our work above, we could use the Real Number Axioms and related theorems to prove $f_n \to 1$

Proof.

i. Let $\epsilon > 0$ (premise)

ii. Let $M = \lceil 1/\epsilon \rceil$. (premise)

iii. Let $n \in \mathbb{N}^+$ so that $n > M$. (premise)

iv. $n > 1/\epsilon$ (O2)

v. $1/n < \epsilon$ (Thm 3.8)

vi. $|-1/n| < \epsilon$ (iv, defn of Absolute Value)

vii. $|(1 - 1/n) - 1| < \epsilon$ (iv, defn of Absolute Value, A2,A3,A4,A5)

viii. $|f_n - 1| < \epsilon$ (defn of f_n)

ix. For every $\epsilon > 0$ we have $|f_n - 1| < \epsilon$ whenever $n > \lceil 1/\epsilon \rceil$. (i,viii)

x. $f_n \to 1$ (ix, defn of $f_n \to 1$)

□

Alternatively, we could write this proof informally assuming that our reader is familiar with the details of how to manipulate algebraically in $\mathbb{R}$.

Proof. Let $\epsilon > 0$, $M = \lceil 1/\epsilon \rceil$, and $n \in \mathbb{N}^+$ so that $n > M$. By construction we have $n > 1/\epsilon$. Manipulating we find

$$\begin{aligned} 1/n &< \epsilon \\ |-1/n| &< \epsilon \\ |1 - 1/n - 1| &< \epsilon \end{aligned}$$

Therefore $|f_n - 1| < \epsilon$ whenever $n > M$. Therefore f_n converges to 1. In other words, $f_n \to 1$. □

Aside. *A common experience for students studying Real Analysis is the fog of confusion arising from* magic number *proofs. In most resources, one can expect the following format for a sample question and solution:*

Example*: Let $f_n = 1 - 1/n$. Show $f_n \to 1$*

Solution: Let $\epsilon > 0$, $M = \lceil 1/\epsilon \rceil$, and $n \in \mathbb{N}^+$ so that $n > M$.

$$\begin{aligned} 1/n &< \epsilon \\ |-1/n| &< \epsilon \\ |1 - 1/n - 1| &< \epsilon \end{aligned}$$

Therefore $|f_n - 1| < \epsilon$ whenever $n > M$. And so $f_n \to 1$.

Without having done all of the setup work to figure out what M should be, this "Solution" does not help a student learn how to solve a problem. If you are looking at other resources and you see choices for constants in the first line or two of a solution you are right to wonder "How would I know to do that?". The answer is that you wouldn't. There is no choice of M that will work for every sequence f_n. We have to first figure out what a reasonable choice for M is before we can show that the sequence converges.

Now that we have the makings of a grasp on convergence, let us consider a sequence that doesn't seem to converge.

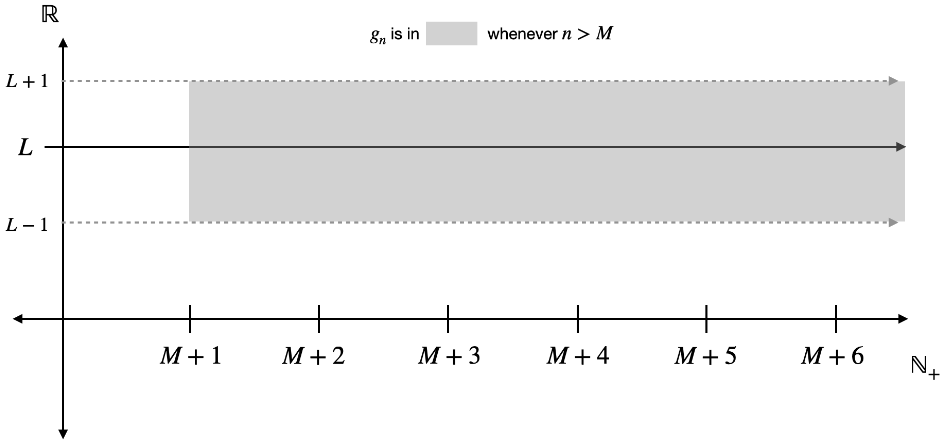
$$(-1, 1, -1, 1, -1, 1, \dots)$$

Let $g_n = (-1)^n$. When we look at the terms of this sequence it doesn't seem as if there is a value L for which the following statement is true

The difference between g_n and L eventually gets smaller (and stays smaller) than any value $\epsilon > 0$.

And so we expect that there is no $L \in \mathbb{R}$ for which g_n converges to L. Let us see if we can verify this by proceeding by contradiction.

Assume there exists $L \in \mathbb{R}$ so that g_n converges to L. By definition, for every $\epsilon > 0$ there exists $M \in \mathbb{N}$ so that $|f_n - L| < \epsilon$ whenever $n > M$. In particular, there exists $M \in \mathbb{N}^+$ so that $|g_n - L| < 1$ whenever $n > M$.



Consider a value $n > M$ so that n is even. Since n is even, we have $g_n = 1$. Therefore

$$|g_n - L| = |1 - L| < 1$$

Since n is even, it must be that $n + 1$ is odd and so $g_{n+1} = -1$. Therefore

$$1 > |g_{n+1} - L| = | - 1 - L| = | - (1 + L)| = |1 + L|$$

Combining these two statements we have

$$|1 - L| + |1 + L| < 2$$

By the Triangle Inequality, we have

$$|1 - L + 1 + L| \leq |1 - L| + |1 + L|$$

Therefore

$$2 = |1 - L + 1 + L| \leq |1 - L| + |1 + L| < 2$$

A contradiction.

For $\epsilon = 1$ there is no value of M for which $|g_n - L| < \epsilon$ whenever $n > M$. Therefore g_n does not converge to L for any $L \in \mathbb{R}$.

As expected, we use the word diverge to describe sequences that don't converge to any value of L.

Definition 4.8. *Let f_n be a sequence. We say f_n diverges when the statement*

$$f_n \text{ converges to } L$$

is false for every $L \in \mathbb{R}$. In other words, we say f_n diverges when for every $L \in \mathbb{R}$ there exists $\epsilon > 0$ so that for every $M \in \mathbb{N}^+$ there exists $n > M$ such that $|f_n - L| \geq \epsilon$.

One way to think about convergence of sequences is as a game between two players. Player 1 is the challenger and player 2 is the responder. At the start of the game the players agree on a value for L. Player 1 calls out any value of ϵ they wish. Player 2 must respond with a value of M for which $|f_n - 1| < \epsilon$ whenever $n > M$. If player 2 cannot respond, then f_n does not converge to L. If player 2 successfully responds, then Player 1 chooses another value for ϵ and the game continues. If Player 2 can respond no matter what value for ϵ Player 1 chooses, then f_n converges to L.

Much like limits of real functions, operations with limits behave as we expect. In the same way that we can define a function as a sum or product of functions, we can define a sequence. For example, if $f_n = n$ and $g_n = 1/n$, then $f_n + g_n$ is the sequence

$$(f_1 + g_1,\ f_2 + g_2, \dots) = (1 + 1,\ 2 + 1/2,\ 3 + 1/3,\ 4 + 1/4, \dots)$$

As we might expect both the sum and product of convergent sequences are convergent.

Theorem 4.9 (Algebra of Limits Theorem)**.** *If $f_n \to \alpha$ and $g_n \to \beta$ are convergent sequences, then*

i. $f_n + g_n \to \alpha + \beta$;

ii. $f_n - g_n \to \alpha - \beta$;

iii. $f_n g_n \to \alpha\beta$; *and*

iv. $f_n/g_n \to \alpha/\beta$ *(provided $\beta \neq 0$ and no term of g_n is 0).*

We prove *i*. Let $f_n \to \alpha$ and $g_n \to \beta$. Let $h_n = f_n + g_n$. We claim

$$h_n \to \alpha + \beta.$$

We use the definition of convergence to verify this fact. To do so, we must show that for every $\epsilon > 0$ there exists $M \in \mathbb{N}^+$ so that $|(f_n + g_n) - (\alpha + \beta)| < \epsilon$ whenever $n > M$.

Consider some fixed value $\epsilon > 0$. And let[4] $\epsilon' = \epsilon/2$. Since $\epsilon' > 0$ and both (f_n) and (g_n) converge, there exists M_f and M_g so that

$$|(f_{n_f}) - \alpha)| < \epsilon' \text{ whenever } n_f > M_f \text{ and } |(g_{n_g}) - \beta)| < \epsilon' \text{ whenever } n_g > M_g$$

Let $M = \max\{M_f, M_g\}$ and let $n \in \mathbb{N}^+$ so that $n > M$. We have

$$|(f_n + g_n) - (\alpha + \beta)| = |f_n - \alpha + g_n - \beta|$$

Applying the triangle inequality yields

$$|f_n - \alpha + g_n - \beta| \leq |f_n - \alpha| + |g_n - \beta|$$

Since $n > M$ and $M \geq M_f$, we have $|f_n - \alpha| < \epsilon'$. Similarly, we have $|g_n - \beta| < \epsilon'$. Therefore

$$|(f_n + g_n) - (\alpha + \beta)| \leq |f_n - \alpha| + |g_n - \beta| < \epsilon' + \epsilon' = \epsilon$$

Therefore for every $\epsilon > 0$ there exists $M \in \mathbb{N}^+$ so that $|(f_n + g_n) - (\alpha + \beta)| < \epsilon$ whenever $n > M$. Therefore $h_n \to \alpha + \beta$.

---

[4] This is a magic number proof. We choose $\epsilon' = \epsilon/2$ because it works.

Test Your Understanding

1. Let $f_n = \frac{2n}{n+1}$. In this exercise we verify $f_n \to 2$.

   (a) Find a value for M so that $p(n, 0.1)$ is true for every $n > M$.

   (b) Using your algebra from the previous part, find a value for M as a function of ϵ so that $p(n, \epsilon)$ is true for every $n > M$.

   (c) Fill in the blanks in the proof below.

   Proof. Let $\epsilon > 0$, $M = \underline{\qquad\quad}$, and $n \in \mathbb{N}^+$ so that $n > M$.

$$\begin{aligned} \underline{\qquad\quad} &< n \\ 2 &< \epsilon(n+1) \\ |-2| &< \epsilon(n+1) \\ |-2+2n-2n| &< \epsilon(n+1) \\ \underline{\qquad\qquad\qquad} &< \epsilon(n+1) \\ \underline{\qquad\qquad\qquad\quad} &< \epsilon \end{aligned}$$

   Therefore $|f_n - 2| < \epsilon$ whenever $n > M$. And so $f_n \to 2$. □

Test Your Understanding - Answers

1. (a) $M > 2/0.1 - 1$.

   (b) $M = \lceil 2/0.1 \rceil - 1$.

   (c) *Proof.* Let $\epsilon > 0$, $M = \underline{\lceil 2/\epsilon \rceil - 1}$, and $n \in \mathbb{N}^+$ so that $n > M$.

$$\begin{aligned}
2/\epsilon - 1 &< n \\
2 &< \epsilon(n+1) \\
|-2| &< \epsilon(n+1) \\
|-2+2n-2n| &< \epsilon(n+1) \\
|2n-2(n+1)| &< \epsilon(n+1) \\
|2n/(n+1)-2| &< \epsilon
\end{aligned}$$

   Therefore $|f_n - 2| < \epsilon$ for all $n > M$. And so $f_n \to 2$. □

4.2.1 Divergence

Our work above that showed $g_n = (-1)^n$ diverges took some cleverness. We had to make a very precise argument using ϵ and L to verify that (g_n) did not satisfy the definition of convergence for any $L \in \mathbb{R}$.

Rather than have to depend on our ability to be clever[5], in this section we develop some criteria we can use to decide that a sequence diverges. For example, our intuition tells us that each of the following sequences does not converge to any L.

i. $(1, 2, 3, 4, 5, \dots)$

ii. $(-1, -2, -3, -4, -5, \dots)$

iii. $(1, 4, 9, 16, 25, 49, \dots)$

iv. $(-2, -4, -8, -16, -32, \dots)$

Equipped only with our $\epsilon - M$ definition of convergence, proving that each of these diverges would be an onerous task. However, the underlying reason why none of these sequences converge is the same: they each eventually get bigger (or smaller) and stay bigger (or smaller) than any $r \in \mathbb{R}$. In other words, they are not *bounded.*

In Section 2 and 3 of the course we spend a lot of time thinking about *boundedness* for sets. Let (f_n) be a sequence and consider the set

$$A = \{f_n \mid n \in \mathbb{N}\}$$

Depending on f_n, this set may or not be bounded (above/below) in $\mathbb{R}$. For example, when $f_n = n$, this set is bounded below but not bounded above. Whereas when $f_n = 4 + (-1)^n \frac{2}{\sqrt{n}}$, this set is bounded both above and below. Looking at our four divergent examples above, none of them are bounded both above and below.

Definition 4.10. *Let (f_n) be a sequence. We say $\underline{f_n \text{ is bounded}}$ when the set $\{f_n \mid n \in \mathbb{N}\}$ is bounded both above and below in $\mathbb{R}$. When (f_n) is not bounded, we say $\underline{(f_n) \text{ is unbounded}}$*

It turns out boundedness plays an important role in convergence.

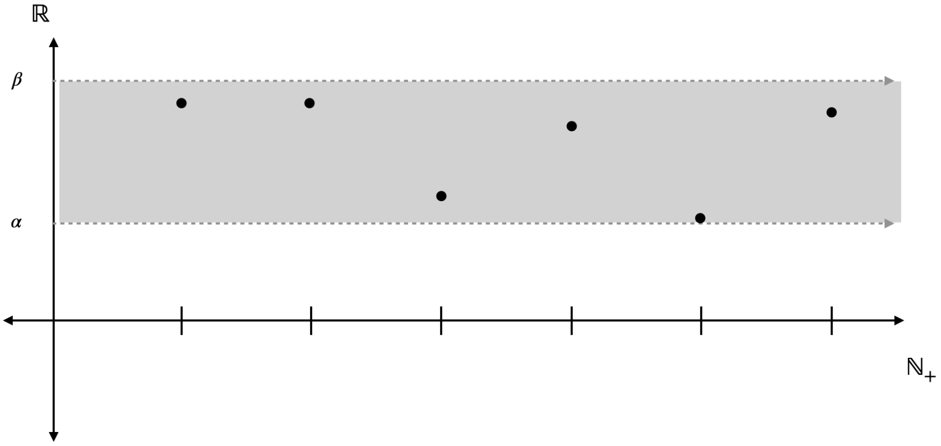
Theorem 4.11. *Let (f_n) be a sequence. If (f_n) is not bounded, then (f_n) does not converge.*

To begin to understand why this theorem is true, let us take a closer look at the concept of boundedness. Let (f_n) be a bounded sequence. Since (f_n) is bounded, the set $\{f_n \mid n \in \mathbb{N}\}$ is bounded above and below in $\mathbb{R}$. In other words, there exists $\alpha, \beta \in \mathbb{R}$ so that for all $n \in \mathbb{N}^+$ we have

$$\alpha \leq f_n \leq \beta$$

Looking at our sequence, the values for f_n are confined to the strip between α and β.

[5] an ability that is in short supply for all of us these past few months

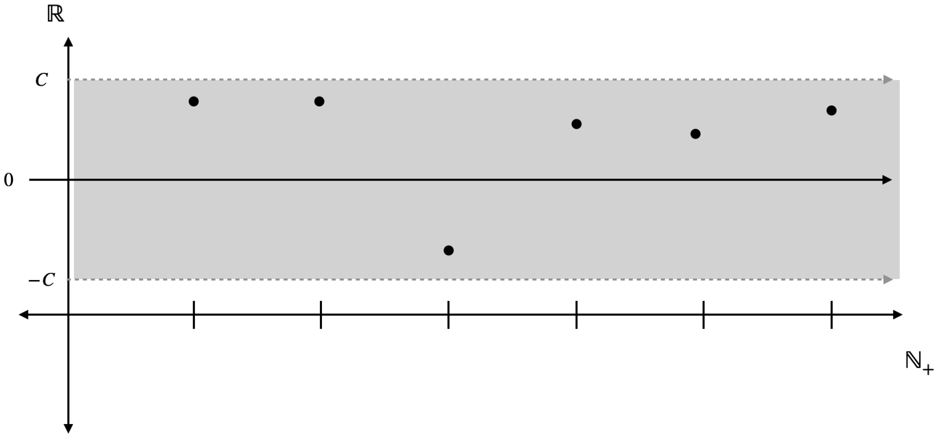


Let $C = \max\{|\alpha|, |\beta|\}$. Using C we can re-write our inequality above:

$$-C \leq f_n \leq C$$

And applying our work on absolute values, we have

$$|f_n| \leq C$$



.

In summary, if (f_n) is bounded, then there exists $C \in \mathbb{R}_{\geq 0}$ so that $|f_n| \leq C$ for all $n \in \mathbb{N}^+$. Consider now the converse of the previous sentence:

> If there exists $C \in \mathbb{R}_{\geq 0}$ so that $|f_n| \leq C$ for all $n \in \mathbb{N}^+$, then (f_n) is bounded.

If $|f_n| \leq C$ for all $n \in \mathbb{N}^+$, then $-C \leq f_n \leq C$ for all $n \in \mathbb{N}^+$. In other words (f_n) is bounded both above and below. The converse is true!

Theorem 4.12. *A sequence (f_n) is bounded if and only if there exists $C \in \mathbb{R}_{\geq 0}$ so that $|f_n| \leq C$ for all $n \in \mathbb{N}^+$.*

With this new characterisation of bounded sequences, let us return to Theorem 4.11. Rather than prove Theorem 4.11 directly, instead we establish the contrapositive.

Theorem 4.13. *Let (f_n) be a sequence. If (f_n) converges, then (f_n) is bounded.*

Proof. Let $(f_n) \to L$. Since (f_n) converges to L, then for any $\epsilon > 0$ there exists $M \in \mathbb{N}^+$ so that $|f_n - L| < \epsilon$ whenever $n > M$. In particular, when $\epsilon = 1$ the following is true for (f_n):

$$\text{there exists } M \in \mathbb{N}^+ \text{ so that } |f_n - L| < 1 \text{ whenever } n > M.$$

Manipulating the inequality $|f_n - L| < 1$ we have

$$L - 1 < f_n < L + 1$$

for all $n > M$. Notice

$$|f_n| \leq \max\{|L-1|, |L+1|\} \tag{1}$$

Let $C = \max\{|f_0|, |f_1|, \dots, |f_M|, |L-1|, |L+1|\}$.

By our choice of r, the following statements are true:

- $|f_0| \leq C, |f_1| \leq C \dots |f_M| \leq C$
- $\max\{|L-1|, |L+1|\} \leq C$.

Combining these two statements with Equation (1) implies $|f_n| \leq C$ for all $n \in \mathbb{N}^+$. In other words, (f_n) is bounded. □

Unfortunately unbounded sequences are not the only flavour of divergent sequence. For example, our sequence

$$(-1, 1, -1, 1, \dots)$$

is certainly bounded, but is not convergent for any L.

One way to look at the sequence $(-1, 1, -1, 1, \dots)$ is an interleaving of two different sequences

i. $(1, 1, 1, 1, 1, \dots)$ and

ii. $(-1, -1, -1, -1, \dots)$

The first of these two sequence converges to 1. The second of these two sequences converges to -1. It turns out that having subsequences that converge to different values is enough to imply that a sequence diverges.

Theorem 4.14. *A sequence (f_n) converges to L if and only if every subsequence of (f_n) converges to L.*

Corollary 4.15. *Let (f_n) be a sequence. If (f_n) has two subsequences each of which converge to a different value, then (f_n) diverges.*

To be able to make sense of these two results, we require a definition for subsequence.

Definition 4.16. *Let (f_n) be a sequence and let*

$$0 < n_1 < n_2 < n_3 \dots$$

be a increasing sequence of integers. Then

$$(f_{n_1}, f_{n_2}, f_{n_3}, \dots)$$

is a *subsequence of (f_n)* *We denote a subsequence of (f_n) as (f_{n_k})*

For example, let $f_n = 1/n$ Consider the increasing sequence

$$1 < 3 < 5 < 7 \cdots$$

The corresponding subsequence is

$$(f_{n_1}, f_{n_2}, f_{n_3}, \dots) = (f_1, f_3, f_5, \dots) = \left(1, \frac{1}{3}, \frac{1}{5}, \frac{1}{7} \cdots\right)$$

The terms of the increasing sequence are the indices of the subsequence.

Returning now to the statement of Theorem 4.14, let us consider a sequence (f_n) that converges to $L \in \mathbb{R}$. By definition, for every $\epsilon > 0$ we can find $M \in \mathbb{N}^+$ so that $|f_k - L| < \epsilon$ whenever $k > M$.

Consider now a subsequence

$$(f_{n_k}) = (f_{n_1}, f_{n_2}, f_{n_3}, f_{n_4}, f_{n_5} \dots)$$

We want to conclude (f_{n_k}) converges to L. And so we are going to have to make some argument about ϵ's and M's.

Our subsequence is built using an increasing sequence of natural numbers:

$$n_1 < n_2 < n_3 \dots$$

Since $n_1 \in \mathbb{N}^+$ we have $n_1 \geq 1$. Since $n_2 > n_1$ and $n_1 \geq 1$, we have $n_2 \geq 2$. Since $n_3 > n_2$ and $n_2 \geq 2$, we have $n_3 \geq 3$.

Using a proof by induction, we could prove

$$n_k \geq k \text{ for every } k \in \mathbb{N}^+$$

It is this observation that is the key to the proof of Theorem 4.14

Proof of Theorem 4.14. Let (f_n) be a sequence.

Assume $(f_n) \to L$. Let (f_{n_k}) be a subsequence of (f_n). We use the definition of convergence to prove $(f_{n_k}) \to L$.

Let $\epsilon > 0$. Since $(f_n) \to L$, there exists $M \in \mathbb{N}^+$ so that $|f_k - L| < \epsilon$ whenever $k > M$. Since $n_k \geq k$ for all $k \in \mathbb{N}^+$, we have $|f_{n_k} - L| < \epsilon$ for all $n_k \geq M$. Therefore $|f_{n_k} - L| < \epsilon$ whenever $k > M$. Therefore, $(f_{n_k}) \to L$.

Assume every subsequence of (f_n) converges to L. Since (f_n) is a subsequence of itself, it follows directly that (f_n) converges to L. □

Theorem 4.14 tells us that any sequence that has two subsequences that converge to different values necessarily diverges. Hiding inside the statement of Theorem 4.14 is the following implication

Every subsequence of a convergent sequence is convergent.

Taking the contrapositive of this observation provides us another divergence criteria.

Corollary 4.17. *Let (f_n) be a sequence. If (f_n) has a subsequence that diverges, then (f_n) diverges.*

For convenience, we combine the statement of our three divergence criteria in to a single theorem.

Theorem 4.18 (Divergence Criteria Theorem)**.** *Let (f_n) be a sequence. If any of the following are true, then (f_n) diverges.*

1. *(f_n) is unbounded.*
2. *(f_n) has two subsequences each of which converge to a different value.*
3. *(f_n) has a subsequence that diverges.*

Before we finish our work in this section, we have one last result to mention. Let us return to our example

$$(f_n) = (-1, 1, -1, 1, \dots)$$

This example convinced us that not every bounded sequence is convergent. For each $n \in \mathbb{N}^+$ we have $|f_n| \leq 1$. And so by Theorem 4.12, the sequence (f_n) is bounded. However, by part ii of the Divergence Criteria Theorem, (f_n) is not convergent. There is a subsequence of (f_n) that converges to 1, and another one that converges to -1. It turns out that hiding inside every bounded sequence (whether convergent or divergent) is a convergent subsequence.

Theorem 4.19 (Bolazano-Weierstrauss Theorem)**.** *Every bounded sequence has a convergent subsequence.*

At this point, such a fact should come entirely as a surprise to us. The fact that this theorem comes with some names attached should suggest to you that it is quite an important result in the annals of Real Analysis. We'll come back to this result in the subsequent section when we take on the problem of finding a classification of convergent sequences.

Test Your Understanding

Prove that the sequences given by the following formula diverge.

1. $e_n = \dfrac{n+1}{\sqrt{n}}$
2. $g_n = 1 + (-1)^n$

Test Your Understanding - Answers

1. N.B: Depending on the assessment, you could be asked to show none or all of this work. This question is *easy*[6] if one is permitted to state (e_n) is unbounded, but tricky if one is asked to verify that (e_n) is unbounded. Can you imagine how long this would be if we had to appeal to Real Number axioms and theorems for every line of the argument!

   We claim $(e_n) = \left(\dfrac{n+1}{\sqrt{n}}\right)$ is unbounded. We proceed with a direct proof. (Though one could also proceed by contradiction if they wished) We show that for every $r \in \mathbb{R}^+$ there exists $n \in \mathbb{N}^+$ so that $e_n > r$.

   For fixed $r \in \mathbb{R}^+$ let[7] $M = \lceil r^2 \rceil$ and $n > M$.

$$\begin{aligned} n &> r^2 \\ n + 2 + 1/n &> r^2 \\ n^2 + 2n + 1 &> nr^2 \\ (n+1)^2 &> nr^2 \\ \frac{(n+1)^2}{n} &> r^2 \\ \frac{n+1}{\sqrt{n}} &> r \end{aligned}$$

   Therefore $e_n > r$ whenever $n > \lceil r^2 \rceil$ Therefore the set

$$\{e_n \mid n \in \mathbb{N}^+\}$$

   is not bounded above. Therefore (e_n) is unbounded. The result follows by the Divergence Criteria Theorem.

2. Notice

$$(g_1, g_3, g_5, \dots) = (0, 0, 0, 0, 0, 0, 0, 0, 0, \dots)$$

   This subsequence converges to 0. However, the subsequence

$$(g_2, g_4, g_6, \dots) = (2, 2, 2, 2, 2, 2, 2, 2, \dots)$$

   converges to 1. The result follows from the Divergence Criteria Theorem.

---

[6] Easy is a relative term. This question is easier if one doesn't have to show the work. None of the work we are going in this course is objectively easy.

[7] Magic number

4.2.2 Sequences and Convergence: Classifying Convergent Sequences

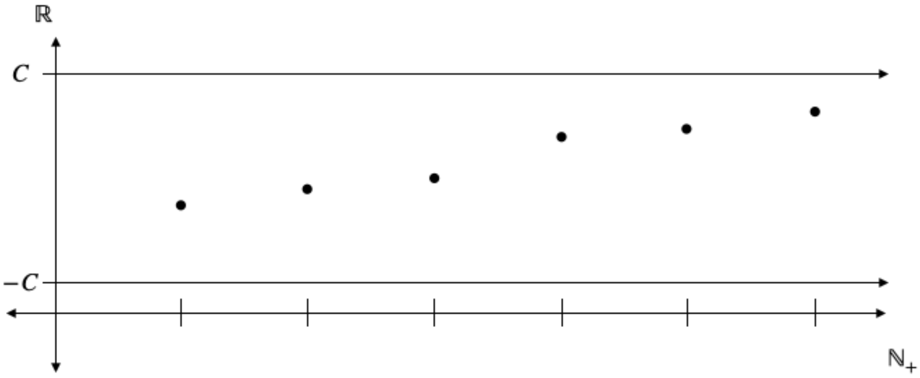
Recall the result from the previous section that relates boundedness and convergence:

Theorem. *Let (f_n) be a sequence. If (f_n) converges, then (f_n) is bounded.*

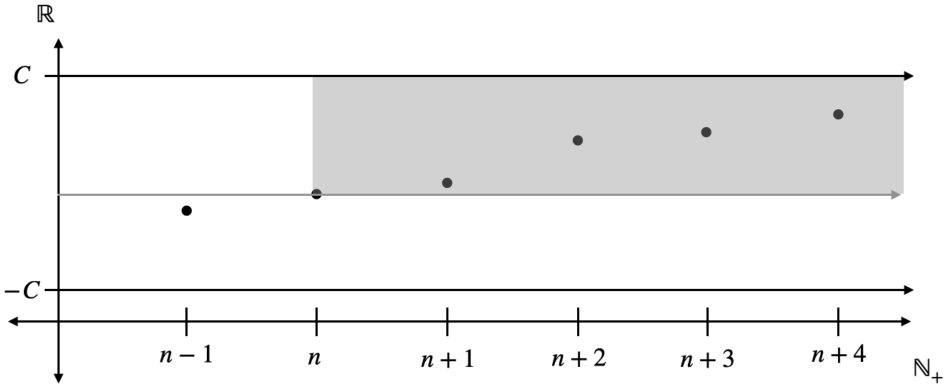
The converse of this theorem is not true in general. For example, the sequence given by $f_n = (-1)^n$ is certainly bounded, but does not converge. Whereas, the sequence given by $g_n = -\frac{1}{n}$ is bounded and also converges.

There are lots of differences between the sequence (f_n) and the sequence (g_n). One striking difference is that the sequence (g_n) is always increasing, whereas the sequence (f_n) alternates between increasing and decreasing. It is this former case we will take an interest in.

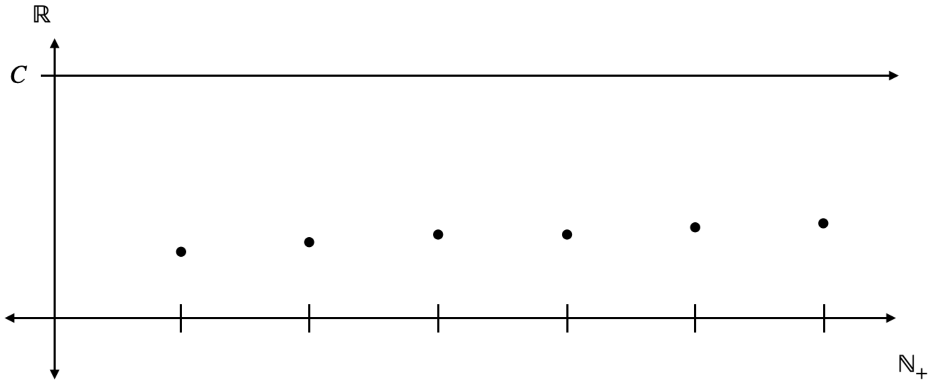
Let (h_n) be a sequence so that for all $n \in \mathbb{N}^+$ we have $h_n \leq h_{n+1}$. If (h_n) is bounded, then we have the following picture:



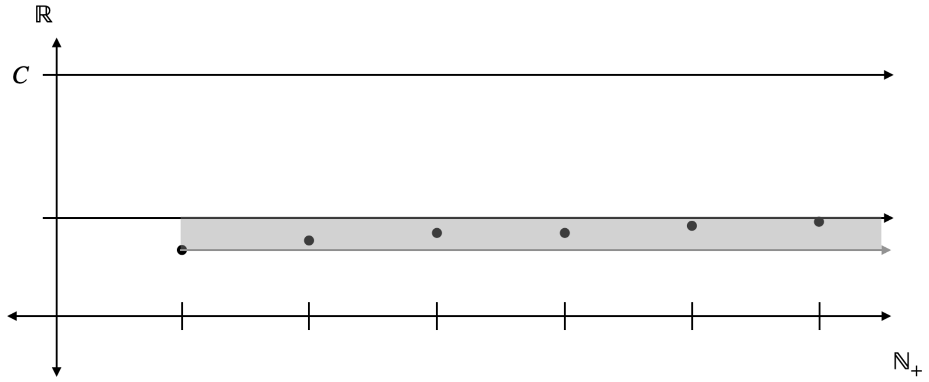
Since h_n is increasing, for any $n \in \mathbb{N}^+$ subsequent values are constrained within the strip between h_n and C.



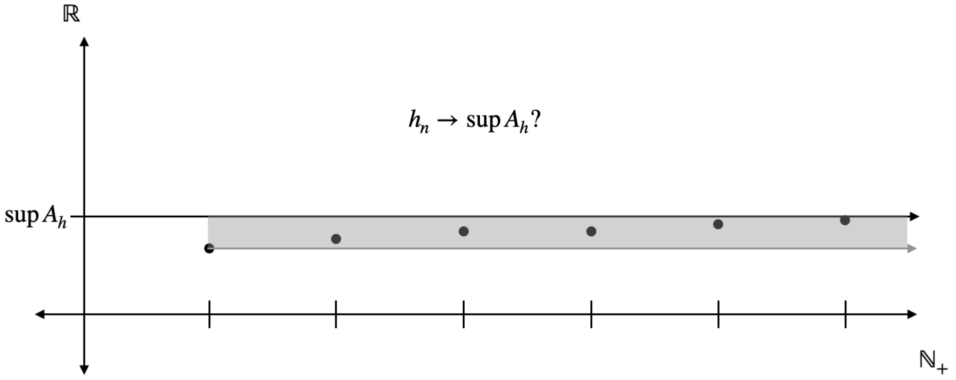
Unfortunately it may not be that the values of (h_n) approach C. In fact, the values may not even get close to C.



However, if not, then perhaps we can find a better upper bound than C.



Let $A_h = \{h_n \mid n \in \mathbb{N}^+\}$. By definition, since (h_n) is bounded, the set A_h is bounded above in $\mathbb{R}$. And so by the Completeness Axiom, this set necessarily has a supremum in $\mathbb{R}$. Since this supremum is the smallest upper bound, it seems reasonable we have $h_n \to \sup A_h$.

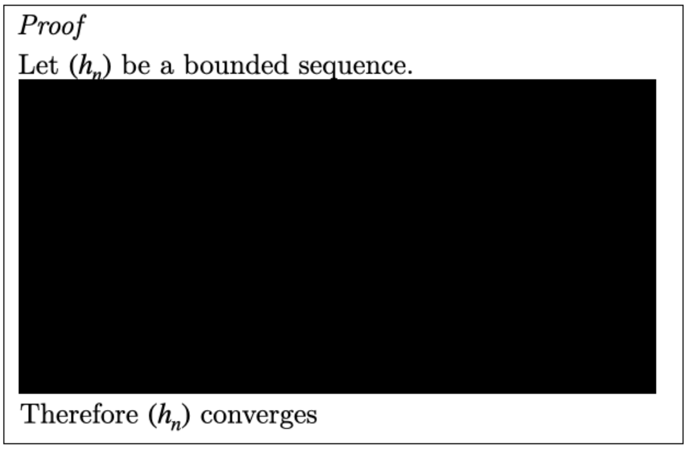


Aside. *Take a moment to marvel in the mathematical density of the notation* $h_n \ \to \ \sup \ A_h$*. There is just so much stuff we first had to make sense of before this notation has meaning for us.*

Equipped with this intuition, perhaps we suspect the following statement is true:

Let (h_n) be a bounded sequence. If for every $n \in \mathbb{N}^+$ we have $h_n \leq h_{n+1}$, then (h_n) converges

Without yet knowing how to prove such a statement, we can imagine the structure of a direct proof of this fact.



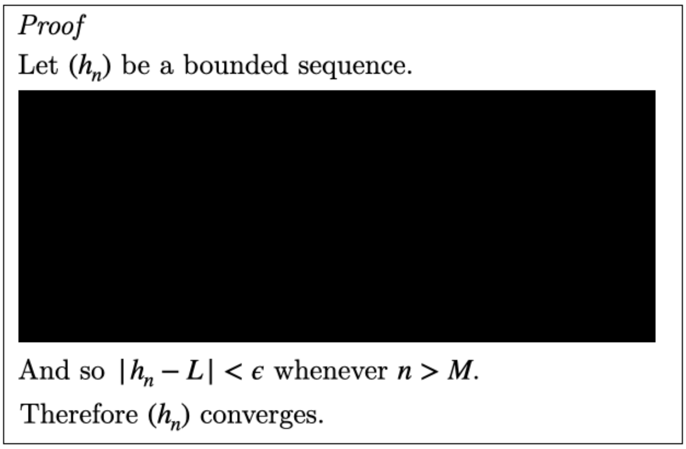
Proof

Let (h_n) be a bounded sequence.

[redacted]

Therefore (h_n) converges

Since the last line of our proof is "(h_n) converges", we can expect the line just before this to resemble the definition of convergence.



Proof

Let (h_n) be a bounded sequence.

[redacted]

And so $|h_n - L| < \epsilon$ whenever $n > M$.

Therefore (h_n) converges.

To get to this second-to-last line, we can imagine having to make a precise argument using ϵ's and M's. In this argument, we can also imagine needing to invoke the following fact about (h_n):

for every $n \in \mathbb{N}^+$ we have $h_n \leq h_{n+1}$

This property seems like a pretty special one for sequences. And we can imagine it coming up again in other contexts. And so let us mint a definition that gives us a short-form way to refer to this property.

Definition 4.20. *Let (f_n) be a sequence. We say* $\underline{(f_n) \text{ is monotone increasing}}$ *when we have $f_n \leq f_{n+1}$ for all $n \in \mathbb{N}^+$. We say* $\underline{(f_n) \text{ is monotone decreasing}}$ *when we have $f_n \geq f_{n+1}$ for all $n \in \mathbb{N}^+$.*

In thinking about (h_n) above we are considering convergence of a monotone increasing sequence. But we can imagine making the same sort of argument for a monotone decreasing sequence. Our ideas would be identical, except for reversing some inequalities and swapping out supremum for infimum. And so to permit us to consider both flavours of monotone sequences simultaneously, we use the following terminology.

Definition 4.21. *Let (f_n) be a sequence. We say (f_n) is* monotone *when (f_n) is monotone increasing or monotone decreasing.*

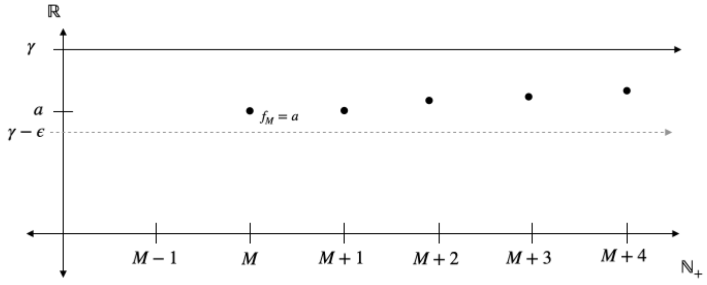
From our reasoning above, we can expect the following to be true:

Lemma 4.22. *Let (f_n) be a bounded sequence. If (f_n) is monotone increasing, then (f_n) converges.*

As we did above, we consider the set $A_f = \{f_n \mid n \in \mathbb{N}^+\}$. We show $f_n \to \sup A_f$. To do this, we recall a theorem[8] we proved on Assignment 2:

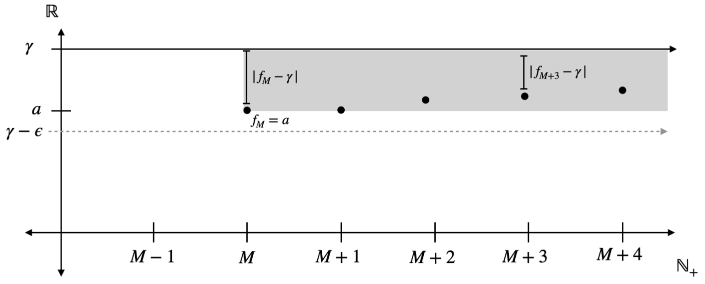
Theorem 4.23. *Let $A \subseteq \mathbb{R}$ be bounded above and let $\gamma \in \mathbb{R}$ be an upper bound of A in $\mathbb{R}$. We have $\sup A = \gamma$ if and only if for every $\epsilon \in \mathbb{R}$ with $\epsilon > 0$ there is an element of A greater than $\gamma - \epsilon$.*

Let $\gamma = \sup A_f$. We appeal to the definition of convergence and show that for any $\epsilon > 0$ there exists $M \in \mathbb{N}^+$ so that $|f_n - \gamma| < \epsilon$. Once we fix a value for ϵ, Theorem 4.23 tells us there exists $a \in A_f$ so that $a > \gamma - \epsilon$. Since $a \in A_f$, there exists $M \in \mathbb{N}^+$ so that $f_M = a$. This gives us the following picture:



The monotone property of (f_n) implies that f_n is contained in the grey strip shown below for all $n > M$. Subsequently, $|f_n - \gamma| < \epsilon$ for all $n > M$.

[8] What? How dare my instructor expect that I took the time to understand the assignment questions/solutions and not just copy them off my friend/the internet!



With this proof sketch in mind, we proceed with our proof of Lemma 4.22.

Proof. Let (f_n) be a bounded monotone increasing sequence. Let $A_f = \{f_n \mid n \in \mathbb{N}^+\}$.

Assume (f_n) is bounded. Since (f_n) is bounded, the set A_f is bounded above in $\mathbb{R}$. Since A_f is bounded above in $\mathbb{R}$ then, by the Completeness Axiom[9], $\sup A_f$ exists. Let $\gamma = \sup A_f$.

We show $f_n \to \gamma$ by appealing to the $\epsilon - M$ definition of convergence. Let $\epsilon > 0$. By Theorem 4.23, there exists $a \in A_f$ so that $a > \gamma - \epsilon$. Since $a \in A_f$, there exists $M \in \mathbb{N}^+$ so that $f_M = a$.

Since (f_n) is monotone increasing, for all $k \geq 1$ we have $\gamma - \epsilon < f_M \leq f_{M+k}$. Therefore

$$\gamma - f_{M+k} < \epsilon$$

for all $k \in \mathbb{N}^+$.

Since $\gamma = \sup A_f$ and $f_{M+k} \in A_f$ we have $\gamma \geq f_{M+k}$. Therefore $\gamma - f_{M+k} \geq 0$. And so $\gamma - f_{M+k} = |\gamma - f_{M+k}| = |f_{M+k} - \gamma|$. Therefore

$$|f_{M+k} - \gamma| < \epsilon$$

for all $k \in \mathbb{N}^+$. Therefore

$$|f_n - \gamma| < \epsilon$$

for all $n > M$. Therefore (f_n) converges. □

One can imagine modifying the proof of Lemma 4.22 for monotone decreasing sequences.

Thinking about the converse of Lemma 4.22, Theorem 4.13 tells us that any convergent (monotone) sequence is bounded. Putting this all together we have the following classification of convergent monotone sequences.

Theorem 4.24. *Let (f_n) be a monotone sequence. The sequence (f_n) converges if and only if (f_n) is bounded.*

[9] If you have been avoiding putting in the time to make sense of the Completeness Axiom, go and do it now. It is not going away. It all likelihood there will be a question on the final exam about it.

The statement of Theorem 4.24 is one of existence: if (f_n) is bounded and monotone, then there exists $L \in \mathbb{R}$ so that $f_n \to L$. The statement of the theorem alone tells us nothing about how to find the limit of a bounded and increasing monotone sequence. However, hiding inside the proof of Lemma 4.22 is a proof that a bounded and monotone increasing/decreasing sequence converges to its supremum/infimum. And so we have the following result.

Corollary 4.25. *Let (f_n) be a bounded sequence and let $A_f = \{f_n \mid n \in \mathbb{N}^+\}$.*

- *If (f_n) is monotone increasing, then $f_n \to \sup A_f$.*
- *If (f_n) is monotone decreasing, then $f_n \to \inf A_f$.*

Aside. *Mathematical truths are often labelled using words like* theorem*,* lemma *and* corollary*. The choice of which word is used can be largely subjective. However, the following guidelines usually apply: Broadly,* theorems *are main/important results.* Lemmas *are small results that are useful in the context of proving a theorem. And* corollaries *are results that are straightforward to prove once a theorem has been established.*

There are lots of exceptions to these guidelines. If your mathematical interests lie in the intersection of discrete mathematics and probability, one day you will encounter the Locász Local Lemma*, (pronounced LOH-VASH) a stunning statement that provides one a tool to assert the existence of a combinatorial object using only a probabilistic argument.*

Thinking back to our goal from the start of the reading, we have a partial solution. Rather than classify all convergent sequences, we have classified all convergent monotone sequences; they are exactly those monotone sequences that are bounded.

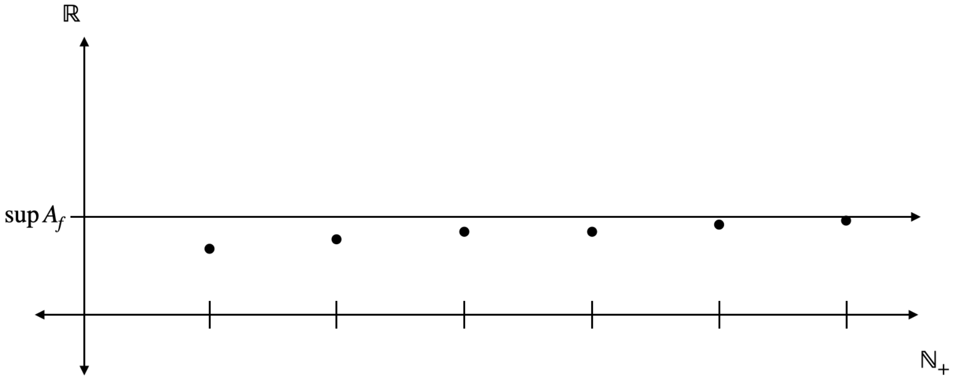
In what is becoming a running theme in Section 4 of the course, it is easy to find examples of convergent and divergent sequences for which our result tells us nothing. For example, the sequence given by $f_n = \frac{(-1)^n}{n}$ certainly converges, but is not monotone. And so the result of Theorem 4.24 does not apply.

By this point we are starting to develop some intuition around convergence. So let us poke a little at our intuition and see what else is lurking.

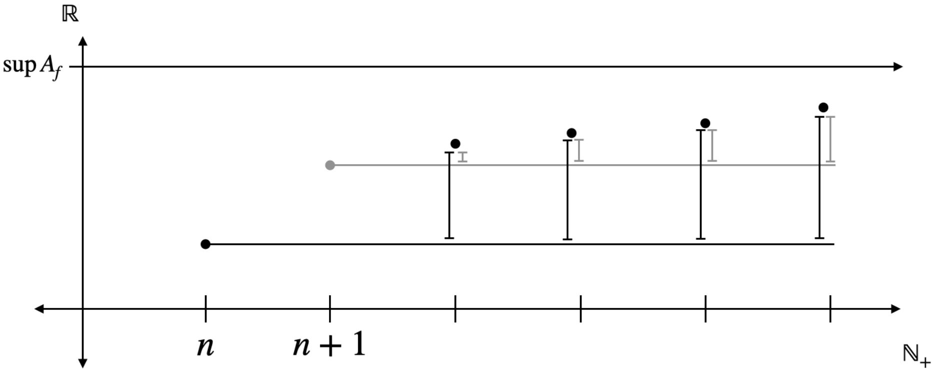
In the bounded increasing monotone case we have

$$\lim_{n\to\infty} f_n = \sup A_f$$

As n gets large, the terms get squeezed closer and closer to $\sup A_f$.



One consequence of this, is that the terms get squeezed closer and closer to each other!



In the picture the distance (in grey) between f_{n+1} and subsequent terms is less than the distance (in black) between f_n and subsequent terms. As the terms approach $\sup A_f$, the distance between the terms eventually gets smaller and stays smaller than any value.

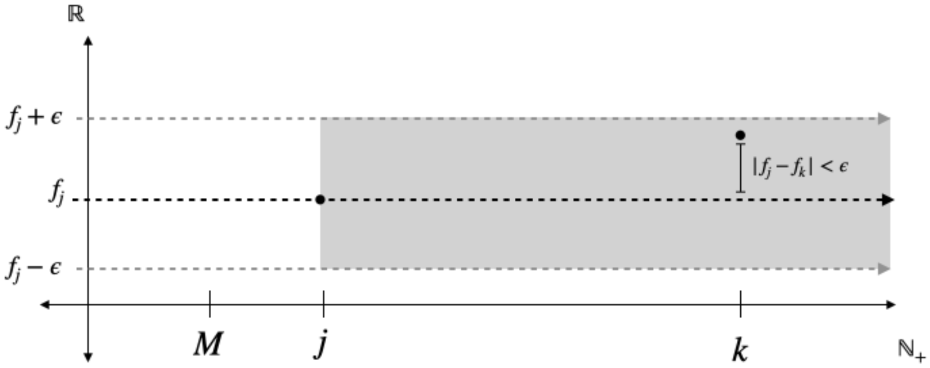
This idea of *gets smaller and stays smaller than any value* is one we are familiar with. But this time, rather than thinking of the distance between the terms of f_n and the limit getting smaller (and staying smaller) than any value, here we are noting that it is the distance is between the terms themselves that is getting (and staying) smaller than any value.

In other words, for a convergent sequence, perhaps we can expect the following statement to be true:

For every $\epsilon > 0$ there exists $M \in \mathbb{N}^+$ so that the distance between f_j and f_k is less than ϵ whenever $j, k > M$

That is,

For every $\epsilon > 0$ there exists $M \in \mathbb{N}^+$ so that $|f_j - f_k| < \epsilon$ whenever $j, k > M$



Let $(f_n) = \left(\frac{(-1)^n}{n}\right)$. This sequence is bounded and convergent, but not monotone. And so the result of Theorem 4.24 doesn't apply. Let us look to see if we can find our *gets smaller and stays smaller* condition among the terms of (f_n).

Let $q(j, k, \epsilon)$ be the following condition:

$$\left|\frac{(-1)^j}{j} - \frac{(-1)^k}{k}\right| < \epsilon$$

Depending on the values of j, k and ϵ the corresponding statement could be true or false. To build some intuition, let us consider $\epsilon = 0.01$.

Using the triangle inequality[10] we have:

$$\left|\frac{(-1)^j}{j} - \frac{(-1)^k}{k}\right| \leq \left|\frac{(-1)^j}{j}\right| + \left|-\frac{(-1)^k}{k}\right| = \frac{1}{j} + \frac{1}{k}$$

And so if $\frac{1}{j} + \frac{1}{k} < 0.01$, then certainly $\left|\frac{(-1)^j}{j} - \frac{(-1)^k}{k}\right| < 0.01$.

By the Archimedean Principle[11], there exists $M \in \mathbb{N}^+$ so that $1/M < 0.005$. And so when $j, k > M$ we have

$$\frac{1}{j} + \frac{1}{k} < 0.01$$

Therefore $p(j, k, 0.01)$ is true whenever $j, k > M$, where M is a natural number satisfying $1/M < 0.005$.

Of course, there is nothing special about 0.01 in this argument; we could substitute 0.01 for any $\epsilon > 0$ and apply the same reasoning. And so for every $\epsilon > 0$ there exists M so that $p(j, k, \epsilon)$ is true whenever $j, k > M$. In particular, given $\epsilon > 0$ we choose $M \in \mathbb{N}^+$ so that $\frac{1}{M} < \frac{\epsilon}{2}$.

It turns out that $p(j, k, \epsilon)$ is exactly the property we need to fully classify all convergent sequences. To ease our discussion of this property, we introduce the following definition.

[10] The triangle inequality seemed so innocuous when we first introduced it. But it is coming up everywhere!

[11] v3, let $y = 1$ and $x = 0.005$

Definition 4.26. *Let (f_n) be a sequence. We say $\underline{(f_n) \text{ is Cauchy}}$ when for each $\epsilon > 0$ there exists $M \in \mathbb{N}^+$ so that $|f_j - f_k| < \epsilon$ whenever $j, k > M$.*

Informally, a sequence is Cauchy if after some point M any two terms of the sequence differ by at most ϵ, for any $\epsilon > 0$ we desire. In other words, after some point M, all terms of the sequence are pairwise closer together than ϵ.

Example 4.27. *Show $(f_n) = \left(\frac{(-1)^n}{n}\right)$ is Cauchy.*

Let $\epsilon > 0$, let $M \in \mathbb{N}^+$ so that $1/M < \frac{\epsilon}{2}$ and let $j, k > M$. By the triangle inequality and the definition of absolute value we have

$$\begin{aligned}
|f_j - f_k| &= \left|\frac{(-1)^j}{j} - \frac{(-1)^k}{k}\right| \\
&\leq \left|\frac{(-1)^j}{j}\right| + \left|-\frac{(-1)^k}{k}\right| \\
&= \frac{1}{j} + \frac{1}{k} \\
&< \frac{1}{M} + \frac{1}{M} \\
&= \frac{2}{M} \\
&< \epsilon
\end{aligned}$$

Therefore $|f_j - f_k| < \epsilon$ whenever $j, k > M$. Therefore (f_n) is Cauchy.

Aside. *Cauchy sequences are named for late 19C mathematician Augustin-Louis Cauchy. Cauchy is widely considered as one of the first mathematicians to recognize a need for careful and precise arguments in the study of real number functions and sequences. He was among the first to use the ϵ-style definitions we are seeing in this course.*

If you are finding this course tedious and are looking for someone to blame, Cauchy isn't a bad choice for you to focus your annoyance on. Though had Cauchy not hit upon these ideas, undoubtedly one of his contemporaries would have and this course would be exactly the same, except this aside would be about some other long-dead mathematician, rather than Cauchy.

It turns out that the Cauchy property is exactly the property we seek that characterises all convergent sequences:

Theorem. *Let (f_n) be a sequence. The sequence (f_n) converges if and only if (f_n) is Cauchy.*

To prove this theorem we need to prove two things:

1. If (f_n) converges, then (f_n) is Cauchy.
2. If (f_n) is Cauchy, then (f_n) converges.

Consider first the case that (f_n) converges. Since (f_n) converges, there exists $L \in \mathbb{R}$ so that $\lim\limits_{n\to\infty} f_n = L$. Looking back at our definition for convergence, this means that for every $\epsilon > 0$ there exists $M \in \mathbb{N}^+$ so that $|f_n - L| \leq \epsilon$ whenever $n > M$.

To show (f_n) is Cauchy, we must show that for every $\epsilon' > 0$, there exists $M' \in \mathbb{N}^+$ so that $|f_j - f_k| < \epsilon'$ whenever $j, k > M'$.

Consider now some fixed value $\epsilon' > 0$ and let[12] $\epsilon = \frac{\epsilon'}{2}$.

Since (f_n) converges, there exists $M \in \mathbb{N}^+$ so that $|f_n - L| \leq \epsilon$ whenever $n > M$. Therefore for every $j, k > M$ we have

$$|f_j - L| < \epsilon \text{ and } |f_k - L| < \epsilon$$

By the Triangle Inequality we have

$$|f_j - f_k| = |f_j - L + L - f_k| \leq |f_j - L| + |-f_k + L| = |f_j - L| + |f_k - L| < 2\epsilon = \epsilon'$$

And so the following statement is true:

$$|f_j - f_k| < \epsilon' \text{ whenever } j, k > M'.$$

Therefore (f_n) is Cauchy.

Theorem 4.28. *Let (f_n) be a sequence. If (f_n) converges, then (f_n) is Cauchy.*

Unfortunately, proving the converse of Theorem 4.28 takes significantly more work. Rather than give the full details here, we instead present some highlights of the argument. To prove the converse of Theorem 4.28 one proves three things:

i. Every bounded sequence has a convergent subsequence

ii. Every Cauchy sequence is bounded.

iii. Every Cauchy sequence with at least one convergent subsequence is convergent.

We recognize part i. of this argument as the statement of the Bolazano-Weierstrauss Theorem, which appeared at the end of Week 5. Let us consider ii.

Let (f_n) be Cauchy. By the definition of bounded, to show (f_n) is bounded it suffices to find real numbers L and U and so that

$$L \leq f_n \leq U$$

for all $n \in \mathbb{N}^+$. We proceed as we did in the proof of Theorem 4.13. We find bounds for (f_n) on two different regions, and then take the maximum/minimum of those two different bounds as our bound for (f_n).

Considering $\epsilon = 1$ in the definition of Cauchy gives the following statement:

There exists $M \in \mathbb{N}^+$ so that $|f_j - f_k| < 1$ whenever $j, k > M$.

[12] Magic number warning. We choose $\epsilon = \frac{\epsilon'}{2}$ because it is what works out for us in the end.

When $k = M + 1$ we have $|f_j - f_{M+1}| < 1$ for all $j > M$. Therefore

$$\begin{aligned} -1 < f_j - f_{M+1} < 1 \\ f_{M+1} - 1 < f_j < 1 + f_{M+1} \end{aligned}$$

In other words, if (f_n) is Cauchy, then there exists $M \in \mathbb{N}^+$ so that

$$f_{M+1} - 1 < f_j < 1 + f_{M+1}$$

for all $j > M$.

Recall that our goal is to show that (f_n) is bounded. Consider the set

$$A = \{f_0, f_1, f_2, \ldots, f_M\}$$

Let U_A be the maximum value in A and let L_A be the minimum value in A. Therefore, the following statement is true:

$$L_A \leq f_t \leq U_A \text{ for all } t \in \{0, 1, \ldots, M\}$$

In summary, the following things are true about (f_n)

- $f_{M+1} - 1 < f_j < 1 + f_{M+1}$ for all $j > M$.
- $L_A \leq f_t \leq U_A$ for all $t \in \{0, 1, \ldots, M\}$

Therefore all terms of (f_n) are bounded above by the maximum of U_A and $1 + f_{M+1}$ and are bounded below by the minimum of L_A and $f_{M+1} - 1$. Thus, by definition, (f_n) is bounded

Lemma 4.29. *Every Cauchy sequence is bounded.*

Together, the Bolazano-Weierstrauss Theorem and Lemma 4.29 complete the first two parts of the argument that every Cauchy sequence converges. To complete the argument that every Cauchy sequence converges we require the following result.

Lemma 4.30. *Every Cauchy sequence with a convergent subsequence is convergent.*

In the interest of brevity, we omit the proof of this result.

Together Theorem 4.28, Lemmas 4.29, 4.30 and the Bolazano-Weierstrauss Theorem give the following result:

Theorem 4.31 (Cauchy Convergence Criterion)**.** *A sequence converges if and only if it is Cauchy.*

Just like the characterisation of invertible matrices using the determinant, the Cauchy Convergence Criterion tells us nothing about the limit of a Cauchy sequence. It merely tells us that such a limit exists.

Aside. *We have worked hard in this course over the last six weeks. Indeed, we have come a long way from our humble beginnings of formal logic in Week 1. The end of Week 6 marks the half-way point in the semester. As the semester 2 non-teaching period (September 20-26) still feels far away, we'll stop here and permit ourselves to take a breath or two before proceeding with the second half of the course.*

Test Your Understanding

1. Using the Cauchy Convergence Criterion, explain how you know the sequence $(f_n) = (\sqrt{n})$ is not Cauchy.

2. Using the definition of Cauchy, show that the sequence given by $f_n = 4 + \frac{(-1)^n}{3n}$ is Cauchy

3. For each item below, give an example or explain why it doesn't exist:

   (a) A divergent sequence with both an monotone increasing subsequence and a monotone decreasing subsequence.

   (b) A bounded monotone sequence that is not Cauchy.

   (c) A Cauchy sequence that is not bounded

   (d) A Cauchy sequence that is not monotone.

   (e) A bounded sequence with a divergent monotone subsequence.

Test Your Understanding - Answers

1. This sequence diverges. And so by the Cauchy Convergence Criterion, (f_n) diverges.

2. Proceed as in Example 4.27. Choose M so that $\frac{1}{M} < \frac{3\epsilon}{2}$. Other choices for M are possible.

3. (a) $(1, -1, 2, -2, 3, -3, \dots)$.

   (b) By the Cauchy Criterion Theorem, such a sequence must not converge. And so by Theorem 4.24, such a sequence must be unbounded. Therefore no example exists.

   (c) Every Cauchy sequence converges and every convergent sequence is bounded. Therefore no example exists.

   (d) $\left(\frac{(-1)^n}{n}\right)$.

   (e) Let (f_n) be a bounded sequence. Therefore there exists $C \in \mathbb{R}$ so that $|f_n| \leq C$ for all $n \in \mathbb{N}^+$. Let (f_{n_k}) be a divergent monotone subsequence. By Theorem 4.24, (f_{n_k}) is unbounded. Therefore there exists $k \in \mathbb{N}^+$ so that $|f_k| > C$, a contradiction. Therefore no such example exists.

4.3 Sequences of Functions

Define convergence for sequences of functions. Look ahead to Taylor Series.

5 Limits and Continuity

Our goal in Section 5 of the course is to understand continuity. Continuity is a concept we have seen many times since our first introduction into the world of Calculus. And so we wonder, what possible things are there left for us to learn about continuity?

To that, we answer: Consider the following function.

$$f(x) = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & x \notin \mathbb{Q} \end{cases}$$

Is this function continuous?

Our intuition for continuous functions as *functions that can be drawn without lifting our pencil off the page*, isn't much help here. It is difficult for us to draw a realistic picture of f in the plane. And, so, devoid of intuition, let us turn to a definition of continuity we have likely seen before.

Definition. *Let $f : \mathbb{R} \to \mathbb{R}$, we say $\underline{f \text{ is continuous}}$ when for every $a \in \mathbb{R}$ we have*

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

Aside. *Don't put this definition of continuity in your notes. It has some shortcomings that we will return to discuss in an upcoming section.*

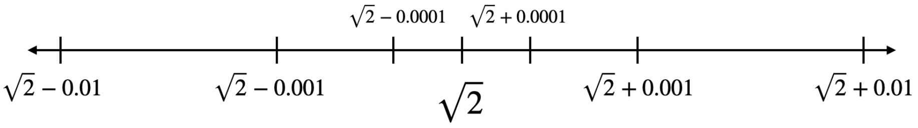
With this definition, our question about continuity has been recast into a question about limits, another topic we have seen many times since our first introduction into the world of Calculus. But again, our intuitive definition of what limits are isn't really much help. Is the following statement true:

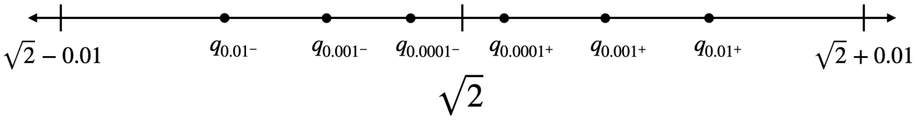
$$\lim_{x \to \sqrt{2}} f(x) = 0$$

Let us revert to a technique we have seen before for computing limits – a table of values. Recall that the sum of an irrational number and a rational number is necessarily irrational.

x	**f(x)**
$\sqrt{2} - 0.01$	0
$\sqrt{2} - 0.001$	0
$\sqrt{2} - 0.0001$	0
$\sqrt{2} + 0.01$	0
$\sqrt{2} + 0.001$	0
$\sqrt{2} + 0.0001$	0

Great! Looking at this table, it seems as if $\lim\limits_{x \to \sqrt{2}} f(x) = 0$. But wait, between any two irrational numbers we can find a rational number. For example, between $\sqrt{2} - 0.01$ and $\sqrt{2}$ we can find a rational number, $q_{0.01^-}$.





And so, we can make another table of values that are *close* to $\sqrt{2}$:

$\mathbf{x}$	$\mathbf{f(x)}$
$q_{0.01^-}$	1
$q_{0.001^-}$	1
$q_{0.0001^-}$	1
$q_{0.01^+}$	1
$q_{0.001^+}$	1
$q_{0.0001^+}$	1

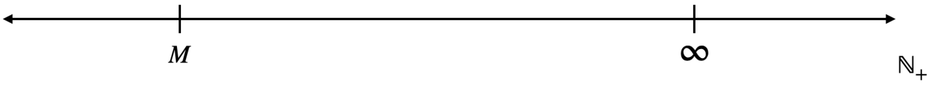
From this example, perhaps it is clear that our understanding of the meaning of limits could stand to be broadened. And so, let us take some time to develop a more precise definition of the notation.

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

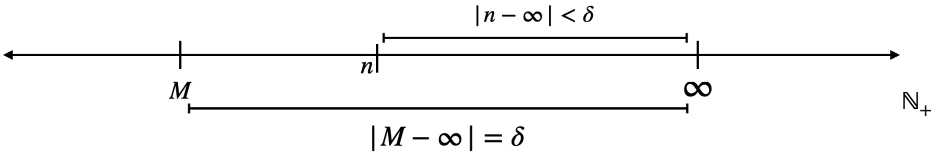
We spent the last two weeks of this subject thinking about convergence and divergence for sequences. In doing so we agreed upon the following definition:

Definition. *Let (f_n) be a sequence and let $L \in \mathbb{R}$. We say $\underline{(f_n) \text{ converges to } L}$ when for every $\epsilon > 0$ there exists $M \in \mathbb{N}^+$ so that $|f_n - L| < \epsilon$ whenever $n > M$. When (f_n) converges to L we write $\lim\limits_{n \to \infty} f_n = L$.*

To get ourselves to a reasonable definition of limits of real functions, let us massage this definition a little bit by dabbling in some absurdist mathematics. Imagine, for the moment, that ∞ appears on our number line.

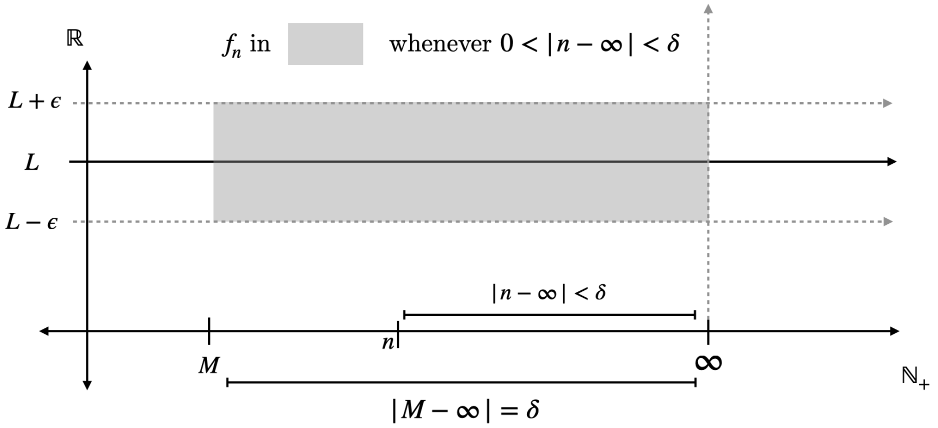


If $f_n \to L$, then for any $\epsilon > 0$ there exists M so that $|f_n - L| < \epsilon$ whenever $n > M$. Let $|M - \infty| = \delta$. We have the following picture:



Again, this is absurd! Infinity is not a number and so we can't measure the distance between M and ∞. However, if we continue to suppress our discomfort, we can re-write our definition above as:

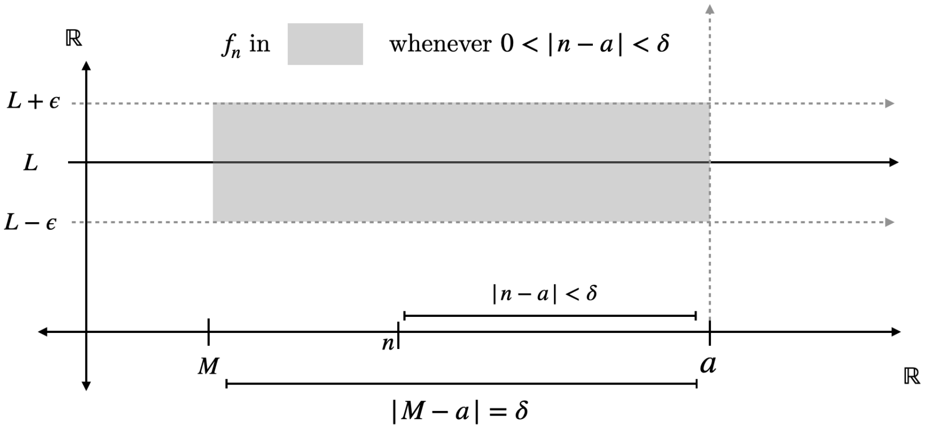
Definition (Absurdist Mathematics Warning! Don't put this definition in your notes!)**.** *Let (f_n) be a sequence and let $L \in \mathbb{R}$. We say $\underline{(f_n) \text{ converges to } L}$ when for every $\epsilon > 0$ there exists $M \in \mathbb{N}^+$ so that $|f_n - L| < \epsilon$ whenever $0 < |n - \infty| < \delta$.*



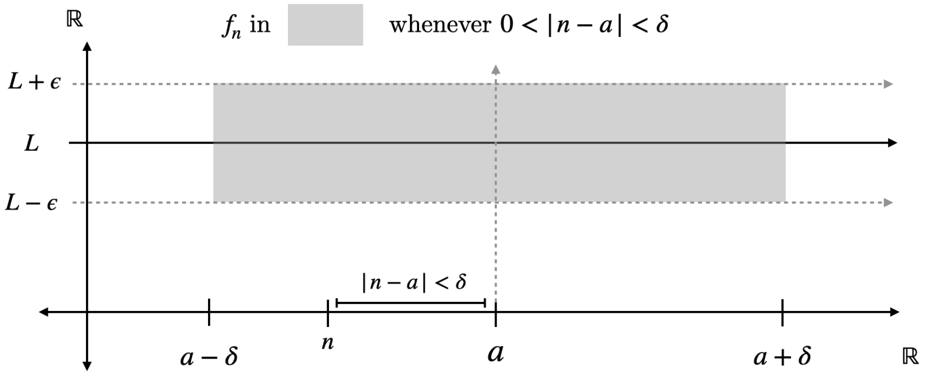
Though we have changed our definition to something absurd, our intuition hasn't changed:

as long as the distance between n and ∞ is less than δ, we are guaranteed that the distance between f_n and L is less than ϵ.

However, when we change our frame of reference to real-valued functions and swap ∞ for $a \in \mathbb{R}$, we arrive at something quite reasonable:



Except now, since $a \in \mathbb{R}$, there are points to the right of a that may satisfy $0 < |x-a| < \delta$

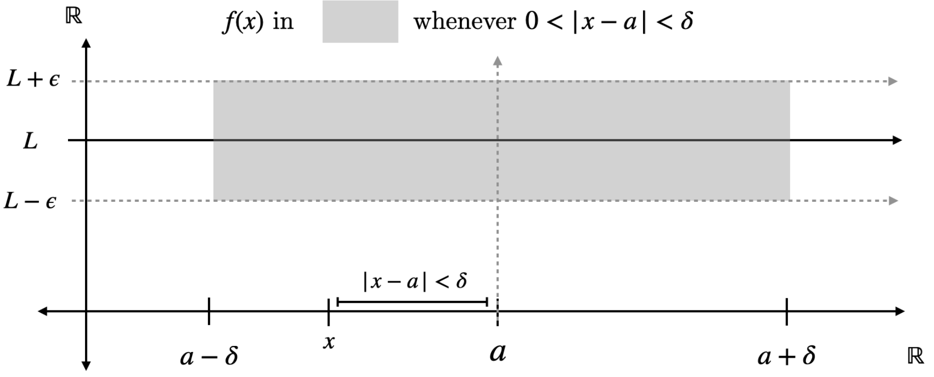


And so we propose the following definition for $\lim_{x\to a} f(x) = L$.

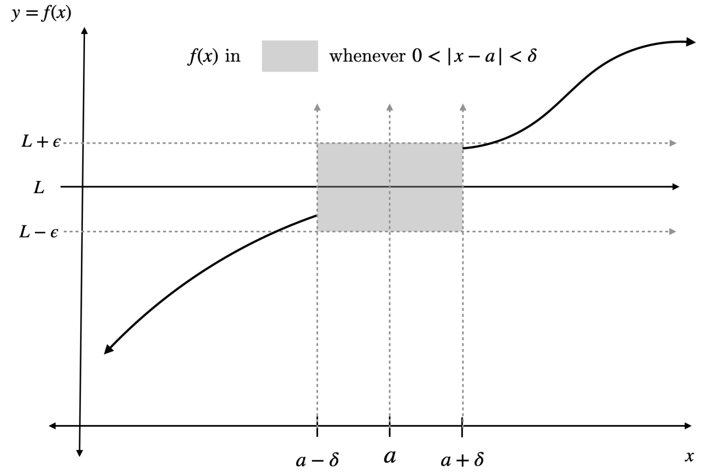
Definition. *Let $f : \mathbb{R} \to \mathbb{R}$ and let $a \in \mathbb{R}$. We say the limit of $f(x)$ as x approaches a is L when the distance between $f(x)$ and L eventually gets smaller (and stays smaller) than any value $\epsilon > 0$ as the distance between x and a gets small.*

More precisely, we have:

Definition (Attempt #1)**.** *Let $f : \mathbb{R} \to \mathbb{R}$, $a \in \mathbb{R}$ and let $L \in \mathbb{R}$. We say the limit of $f(x)$ as x approaches a is L when for every $\epsilon > 0$ there exists $\delta > 0$ so that $|f(x) - L| < \epsilon$ whenever $0 < |x - a| < \delta$. When the limit of $f(x)$ as x approaches a is L we write $\lim_{x\to a} f(x) = L$.*



Let us zoom out a little, and imagine an actual picture of a function in the plane. We say $\lim_{x\to a} f(x) = L$ when for every $\epsilon > 0$ we can find a value $\delta > 0$ so that when the distance between x and a is less than δ, we are guaranteed that the distance between $f(x)$ and L is less than ϵ. In other words, when x is between $a - \delta$ and $a + \delta$ and $x \neq a$, we are guaranteed that $f(x)$ is between $L - \epsilon$ and $L + \epsilon$.



Aside. *Don't put this definition in to your notes quite yet. This definition has some shortcomings that we will discuss at the start of the next section.*

The collection of ideas in the last few pages form the basis of our definition for limits of real functions. There are a few technical details still to take care of, but from this point forward the broad idea of the meaning of the notation $\lim_{x \to a} f(x) = L$ is not going to change: We write $\lim_{x \to a} f(x) = L$ when for every $\epsilon > 0$ we can find a distance $\delta > 0$ so that the distance between $f(x)$ and L is less than ϵ whenever the distance between x and a is less than δ.

5.1 Limits of Real Functions

At the end of our introductory material on limits, we arrived upon the following definition for the limit of a real function:

Definition (Attempt #1)**.** *Let $f : \mathbb{R} \to \mathbb{R}$, $a \in \mathbb{R}$ and let $L \in \mathbb{R}$. We say the limit of $f(x)$ as x approaches a is L when for every $\epsilon > 0$ there exists $\delta > 0$ so that $|f(x) - L| < \epsilon$ whenever $0 < |x - a| < \delta$. When the limit of $f(x)$ as x approaches a is L we write $\lim\limits_{x \to a} f(x) = L$.*

Let us play around with this definition to see if it agrees with our computational experience.

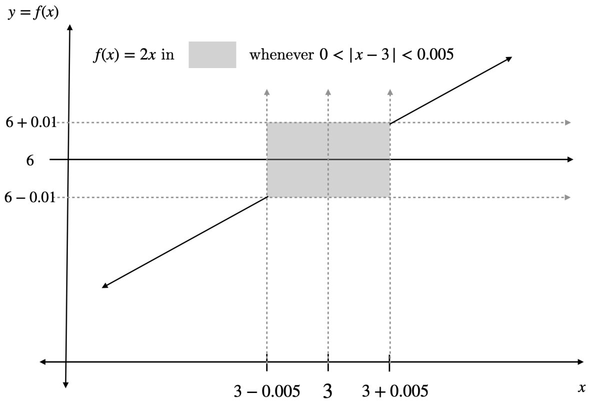
Consider first the statement $\lim\limits_{x \to 3} 2x = 6$. Our work from previous Calculus courses justifies the truth of this statement with some self-assured statements about continuous functions[1]. Let us see, however, if we can verify this fact using our definition of the limit above.

To prove $\lim\limits_{x \to 3} 2x = 6$ we must prove that for any $\epsilon > 0$, we can find $\delta > 0$ so that $|2x - 6| < \epsilon$ whenever x satisfies $0 < |x - 3| < \delta$.

To build some intuition, let us consider $\epsilon = 0.01$. Algebraically, we find

$$\begin{aligned} |2x - 6| &< 0.01 \\ |2(x - 3)| &< 0.01 \\ |x - 3| &< 0.005 \end{aligned}$$

And so we find $|2x - 6| < 0.01$ whenever $|x - 3| < 0.005$.



Using our work here, we give a proof that $\lim\limits_{x \to 3} 2x = 6$:

[1] If a function is continuous, then $\lim\limits_{x \to a} = f(a)$. Since $f(x) = 2x$ is continuous, we have $\lim\limits_{x \to 3} 2x = 6$ ....right...?

Let $\epsilon > 0$ and let $\delta = \frac{\epsilon}{2}$. Let $x \in \mathbb{R}$ so that $0 < |x-3| < \delta$. Notice

$$\begin{aligned} 0 < |x-3| &< \delta \\ 0 < |x-3| &< \frac{\epsilon}{2} \\ 0 < |2(x-3)| &< \epsilon \\ 0 < |2x-6| &< \epsilon \end{aligned}$$

Therefore $|2x-6| < \epsilon$ whenever $0 < |x-3| < \delta$. And so we conclude $\lim\limits_{x \to 3} 2x = 6$.

So far, so good. Our definition above seems fine, however there are still minor technicalities to deal with. Looking back at our proposed definition for this notation above, we see a problem.

Consider now the statement

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

Our definition above is stated for functions $f : \mathbb{R} \to \mathbb{R}$. But the domain of $\sqrt{x}$ is $\{x \in \mathbb{R} \mid x \geq 0\}$.

So that our definition of the limit is useful for functions whose domain is not $\mathbb{R}$, we modify the definition of limit so restrict the domain of the function of interest. We use $E \subseteq \mathbb{R}$ to denote the domain of the function we are considering.

Definition (Attempt #2)**.** *Let $E \subseteq \mathbb{R}$, let $f : E \to \mathbb{R}$, $a \in \mathbb{R}$ and let $L \in \mathbb{R}$. We say the limit of $f(x)$ as x approaches a is L when for every $\epsilon > 0$ there exists $\delta > 0$ so that $|f(x) - L| < \epsilon$ whenever $0 < |x-a| < \delta$. When the limit of $f(x)$ as x approaches a is L we write $\lim\limits_{x \to a} f(x) = L$.*

Aside. *Almost there! But still, this version isn't the one we will settle on. Don't put it in your notes.*

Continuing with the example $f(x) = \sqrt{x}$, what can we make of the statement

$$\lim_{x \to -5} \sqrt{x} = L$$

There is nothing in our current version of the definition that prohibits us from considering this limit, but again we have landed back in the world of absurdist mathematics. It is impossible to consider the function $\sqrt{x}$ as x approaches -5 as the domain of $\sqrt{x}$ doesn't include any values that *approach* -5.

One possible fix is to restrict a to be an element of E, but that too leads to a problem. For example our definition of the limit should permit us to verify

$$\lim_{x \to 2} \frac{x^2 - 4}{x - 2} = 4$$

However, in this case, $x = 2$ is not in the domain of $f(x) = \frac{x^2-4}{x-2}$.

In thinking about our intuition for the statement

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

we are thinking about the behaviour of the function $f : E \to \mathbb{R}$ for values that are *very close* to a, but not necessarily at $x = a$. For this to make sense there must be some values of E that are *close* to a. Further, there must be some values of E that are close to a no matter how *close* we specify. In other words, it only makes sense to permit us to take limits approaching those values $a \in \mathbb{R}$ where for every $\delta > 0$ there exists at least one $x \in E$ so that $0 < |x - a| < \delta$.

Definition 5.1. *Let $E \subseteq \mathbb{R}$ and let $a \in \mathbb{R}$. We say a is a limit point of E when for every $\delta > 0$ there exists $x \in E$ so that $0 < |x - a| < \delta$.*

For example, consider the function $f(x) = \log x$. The domain of this function is $E = \{x \in \mathbb{R} \mid x > 0\}$. When we write the statement

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

we can imagine a taking on the value of any real number in the interval $[0, \infty)$. And so we expect that the set of limit points of E is the set $P = \{a \in \mathbb{R} \mid a \geq 0\}$.

Consider $a \in \mathbb{R}$ so that $a \geq 0$. To prove a is a limit point of E, we must prove that for each $\delta > 0$ there exists at least one $x \in E$ so $x \neq a$ and

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

Re-arranging this inequality, we have

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

Since $\delta \neq 0$, we have

$$a < \frac{2a + \delta}{2} < a + \delta$$

Therefore $x = \frac{2a+\delta}{2}$ satisfies -$\delta < x - a < \delta$. Since $a \geq 0$ and $\delta > 0$, we have $x > 0$. Therefore $x \in E$. Therefore there exists $x \in E$ so that $0 < |x - a| < \delta$. Therefore a is a limit point of E.

However, not every set is brimming with limit points. Consider the domain $E = \mathbb{N}$. For any natural number k, there is no other natural number within distance $\delta = \frac{1}{2}$ of k. For the domain $E = \mathbb{N}$ the set

$$\{x \in \mathbb{R} \mid x \text{ is a limit point of } E\}$$

is empty[2]

Correcting all of our shortcomings in our definition for the limit of $f(x)$ as x approaches a is L, we finally reach our complete definition of the limit.

[2] It would be natural to have said here *E has no limit points.* But, for reasons that we will return to near the end of 5.1, it is not completely correct to say *E has no limit points.*

Definition 5.2. *Let $E \subseteq \mathbb{R}$, let a be a limit point of E, let $f : E \to \mathbb{R}$ and let $L \in \mathbb{R}$ We say the limit of $f(x)$ as x approaches a is L when for every $\epsilon > 0$ there exists $\delta > 0$ so that $|f(x) - L| < \epsilon$ whenever $0 < |x - a| < \delta$. When the limit of $f(x)$ as x approaches a is L we write $\lim\limits_{x \to a} f(x) = L$ and we say $f(x)$ converges to L as x approaches a.*

Aside. *Many introductory materials in Calculus use Zeno's Paradoxes as a motivation for considering the concept of the limit. It is worth noting that Zeno's Paradoxes date from approximately 500 BCE. It took over 2000 years of human scientific exploration to come up with formal methods to resolve these paradoxes. A version of this definition first appears in the early 19th century in the work of Bolzano.*

Using this definition we prove

$$\lim_{x \to 2} \frac{x^2 - 4}{x - 2} = 4$$

To do this we want to show that for every $\epsilon > 0$ there exists $\delta > 0$ so that $|\frac{x^2-4}{x-2} - 4| < \epsilon$ whenever x satisfies $0 < |x - 2| < \delta$. To find our value for δ we start with the statement $|\frac{x^2-4}{x-2} - 4| < \epsilon$ and algebraically manipulate it to $|x - 2| < \delta$.

$$\begin{aligned} |\frac{x^2 - 4}{x - 2} - 4| &< \epsilon \\ |(x + 2) - 4| &< \epsilon \\ |(x - 2)| &< \epsilon \end{aligned}$$

If we choose $\delta = \epsilon$, then we will have $|\frac{x^2-4}{x-2} - 4| < \epsilon$ whenever $0 < |x - 2| < \delta$. Now that we have figured out our magic number, we give our proof that $\lim\limits_{x \to 2} \dfrac{x^2 - 4}{x - 2} = 4$.

Let $\epsilon > 0$, let $\delta = \epsilon$ and let x satisfy $0 < |x - 2| < \delta$. Algebraically we find

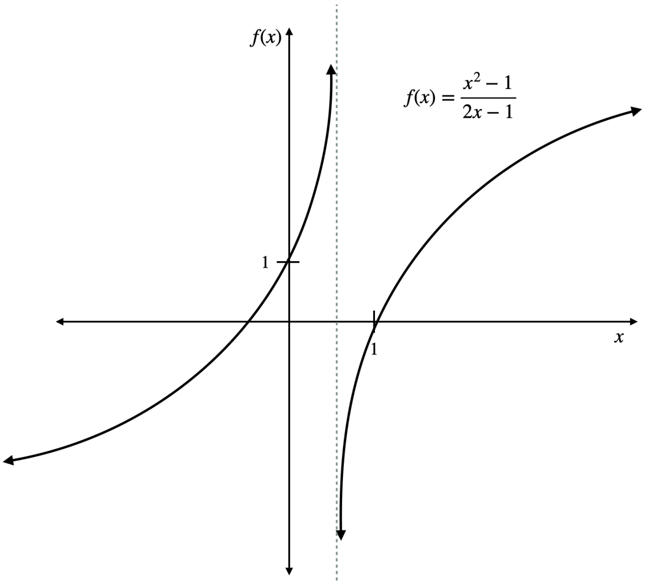
$$\begin{aligned} 0 < |x - 2| &< \delta \\ 0 < |x + 2 - 4| &< \delta \\ 0 < \left|\frac{x - 2}{x - 2}(x + 2) - 4\right| &< \delta \\ 0 < \left|\frac{x^2 - 4}{x - 2} - 4\right| &< \delta \\ 0 < \left|\frac{x^2 - 4}{x - 2} - 4\right| &< \epsilon \end{aligned}$$

Therefore $|\frac{x^2-4}{x-2} - 4| < \epsilon$ whenever x satisfies $0 < |x - 2| < \delta$. And so we conclude $\lim\limits_{x \to 2} \dfrac{x^2 - 4}{x - 2} = 4$.

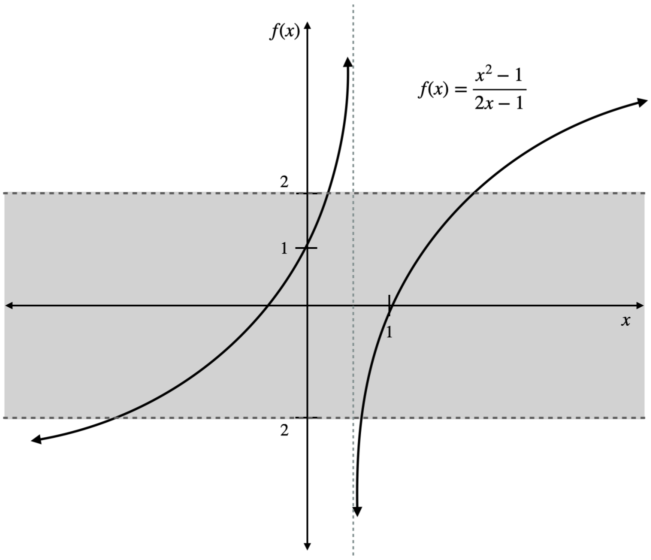
Aside. *Imagine how long this proof would be if we insisted on going back to the Real Number Axioms to justify every algebraic step!*

In all of the examples we have seen so far, our function $f(x)$ was *well-behaved* near a. As x approached a from either side, we didn't encounter odd behaviour as the distance between x and a became small. As the world of real functions is big and scary, such behaviour is not guaranteed. For example, consider the following statement

$$\lim_{x \to 1} \frac{x^2 - 1}{2x - 1} = 0$$



Working as we did above, we want to find a value for δ, as a function of ϵ, so that $\left|\frac{x^2-1}{2x-1} - 0\right| < \epsilon$ whenever x satisfies $0 < |x - 1| < \delta$. Consider the case $\epsilon = 2$ and the picture below.

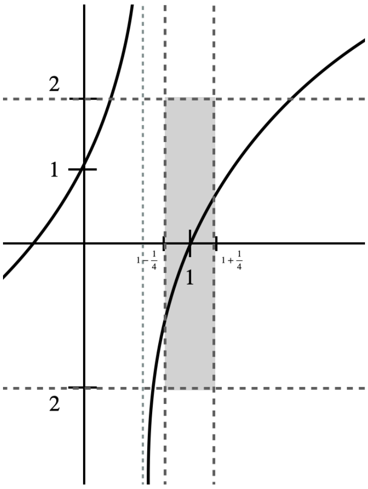


When x is near $\frac{1}{2}$, $f(x)$ does not satisfy $\left|\frac{x^2-1}{2x-1} - 0\right| < 2$.

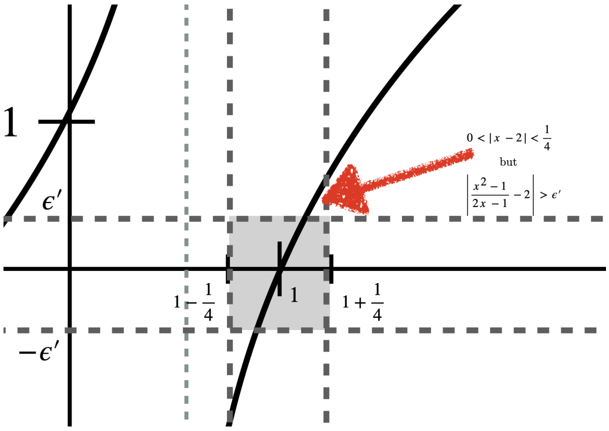
In fact, no matter which ϵ we consider, our choice of δ cannot permit choices for x that are *close to* $x = \frac{1}{2}$. As otherwise, we may be able to find x so that $\left|\frac{x^2-1}{2x-1} - 0\right| > \epsilon$. To avoid this problem, we can consider choosing δ more carefully.

Given a value $\epsilon > 0$, rather than choosing δ as a function of ϵ, let us first consider if choosing, say, $\delta = \frac{1}{4}$ will suffice. If so, then we will choose $\delta = \frac{1}{4}$. For example, when $\epsilon = 2$, choosing $\delta = \frac{1}{4}$ permits us to algebraically verify:

$$\left|\frac{x^2-1}{2x-1} - 0\right| < 2 \text{ whenever } 0 < |x-1| < \frac{1}{4}$$



However, for small choices of ϵ, choosing $\delta = \frac{1}{4}$ may not suffice[3]. Let ϵ' be a choice of ϵ for which there is a value x so that $0 < |x-1| < \frac{1}{4}$, but $\left|\frac{x^2-1}{2x-1} - 0\right| > \epsilon'$



When $|x-1| < \frac{1}{4}$, one can show $|2x-1| > \frac{1}{2}$. Therefore

$$\left|\frac{x^2-1}{2x-1} - 0\right| < 2|x^2-1| = 2|x+1| \cdot |x-1|$$

[3] ... but, how do we know to choose $\frac{1}{4}$? Unfortunately, we don't. But, we will see, that we could choose any other small value for δ and the argument would work out similarly

Notice

$$|x+1| = |x-1+2| \leq |x-1| + 2 < \frac{1}{4} + 2 = \frac{9}{4}$$

Therefore

$$\left|\frac{x^2-1}{2x-1} - 0\right| < 2 \cdot \frac{9}{4} \cdot |x-1| = \frac{9}{2}|x-1|.$$

Consider $\delta = \frac{2}{9}\epsilon'$. If $0 < |x-1| < \delta$, then

$$\left|\frac{x^2-1}{2x-1} - 0\right| < 2 \cdot \frac{9}{4} \cdot |x-1| = \frac{9}{2}|x-1| < \frac{9}{2} \cdot \frac{2}{9}\epsilon' = \epsilon'$$

Therefore when $\delta = \frac{2}{9}\epsilon'$, we have $\left|\frac{x^2-1}{2x-1} - 0\right| < \epsilon'$.

We are still trying to verify

$$\lim_{x\to 1} \frac{x^2-1}{2x-1} = 0$$

Let $\epsilon > 0$. From our work above, if choosing $\delta = \frac{1}{4}$ does not imply

$$\left|\frac{x^2-1}{2x-1} - 0\right| < \epsilon \text{ whenever } |x-1| < \delta$$

then choosing $\delta = \frac{2}{9}\epsilon$ implies

$$\left|\frac{x^2-1}{2x-1} - 0\right| < \epsilon \text{ whenever } |x-1| < \delta$$

Therefore for every $\epsilon > 0$ there exists $\delta > 0$ so that

$$\left|\frac{x^2-1}{2x-1} - 0\right| < \epsilon \text{ whenever } |x-1| < \delta$$

In other words,

$$\lim_{x\to 1} \frac{x^2-1}{2x-1} = 0$$

Aside. *Proving* $\lim_{x\to 1} \frac{x^2-1}{2x-1} = 0$ *is a difficult exercise. There is no reason to think that any student in this course would have been able to do this without first having been shown how. We'll get some more practice on these sorts of limits in Tutorial 7 and in future assignments.*

As we see from our examples, above, even computing *easy* limits can take lots of computational carefulness. Fortunately, much like our work for sequences, we have an Algebra of Limits Theorem, which eases our burden in computing limits.

Theorem 5.3 (Algebra of Limits Theorem)**.** *Let* $E \subseteq \mathbb{R}$*, let* a *be limit point of* E *and let* $f, g : E \to \mathbb{R}$*. If* $\lim_{x\to a} f(x) = \alpha$ *and* $\lim_{x\to a} g(x) = \beta$*, then* *i.* $\lim_{x\to a} f(x) + g(x) = \alpha + \beta$*;*

ii. $\lim_{x\to a} f(x) \cdot g(x) = \alpha \cdot \beta$*; and*

iii. $\lim_{x\to a} \dfrac{f(x)}{g(x)} = \alpha/\beta$*, provided* $\beta \neq 0$*.*

When we considered the Algebra of Limits Theorem for sequences, we took the time to carefully prove the additive law for limits of sequences. We consider the proof for i in the *Test Your Understanding* section below.

This concludes our first look at limits of real functions. However, there are still a few more topics to consider before we move on completely. As we recall from our discussion on sequences there is still a matter of divergence to consider. Further, from Calculus 2, we recall that there are techniques for thinking about limits as x approaches infinity. We consider these two topics in turn in the coming sections.

Test Your Understanding

1. What are the limit points of the set $\mathbb{R} \setminus \{2, -2\}$

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

3. In this exercise we prove $\lim_{x\to 2} x^2 = 4$. Let $\epsilon > 0$. Notice $|x^2 - 4| < \epsilon$ if and only if $|x + 2| \cdot |x - 2| < \epsilon$.

   (a) Prove that if $|x - 2| < 1$, then $|x + 2| < 5$.

   (b) Prove that if $|x - 2| < 1$, then $|x^2 - 4| < 5|x - 2|$.

   (c) Let $\delta = \min\{1, \epsilon/5\}$. Prove that if $0 < |x - 2| < \delta$, then $|x^2 - 4| < \epsilon$. Proceed in two cases based on the value of ϵ: Case I: $\epsilon > 5$, Case II: $\epsilon \leq 5$.

4. Using the method from the previous question, prove $\lim_{x\to 3} x^2 = 9$.

5. Using the Algebra of Limits Theorem for real functions and the previous question, prove $\lim_{x\to 3} x^4 = 81$.

6. In this exercise we construct[4] the proof of part i of the Algebra of Limits Theorem for real functions.

   (a) Go and read the proof of part i of the Algebra of Limits Theorem for sequences. Don't come back until you understand the proof.

   (b) Write out the definition of the notation $\lim_{x\to a} f(x) = \alpha$. Use ϵ' instead of ϵ and δ_f instead of δ.

[4] The choice of the verb *to construct* is intentional here. As we move from one line explanations to multi-step proofs, we can think of our proofs as having a sense of structure, which we must first design before we fill in the details.

(c) Write out the definition of the notation $\lim_{x \to a} g(x) = \beta$. Use ϵ' instead of ϵ and δ_g instead of δ.

(d) Let $h(x) = f(x) + g(x)$ and let $\epsilon > 0$. Let $\epsilon' = \frac{\epsilon}{2}$. Let $\delta = \min\{\delta_f, \delta_g\}$. Prove $|h(x) - (\alpha + \beta)| < \epsilon$ whenever $0 < |x - a| < \delta$. If this seems very difficult, go back and do part (a) again.

Test Your Understanding - Answers

1. $\mathbb{R}$

2. Let $\delta = \frac{\epsilon}{2}$.

3. In this exercise we prove $\lim\limits_{x \to 2} x^2 = 4$. Let $\epsilon > 0$. Notice $|x^2 - 4| < \epsilon$ if and only if $|x + 2| \cdot |x - 2| < \epsilon$.

   (a) We can re-write $|x - 2| < 1$ as

   $$-1 < x - 2 < 1$$

   Adding 4 to everything yields

   $$3 < x + 2 < 5$$

   Since $-5 < 3$ we have

   $$-5 < x + 2 < 5$$

   Therefore $|x + 2| < 5$.

   (b) Notice

   $$|x^2 - 4| = |x + 2| \cdot |x - 2|$$

   By the previous part we have $|x + 2| < 5$. Therefore

   $$|x^2 - 4| = |x + 2| \cdot |x - 2| < 5|x - 2|$$

   (c) Let $\epsilon > 0$. If $\epsilon > 5$, then $\min\{1, \epsilon/5\} = 1$. Let $\delta = 1$. Let $x \in \mathbb{R}$ so that $|x - 2| < 1$. From our work above, we have $|x + 2| < 5$. Therefore

   $$|x^2 - 4| = |x - 2| \cdot |x + 2| < 5 < \epsilon$$

   If $\epsilon \leq 5$, then $\min\{1, \epsilon/5\} = \epsilon/5$. Let $\delta = \epsilon/5$. Let $x \in \mathbb{R}$ so that $|x - 2| < \epsilon/5$. Since $\epsilon/5 < 1$ we have $|x - 2| < 1$. Therefore $|x + 2| < 5$. Therefore

   $$|x^2 - 4| = |x - 2| \cdot |x + 2| < 5\delta = \epsilon$$

   Therefore we have $|x^2 - 4| < \epsilon$ whenever $|x - a| < \delta$. Therefore $\lim\limits_{x \to 2} x^2 = 4$

4. Let $\delta = \min\{1, \epsilon/7\}$. Proceed as in the previous question.

5. Notice $x^4 = x^2 \cdot x^2$. Therefore $\lim\limits_{x \to 3} x^4 = (\lim\limits_{x \to 3} x^2) \cdot (\lim\limits_{x \to 3} x^2) = 81$.

6. (a)

   (b) For every $\epsilon' > 0$ there exists $\delta_f > 0$ so that $|f(x) - \alpha| < \epsilon$ whenever $0 < |x - a| < \delta_f$.

   (c) For every $\epsilon' > 0$ there exists $\delta_g > 0$ so that $|g(x) - \beta| < \epsilon$ whenever $0 < |x - a| < \delta_g$.

(d) Since $\delta < \delta_f$, if x satisfies $0 < |x - a| < \delta$, then x satisfies $0 < |x - a| < \delta_f$. Similarly, if x satisfies $0 < |x - a| < \delta$, then x satisfies $0 < |x - a| < \delta_g$.

We have

$$|(f(x) + g(x)) - (\alpha + \beta)| = |f(x) - \alpha + g(x) - \beta|$$

Applying the triangle inequality yields

$$|f(x) - \alpha + g(x) - \beta| \leq |f(x) - \alpha| + |g(x) - \beta|$$

Let $x \in \mathbb{R}$ so that $0 < |x - a| < \delta$. Therefore $0 < |x - a| < \delta_f$. and so $|f(x) - \alpha| < \epsilon'$. Similarly, $|g(x) - \beta| < \epsilon'$. Therefore

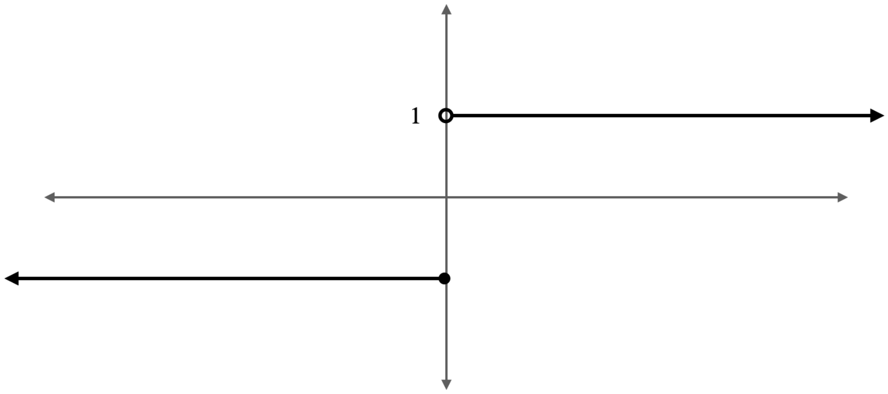
$$|(f(x) + g(x)) - (\alpha + \beta)| \leq |f(x) - \alpha| + |g(x) - \beta| < \epsilon' + \epsilon' = \epsilon$$

Therefore for every $\epsilon > 0$ there exists $\delta > 0$ so that $|(f(x)+g(x))-(\alpha+\beta)| < \epsilon$ whenever $0 < |x - a| < \delta$ Therefore $\lim_{x \to a} h(x) = \alpha + \beta$.

5.1.1 Divergence as x Approaches a

Consider the following familiar example:

$$f(x) = \begin{cases} 1 & x > 0 \\ -1 & x \leq 0 \end{cases}$$



Our work in previous Calculus courses tell us that the limit of $f(x)$ as x approaches 0 does not exist by virtue of examining the left-side and right-side limits:

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

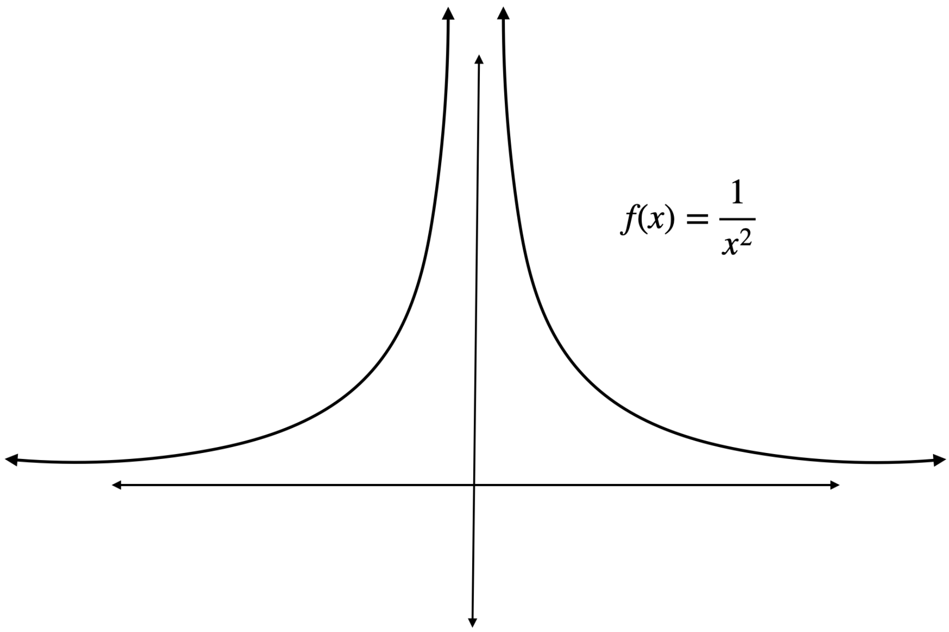
From this we would conclude that this limit does not exist as x approaches a. As we did in Section 4, we define divergence to mean *not converge to any* $L \in \mathbb{R}$.

Definition 5.4. *Let $E \subseteq \mathbb{R}$, let a be a limit point of E and let $f : E \to \mathbb{R}$. We say $\underline{f(x) \text{ diverges as } x \text{ approaches } a}$ when the statement*

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

is false for all $L \in \mathbb{R}$

Just as we saw for sequences, divergence comes in different flavours, The function $f(x)$ above does not converge to any L as x approaches 0, and so by definition it diverges. Similarly, the function $g(x) = \frac{1}{x^2}$ does not converge to any L as x approaches 0, and so by definition it diverges.

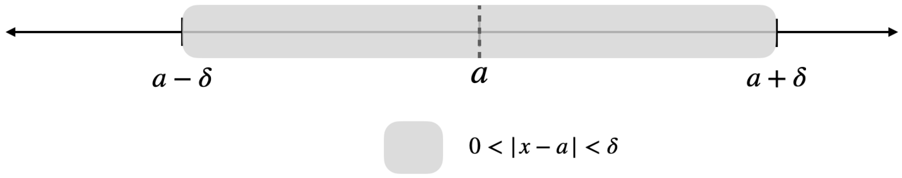


However, our intuition for why each of these two functions diverge is completely different. In the former case, the function behaves differently on the left and right side of 0. Whereas, in the latter case, the function goes off towards infinity as x approaches 0. We deal with these possibilities[5] in turn:

One Sided Limits Looking back at our definition of the limit, let us focus our attention on statement:

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

This part of the definition is specifying a region in which we are to take our values for x:



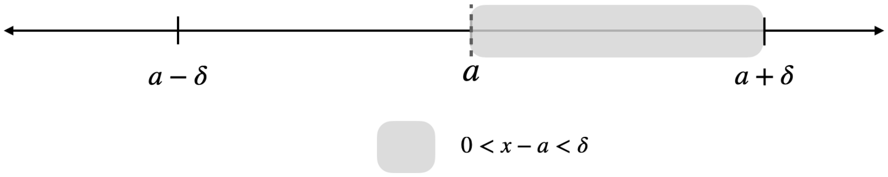
Recalling our definition of absolute value, the statement above is equivalent to:

$$-\delta < x - a < \delta \text{ and } x \neq a$$

To restrict our attention to only the right side of a (i.e $x > a$), we require the following condition:

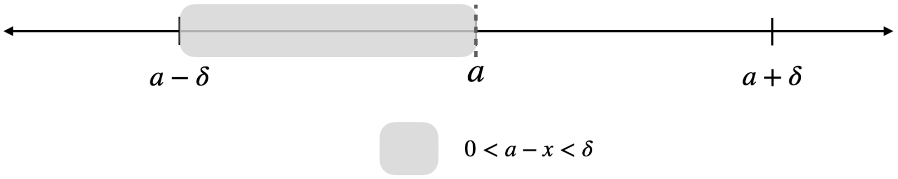
$$0 < x - a < \delta$$

[5] These are not the only reasons by a function might diverge. Look back at the example from the Introductory section. $f(x)$ diverges as x approaches $\sqrt{2}$



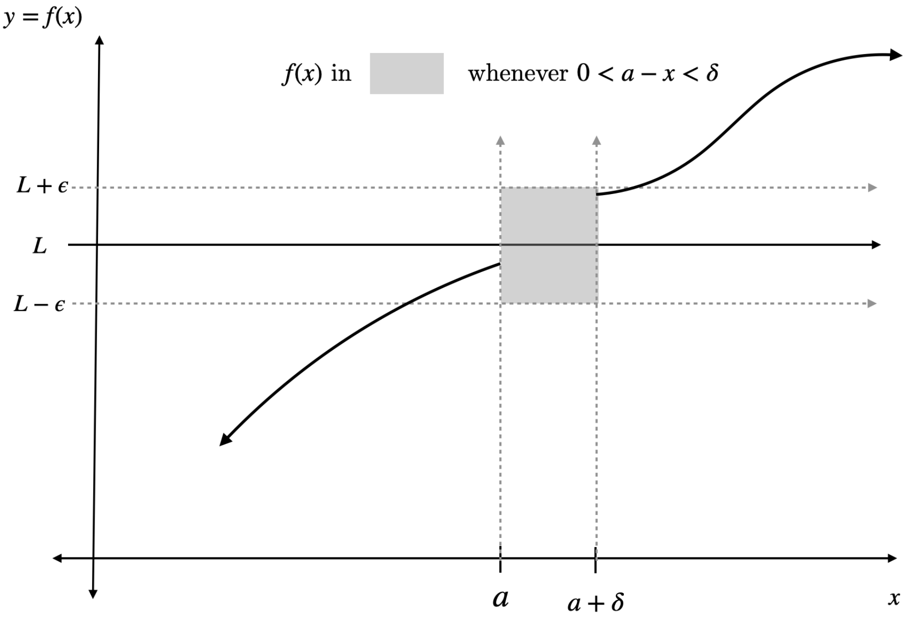
Whereas to restrict to the left side of a (i.e., $x < a$) we require the following condition:

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

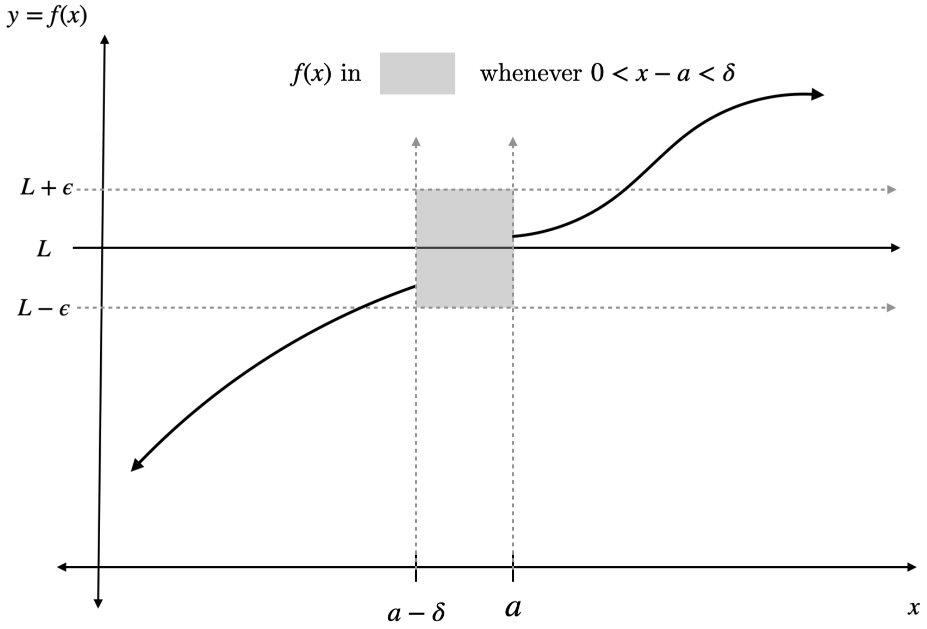


And so we arrive at the following definitions for *one-sided* limits.

Definition 5.5. *Let $E \subseteq \mathbb{R}$, let a be a limit point of E, let $f : E \to \mathbb{R}$ and let $L \in \mathbb{R}$. We say the limit of $f(x)$ as x approaches a from the right is L when for every $\epsilon > 0$ there exists $\delta > 0$ so that $|f(x) - L| < \epsilon$ whenever $0 < x - a < \delta$. When the limit of $f(x)$ as x approaches a from the right is L we write $\lim\limits_{x \to a^+} f(x) = L$.*



Definition 5.6. *Let $E \subseteq \mathbb{R}$, let a be a limit point of E, let $f : E \to \mathbb{R}$ and let $L \in \mathbb{R}$ We say the limit of $f(x)$ as x approaches a from the left is L when for every $\epsilon > 0$ there exists $\delta > 0$ so that $|f(x) - L| < \epsilon$ whenever $0 < a - x < \delta$. When the limit of $f(x)$ as x approaches a from the left is L we write $\lim\limits_{x \to a^-} f(x) = L$.*



With these definitions in place, we state a familiar theorem that relates our three flavours of limit:

Theorem 5.7. *Let $E \subseteq \mathbb{R}$, let a be limit point of E, let $f : E \to \mathbb{R}$ and let $L \in \mathbb{R}$. We have $\lim\limits_{x \to a} f(x) = L$ if and only if $\lim\limits_{x \to a^-} f(x) = L$ and $\lim\limits_{x \to a^+} f(x) = L$.*

The proof of Theorem 5.7 is an exercise in unravelling definitions, whose essence is captured by the following statement in formal logic:

$$[(0 < a - x < \delta) \vee (0 < x - a < \delta)] \Leftrightarrow (0 < |x - a| < \delta)$$

If $\lim\limits_{x \to a} f(x) = L$, then for any $\epsilon > 0$ there exists $\delta > 0$ so that $|f(x) - L| < \epsilon$ whenever $0 < |x - a| < \delta$. By definition of absolute value we can re-write $0 < |x - a| < \delta$ as:

$$0 < x - a < \delta \text{ or } 0 < a - x < \delta$$

Thus, if $\lim\limits_{x \to a} f(x) = L$, then for any $\epsilon > 0$ there exists $\delta > 0$ so that $|f(x) - L| < \epsilon$ whenever $0 < x - a < \delta$ and whenever $0 < a - x < \delta$. In other words $\lim\limits_{x \to a^-} f(x) = L$ and $\lim\limits_{x \to a^+} f(x) = L$.

We leave the argument for the converse as a task in Test Your Understanding

Returning to our motivating example above:

$$f(x) = \begin{cases} 1 & x > 0 \\ -1 & x \leq 0 \end{cases}$$

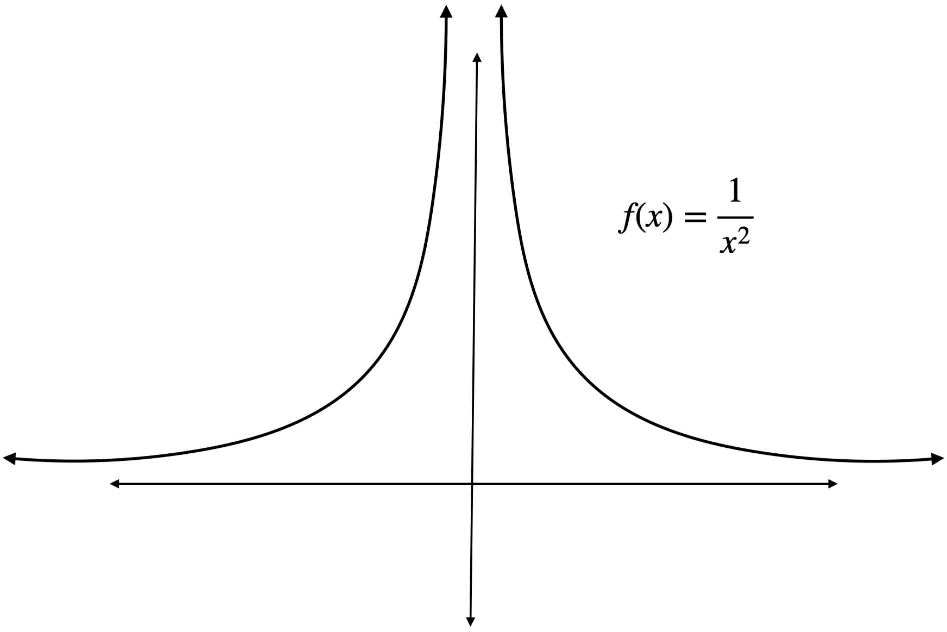
Theorem 5.7 implies that there can be no $L \in \mathbb{R}$ for which the statement $\lim\limits_{x \to 0} f(x) = L$ is true. As otherwise we would have $\lim\limits_{x \to 0^-} f(x) = \lim\limits_{x \to 0^+} f(x) = L$. And so Theorem 5.7 gives us the following corollary for proving divergence.

Corollary 5.8. *Let $E \subseteq \mathbb{R}$, let a be a limit point of E and let $f : E \to \mathbb{R}$. If*

$$\lim_{x \to a^-} f(x) \neq \lim_{x \to a^+} f(x),$$

then $f(x)$ diverges as x approaches a.

Exploding Limits As with limits of sequences, there is another flavour of divergence to consider. Consider the function $f(x) = \frac{1}{x^2}$. Though $x = 0$ is not in the domain of f, it is a limit point of the domain of f. And so it is reasonable to consider the limit of f as x approaches 0.



Using the language of our work in sequences, we would describe the behaviour of this function as *unbounded above* near $x = 0$. Here the word *unbounded* is describing the set of values of $f(x)$ whenever x is close to 0.

As we delve into this thought, let us take a moment to remind ourselves of a definition from Assignment 1:

Definition. *Let $f : \mathbb{R} \to \mathbb{R}$ be a function. We say the limit of $f(x)$ as x goes to ∞ is ∞ when for every $r \in \mathbb{R}$ there exists $k \in \mathbb{R}$ so that $f(x) > r$ for every $x > k$.*

Our intuition for the the limit of $f(x)$ as x goes to ∞ is ∞ is that $f(x)$ gets larger and stays larger than any value $r \in \mathbb{R}$. In other words, for every $r \in \mathbb{R}$ we can find $k \in \mathbb{R}$ so that $f(x) > r$ whenever $k > x$.

As we turn our attention to limits approaching a value a, this intuition still applies.

Definition 5.9. *Let $E \subseteq \mathbb{R}$, let a be a limit point of E and let $f : E \to \mathbb{R}$. We say $\underline{f(x) \text{ diverges to infinity as } x \text{ approaches } a}$ when for every $r \in \mathbb{R}$ there exists $\delta > 0$ so that $f(x) > r$ whenever $0 < |x - a| < \delta$. When $f(x)$ diverges to infinity as x approaches a we write $\lim\limits_{x \to a} f(x) = \infty$.*

Using this definition we show

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

Let $r \in \mathbb{R}$. We want to find δ so that $f(x) > r$ whenever x satisfies $0 < |x - 0| < \delta$. As is customary in our work, we first consider a particular value of r to build some intuition. Let $r = 100$. We want to find all values x so that $\frac{1}{x^2} > 100$.

$$\begin{aligned} \frac{1}{x^2} &> 100 \\ \frac{1}{100} &> x^2 \\ \frac{1}{10} &> |x| \\ |x - 0| &< \frac{1}{10} \end{aligned}$$

Therefore $\frac{1}{x^2} > 100$ whenever $0 < |x - 0| < \frac{1}{10}$. Similarly, we expect $\frac{1}{x^2} > r$ whenever $0 < |x - 0| < \frac{1}{\sqrt{|r|}}$.

Using this intuition we give a proof that $\lim\limits_{x \to 0} \dfrac{1}{x^2} = \infty$. Let $r \in \mathbb{R}$, let $\delta = \frac{1}{\sqrt{|r|}}$ and let x satisfy $0 < |x - 0| < \delta$.

$$\begin{aligned} |x - 0| &< \delta \\ |x - 0| &< \frac{1}{\sqrt{|r|}} \\ \frac{1}{x^2} &> |r| \geq r \end{aligned}$$

Therefore $\frac{1}{x^2} > r$ whenever $0 < |x - 0| < \frac{1}{\sqrt{|r|}}$. And so we conclude $\lim\limits_{x \to 0} \dfrac{1}{x^2} = \infty$.

Test Your Understanding

1. Let
$$f(x) = \begin{cases} 1 & x > 0 \\ -1 & x \leq 0 \end{cases}$$
Using $\epsilon - \delta$ arguments, prove
$$\lim_{x \to 0^-} f(x) = -1 \text{ and } \lim_{x \to 0^+} f(x) = 1$$

2. Prove $\lim_{x \to 0} \frac{1}{x^4} = \infty$

3. Give a reasonable definition for the following notations:
   (a) $\lim_{x \to a} f(x) = -\infty$
   (b) $\lim_{x \to a+} f(x) = \infty$
   (c) $\lim_{x \to a-} f(x) = \infty$

4. Let $E \subseteq \mathbb{R}$, let $f : E \to \mathbb{R}$, let a be limit point of E and let $L \in \mathbb{R}$. Using an ϵ-δ argument, prove that if $\lim_{x \to a^-} f(x) = \lim_{x \to a^+} f(x) = L$, then $\lim_{x \to a} f(x) = L$.

Test Your Understanding - Answers

1. In both cases we have $|f(x) - L| = 0$. And so we can choose any value of δ we would like. For example, we can let $\delta = 1$.

2. Let $r \in \mathbb{R}$. We want to find $\delta > 0$ so that $f(x) > r$ whenever $0 < |x - a| < \delta$. Choose $\delta = \frac{1}{\sqrt[4]{|r|}}$.

3. (a) for every $r \in \mathbb{R}$ there exists $\delta > 0$ so that $f(x) < r$ whenever $0 < |x - a| < \delta$.

   (b) for every $r \in \mathbb{R}$ there exists $\delta > 0$ so that $f(x) > r$ whenever $0 < x - a < \delta$.

   (c) for every $r \in \mathbb{R}$ there exists $\delta > 0$ so that $f(x) > r$ whenever $0 < a - x < \delta$.

5.1.2 Limits to Infinity

Our discussion about limits going to infinity to be a quick one because, without realising it, we have have already spent time thinking about the main ideas here.

Recall the definition of convergence for sequences:

Definition. *Let (f_n) be a sequence and let $L \in \mathbb{R}$. We say (f_n) converges to L when for every $\epsilon > 0$ there exists $M \in \mathbb{N}^+$ so that $|f_n - L| < \epsilon$ whenever $n > M$. When (f_n) converges to L we write $f_n \to L$ or $\lim\limits_{n\to\infty} f_n = L$.*

The intuition for this idea is that the distance between (f_n) and L gets smaller and stays smaller than any $\epsilon > 0$. In our work we looked at the sequence given by $f_n = 1 + \frac{1}{n}$. We showed that for any particular value of ϵ, we have $|f_n - 1| < \epsilon$ whenever $n > 1/\epsilon$.

Very little about these ideas needs to change as we transition from sequences to functions:

Definition (Attempt #1)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ We say $f(x)$ converges to L as x goes to ∞ when for every $\epsilon > 0$ there exists $M \in \mathbb{R}$ so that $|f(x) - L| < \epsilon$ whenever $x > M$.*

From our discussion at the start of this section, we can immediately notice a shortcoming of this definition: this definition requires our function to have domain $\mathbb{R}$. In our work above, we replaced $\mathbb{R}$ with E, a subset of $\mathbb{R}$. However here, just a straight swap does not suffice.

For example, consider the function $f(x) = \frac{1}{x}$ on the domain $E = (3, 7)$. Since the domain does not contain values that permit us to *approach* infinity, the statement $\lim\limits_{x\to\infty} f(x) = \infty$ doesn't make a lot of sense. To consider the value of a function as x goes to infinity, we require the specified domain of that function to contain values that *approach infinity.* In other words, the specified domain must not be bounded above in $\mathbb{R}$.

Definition 5.10. *Let $E \subseteq \mathbb{R}$. We say E has a limit point at ∞ when E is not bounded above in $\mathbb{R}$.*

To consider the limit of a function as x goes to infinity, we require the specified domain to have a limit point[6] at ∞.

Definition 5.11. *Let E be a subset of $\mathbb{R}$ so that E has a limit point at ∞ and let $f : E \to \mathbb{R}$. We say $f(x)$ converges to L as x goes to ∞ when for every $\epsilon > 0$ there exists $M \in \mathbb{R}$ so that $|f_n - L| < \epsilon$ whenever $n > M$. When $f(x)$ converges to L as x goes to ∞ we write $\lim\limits_{x\to\infty} f(x) = L$.*

As with sequences we define divergence accordingly:

Definition 5.12. *Let E be a subset of $\mathbb{R}$ so that E has a limit point at ∞ and let $f : E \to \mathbb{R}$. We say $f(x)$ diverges as x goes to ∞ when the statement*

$$f(x) \text{ converges to } L \text{ as } x \text{ goes to } \infty$$

is false for every $L \in \mathbb{R}$.

[6] See footnote 2. The set $\mathbb{N}$ has a limit point at ∞.

Just as we saw with sequences, we can notice different flavours of divergence. Consider $g(x) = sin(x)$ and $f(x) = 2x + 1$ each on the domain $\mathbb{R}$.

Using the definition of divergence, we prove $g(x)$ diverges as x goes to ∞. To do so, we proceed by contradiction.

Aside. *This argument will be almost identical to the argument from 4.1 that proves that the sequence* $(1, -1, 1, -1, \dots)$ *diverges. To ease what is about to come, take a moment to find this argument and remember how it went.*

Assume there exists $L \in \mathbb{R}$ so that $g(x) = \sin(x)$ converges to L as x goes to ∞. By definition, for every $\epsilon > 0$ there exists $M \in \mathbb{R}$ so that $|g(x) - L| < \epsilon$ whenever $x > M$. In particular, there exists $M \in \mathbb{R}$ so that $|g(x) - L| < 1$ whenever $x > M$.

Consider a value $x > M$ so that $\sin(x) = 1$

Therefore

$$|g(x) - L| = |1 - L| < 1$$

Let $x' = x + \pi$. By construction we have $\sin(x') = -1$ and thus

$$1 > |g(x') - L| = |-1 - L| = |-(1 + L)| = |1 + L|$$

Combining these two statements we have

$$|1 - L| + |1 + L| < 2$$

By the Triangle Inequality, we have

$$|1 - L + 1 + L| \leq |1 - L| + |1 + L|$$

Therefore

$$2 = |1 - L + 1 + L| \leq |1 - L| + |1 + L| < 2$$

A contradiction.

For $\epsilon = 1$ there is no value of M for which $|g(x) - L| < \epsilon$ whenever $n > M$. Therefore $g(x)$ does not converge to L for any $L \in \mathbb{R}$.

Consider now $f(x) = 2x + 1$. Recall the following definition from Assignment 1:

Definition. *Let* $f : \mathbb{R} \to \mathbb{R}$ *be a function. We say the limit of* $f(x)$ *as* x *goes to* ∞ *is* ∞ *when for every* $r \in \mathbb{R}$ *there exists* $M \in \mathbb{R}$ *so that* $f(x) > r$ *for every* $x > M$.

From our discussion above, we immediately notice a shortcoming of this definition: this definition requires our function to have domain $\mathbb{R}$. To consider the limit of a function as x goes to infinity, we require the specified domain to have a limit point at ∞.

Definition 5.13. *Let* E *be a subset of* $\mathbb{R}$ *so that* E *has a limit point at* ∞ *and* $f : E \to \mathbb{R}$. *We say the limit of* $f(x)$ *as* x *goes to* ∞ *is* ∞ *when for every* $r \in \mathbb{R}$ *there exists* $M \in \mathbb{R}$ *so that* $f(x) > r$ *whenever* $x > M$. *When the limit of* $f(x)$ *as* x *goes to* ∞ *is* ∞ *we write*

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

From our work on Assignment 1, for $r \in \mathbb{R}$, we have $f(x) = 2x+1 > r$ whenever $x > \frac{r-1}{2}$. Therefore

$$\lim_{x \to \infty} 2x + 1 = \infty$$

Test Your Understanding

1. Determine if each set below as a limit point at ∞
   (a) $(3, 7)$
   (b) $[3, 7]$
   (c) $(3, \infty)$
   (d) $(-\infty, 3]$
   (e) $\mathbb{Q}$
   (f) $\mathbb{R} \setminus \mathbb{N}$.
2. What is a reasonable definition for the phrase limit point at -∞?

Test Your Understanding - Answers

1. Determine if each set below as a limit point at ∞
   (a) No
   (b) No
   (c) Yes
   (d) No
   (e) Yes
   (f) Yes
2. **Definition.** *Let $E \subseteq \mathbb{R}$. We say $\underline{E \text{ has a limit point at } -\infty}$ when E is not bounded below in $\mathbb{R}$.*

5.2 Continuity

Let us return to the definition of continuity we saw from the start of Section 5:

Let $f : \mathbb{R} \to \mathbb{R}$, we say $\underline{f \text{ is continuous on } \mathbb{R}}$ when for every $a \in \mathbb{R}$ we have

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

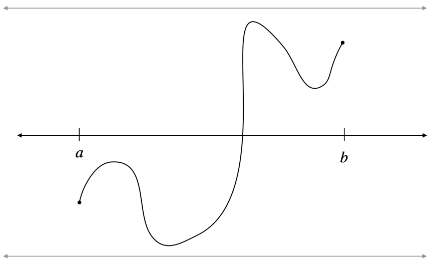
Recalling our thoughts from last section on the definition of the notation $\lim\limits_{x \to a} f(x) = L$, perhaps we can start to see some shortcomings of this definition. For one, we want to be able to consider continuity on domains other than $\mathbb{R}$. And, in doing so, we may need to restrict a to be a limit point of the domain. What's further, given that we will be evaluating f at a, we need a to be in the domain of f.

Our work in 5.2 proceeds as follows. We begin by refining our definition of continuity above. From our work in 5.1 we can unpack the notation $\lim\limits_{x \to a} f(x) = f(a)$ into a statement about ϵ's and δ's. In doing so, we will realize that we can relax some of the considerations that arise in the $\epsilon - \delta$ definition of the limit.

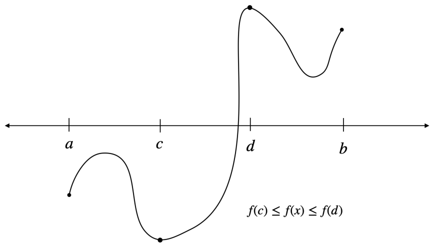
Once we have done this, we will spend some time developing results that let us prove functions are continuous without having to resort to an $\epsilon - \delta$ style argument.

Our study of continuity will culminate with the proofs of three *obvious* mathematical facts:

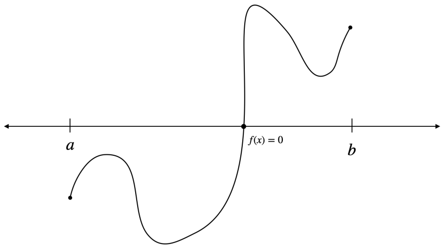
1. If a function is continuous on a closed interval, then it is bounded on that interval



2. If a function is continuous on a closed interval, then it attains a maximum/minimum value on that interval



3. If a continuous function is below the x-axis at one point, and above the x axis at another point, then it must cross the x-axis at some point.



Each of these facts are *obvious* to us based on our intuition for the meaning of continuity. If our definition of continuity fully represents our intuition for the concept, then we should be able to give mathematical proofs for these facts.

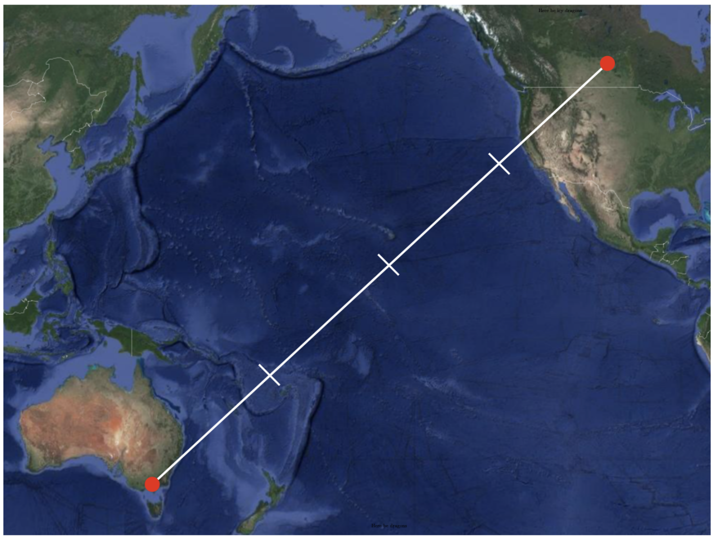
Before we start our work, however, it is reasonable for us to wonder – what is the point of all of this? Why is it important that we delve so deeply into limits and continuity? We know already know what these things are and how to use them, why does it matter than we can prove things about them?

Beyond the vague assurances that these concepts will come up again in applied contexts during future studies, permit me to make a different sort of argument for the value of studying abstract mathematics. Much like literature or art, learning abstract mathematics equips you with a cultural literacy.

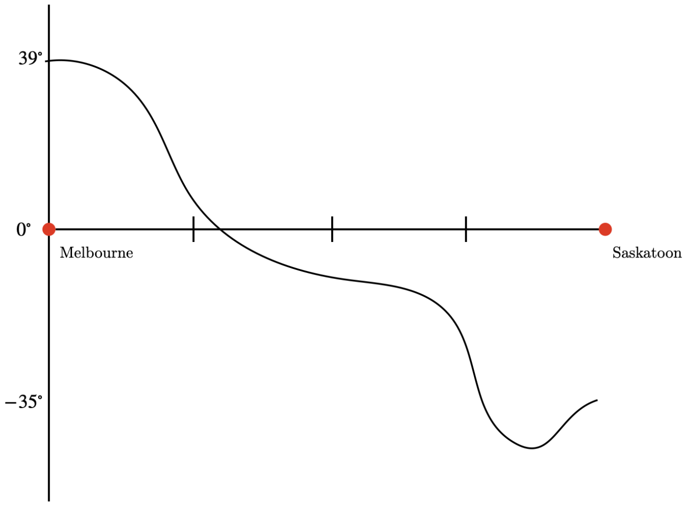
In a purely economic sense, there is little *value* in knowledge of the arts – none of us is financially better off for our ability to recognize modern and historical music, to recite a few lines of poetry, or to appreciate sculpture. However, as a society, we ascribe value to these things not because of their potential for economic output, but because we recognize that the study of literature and art permits us avenues to fully express our humanity.

Through the study of abstract mathematics we equip ourselves with tools we can use to satisfy our curiousity about the world around us. Once we have modeled a physical phenomenon in mathematical terms, no longer are we bound only to conclusions we can draw with our senses. By learning to reason about objects we cannot physically sense, we are no longer bound to draw conclusions based only on our physical experiences.

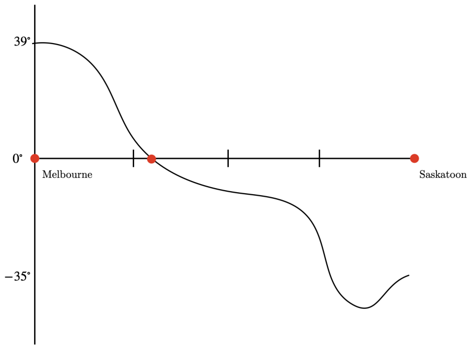
Consider the following concrete example. On Monday 25 January 2021, the temperature hit 39°C in Melbourne. Simultaneously, the temperature hit a low of -35°C in Saskatoon, Canada. The line on the map below runs from Melbourne to Saskatoon.



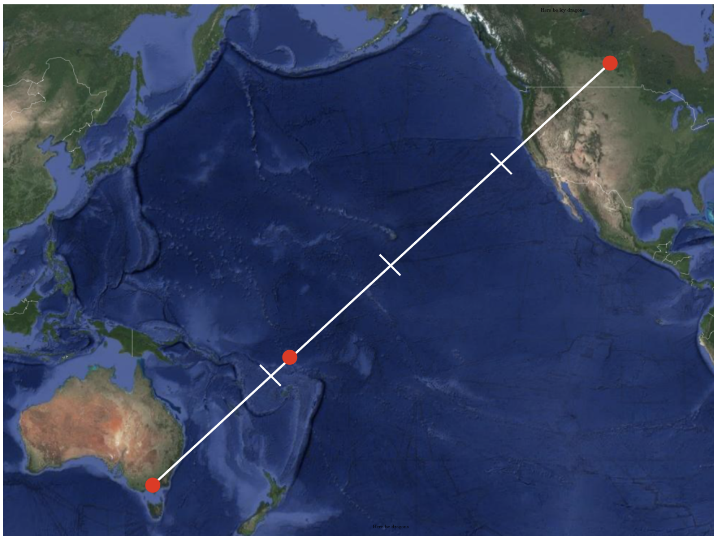
Consider measuring the temperature at each point on the line at the moment it was -35°C in Saskatoon and 39°C in Melbourne. Since temperature is, in some sense, *continuous*, the temperature line between -35°C in Saskatoon and 39°C in Melbourne would be unbroken. It would look something like this:



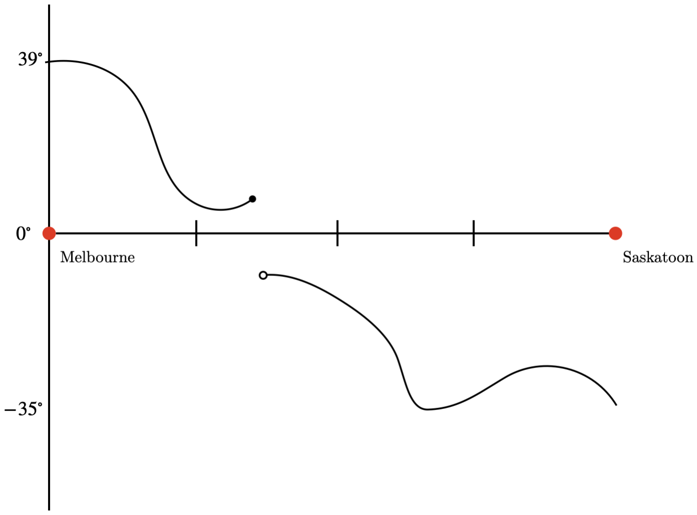
Since the temperature was below 0 in Saskatoon and above 0 in Melbourne, there must have been some point on the line for which the temperature is exactly 0.



Without having to actually visit Saskatoon or any of the points on the line between Saskatoon and Melbourne we can conclude that there was some location between Melbourne and Saskatoon for which the temperature was exactly $0°C$.



This argument hangs on out intuition of the word *continuous.* If temperature were not continuous, if it could jump instantly from one temperature in one location to a vastly different temperature in an adjacent location, then our conclusion is not assured.



And so let us spend some time studying the mathematical meaning of the word *continuous.*

Let us return to our naive definition of continuity from Calculus 2

Definition (Attempt #1)**.** *Let $E \subseteq \mathbb{R}$ Let $f : E \to \mathbb{R}$, we say f is continuous on $\mathbb{R}$ when for every $a \in \mathbb{R}$ we have*

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

Looking at this definition a little more carefully, we notice something odd. Often we think of continuity as a *global* property of a function. However, our definition for continuity relies on drawing a conclusion about $f(x)$ for each single point in its domain. It is from these *local* properties of continuity that the global property of continuity occurs; a function is *continuous* when something convenient is happening at each point in its domain.

As our naive definition of continuity is statement about convergence at every element of the domain, our newly found knowledge of limits helps us understand a little better the meaning the statement $\lim_{x \to a} f(x) = f(a)$.

Having $\lim_{x \to a} f(x) = f(a)$ means that for every $\epsilon > 0$ there exists $\delta > 0$ so that $|f(x) - f(a)| < \epsilon$ whenever $0 < |x - a| < \delta$. Since a is an element of the domain of E (i.e., $f(a)$ is defined) we no longer need to worry about the condition $|x - a| > 0$. When $x = a$, the statement $|f(x) - f(a)| < \epsilon$ is always true: $f(a) - f(a) = 0$. And so we can slightly relax our requirements on x to permit $x = a$. Since we permit a in the interval $|x - a| < \delta$, no longer do we require there to exist values in the domain that *approach* a. When a is in the domain of f, the set of solutions for $|x - a| < \delta$ is always non-empty.

These thoughts permit us to refine our definition of continuity.

Definition 5.14 ($\epsilon - \delta$ Definition of Continuity)**.** *Let $E \subseteq \mathbb{R}$, let $f : E \to \mathbb{R}$ and let $a \in E$. We say $\underline{f \text{ is continuous at } a}$ when for every $\epsilon > 0$ there exists $\delta > 0$ so that $|f(x) - f(a)| < \epsilon$ whenever $x \in E$ satisfies $|x - a| < \delta$. We say $\underline{f \text{ is continuous on } E}$ when f is continuous at a for all $a \in E$.*

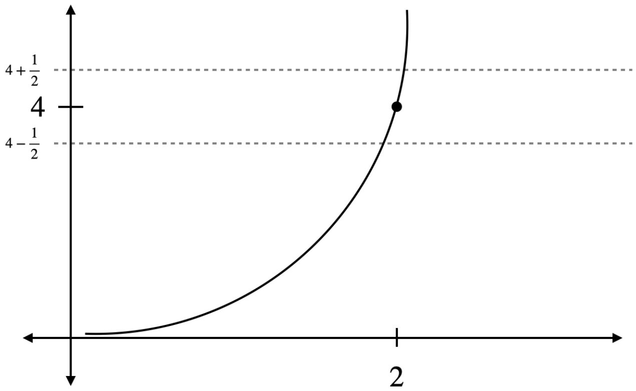
Aside. *There is no difference between the phrase f* is continuous at $x = a$ *and the f* is continuous at a. *Feel welcome to use either.*

To say $\underline{f \text{ is continuous at } a}$ means that a small change from $f(a)$ to something in the interval $(f(a) - \epsilon, f(a) + \epsilon)$ requires only a change in a to something in the interval $(a - \delta, a + \delta)$. No matter how small of an ϵ we choose, we can always find δ so that changing a by less than δ results in a change in $f(a)$ of less than ϵ. No matter how close we look to a we will always find a point, x, *close* to a so that $f(x)$ is *close* to $f(a)$.

Let's look at a function we *know* to be continuous. Consider $f(x) = x^2$ at $x = 2$.

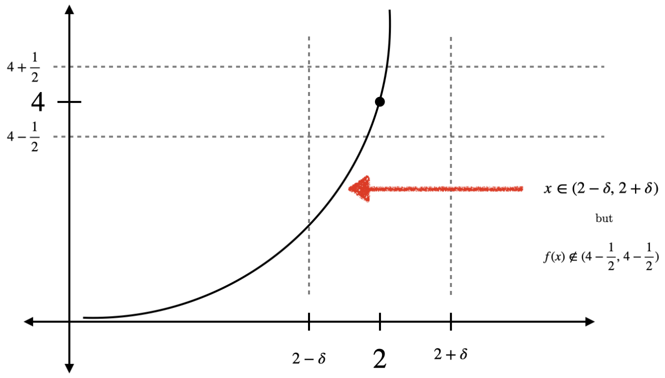
Since $f(x)$ is continuous on $\mathbb{R}$, then it is continuous at 2. This means, that no matter what choice of $\epsilon > 0$ we declare, we can find a value for δ that so whenever x is constrained within $2 - \delta$ and $2 + \delta$, then $f(x)$ is constrained within $2^2 - \epsilon$ and $2^2 + \epsilon$.

For example, consider $\epsilon = \frac{1}{2}$

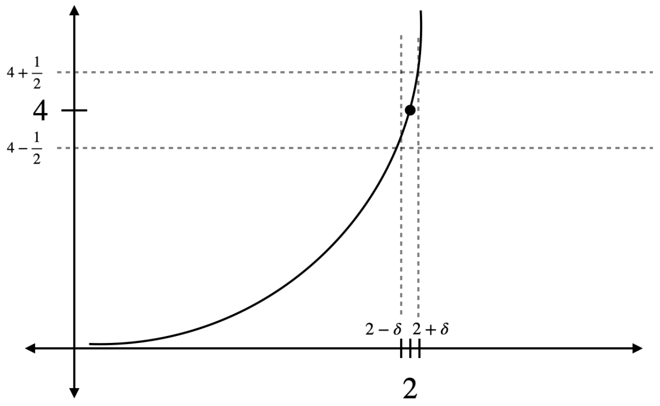


Since f is continuous at 2 we can find δ so that if x is constrained within $2-\delta$ and $2+\delta$, then $f(x)$ is constrained within $4-\frac{1}{2}$ and $4+\frac{1}{2}$.

The choice of δ shown below is a bad choice. It is possible to have $x \in (2-\delta, 2+\delta)$ but have $f(x) \notin (4+\frac{1}{2}, 4-\frac{1}{2})$.



On the other hand, the choice of δ shown below is a good choice. We observe that whenever $x \in (2-\delta, 2+\delta)$ we have $f(x) \in (4+\frac{1}{2}, 4-\frac{1}{2})$.



The statement

$$f(x) = x^2 \text{ is continuous at 2}$$

means that we can choose any $\epsilon > 0$ we want, and be certain we can find a choice of δ so that having $x \in (2-\delta, 2+\delta)$ guarantees $f(x) \in (4+\frac{1}{2}, 4-\frac{1}{2})$.

This discussion suggests to us the work we must do in order to prove that a function is continuous at point. Given any $\epsilon > 0$ we must find $\delta > 0$ so that $f(x) \in (f(a)-\epsilon, f(a)+\epsilon)$ whenever $x \in (a-\delta, a+\delta)$.

Let $f(x) = x$. We prove f is continuous at 2.

Let $\epsilon > 0$. To prove f is continuous at 2 we want to find a value for δ so that $|f(x)-f(2)| < \epsilon$ whenever $|x-2| < \delta$. Substituting in $f(2) = 4$, we want to find a value for δ so that $|f(x)-4| < \epsilon$ whenever $|x-2| < \delta$.

Let $\delta = \epsilon$. Directly it follows that if $|x-2| < \delta$, then $|f(x)-4| < \epsilon$. Therefore f is continuous at 2.

A similar argument shows that f is continuous at a for all $a \in \mathbb{R}$. Therefore f is continuous on $\mathbb{R}$.

Example 5.15. Let $f(x) = x$. Prove f is continuous on $\mathbb{R}$.

Solution: *To prove f is continuous on $\mathbb{R}$ we must prove f is continuous at a for all $a \in \mathbb{R}$.*

Let $a \in R$. To prove f is continuous at a we want to find a value for δ so that $|f(x) - f(a)| < \epsilon$ whenever $|x-a| < \delta$. Substituting in $f(a) = a$, we want to find a value for δ so that $|f(x)-a| < \epsilon$ whenever $|x-a| < \delta$. Let $\delta = \epsilon$. Directly it follows that if $|x-a| < \delta$, then $|f(x)-a| < \epsilon$. Therefore f is continuous at a. Since f is continuous at a for all $a \in \mathbb{R}$, f is continuous on $\mathbb{R}$.

Similarly, we can prove the function $f(x) = 1$ is continuous on $\mathbb{R}$.

Example 5.16. Let $f(x) = 1$. Prove f is continuous on $\mathbb{R}$.

Solution: *To prove f is continuous on $\mathbb{R}$ we must prove f is continuous at a for all $a \in \mathbb{R}$.*

Let $a \in \mathbb{R}$. To prove f is continuous at a we want to find a value for δ so that $|f(x) - f(a)| < \epsilon$ whenever $|x - a| < \delta$. Since $f(x) = 1$ for all $x \in \mathbb{R}$ we have $|f(x) - f(a)| = 0$

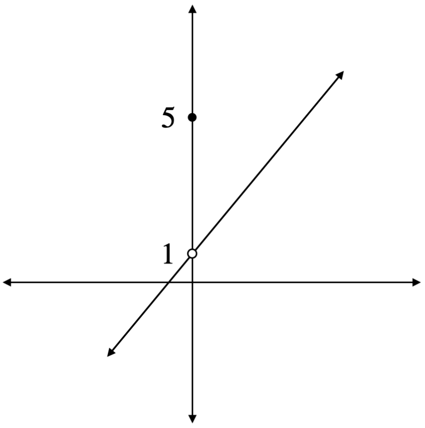
Let $\delta = 1$. Since $|f(x) - f(a)| = 0 < \epsilon$ for all $x \in \mathbb{R}$, then $|f(x) - f(a)| < \epsilon$ for all x that satisfy $|x - a| < \delta$. Therefore f is continuous at a. Since f is continuous at a for all $a \in \mathbb{R}$, f is continuous on $\mathbb{R}$

To broaden our understanding, we consider an example of a function that is not continuous on $\mathbb{R}$.

Example 5.17. Let

$$f(x) = \begin{cases} 2x + 1 & x \neq 0 \\ 5 & x = 0 \end{cases}$$

Prove f is not continuous at 0.



Solution*:*

To prove that f is not continuous at 0 it suffices to find an $\epsilon > 0$ so that for every $\delta > 0$ there exists $x \in \mathbb{R}$ so that $|x - 0| < \delta$ but $|f(x) - f(0)| \geq \epsilon$.

Consider $\epsilon = 1$ and $\delta > 0$. Let[7] *$x = -\delta/2$.*

We have $|f(x) - f(0)| = |-\delta + 1 - 5| = |-\delta - 4| = |4 + \delta| = 4 + \delta > 1$

Therefore there is no value $\delta > 0$ for which $|f(x) - f(0)| < 1$ whenever $|x - 0| < \delta$. Therefore f is not continuous at 0.

Our process for testing if a function is continuous at a point feels a lot like verifying a limit. Given our $\epsilon - \delta$ definition of continuity, this perhaps shouldn't be too surprising.

[7] We want to find x so that $|x| < \delta$ and $|f(x) - 5| > 1$. Notice $|f(x) - 5| = |5 - f(x)|$. If we choose x so that $2x < 0$ we will have $|5 - (2x + 1)| = |4 - 2x| > 4 > 1$. Taking $x = -\delta/2$ satisfies $2x < 0$ and makes the algebra easy.

Theorem 5.18. *Let $E \subseteq \mathbb{R}$, let $a \in E$ so that a is a limit point of E. Let $f : E \to \mathbb{R}$. The function f is continuous at a if and only if*

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

Aside. *There is something subtle happening here that isn't worth spending any time on, but it worth pointing out. Our definition of continuity at a doesn't require a to be a limit point of E. Since limits only make sense when we take them approaching a limit point, the statement of Theorem 5.18 requires that we have a be a limit point of E.*

Looking back at our definition of limit point, if a is not a limit point, then there exists $\delta > 0$ so that $x = a$ is the only point that satisfies $|x - a| < \delta$. When $x = a$ we have $|f(x) - f(a)| = 0$. Therefore for any $\epsilon > 0$, it follows that $|f(x) - f(a)| < \epsilon$ whenever $|x - a| < \delta$. And so we conclude that if a is not a limit point of E, but is an element of E, then f is continuous at a.

Since you didn't know about limit points in Calculus 2, this distinction couldn't have been made anytime before now.

Our definition of continuity requires us to make precise arguments with ϵ's and δ's, and our experience from Section 5.1 tells us that making these arguments can be difficult. And so we are pleased to see that like limits for sequences and real functions, continuity also admits an algebra theorem:

Theorem 5.19 (Algebra of Continuous Functions Theorem)**.** *Let $E \subseteq \mathbb{R}$, let $f, g : E \to \mathbb{R}$, and let $a \in E$. If $f(x)$ and $g(x)$ are both continuous at a, then*

i. *$k \cdot f(x)$ is continuous at a for all $k \in \mathbb{R}$*

ii. *$f(x) + g(x)$ is continuous at a;*

iii. *$f(x) \cdot g(x)$ is continuous at a; and*

iv. *$f(x)/g(x)$ is continuous at a; provided $g(a) \neq 0$.*

In the interest of brevity, we omit the proof of this theorem. We note, however, that these results follow almost directly from The Algebra of Limits Theorem and Theorem 5.18.

The Algebra of Continuous Real Functions Theorem is an important tool in the analyst's toolkit. Above, we proved that $f(x) = x$ is continuous. By being clever with the Algebra of Continuous Real Functions Theorem we can prove, for example, that every polynomial is continuous on $\mathbb{R}$.

Imagine we wanted to give a proof that a polynomial of degree 4 was continuous at 0.

$$h(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4$$

We can express $h(x)$ as: $h(x) = a_0 \cdot 1 + x g(x)$, where $g(x)$ is a polynomial of degree 3.

If $g(x)$ is continuous, then by part *ii* above, $xg(x)$ is continuous.

Further, from our work above, and part *i* above, we know that $a_0 \cdot 1$ is continuous. (Let $k = a_0$ and $f(x) = 1$).

Thus to verify $h(x)$ is continuous, we need only prove that $g(x)$, polynomial of degree 3, is continuous at a.

By applying the same reasoning, we realize that to prove a polynomial of degree 3 is continuous at a, then it suffices to prove that polynomials of degree 2 are continuous at a.

Applying the same reasoning again, we realize that to prove a polynomial of degree 2 is continuous at a, then it suffices to prove that polynomials of degree 1 are continuous at a.

In general, to prove that a polynomial of degree $k+1$ is continuous it suffices to know that polynomials of degree k are continuous. Induction is lurking!

Theorem 5.20. *Let $a \in \mathbb{R}$. For every $n \in \mathbb{N}^+$, every polynomial of degree n is continuous at a.*

Proof. Let $a \in \mathbb{R}$. We proceed by induction to prove every polynomial of degree $n \geq 1$ is continuous at a.

Let $h(x)$ be a polynomial of degree 1. Therefore there exists $k \in \mathbb{R}$ so that $h(x) = kx$. Let $f(x) = x$. From our work above, $f(x)$ is continuous at a. And so by part *i* of the Algebra of Continuous Real Functions Theorem, it follows that $h(x)$ is continuous.

Let $n = k$. Assume all polynomials of degree k are continuous at a.

Consider $h(x)$, a polynomial of degree $k+1$. Since $h(x)$ is a polynomial of of degree $k+1$, there exists coefficients $a_0, a_1, \ldots, a_{k+1} \in \mathbb{R}$ so that

$$h(x) = a_0 + a_1 x + \cdots + a_{k+1} x^{k+1}$$

Let $f(x) = 1$ and $g(x) = a_1 + a_2 x + \cdots + a_{k+1} x^k$. Notice

$$h(x) = a_0 f(x) + x g(x)$$

From our work above, $f(x)$ is continuous at a. And so by part *i* of the Algebra of Continuous Real Functions Theorem, it follows that $a_0 f(x)$ is continuous at a.

Since $g(x)$ is a polynomial of degree at most k, by induction it follows that $g(x)$ is continuous at a. From our work above and part *iii* of the Algebra of Continuous Real Functions Theorem, it follows that $xg(x)$ is continuous at a.

Since $a_0 f(x)$ is continuous at a and $xg(x)$ is continuous at a, it follows from part *ii* of the Algebra of Continuous Real Functions Theorem that $a_0 f(x) + xg(x)$ is continuous at a. Therefore $h(x)$ is continuous at a.

The result now follows from the Principle of Mathematical Induction. □

Since every polynomial is continuous at a for all $a \in \mathbb{R}$, the definition of continuity gives us the following result:

Corollary 5.21. *For every $n \in \mathbb{N}^+$, every polynomial of degree n is continuous on $\mathbb{R}$.*

Even appealing to the Algebra of Continuous Functions Theorem can make for lengthly work in proving that a function is continuous. Fortunately the community of mathematicians who have come before us have taken the time to prove that all of the functions we *expect* to be continuous on their domains indeed satisfy the $\epsilon - \delta$ definition of continuity.

Theorem 5.22. *The following functions are continuous on the stated domains:*

1. $e^x, \cos(x)$ *and* $\sin(x)$ *for* $E = \mathbb{R}$
2. $log(x)$ *for* $E = \mathbb{R}^+$
3. $x^{1/p}$ *for* $E = \mathbb{R}^+$ *and* $p = \mathbb{N}^+$
4. $a_0 + a_1 x + \cdots + a_n x^n$ *for* $E = \mathbb{R}$

Our goal in the remainder of 5.2 is to prove our three *obvious* facts about continuous functions.

1. If a function is continuous on a closed interval, then it is bounded on that interval.
2. If a function is continuous on a closed interval, then it attains a maximum/minimum value on that interval
3. If a continuous function is below the x-axis at one point, and above the x axis at another point, then it must cross the x-axis at some point.

Before we dive in to proving our results we require some terminology that will ease our discussion: Let us think a little more about the statement:

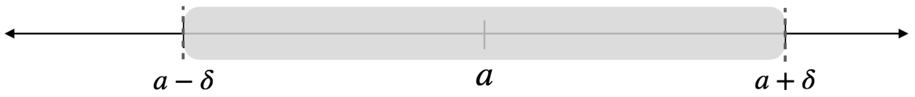
for every $\epsilon > 0$ there exists $\delta > 0$ so that $|f(x) - f(a)| < \epsilon$ whenever $|x - a| < \delta$.

We can re-write this second inequality as

$$\begin{aligned} -\delta < x - a < \delta \\ a - \delta < x < a + \delta \\ x \in (a - \delta, a + \delta) \end{aligned}$$

This is an interval we talk a lot about in the coming sections. And so let us introduce the following terminology.

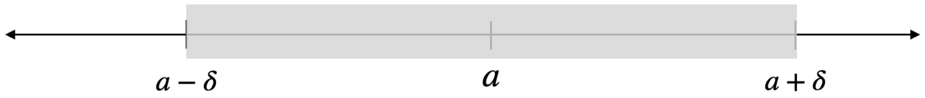
Definition 5.23. *Let $E \subseteq \mathbb{R}$, let $a \in \mathbb{R}$ and let $\delta > 0$. The open δ-neighbourhood of a is the interval $(a - \delta, a + \delta)$.*



An open $\delta-$neighbourhood of a (endpoints not included)

For reasons that will become clear in the following reading, we will also be interested in the interval $[a-\delta, a+\delta]$.

Definition 5.24. *Let $E \subseteq \mathbb{R}$, let $a \in \mathbb{R}$ and let $\delta > 0$. The closed δ-neighbourhood of a is the interval $[a-\delta, a+\delta]$.*



A closed $\delta-$neighbourhood of a (endpoints included)

Aside. *If we were only ever going to write mathematics, these two definitions probably wouldn't make our work any easier. Writing $x \in (a-\delta, a+\delta)$ is just as easy as writing* x is the the open δ-neighbourhood of a.

However, the practice of mathematics is more than just writing down mathematical statements, but communicating to others about mathematical statements. When communicating verbally about mathematics it is much easier to understand `x is the the open delta neighbourhood of a` *than it is to understand* `x is in the open interval whose end points are a minus delta and a plus delta`*.*

Test Your Understanding

1. Let f and g be continuous at a. Using only the Definition 5.14, prove that $f+g$ is continuous at a. (Hint: Proceed similarly to the proof of part i of the Algebra of Limits Theorem.)
2. Using the $\epsilon - \delta$ definition of continuity and Definition 5.14, prove Theorem 5.18 for functions with domain $\mathbb{R}$.

Test Your Understanding - Answers

1. Assume f and g are continuous at a. Let $\epsilon > 0$ and let $\epsilon' = \epsilon/2$. Since f and g are continuous at a, there exists $\delta_f, \delta_g > 0$ so that

   - $|f(x) - f(a)| < \epsilon'$ whenever $x \in (a - \delta_f, a + \delta_f)$; and
   - $|g(x) - f(a)| < \epsilon'$ whenever $x \in (a - \delta_g, a + \delta_g)$.

   Let $\delta = \min\{\delta_f, \delta_g\}$.

   Notice that if $x \in (a-\delta, a+\delta)$, then $x \in (a-\delta_f, a+\delta_f)$ and $x \in (a-\delta_g, a+\delta_g)$. Let $x \in (a - \delta, a + \delta)$ and consider $|(f(x) + g(x)) - (f(a) + g(a))|$. Using the Triangle Inequality we find

   $$\begin{aligned} |(f(x) + g(x)) - (f(a) + g(a))| &= |f(x) - f(a) + g(x) - g(a)| \\ &\leq |f(x) - f(a)| + |g(x) - g(a)| \\ &< \epsilon' + \epsilon' \\ &= \epsilon \end{aligned}$$

   Therefore $|(f(x) + g(x)) - (f(a) + g(a))| < \epsilon$ whenever $x \in (a - \delta, a + \delta)$. And so we conclude $f + g$ is continuous at a.

2. Hint: When $x = a$, the statement $|f(x) - f(a)| < \epsilon$ is always true. And so:

   there exists $\delta > 0$ so that $|f(x) - f(a)| < \epsilon$ whenever $0 < |x - a| < \delta$

   if and only if

   there exists $\delta > 0$ so that $|f(x) - f(a)| < \epsilon$ whenever $|x - a| < \delta$

5.2.1 Continuity Implies Boundedness

As we did in our study of convergent sequences, we consider the implication of continuity on boundedness of a function. To do so we first define boundedness to mean what we expect it to mean[8].

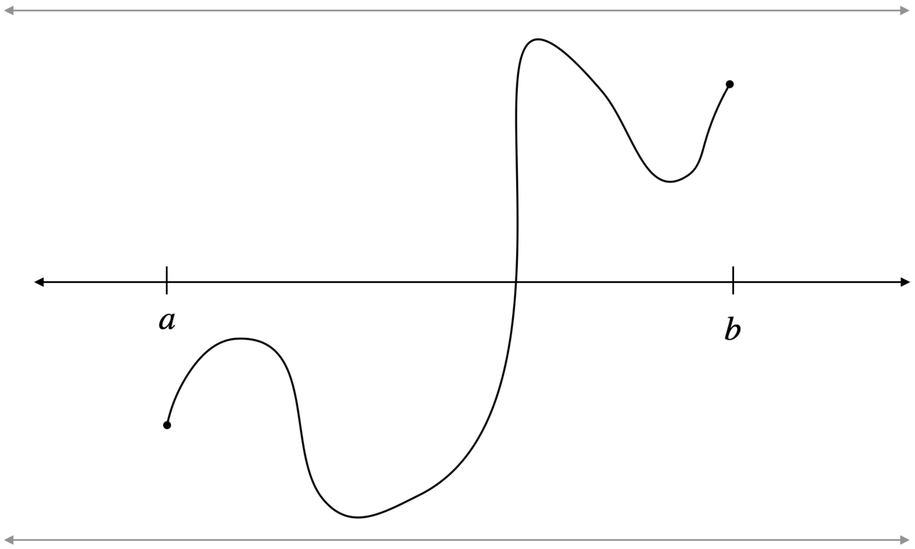
Definition 5.25. *Let $f : E \to \mathbb{R}$ and let $a, b \in E$ with $a \leq b$. Let $I \subseteq E$ be one of the following intervals: $(a, b), [a, b], (a, \infty), [a, \infty), (\infty, b), (-\infty, b]$ or $\mathbb{R}$. Let $A_f = \{f(x) \mid x \in I\}$. We say $\underline{f \text{ is bounded above on } I}$, when A_f is bounded above in $\mathbb{R}$. We say $\underline{f \text{ is bounded below on } I}$, when A_f is bounded below in $\mathbb{R}$. We say $\underline{f \text{ is bounded on } I}$, when A_f is bounded above and bounded below in $\mathbb{R}$.*

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

- $f(x)$ is bounded below on the interval $(0, \infty)$
- $f(x)$ is not bounded above on the interval $(0, \infty)$
- $f(x)$ is bounded on the interval $[1, 10]$.

Our goal in this section is to prove the following theorem:

Theorem 5.26. *Let $f : [a, b] \to \mathbb{R}$. If f is continuous on $[a, b]$, then f is bounded on $[a, b]$*



Before we dive in to our proof, let us take a moment to consider the relationship between continuity and boundedness.

Consider $f : [a, b] \to \mathbb{R}$ so that f is continuous on $[a, b]$. And consider $c \in (a, b)$. Our definition of continuity tells us that f is continuous at c; to be continuous on an interval means being continuous at every point in the interval.

[8] If boundedness means what we expect it to mean, why take the time to write down a definition? Having a definition of boundedness is essential if we want to prove that something is bounded. We prove something is bounded by proving it satisfies the definition of boundedness.

Since f is continuous at c, for every $\epsilon > 0$ there exists $\delta > 0$ so that $|f(x) - f(c)| < \epsilon$ whenever $|x - c| < \delta$. In particular, there exists $\delta > 0$ so that $|f(x) - f(c)| < 1$ for every $x \in (c - \delta, c + \delta)$. Rearranging this first inequality permits us to conclude

$$f(x) < 1 + f(c)$$

for all $x \in (c - \delta, c + \delta)$.

In other words, if f is continuous at c, then there exists an open δ-neighbourhood of c on which f is bounded above[9]. What's more, as f is continuous at $c - \delta$, then f is bounded above at $x = c - \delta$ by $f(c - \delta)$. Similarly, as f is continuous at $c + \delta$, then f is bounded above at $x = c + \delta$ by $f(c + \delta)$.

Taking the maximum of these three upper bounds (the one for the interval $(c - \delta, c + \delta)$, the one for the interval $[c - \delta, c - \delta]$ and the one for the interval $[c + \delta, c + \delta]$) gives an upper bound for f on the interval $[c - \delta, c + \delta]$.

Therefore if f is continuous at c, then there exists $\delta > 0$ so that f is bounded above on the interval $[c - \delta, c + \delta]$. A similar argument shows f is bounded below on the interval $[c - \delta, c + \delta]$. Therefore if f is continuous at c, then there exists $\delta > 0$ so that f is bounded on the interval $[c - \delta, c + \delta]$.

This seems useful. Let's write it down as a lemma.

Lemma 5.27. *Let $f : [a, b] \to \mathbb{R}$ so that f is continuous on $[a, b]$. For every $c \in (a, b)$ there exists $\delta > 0$ so that f is bounded on the closed δ-neighbourhood of c.*

Notably, this lemma requires $c \in (a, b)$, rather than $c \in [a, b]$. To see why, let us construct the same line of reasoning for an endpoint of the interval.

Consider $f : [a, b] \to \mathbb{R}$ so that f is continuous on $[a, b]$. Our definition of continuity tells us f is continuous at a; to be continuous on an interval means being continuous at every point in the interval.

Since f is continuous at a, for every $\epsilon > 0$ there exists $\delta > 0$ so that $|f(x) - f(a)| < \epsilon$ whenever $|x - a| < \delta$. Since the domain of f includes no values less than a, for all $x \in [a, b]$ we have $|x - a| = x - a$.

Therefore for every $\epsilon > 0$ there exists $\delta > 0$ so that $|f(x) - f(a)| < \epsilon$ whenever $x - a < \delta$. In particular, there exists $\delta > 0$ so that $|f(x) - f(a)| < 1$ whenever $x - a < \delta$. Rearranging this first inequality permits us to conclude

$$f(x) < 1 + f(a)$$

for all $x \in (a, a + \delta)$.

Proceeding as in the argument for Lemma 5.27, we conclude there exists $\delta > 0$ so that f is bounded on $[a, a + \delta]$. A similar argument proves that there exists $\delta > 0$ so that f is bounded on $[b - \delta, b]$

[9] This sentence has three terms we defined in this section. Take a moment to make sure you fully understand what it is saying.

Lemma 5.28. *Let $f : [a, b] \to \mathbb{R}$ so that f is continuous on $[a, b]$. There exists $\delta_a, \delta_b > 0$ so that f is bounded on the intervals $[a, a + \delta_a]$ and $[b - \delta_b, b]$*

Let us return now to our original goal of proving Theorem 5.26.

Consider the set B of elements $x \in [a, b]$ so that f is bounded on $[a, x]$. For example, by Lemma 5.28, there exists $\delta > 0$ so that $a + \delta \in B$. Our goal is to prove that b is an element of this set. From this it follows f is bounded on $[a, b]$.

The set B is bounded above by b; every element of B is an element of $[a, b]$. By the Completeness Axiom, the set B has a supremum, γ.

We will prove two things:

1. f is bounded on $[a, \gamma]$.
2. $\gamma = b$

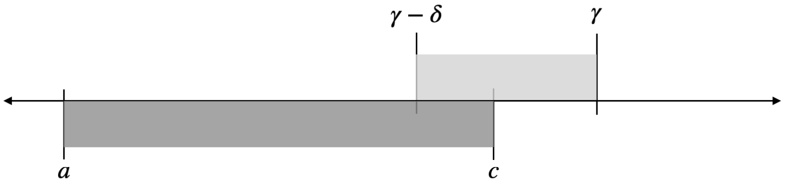
In our proof, we restrict our attention to proving that f is bounded above on the interval $[a, b]$. The proof that f is bounded below on this interval follows similarly.

To achieve both 1 and 2 we must recall the following result Section 2

Theorem (Theorem 2.20)**.** *Let $A \subseteq \mathbb{R}$ be bounded above and let $\gamma \in \mathbb{R}$ be an upper bound of A in $\mathbb{R}$. We have $\sup A = \gamma$ if and only if for every $\delta > 0$ there is an element of A greater than $\gamma - \delta$.*

We achieve 1 by first applying Lemma 5.27 to γ. Using the resulting δ we apply Theorem 4.23 to find $c \in B$ so that $c > \gamma - \delta$.

Since $c > \gamma - \delta$, the intervals $[a, c]$ and $[\gamma - \delta, \gamma]$ overlap. Since $c \in B$, f is bounded above on $[a, c]$. By Lemma 5.27, f is bounded above on $[\gamma - \delta, \gamma]$. Since f is bounded above on both intervals, then b is bounded above on the union of the intervals. In other words, f is bounded above on $[a, \gamma]$.



We achieve 2 by again applying Theorem 4.23. We proceed by contradiction (i.e., we assume $\gamma < b$) and show that f is bounded on the interval $[a, \gamma + \delta]$ for some $\delta > 0$. This contradicts that γ is the supremum of B.

With our strategy established, off we go!

Proof of Theorem 5.26. Let $f : [a, b] \to \mathbb{R}$ so that f is continuous on $[a, b]$. We prove f is bounded above on $[a, b]$. Let $B = \{x \in [a, b] \mid f \text{ is bounded above on } [a, x]\}$.

By Lemma 5.28, we have $a \in B$.

Since b is an upper bound of B in $\mathbb{R}$, the Completeness Axiom implies B has a supremum. Let $\sup B = \gamma$. Since b is an upper bound for $[a, b]$ it cannot be $\gamma > b$.

By Lemma 5.27, there exists δ so that $a + \delta \in B$. Therefore $\gamma > a$. And so we conclude $\gamma \in (a, b]$.

Claim 1: f is bounded above on $[a, \gamma]$.

Since $\gamma \in (a, b]$, then by Lemma 5.27 or Lemma 5.28, there exists $\delta > 0$ so that f is bounded above on $[\gamma - \delta, \gamma]$.

(It is certainly possible that $\gamma = b$ and so at this point we are not sure if we are applying Lemma 5.27 or Lemma 5.28 to conclude f is bounded above on $[\gamma - \delta, \gamma]$.)

By Theorem 4.23, there exists $c \in B$ so that $c > \gamma - \delta$. Since γ is an upper bound for B, and $c \in B$ we have $c \leq \gamma$. Therefore $c \in [\gamma - \delta, \gamma]$.

Since $c \in B$, f is bounded above on $[a, c]$. The intervals $[a, c]$ and $[\gamma - \delta, \gamma]$ overlap. Therefore f is bounded above on their union, Therefore f is bounded above on $[a, \gamma]$.

Claim 2: $\gamma = b$

Since $\gamma \in (a, b]$, to prove $\gamma = b$ it suffices to prove that the statement $\gamma < b$ is false. We proceed by contradiction. Assume $\gamma < b$ is true.

Since $\gamma \in (a, b)$, f is continuous at γ. By Lemma 5.27 there exists δ so that f is bounded above on $[\gamma - \delta, \gamma + \delta]$. However, we notice that the intervals $[a, \gamma]$ and $[\gamma - \delta, \gamma + \delta]$ overlap. Therefore f is bounded above on $[a, \delta + \gamma]$. Therefore $\delta + \gamma \in B$, which contradicts that γ is an upper bound of B. Therefore the statement $\gamma < b$ is false.

By Claim 1, f is bounded above on $[a, \gamma]$. By Claim 2, $\gamma = b$. Therefore f is bounded above on $[a, b]$. □

That our boundedness theorem is a statement about a closed interval matters. Consider $f(x) = \frac{1}{x}$ on the interval $(0, 1]$. By part *iv* of the Algebra of Continuous Functions Theorem, f is continuous on the interval $(0, 1]$. However, f is not bounded on this interval. Our proof above fails in our construction of the set B. Since f is not continuous at $x = 0$, it is not continuous on any interval $[a, x]$.

Test Your Understanding

1. Consider $f(x) = \frac{1}{|x|}$ on the interval $[-1, 0) \cup (0, 1]$. Define the set B as in the proof of Theorem 5.26. Determine the value of $\sup B$, if it exists.

2. At the start of Claim 1, the following sentence appears:

   > *Since $\gamma \in (a, b]$, then by Lemma 5.27 or Lemma 5.28, there exists $\delta > 0$ so that f is bounded above on $[\gamma - \delta, \delta]$.*

   If it possible that $\gamma \neq b$ and thus $\gamma \in (a, b)$, why do we conclude f is bounded above on $[\gamma - \delta, \gamma]$ instead of f is bounded above on $[\gamma - \delta, \delta + \gamma]$ (as stated in Lemma 5.27)?

3. A key component to the proof in this section is the following statement:

   > Let I and I' be intervals. If f is bounded above on I and f is bounded above on I', then f is bounded above on $I \cup I'$.

   Using the definition of bounded, prove this statement is true.

Test Your Understanding - Answers

1. $\sup B = 0$.

2. It is indeed true that f is bounded above on $[\gamma - \delta, \gamma + \delta]$ in $\gamma \in (a, b)$. If f is bounded above on $[\gamma - \delta, \delta + \gamma]$, then we can conclude f is bounded above on $[\gamma - \delta, \gamma]$. And so we conclude f is bounded above on $[\gamma - \delta, \gamma]$ no matter if $\gamma = b$ or $\gamma < b$.

3. Let I and I' be intervals so that f is bounded above on both. Therefore there exists b_I and $b_{I'}$ so that the following statements are true:

   - $f(x) < b_I$ for all $x \in I$
   - $f(x') < b_{I'}$ for all $x' \in I'$.

   Let $b = \max\{b_I, b_{I'}\}$. For all $z \in I \cup I'$ we have $f(z) \leq b$. Therefore f is bounded above by b on $I \cup I'$.

5.2.2 Extreme Value Theorem

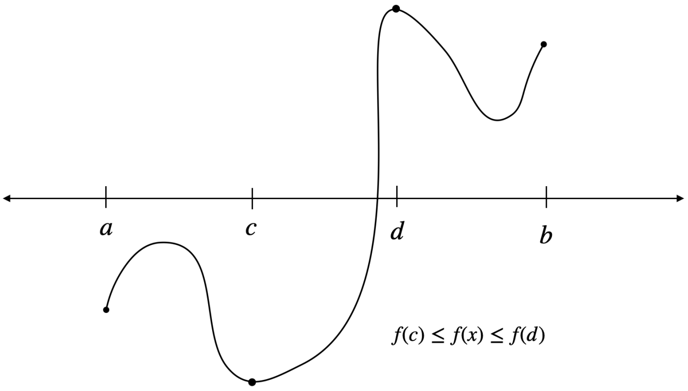
Our proof in the previous tells us that a continuous function on a closed interval is bounded on that interval. Here we prove we can find a maximum and minimum for f on a closed interval.

Notice how this statement is not true when we swap closed for open. Though function $f(x) = \frac{1}{x}$ is continuous on $(0, 1)$ it attains no maximum on this interval.

Theorem 5.29 (The Extreme Value Theorem)**.** *Let $f : [a, b] \to \mathbb{R}$. If f is continuous on $[a, b]$, then there exists $c, d \in [a, b]$ so that*

$$f(c) \leq f(x) \leq f(d)$$

for all $x \in [a, b]$



Let us take some time to strategise before we dive in to our proof. As in the previous section, we prove only the existence of the upper bound, $f(d)$.

Let us consider the range of f on the domain $[a, b]$

$$A_f = \{f(x) \mid x \in [a, b]\}$$

By Theorem 5.26, this set is bounded above in $\mathbb{R}$. And so, by the Completeness Axiom, it has a supremum in $\mathbb{R}$. Our goal is to prove there exists $d \in [a, b]$ so that $f(d) = \sup A_f$. The result then follows from the definition of supremum.

For $x \in [a, b]$ let us consider the value

$$\sup A_f - f(x)$$

By definition of supremum, we have $\sup A_f - f(x) \geq 0$ for all $x \in [a, b]$.

We will proceed by contradiction. If there is no $d \in [a, b]$ so that $f(d) = \sup A_f$, then $\sup A_f - f(x) > 0$ for all $x \in [a, b]$.

Let

$$g(x) = \frac{1}{\sup A_f - f(x)}$$

Let us take a moment to think about the range of g. When the difference between $\sup A_f$ and $f(x)$ is very small, then $g(x)$ will be very large. In fact, as the difference between $\sup A_f$ and $f(x)$ goes towards 0, the function $g(x)$ explodes towards infinity. So, then, perhaps we conclude that $g(x)$ is not bounded above on $[a, b]$.

On the other hand part *iv* of Algebra of Continuous Functions Theorem tells us $g(x)$ is continuous on $[a, b]$. And then the Extreme Value Theorem tells us $g(x)$ is bounded above on $[a, b]$.

It cannot be that $g(x)$ is bounded above on $[a, b]$ and not bounded above on $[a, b]$. It seems we have arrived at a contradiction. With our sights set on this contradiction ($g(x)$ is both bounded and unbounded on $[a, b]$) we proceed with our proof.

Proof of Extreme Value Theorem. Let $f : [a, b] \to \mathbb{R}$ so that f is continuous on $[a, b]$. Let $A_f = \{f(x) \mid x \in [a, b]\}$. By Theorem 5.26, the set A_f is bounded above. By the Completeness Axiom, A_f has a supremum. Let $\sup A_f = \gamma$.

We claim there exists $d \in [a, b]$ such that $f(d) = \gamma$. We proceed by contraction.

Assume that for every $x \in [a, b]$ we have $f(x) < \gamma$. Let $g(x) = \frac{1}{\gamma - f(x)}$.

Claim 1: The function $g(x)$ is bounded above on $[a, b]$.

Since $\gamma - f(x) \neq 0$ for all $x \in [a, b]$, it follows from part *iv* of the Algebra of Continuous Functions Theorem that $g(x)$ is continuous on $[a, b]$. By Theorem 5.26, $g(x)$ is bounded above on $[a, b]$.

Claim 2: The function $g(x)$ is not bounded above on $[a, b]$.

Let $r \in \mathbb{R}$. To show $g(x)$ is not bounded above on $[a, b]$, it suffices to find $x \in [a, b]$ so that $g(x) > r$.

Let $\epsilon = \frac{1}{r}$. By Theorem 4.23, there exists $a \in A_f$ so that $a > \gamma - \epsilon$. By construction, there exists $x' \in [a, b]$ so that $f(x') = a$. Therefore there exists $x' \in [a, b]$ so that $f(x') > \gamma - \epsilon$. Rearranging, we have $\gamma - f(x') < \epsilon$. Therefore

$$\frac{1}{\gamma - f(x')} > \frac{1}{\epsilon} = r$$

Therefore $g(x)$ is not bounded above on $[a, b]$.

This is a contradiction, as it cannot be that $g(x)$ is bounded above on $[a, b]$ and $g(x)$ is not bounded above on $[a, b]$. Therefore there exists $d \in [a, b]$ such that $f(d) = \gamma$. And so, there exists $d \in [a, b]$ such that $f(x) \leq f(d)$ for all $x \in [a, b]$. □

Test Your Understanding

1. Consider $f(x) = \frac{1}{|x|}$ on the interval $[-1, 0) \cup (0, 1]$. Let $A_f = \{f(x) | |x \in [-1, 0) \cup (0, 1]\}$. Determine the value of $\sup A_f$, if it exists.

2. At the start of Claim 2, the following sentence appears:

> *To show $g(x)$ is not bounded above on $[a, b]$, it suffices to find $x \in [a, b]$ so that $g(x) > r$.*

Why is this true? ______________________________

Test Your Understanding - Answers

1. $\sup A_f$ does not exist because f is not bounded on the interval $[-1, 0) \cup (0, 1]$.
2. Recall the definition of bounded above on I. If g is not bounded above on $[a, b]$ then for every $r \in \mathbb{R}$, the statement

   r is an upper bound for g on $[a, b]$ is false.

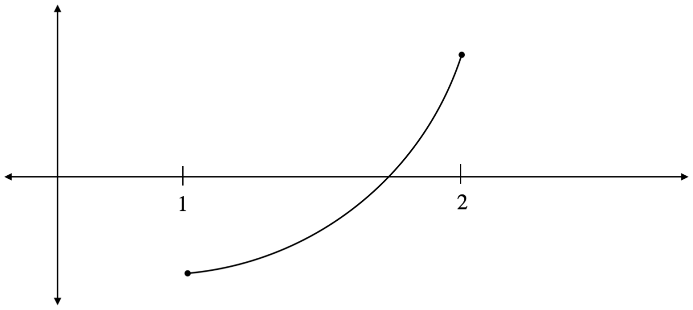
   This statement is true if and only if there exists $x \in [a, b]$ so that $g(x) > r$.

5.2.3 Intermediate Value Theorem

On to our third *obvious* fact:

If a continuous function is below the x-axis at one point, and above the x axis at another point, then it must cross the x-axis at some point.

Rather than prove this statement for an arbitrary continuous function, let us consider a function that satisfies the hypothesis of the statement above: $g(x) = x^2 - 2$.



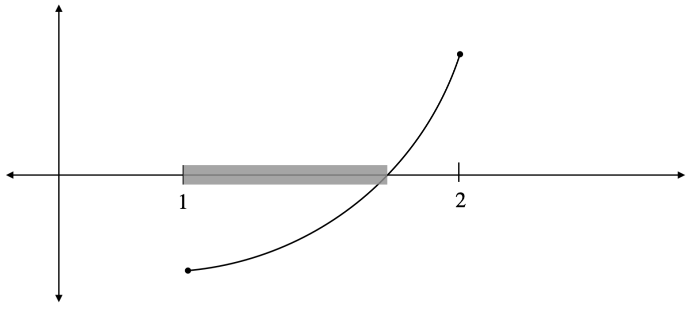
Theorem 5.30. *There exists $x \in \mathbb{R}$ so that $x^2 - 2 = 0$.*

Aside. *Before we even set out, we know this theorem is true as $\sqrt{2} \in \mathbb{R}$! However, let us humour ourselves for the moment.*

As we have done in the previous two proofs, we will cleverly construct a set and then prove that the supremum of that set has the property we seek. Consider the set X of points in $[1, 2]$ for which g is bounded above by 0 on the interval $[1, x]$

For example, $g(x) \leq 0$ for all $x \in [1, 1.2]$, therefore $1.2 \in X$. On the otherhand, there exists x in the interval $[1, 1.5]$ so that $g(x) > 0$. Therefore $1.5 \notin X$.

The set X is those values in $[1, 2]$ that occur between 1 and the first point that $g(x)$ crosses the x-axis:

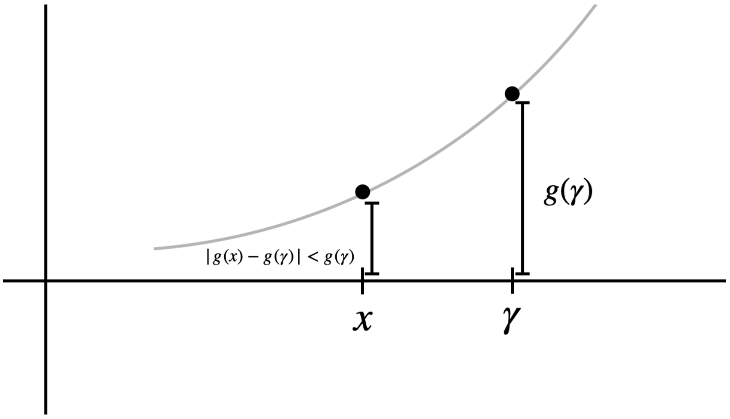


As X is a subset of $[1, 2]$, 2 is an upper bound for X in $\mathbb{R}$. And so, by the Completeness Axiom, this set has an supremum in $\mathbb{R}$. Let $\sup X = \gamma$.

We want to show $g(\gamma) = 0$.

If $g(\gamma) > 0$, then perhaps we can find $x \in [1, 2]$ so that $x < \gamma$ but $g(x) > 0$. The existence of x would contradict that γ is the supremum of X, as x would be a smaller upper bound for X.

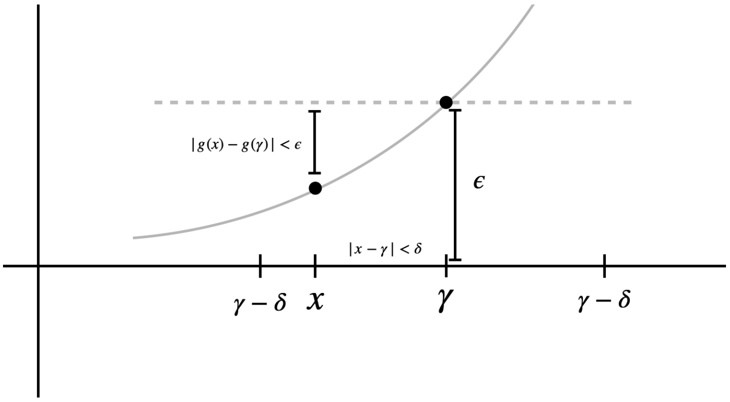
As the subject of 5.2 is continuity, it seems as if we should need to exploit the fact that f is continuous on $[1, 2]$. As $\gamma \in (1, 2)$, then f is continuous at γ. Therefore for every $\epsilon > 0$ there exists $\delta > 0$ so that $|g(x) - g(\gamma)| < \epsilon$ whenever $|x - \gamma| < \delta$. The key here is making a choice of ϵ that implies the existence of $x \in [a, b]$ so that $x < \gamma$ and $g(x) > 0$.



From our picture, it seems as if choosing $\epsilon = g(\gamma)$ should do the trick. Since f is continuous at γ there exists $\delta > 0$ so that $|g(x) - g(\gamma)| < g(\gamma)$ whenever $|x - \gamma| < \delta$. Unpacking this first inequality, we have

$$-g(\gamma) < g(x) - g(\gamma) < g(\gamma)$$
$$0 < g(x) < 2g(\gamma)$$

for all $x \in (\gamma - \delta, \gamma + \delta)$. In particular, there exists $x \in (\gamma - \delta, \gamma)$ so that $g(x) > 0$.



However if $x \in (\gamma - \delta, \gamma)$, then $x < \gamma$. Further, as $g(x) > 0$, x is an upper bound for X. And so γ is not a supremum of X. Therefore the statement $g(\gamma) > 0$ is false.

Using a similar approach, we show that the statement $g(\gamma) < 0$ is false.

With these ideas in mind, we proceed on to our proof of Theorem 5.30

Proof of Theorem 5.30. Let $g(x) = x^2 - 2$. By Theorem 5.22, $g(x)$ is continuous on $\mathbb{R}$. Therefore $g(x)$ is continuous on $[1, 2]$. Notice $g(1) < 0$ and $g(2) > 0$.

Let $X = \{x \in [1, 2] \mid g(c) \leq 0 \text{ for all } c \in [1, x]\}$. Since $g(2) > 0$ we have that $2 \notin X$. Therefore X is bounded above in $\mathbb{R}$. Let $\sup X = \gamma$. We claim $g(\gamma) = 0$.

Since $1 \in X$ and γ is the supremum of X, we have $\gamma \geq 1$. Since 2 is an upper bound of X in $\mathbb{R}$ and γ is the supremum X, we have $\gamma \leq 2$. Therefore $\gamma \in [1, 2]$. Therefore $g(x)$ is continuous at γ.

We proceed by contradiction to rule out the possibility $g(\gamma) > 0$ and $g(\gamma) < 0$. From this it will follow that $g(\gamma) = 0$

Assume $g(\gamma) > 0$. Let $\epsilon = g(\gamma)$. Since g is continuous at γ, there exists $\delta > 0$ so that $|g(x) - g(\gamma)| < \epsilon$ whenever $|x - \gamma| < \delta$. Since $g(\gamma) = \epsilon$ for all $x \in (\gamma - \delta, \gamma + \delta)$ we have

$$\begin{aligned}
&|g(x) - g(\gamma)| < \epsilon \\
&-\epsilon < g(x) - \epsilon < \epsilon \\
&0 < g(x) < 2\epsilon
\end{aligned}$$

Therefore $g(x) > 0$ for all $x \in (\gamma - \delta, \gamma + \delta)$. In particular, $g(\gamma - \delta/2) > 0$. Therefore $\gamma - \delta/2 \notin X$. However, this then implies that $\gamma - \delta/2$ is an upper bound for X. This contradicts that γ is the least upper bound of X. Therefore $g(\gamma) > 0$ is false.

Assume $g(\gamma) < 0$. Let $\epsilon = -g(\gamma)$. Since g is continuous at γ, there exists $\delta > 0$ so that $|g(x) - g(\gamma)| < \epsilon$ whenever $|x - \gamma| < \delta$. Since $g(\gamma) = \epsilon$, for all $x \in (\gamma - \delta, \gamma + \delta)$ we have

$$\begin{aligned}
&|g(x) - g(\gamma)| < \epsilon \\
&-\epsilon < g(x) + \epsilon < \epsilon \\
&2\epsilon < g(x) < 0
\end{aligned}$$

Therefore $g(x) < 0$ for all $x \in (\gamma - \delta, \gamma + \delta)$. And so, $g(x) < 0$ for all $x \in [1, \gamma + \delta)$. Therefore for all $[1, \gamma + \delta)$ we have $x \in X$.

Since $\gamma + \delta/2 \in [1, \gamma + \delta)$, we have $\gamma + \delta/2 \in X$. Therefore γ is not an upper bound of X, a contradiction. Therefore $g(\gamma) < 0$ is false.

Since $g(\gamma) < 0$ is false and $g(\gamma) > 0$ is false, it must be that $g(\gamma) = 0$.

Therefore there exists $x \in [1, 2]$ so that $g(x) = 0$. □

Looking back at our argument, what is special about $g(x) = x^2 - 2$? For our proof to work we needed g to be continuous on $[a, b]$ and for there to be $a, b \in \mathbb{R}$ so that $f(a) < 0$ and $f(b) > 0$. These are exactly the conditions we saw in the statement of our obvious fact above.

Looking at our proof above, consider making the following changes:

- replace $g(x) = x^2 - 2$ with any continuous function;
- replace 1 with a value $a \in \mathbb{R}$ such that $f(a) < 0$; and

- replace 2 with a value $b \in \mathbb{R}$ such that $f(b) > 0$.

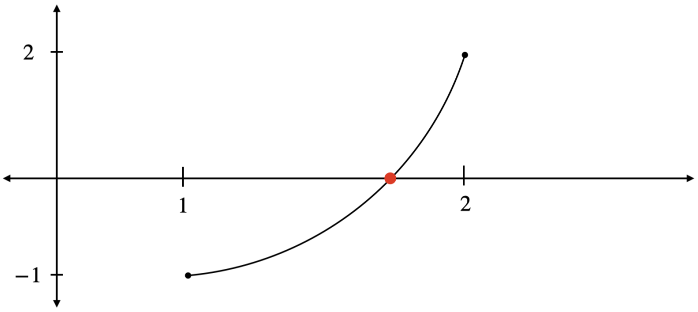
It turns out that by doing so, nothing in our argument needs to change! And so the argument above gives a proof of the following result:

Lemma 5.31. *Let f be continuous on $[a, b]$. If $f(a) < 0$ and $f(b) > 0$, then there exists $c \in [a, b]$ so that $f(c) = 0$.*

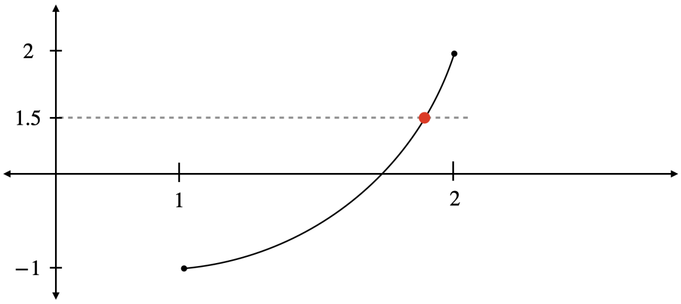
But wait, why is this stated as a lemma and why is there still another page of reading to go? We have achieved our goal of proving our *obvious* fact.

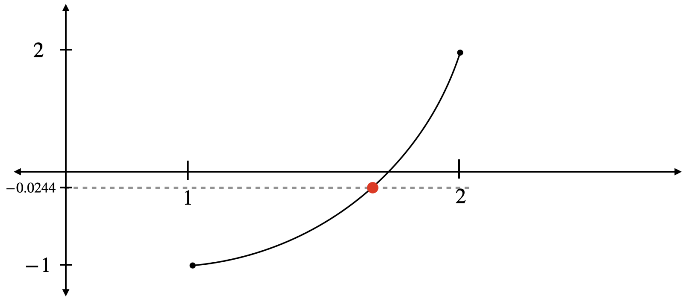
Let us reach slightly higher. Returning back to our argument in Theorem 5.30 we proved the following fact:

$$g(1) = -1,\ g(2) = 2 \text{ and there exists } c \in [1, 2] \text{ so that } g(c) = 0.$$



Looking at our picture, changing 0 to another value in the interval $[-1, 2]$ should yield the same result. For example, there should be $c \in [1, 2]$ so that, say, $g(x) = 1.5$. Similarly, there should be $c \in [1, 2]$ so that, say, $g(x) = -.0244$.





And so we expect the following to be true:

$$\text{For all } y \in [-1,2] \text{ there exists } c \in [1,2] \text{ so that } g(c) = y.$$

And again, there is nothing particularly special about $g(x) = x^2 - 2$ other than its continuity on the interval $[1, 2]$. And so we have the following result:

Theorem 5.32 (The Intermediate Value Theorem)**.** *Let f be continuous on $[a, b]$ so that $f(a) < f(b)$ and let $y \in \mathbb{R}$. If $f(a) < y < f(b)$, then there exists $c \in [a, b]$ so that $f(c) = y$.*

We worked awfully hard to prove Lemma 5.31. It seems a shame for all of that hard work to go to waste. And so let us exploit the statement of Lemma 5.31 to prove the Intermediate Value Theorem.

Consider a continuous function f on the interval $[a, b]$ so that $f(a) < f(b)$. Consider $y \in (f(a), f(b))$. Notice that translating f by $-y$ results in $f(a) < 0$ and $f(b) > 0$.

Let $h(x) = f(x) - y$. From our observation in the previous sentence, we have $h(a) < 0$ and $h(b) > 0$.

By Lemma 5.31 there exists $c \in [a, b]$ so that $h(c) = 0$. And so

$$\begin{aligned} h(c) &= f(c) - y \\ 0 &= f(c) - y \\ h(c) &= y \end{aligned}$$

Therefore there exists $c \in [a, b]$ so that $h(c) = y$.

Proof of the Intermediate Value Theorem. Let $f : [a, b] \to \mathbb{R}$ be a continuous function so that $f(a) < f(b)$ and let $y \in \mathbb{R}$. Assume $f(a) < y < f(b)$. Let $h(x) = f(x) - y$. Since $f(a) < y < f(b)$ we have $h(a) < 0$ and $h(b) > 0$. By Lemma 5.31, there exists $c \in [a, b]$ so that $h(c) = 0$. Notice

$$\begin{aligned} h(c) &= f(c) - y \\ 0 &= f(c) - y \\ f(c) &= y \end{aligned}$$

Therefore there exists $c \in [a, b]$ so that $f(c) = y$. □

Test Your Understanding

1. Both Lemma 5.31 and the Intermediate Value Theorem are stated with the hypothesis $f(a) < f(b)$. State the versions of these results for the case that $f(a) > f(b)$.

2. Let $f(x) = x^3 - 5x^2 + 7x - 9$. Prove that there is a real number c such that $f(c) = 100$.

3. Let $f(x) = x^5 - 3x^4 - 2x^3 - x + 1$. Proof that f has at least one root between 0 and 1.

Test Your Understanding - Answers

1.

   Lemma. *Let f be continuous on $\mathbb{R}$. If $f(a) > 0$ and $f(b) < 0$, then there exists $c \in [a, b]$ so that $f(c) = 0$.*

   Theorem (The Intermediate Value Theorem)**.** *Let f be continuous on $[a, b]$ so that $f(a) > f(b)$ and let $y \in \mathbb{R}$. If $f(a) > y > f(b)$, then there exists $c \in [a, b]$ so that $f(c) = y$.*

2. Since f is a polynomial, f is continuous at a for all $a \in \mathbb{R}$.

   Notice $f(0) = -9 < 100$ and $f(10) = 561 > 100$.

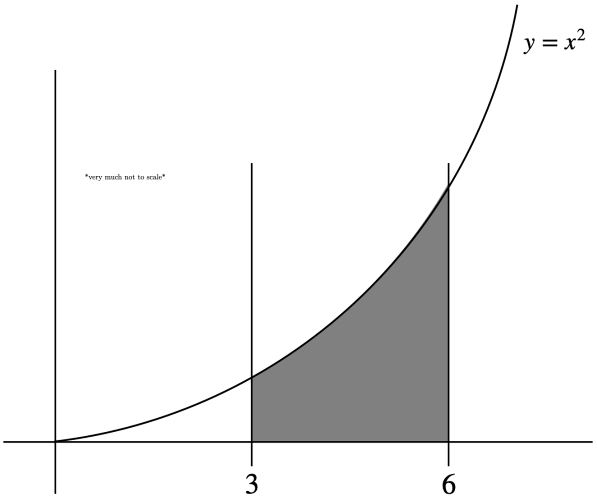
   Therefore, by the Intermediate Value Theorem, there exists $c \in (0, 10)$ such that $f(c) = 100$.

3. Find $f(0)$ and $f(1)$. Notice they have opposite signs.

6 Differentiation and Integration

Consider the following problem and its solution:

Find the area of the region bounded by the x-axis, $x = 3$, $x = 6$ and the curve $f(x) = x^2$



Solution We compute

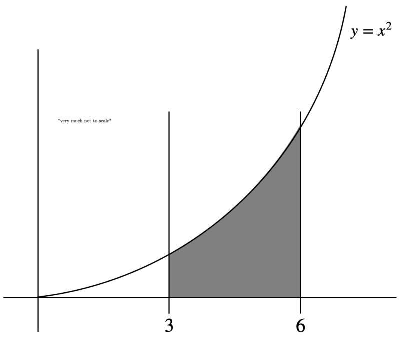
$$\int_3^6 x^2\, dx = F(6) - F(3)$$

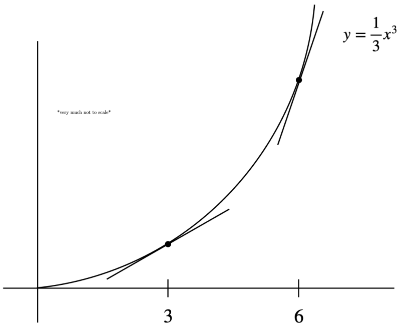
where $F'(x) = x^2$. Since $\frac{d}{dx}\left(\frac{1}{3}x^3\right) = x^2$, we have $F(x) = \frac{1}{3}x^3$. Therefore

$$\int_3^6 x^2\, dx = F(6) - F(3) = \frac{1}{3}6^3 - \frac{1}{3}3^3 = 63.$$

Therefore the area of the region is 63 units squared.

This is work that perhaps we have been able to do for years. Except, wait... what? None of this makes any sense at all! The notation $\int_3^6 x^2\, dx$ refers to an area and the notation $F'(x) = x^2$ refers to the slope of a line tangent to a curve at a particular point. How could these two concepts possibly be related? There is no reason at all to suspect that the study of differentiation should be any help at all in the study of integration.





What's more, how do we know that the slope of a line tangent to a point on the curve $F(x) = \frac{1}{3}x^3$ is given by the formula $f(x) = x^2$? The answer of *well... umm... that's what the power rule for derivatives tells us* is deeply unsatisfying. If the power rule for derivatives, i.e., $\frac{d}{dx}\left(x^k\right) = kx^{k-1}$, is true[1], then there should be some theorem that states this mathematical fact.

Theorem. *Let $k \in \mathbb{N}^+$. If $f(x) = x^k$, then $f'(x) = kx^{k-1}$.*

Even the mere act of writing down the statement of the theorem forces us to reckon with considering its proof. But without a careful treatment on the precise mathematical meaning of differentiation, we are at a loss for how to proceed.

The same goes for the relationship between integrals and derivatives (i.e., between areas and slopes). That these two concepts should be related at all comes to us via the Fundamental Theorem of Calculus.

Theorem (The Fundamental Theorem of Calculus Part I)**.** *If $f : [a, b] \to \mathbb{R}$ is integrable and $F : [a, b] \to \mathbb{R}$ satisfies $F'(x) = f(x)$ for all $x \in [a, b]$, then*

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

Theorem (The Fundamental Theorem of Calculus Part II)**.** *Let $g : [a, b] \to \mathbb{R}$ be integrable and let $G : [a, b] \to \mathbb{R}$ so that*

$$G(x) = \int_a^x g(u)\,du$$

The following statements are true:

i The function G is continuous on $[a, b]$.

ii If g is continuous at $c \in [a, b]$, then G is differentiable at c and $G'(c) = g(c)$

[1] Take a moment to bask in how astounding this statement is. Other than the fact that *someone told us it is true*, there is no reason at all to suspect that the formula slope of a tangent line of a point on a polynomial curve should be given by another polynomial. This is amazing!

If we are comfortable with the notation $F'(x) = f(x)$ and $\int_a^b f(x)\ dx$, then Part I of the Fundamental Theorem of Calculus is fairly straightforward: one can compute a definite integral of a function by finding an anti-derivative of the function.

However, simplicity of Part I of the Fundamental Theorem of Calculus belies a much more interesting conceptual result. The notation $\int_a^b f(x)\ dx$ refers to an area. And, in the hypothesis of the Part I of the Fundamental Theorem of Calculus, the notation $F(b) - F(a)$ is a difference in the value of a function that is related to $f(x)$ by way of differentiation. Based on our geometric intuition for tangent lines and areas, there is no reason at all to suspect that slopes and areas should be related in any way. That they are is astounding.

As compared the elegance of Part I of the Fundamental Theorem of Calculus, Part II of the Fundamental Theorem of Calculus is an inscrutable mess. In Part II of the Fundamental Theorem of Calculus we are defining a function whose value depends on the upper bound of integration of some other function. Part II of the Fundamental Theorem of Calculus is telling us that knowing something about the function we are integrating to find $G(x)$ tells us something about $G(x)$. Why this is interesting or useful to us, who knows?!

Our broad goal in Section 6 of the course is to understand the intuition behind the Fundamental Theorem of Calculus, and, using that intuition, give a proof for the theorem. As we have seen many times so far in this course, mathematical proofs consist of proving that some mathematical object satisfies some particular criteria. And so before we can even begin our discussion on the Fundamental Theorem of Calculus, we must first take the time to precisely define all of the terms and notation that appear as part of their statements.

The study of differentiation and integration has been a central component of our mathematical upbringing. Prior to this course, our work in this area was focussed on answering the questions whose interrogative was *how.* As it has been throughout the course, our focus in Section 6 of the course is one that seeks to answer *what* and *why* questions about differential and integral calculus.

If all one ever wanted to do was apply calculus methods to problems in industry and scientific research, then the applied methods with which you are already intimately familiar would likely suffice. However, danger lurks when we apply tools and techniques without understanding, at the minimum, why they *ought* to be true. For example, the global financial crisis in 2007 can be partially attributed to a stock traders applying a formula in conditions for which the hypotheses of the formula did not hold[2].

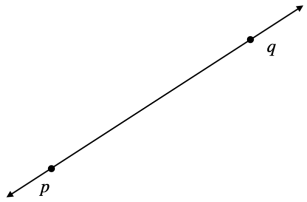
Looking back at the statement of the Fundamental Theorem of Calculus, we have a ways to go in defining terms and notation before we can fully come to grips with that this theorem is telling us. We begin by revisiting a topic from our first introduction to calculus: differentiation.

[2] see `https://www.theguardian.com/science/2012/feb/12/black-scholes-equation-credit-crunch`

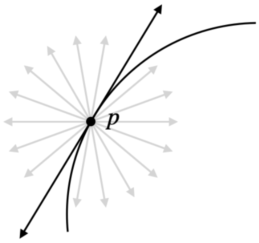
6.1 Differentiation

Unsurprisingly, our discussion about differentiation begins with lines and slopes. Recall the following geometric fact

For every pair of distinct points p and q in the plane, there is exactly one line in the plane that passes through p and q.



A consequence of this fact is that for a point p, there are infinitely many lines in the plane that pass through p. When p lies on a curve, the tangent line through point p is exactly one of these lines.



For lots of applied reasons that you have learned about in other courses, a principle interest is the slope of this tangent line.

Recall that for a line ℓ, one can compute the slope of ℓ using a pair of distinct points on ℓ. If $p = (p_1, p_2)$ and $q = (q_1, q_2)$ are both points on ℓ, then the slope of ℓ is given by

$$m_\ell = \frac{q_2 - p_2}{q_1 - p_1}$$

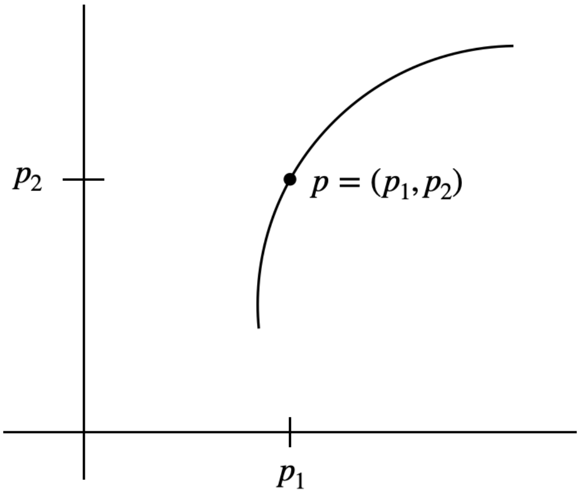
Using these ideas, we start our journey in finding the slope of a line tangent to a curve.

Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous[3] function and consider a point $p = (p_1, p_2)$ so that p satisfies f. In other words, $f(p_1) = p_2$

As mentioned above, there are infinitely many lines that pass through p; we are interested

[3] Continuity isn't strictly necessary here, but continuous functions make for nicer intuitive pictures

in exactly one of them: the one that is tangent[4] to the curve f at p.



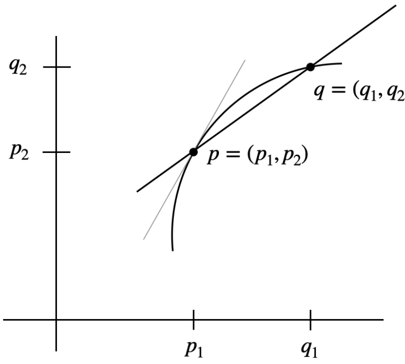
Rather than try to find the slope of this tangent line directly, let us instead consider a method to estimate this value. Consider a point $q = (q_1, q_2)$ so that q_1 is *close* to p_1. Provided we choose q_1 to be close enough to p_2, the slope of the line through p and q approximates the slope of the tangent line at p. This approximate slope is given by

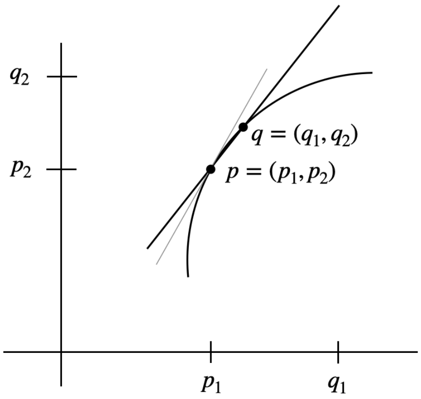
$$m_{approx} = \frac{q_2 - p_2}{q_1 - p_1}$$

As p and q satisfy f, we can re-write q_2 as $f(q_1)$ and p_2 as $f(p_1)$.

$$m_{approx} = \frac{f(q_1) - f(p_1)}{q_1 - p_1}$$

Our choice of q impacts to quality of our approximation. Intuitively, as we choose q *closer and closer* to p, this approximation improves.





[4] If we were so inclined we could precisely define tangent. But this would require more time and energy than is worth spending on the exercise.

The language of *closer and closer* points towards the idea of a limit.

Consider the following limit:

$$\lim_{x\to p_1} \frac{f(x)-f(p_1)}{x-p_1} = m$$

We claim that limit gives the slope of the tangent line at $x = p_1$. We won't prove this fact, as it would require a careful definition of the word *tangent.* However, we'll take a moment to consider a concrete example to ensure that this definition agrees with our computational experiences.

Consider the curve $f(x) = x^2$ and $p_1 = 3$. From our work in previous studies, we expect to compute that the slope of the line tangent to f through $(3, 9)$ is $m = 6$. And so we want to verify:

$$\lim_{x\to 3} \frac{f(x)-f(3)}{x-3} = 6$$

Notice

$$\frac{f(x)-f(3)}{x-3} = \frac{x^2-9}{x-3} = x+3$$

and

$$\lim_{x\to 3}(x+3) = 6$$

We could verify this last statement by taking $\delta = \epsilon$ in the definition of the notation $\lim\limits_{x\to 3} x+3 = 6$.

From our work in past courses, we know that not every continuous function is differentiable at every point. For example $g(x) = |x|$ has no derivative at $p_1 = 0$. Consequently, the function $\frac{g(x)-g(0)}{x-0}$ diverges as x approaches 0. One can prove this fact by examining the left and right side limits:

$$\lim_{x\to 0^-} \frac{g(x)-g(0)}{x-0} = -1 \text{ and } \lim_{x\to 0^+} \frac{g(x)-g(0)}{x-0} = 1$$

With these thoughts in mind, we define the term f is differentiable at c.

A note about this definition, and others to follow: In our study of differentiation and integration will we consider domains of the form: $(a, b), [a, b], (a, b], [a, b), (a, \infty), [a, \infty), (\infty, b), (-\infty, b]$ or $\mathbb{R}$. So as to keep our definitions from being needlessly cumbersome, we use the notation I to refer to any one of these domains.

Definition 6.1. *Let $f : I \to \mathbb{R}$ and let $c \in I$. We say f is differentiable at c when there exists $L \in \mathbb{R}$ so that*

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

We call L the derivative of f at c. When f differentiable at c for all $c \in I$ we say f is differentiable on I.

Aside. *One can alternatively consider the limit* $\lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = L$*. The two limits have the same behaviour. This latter expression is referred to as the* difference quotient.

So far our discussion in differentiation has focussed solely on the slope of a tangent line at a single point. However, our experience with derivatives is in computing derivatives of a function, as opposed to the slope of a tangent line at a single point.

Let f be differentiable on I. Therefore, for every $c \in I$ there exist $L_c \in \mathbb{R}$ so that

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

By definition, the value L_c is the derivative of f at c. Consider now the following function

$$f' = \{(c, L_c) \mid c \in I\}$$

The ordered pairs of f' are of the form (c, L_c). In other words, the image of c is L_c. Therefore $f'(c) = L_c$.

Aside. *If this is confusing, go back and review Definition 4.2.*

So as to agree with our use from previous Calculus courses, we refer to $f'(x)$ as the derivative of $f(x)$. When convenient, we use $\frac{d}{dx}$ as we expect.

Aside. *This previous sentence hurts to write without some further discussion about differentials and* Leibniz's notation. *But, time is tight and sacrifices must be made. We'll come back to this thought when we define the notation* $\int f(x)\, dx$*.*

For example, let us return to $f(x) = x^2$. Above we computed:

$$\lim_{x\to 3} \frac{f(x)-f(3)}{x-3} = 6$$

Therefore $f'(3) = 6$. The function $f' : \mathbb{R} \to \mathbb{R}$ maps $c \in \mathbb{R}$ to the slope of the tangent line through c, L_c. We know from our previous work in calculus that there is an explicit formula we can use to compute $f'(c)$ for all $c \in \mathbb{R}$

$$f'(c) = 2c$$

If this course had more time and more space, we would take the next few pages to prove that all of our computational techniques for working with derivatives work as we expect. For example, using the Definition 6.1, and the Algebra of Limits Theorem for functions, one can prove the *product rule*

$$(f \cdot g)'(x) = f'(x)g(x) + f(x)g'(x)$$

Translating this statement into the language of limits gives:

For all $c \in I$ *we have*

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

Using the *product rule* and induction, we can then prove

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

Proceeding by induction, we can write $h(x) = x^{k+1}$ as $h(x) = f(x)g(x)$, where $f(x) = x$ and $g(x) = x^k$. Applying our *product rule* above gives

$$\begin{aligned} h'(x) &= f'(x)g(x) + f(x)g'(x) \\ &= (1)(x^k) + (x)(kx^{k-1}) \\ &= (k+1)x^k \end{aligned}$$

Verifying all of the computational things we *know* about derivatives is no small feat! There is little to be gained for our conceptual understanding of these topics in taking the time to painstakingly prove that our computational techniques are valid. There is however one result that will come in handy as we work through the material in Section 6. Consider the following computation

$$\begin{aligned} \frac{d}{dx}(3x^2 + 6x) &= \frac{d}{dx}(3x^2) + \frac{d}{dx}(6x) \\ &= 3\frac{d}{dx}(x^2) + 6\frac{d}{dx}(x) \\ &= 3(2x) + 6(1) \\ &= 6x + 6 \end{aligned}$$

Looking at our definition of differentiation above, there is nothing that justifies the first equality. Nothing in our definition of differentiable tells us

the derivative of the sum is equal to the sum of the derivatives

Nor is there anything in definition of the derivative that tells us we can *factor out* constants when we compute derivatives, as we do on the second line. Both of these facts arise due to the linearity of the derivative.

Theorem 6.2 (Linearity of the Derivative)**.** *Let $f, g : I \to \mathbb{R}$ so that f and g are differentiable on I and let $\lambda \in \mathbb{R}$. The following statements are true:*

i $(f + g)'(x) = f'(x) + g'(x)$

ii $\lambda(f'(x)) = (\lambda f)'(x)$.

Since differentiation is a statement about limits, the two parts of the this theorem are statements about limits is disguise. And thus their proofs follow from the Algebra of Limits Theorem for functions.

Aside. *Linearity is a property that comes up in many places in mathematics. We will see it again when we discuss integration. If you have taken any studies in probability theory, you are probably familiar with the phrase* linearity of expectation*.*

We examined linearity in great detail in our past study of vector spaces. Usually, when linearity is lurking so are vector spaces.

Let $\mathcal{P}_3$ be the set of polynomials of degree at most 3. This set forms a vector space with respect to polynomial addition and scalar multiplication. One basis of this vector space is $B = \{1, x, x^2, x^3\}$. This vector space has dimension 4.

Let $D : \mathcal{P}_3 \to \mathcal{P}_3$ so that $D(f(x)) = f'(x)$.

By Theorem 6.2, D is linear. The image of this linear function is $\mathcal{P}_2$. Our basis $\{1, x, x^2, x^3\}$ generates the basis $\{1, 2x, 3x^2\}$. And so $\mathcal{P}_2$ has dimension 3.

By the rank-nullity Theorem, the null space of this linear transformation should have dimension 1. The null space of D is the set of polynomials in $f \in \mathcal{P}_3$ so that $D(f) = 0$. In otherwords, the null space of D is the set of polynomials whose derivative is equal to 0. Therefore null D is the set of constant polynomials: $f(x) = c$, which is spanned by $\{1\}$.

Since this linear transformation is between finite dimensional vector spaces, it can be represented as a matrix. With respect to the basis B, the matrix is

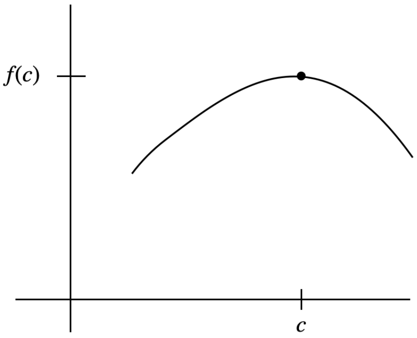
$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

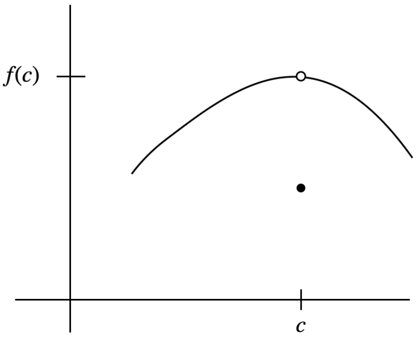
Notably this matrix is not invertible, which makes sense as differentiation is not bijective. The linear transformation D is not invertible: there are many elements of $\mathcal{P}_3$ that have the same image under D. (i.e., they have the same derivative). For example $D(2x+1) = D(2x)$.

Coming back to some familiar footing, a common use for the derivative is to locate maximums (and minimums) of functions. This application follows from the following observation:

If a function is increasing on the left of $x = c$ and decreasing on the right of $x = c$, then f attains a local maximum at $x = c$

Of course, such an observation only holds when f is continuous *near* c.





Looking back at Definition 6.1, any mention of continuity is conspicuously absent. In fact, the following is true:

Theorem 6.3. *Let $f : I \to \mathbb{R}$ and let $c \in I$. If f is differentiable at c, then f is continuous at c.*

Before we dive in to our proof, let us take a moment to strategise. We want to prove f is continuous at c. Since $c \in I$ and I is one of the following intervals:

$$(a, b), [a, b], (a, b], [a, b), (a, \infty), [a, \infty), (\infty, b), (-\infty, b] \text{ or } \mathbb{R}$$

then c is a limit point of I. And so, rather than futz around with ϵ's and δ's to prove continuity at c, we can apply Theorem 5.18.

We want to prove that if f is differentiable at c, then

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

By adding $-f(c)$ to both sides, we can write this as

$$\left(\lim_{x \to c} f(x)\right) - f(c) = 0$$

Since $\lim\limits_{x \to c} f(c) = f(c)$ we can use part II of the Algebra of Limits theorem for functions to express our equality as:

$$\lim_{x \to c} (f(x) - f(c)) = 0$$

And so, to prove $\lim\limits_{x \to c} f(x) = f(c)$ it suffices to prove $\lim\limits_{x \to c} (f(x) - f(c)) = 0$

We proceed using the definition of differentiable at c. By definition, since f is differentiable at c, there exists $L \in \mathbb{R}$ so that

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

Since $\lim\limits_{x\to c}(x-c) = 0$, we have

$$\left(\lim_{x\to c} \frac{f(x)-f(c)}{x-c}\right) \cdot \left(\lim_{x\to c} x-c\right) = L \cdot 0 = 0$$

Simplifying on the left using part II of the Algebra of Limits Theorem for functions gives

$$\begin{aligned}\left(\lim_{x\to c} \frac{f(x)-f(c)}{x-c}\right) \cdot \left(\lim_{x\to c} x-c\right) &= \lim_{x\to c}\left(\frac{f(x)-f(c)}{x-c} \cdot (x-c)\right)\\ &= \lim_{x\to c}(f(x)-f(c))\end{aligned}$$

Proof. Let $c \in I$ and let $f : I \to \mathbb{R}$ so that f is differentiable at c. By definition, there exists $L \in \mathbb{R}$ so that

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

We want to show f is continuous at c. Since $c \in I$, c is a limit point of I. And so by Theorem 5.18, it suffices to prove

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

By the Algebra of Limits Theorem for functions we have:

$$\begin{aligned}0 &= L \cdot 0\\ &= \left(\lim_{x\to c} \frac{f(x)-f(c)}{x-c}\right) \cdot \left(\lim_{x\to c} x-c\right)\\ &= \lim_{x\to c}\left(\frac{f(x)-f(c)}{x-c}(x-c)\right)\\ &= \lim_{x\to c}\left(f(x)-f(c)\right)\end{aligned}$$

Therefore $\lim\limits_{x\to c} f(x) = f(c)$. And so it follows from Theorem 5.18, that f is continuous at c. □

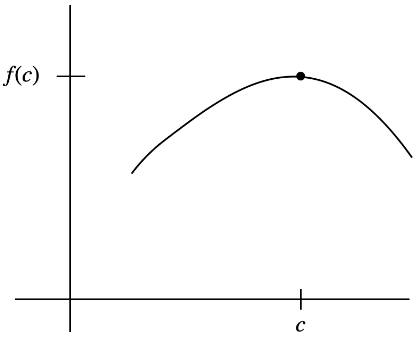
Aside. *Surprisingly, the converse of this theorem is not true. It is possible for a function to be continuous at every point but differentiable nowhere. Do an internet search for* `Weierstrauss Function` *for more details.*

We return now to our discussion on maxima and minima, assured that discontinuities won't ruin our day when we apply our knowledge of the derivative to find maxima and minima of functions.

The relationship between derivatives and extrema comes from the following two ideas:

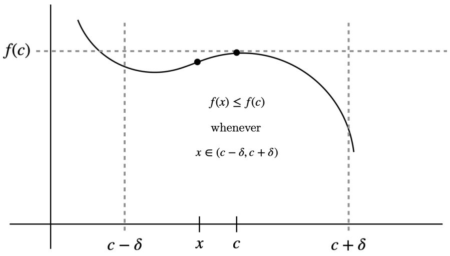
If a function is increasing on the left of $x = c$ and decreasing on the right of $x = c$, then f attains a local maximum at $x = c$

If a function is decreasing on the left of $x = c$ and increasing on the right of $x = c$, then f attains a local minimum at $x = c$



To prove that the derivative being equal to 0 corresponds to a local maximum/minimum, we first need a definition for local maximum/minimum.

Intuitively, for $f(c)$ to be a local *maximum* should mean $f(c) \geq f(x)$ for all values of x that are *close* (i.e., local) to c. This inequality should hold whenever x is *sufficiently close to c*. In other words, there should exist some threshold $\delta > 0$ so that $f(c) \geq f(x)$ whenever $|x - c| < \delta$.



Definition 6.4. *Let $f : I \to \mathbb{R}$ and let $c \in [a, b]$. We say $f(c)$ is a local maximum when there exists $\delta > 0$ so that $f(c) \geq f(x)$ whenever $|x - c| < \delta$. We say $f(c)$ is a local minumum when there exists $\delta > 0$ so that $f(c) \leq f(x)$ whenever $|x - c| < \delta$.*

Using this definition, we can formalise our observation about local maxima and local minima and the derivative at those points.

Lemma 6.5. *Let f be differentiable on $[a, b]$ and let $c \in [a, b]$. If $f(c)$ is a local maximum or a local minimum, then $f'(c) = 0$.*

We won't give a full proof of this theorem as there are some finicky details to overcome, but instead let us talk a little about how we could prove this lemma.

Since f is differentiable at c, the function $\frac{f(x)-f(c)}{x-c}$ converges as x approaches c. By

Theorem 5.7, we have

$$\lim_{x\to 0^-}\frac{f(x)-f(c)}{x-c}=\lim_{x\to 0}\frac{f(x)-f(c)}{x-c}=\lim_{x\to 0^+}\frac{f(x)-f(c)}{x-c}$$

If $f(x)$ is a local maximum then there exists $\delta > 0$ so that $f(x) - f(c) \leq 0$ whenever $|x - c| < \delta$. Approaching c from the right we have $x - c > 0$ and, once x is contained in the open δ-neighbourhood of c, we have $f(x) - f(c) \leq 0$. Therefore $\frac{f(x)-f(c)}{x-c} \leq 0$ for all $x \in (c, c + \delta)$. And so[5],

$$\lim_{x\to 0^+}\frac{f(x)-f(c)}{x-c}\leq 0$$

A similarly-styled argument gives

$$\lim_{x\to 0^-}\frac{f(x)-f(c)}{x-c}\geq 0$$

Therefore

$$0\leq\lim_{x\to 0}\frac{f(x)-f(c)}{x-c}\leq 0$$

And so we conclude

$$\lim_{x\to 0}\frac{f(x)-f(c)}{x-c}=0$$

Therefore $f'(c) = 0$.

Notice that the converse of this lemma is not true. The function $f(x) = x^3$ has $f'(0) = 0$, but $f(0)$ is not a local maximum or local minimum.

Test Your Understanding

1. Using Definition 6.3 prove $f(x) = 2x + 1$ is differentiable at 4 and $f'(4) = 2$
2. By comparing left and right side limits, prove $g(x) = |x + 1|$ is not differentiable at -1
3. Using the definition of local minimum, prove $h(0)$ is a local minimum of $h(x) = x^2$

---

[5] This is the vague part of the argument. To prove this we'd have to make an argument about ϵ's and δ's

Test Your Understanding - Answers

1. We prove $\frac{f(x)-f(4)}{x-4}$ converges at x approaches 4. Notice

$$\begin{aligned}\frac{f(x)-f(4)}{x-4} &= \frac{2x+1-(9)}{x-4}\\ &= \frac{2x-8}{x-4}\\ &= 2\end{aligned}$$

   Therefore
$$f'(4) = \lim_{x\to 4}\frac{f(x)-f(4)}{x-4} = 2$$
   and so $f'(4) = 2$.

2. For $x < 1$ we have $g(x) = -x+1$. We have

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

   Similarly,
$$\lim_{x\to -1^+}\frac{g(x)-g(-1)}{x+1} = 1$$

   Therefore $\frac{g(x)-g(-1)}{x+1}$ does not converge as x approaches -1. Therefore $g(x) = |x+1|$ is not differentiable at -1.

3. We prove that there exists $\delta > 0$ so that $h(x) - h(0) \geq 0$ whenever $x \in (-\delta, \delta)$.

   Let $\delta = 1$. For all $x \in (-\delta, \delta)$ we have $0 \leq h(x) \leq 1$. Since $h(0) = 0$ we have $h(x) - h(0) \geq 0$ whenever $x \in (-\delta, \delta)$. Therefore $h(0)$ is a local minimum of $h(x) = x^2$.

6.1.1 The Mean Value Theorem

Recall the statement of the Intermediate Value Theorem:

Theorem (The Intermediate Value Theorem)**.** *Let f be continuous on $[a,b]$ so that $f(a) < f(b)$ and let $y \in \mathbb{R}$. If $f(a) < y < f(b)$, then there exists $c \in [a,b]$ so that $f(c) = y$.*

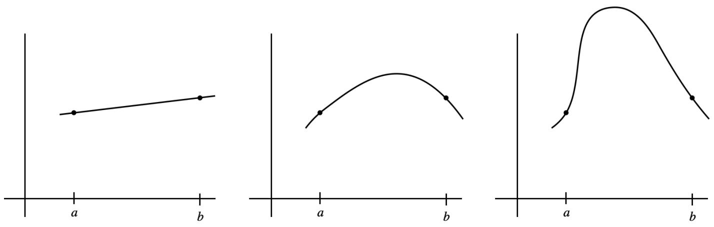
When $f(a) < f(b)$ the Intermediate Value Theorem tells us that a continuous function takes on all possible values in the interval $(f(a), f(b))$ when x is in the interval (a, b). The Intermediate Value Theorem answers the question

If f is continuous on $[a,b]$, what values is f guaranteed to take on this interval?

As differentiation and continuity are closely related (see Theorem 6.3), we consider the same question for differentiable functions:

If f is differentiable on $[a,b]$, what values is f' guaranteed to take on this interval?

If f' is continuous on $[a,b]$, then the Intermediate Value Theorem provides us with an answer: If $f'(a) < f'(b)$, then f' takes on all values in the interval $[f'(a), f'(b)]$. However, even if a function is continuous and differentiable on $[a,b]$ there is no guarantee that f' is continuous[6] on $[a,b]$. And so, the Intermediate Value Theorem is of no help here. To see what values f' is guaranteed to take on the interval $[a,b]$ consider the following three pictures:



On the far left, all tangent lines to this curve would have the same slope. For the middle picture, tangent lines would have slopes ranging between slightly negative and slightly positive. Where as in the third picture, tangent lines could have extremely positive slopes, as well as negative slopes.

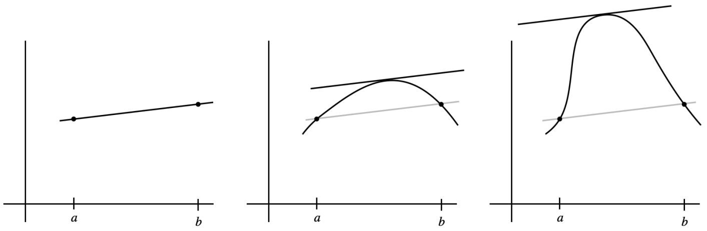
Consider the line ℓ between $(a, f(a))$ and $(b, f(b))$. The line has slope

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

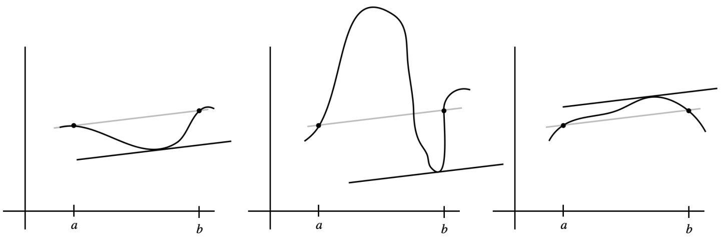
Though the three pictures are of completely different functions, in each picture, we can

[6] We'll look at an example of this in Tutorial 9.

find a tangent line parallel (i.e., with the same slope) to ℓ .

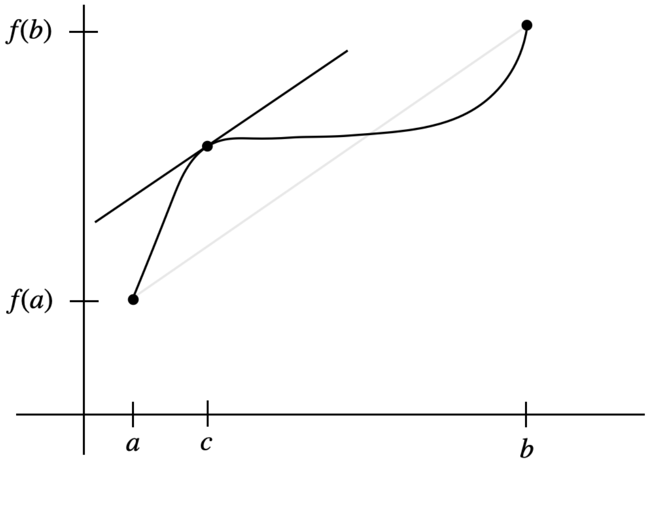


In fact, no matter what differentiable curve we draw between a and b, there will always be at least one point on the curve whose tangent line has slope exactly $\frac{f(b)-f(a)}{b-a}$.



Theorem 6.6 (Mean Value Theorem)**.** *Let $f : [a,b] \to \mathbb{R}$ be continuous on $[a,b]$ and differentiable on (a,b). There exists $c \in (a,b)$ so that*

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



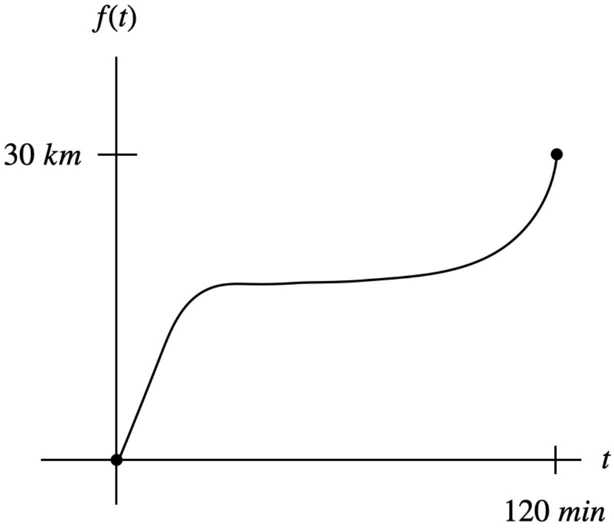
Aside. *This might not seem very interesting or useful to us right now. Fair enough. However, this fact is going to be necessary for us to prove the Fundamental Theorem of Calculus, a fact that is very interesting and useful!*

So far our discussion in this section has been fairly abstract. Before we think about proving this theorem, let us think more about what it is saying.

The word *mean* in the Mean Value Theorem should be taken to mean *average*. Consider the following example:

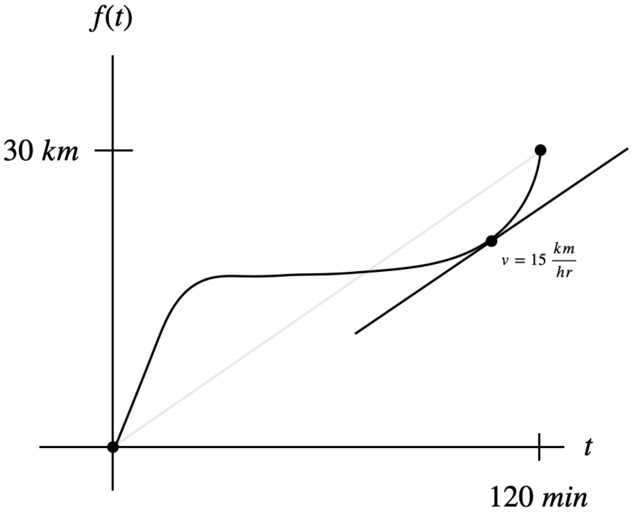
If a bike travels 30km in two hours, then the average speed of the bike was 15km/hr.

Let $f : [0, 120] \to [0, 30]$ so that $f(t)$ is the distance (in km) the bike has traveled after t minutes.



In this context, the derivative of f at time t is the *instantaneous rate of change* of the bike at time t. If $f(0) = 0$ and $f(120) = 30$, then Mean Value Theorem is telling us that there was at least one moment during the journey at which the speed (i.e., instantaneous rate of change) of the bike was

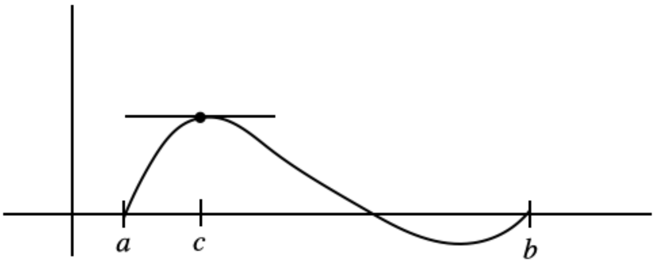
$$\frac{f(120) - f(0)}{120 - 0} = \frac{120\ km}{30\ min} = \frac{1\ km}{4\ min} = 15\ km/hr$$



We spend the remainder of this section proving the Mean Value Theorem. In the interest of brevity, we won't give a full proof the Mean Value Theorem. We will, however, highlight how the proof proceeds.

The proof of the Mean Value Theorem depends on the following result:

Theorem 6.7 (Rolle's Theorem)**.** *Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). If $f(a) = f(b) = 0$, then there exists $c \in (a, b)$ so that $f'(c) = 0$.*



When $f(a) = f(b) = 0$, the statement of the Mean Value Theorem is exactly Rolle's Theorem is disguise. (Take a moment to convince yourself of this fact.)

By the Extreme Value Theorem, since f is continuous on $[a, b]$, f attains a maximum/minimum on this interval. By Lemma 6.5, this maximum/minimum satisfies $f'(c) = 0$.

Proof. Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b) so that$f(a) = f(b) = 0$.

By the Extreme Value Theorem, there exists $k, \ell \in [a, b]$ so that $f(k) \leq f(x) \leq f(\ell)$ for all $x \in [a, b]$. Notice that $f(k)$ is a local minimum and $f(\ell)$ is a local maximum.

We proceed based on the values of $f(k)$ and $f(\ell)$.

If $f(k) = f(\ell) = 0$, then $f(x) = 0$ for all $x \in [a, b]$. Since f is a constant function, the slope of the tangent to any point is 0. Therefore $f'(x) = 0$ for all $x \in [a, b]$. Therefore there exists $c \in (a, b)$ so that $f'(c) = 0$.

Consider now the possibility $f(k) \neq 0$ or $f(\ell) \neq 0$.

If $f(k) \neq 0$, then since $f(a) = f(b) = 0$ we have $k \in (a, b)$. Since $f(k)$ is a local minimum, by Lemma 6.5, it follows that $f'(k) = 0$.

Similarly, if $f(\ell) \neq 0$, then $f'(\ell) = 0$.

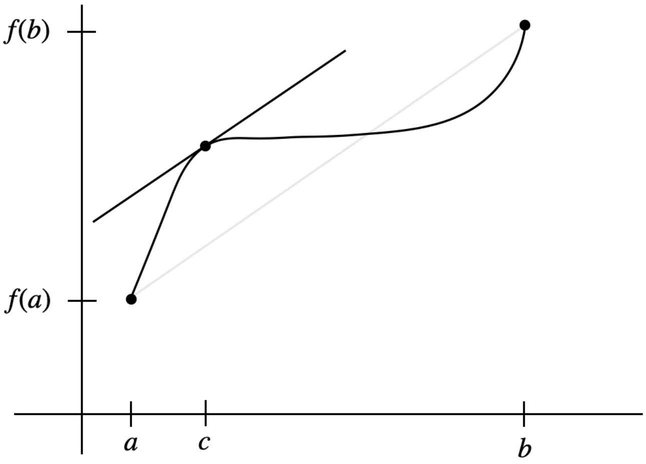
In either case, we have found $c \in (a, b)$ so that $f'(c) = 0$ □

Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). To prove the Mean Value Theorem we use $f(x)$ to define a function $h(x)$ so that

1. $h(x)$ satisfies the hypotheses of Rolle's Theorem, and
2. for every c so that $h'(c) = 0$ we have $f'(c) = \frac{f(b)-f(a)}{b-a}$.

Rolle's Theorem implies that such a c must exist, and it will follow from our construction that

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



The difficult part of this approach is figuring out how we should define $h(x)$. Rather than try to be clever and figure out what $h(x)$ should be, we'll take a look at the answer and see if we can figure out where it came from.

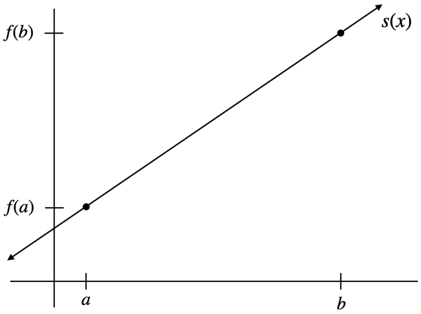
$$h(x) = f(x) - \left(\frac{f(b) - f(a)}{b - a}(x - a) + f(a)\right)$$

Our function $h(x)$ is defined as a difference of two things. Let $s(x) = \frac{f(b)-f(a)}{b-a}(x-a)+f(a)$

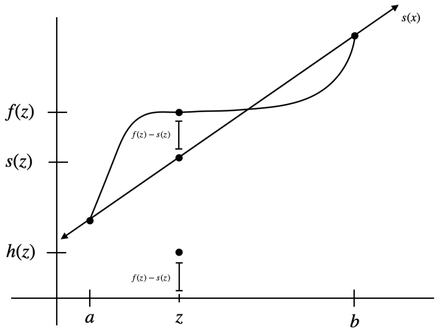
To compute $h(x)$ we would first compute $f(x)$ and then subtract $s(x)$. The function $f(x)$ is the differentiable function defined in the hypothesis of the Mean Value Theorem. But, what is $s(x)$?

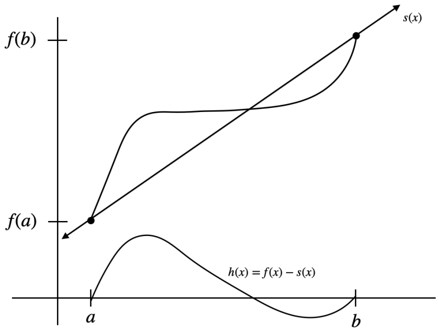
$$s(x) = \frac{f(b) - f(a)}{b - a}(x - a) + f(a)$$

Recalling the various formulas we have to express the equation of a line, we notice $s(x)$ is a line. This line has slope $\frac{f(b)-f(a)}{b-a}$ and it passes through the point $(a, f(a))$.

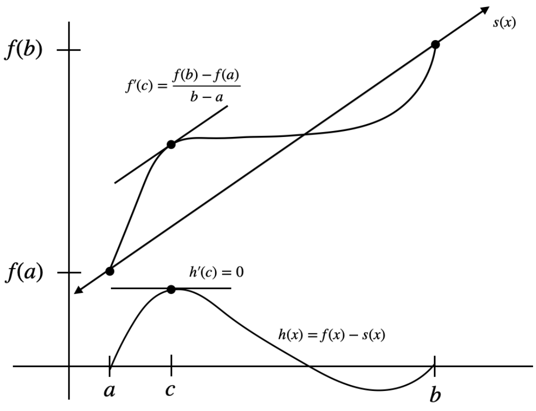


The function $s(x)$ is exactly the line through $(a, f(a))$ and $(b, f(b))$. Therefore the function $h(x)$ is what we get when we subtract the line $s(x)$ from $f(x)$.





Looking at this picture, it appears as if $h(x)$ satisfies the hypotheses of Rolle's Theorem. And furthermore, c, whose existence is guaranteed by the statement of Rolle's Theorem, looks to correspond to an a point on f where the slope of the tangent line is exactly $\frac{f(b)-f(a)}{b-a}$. In other words $f'(c) = \frac{f(b)-f(a)}{b-a}$.



Therefore there exists $c \in (a, b)$ so that

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

The conclusion of the Mean Value Theorem holds.

With all of this work in place, the proof of the Mean Value Theorem isn't terribly interesting. It proceeds using the following steps.

1. Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b).
2. Let $s(x) = \frac{f(b)-f(a)}{b-a}(x - a) + f(a)$.
3. Let $h(x) = f(x) - s(x)$.
4. Prove $h(x)$ satisfies the hypothesis of Rolle's Theorem.
5. Apply Rolle's Theorem to show there exists c so that $c \in (a, b)$ so that $h'(c) = 0$.
6. Use algebraic techniques to show $f'(c) = \frac{f(b)-f(a)}{b-a}$.

This is one of these cases where the written proof doesn't at all do justice to the insight that was needed to get to the proof. If you read the proof of the Mean Value Theorem in a standard Real Analysis textbook, it may eschew all of the intuitive geometry that came from our discussion of $h(x)$. Instead it will define $h(x)$ and proceed with the proof without discussing any of the geometric insight.

Before we end this section, let us take a quick preview of something to come. Recall the first part of the Fundamental Theorem of Calculus:

Theorem (The Fundamental Theorem of Calculus Part I)**.** *If $f : [a, b] \to \mathbb{R}$ is integrable and $F : [a, b] \to \mathbb{R}$ satisfies $F'(x) = f(x)$ for all $x \in [a, b]$, then*

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

The right side of the equals sign, $F(b) - F(a)$ looks very similar to the expression we see if we applied the Mean Value Theorem to $F(x)$

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

The Mean Value Theorem will play an important role in our proof of the Fundamental Theorem of Calculus Part I.

Test Your Understanding

1. Apply the Mean Value Theorem to $f(x) = x^2$ on the interval $[-1, 4]$.
2. Consider $f(x) = 3x^2 + 2x + 1$ on the interval $[1, 3]$.
   (a) Find the equation of the line $s(x)$ as defined in the proof outline of the Mean Value Theorem

(b) Compute $h(2)$ as defined in the proof outline of the Mean Value Theorem.

3. The proof of Rolle's Theorem makes the following claim:

   Since $f(k)$ is a local minimum, by Lemma 6.5, it follows that $f'(k) = 0$.

   However, nowhere in the proof do we prove that $f(k)$ is in a local minimum. Using the definition of local minimum prove $f(k)$, as defined in the proof of Rolle's Theorem, is a local minimum.

Test Your Understanding - Answers

1. There exists $c \in (-1, 4)$ so that $f'(c) = 3$

2. Consider $f(x) = 3x^2 + 2x + 1$ on the interval $[1, 3]$.

   (a) $s(x) = \frac{f(3)-f(1)}{3-1}(x-1) + f(1) = \frac{34-6}{2}(x-1) + 6 = 14(x-1) + 6.$

   (b) $h(2) = f(2) - s(2) = 17 - 20 = -3$

3. By definition, for all $x \in [a, b]$ we have $f(x) \geq f(k)$. Thus we need only to ensure our choice of δ guarantees $(k - \delta, k + \delta) \subseteq [a, b]$.

   Let $\delta = \min\{|k - a|, |k - b|\}$. Therefore $k - \delta \geq a$ and $k + \delta \leq b$. Therefore $(k - \delta, k + \delta) \subseteq [a, b]$. Therefore $f(x) \geq f(k)$ for all $x \in (k - \delta, k + \delta)$. Therefore $f(k)$ is a local minimum.

6.2 Partitions and Riemann Sums

Recall our goal in Section 6 is to prove the Fundamental Theorem of Calculus. After spending some time in 6.1 thinking about differentiation, we now turn to integration. Recall our familiar notation for definite integrals

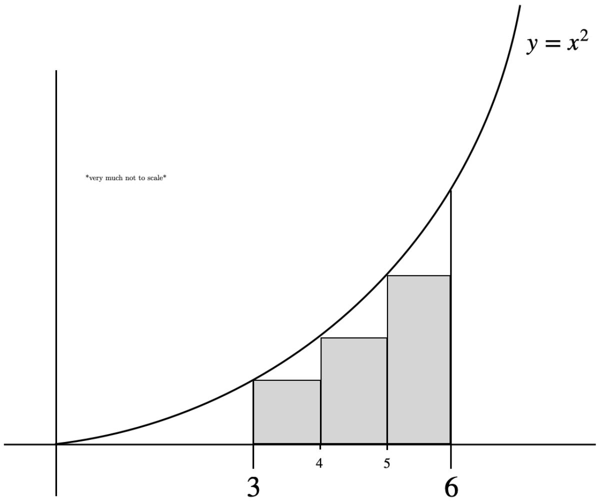
$$\int_a^b f(x)\, dx$$

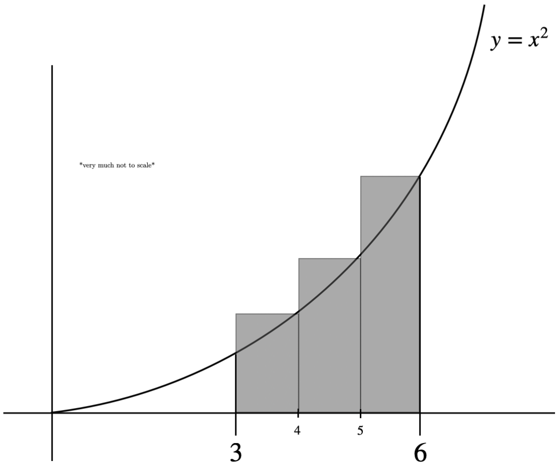
Our introduction to this notation was likely though a discussion about *area under a curve.* This seems like a good place to start.

Consider the definite integral

$$\int_3^6 x^2\, dx$$

Though we have the tools (i.e., the Fundamental Theorem of Calculus) to find the prescribed area exactly, let us instead take a more naive approach. Rather than find the area exactly, we will find upper and lower bounds for this area. To do so, we model the area with a series of rectangles by breaking the interval $[3, 6]$ into sub-intervals.



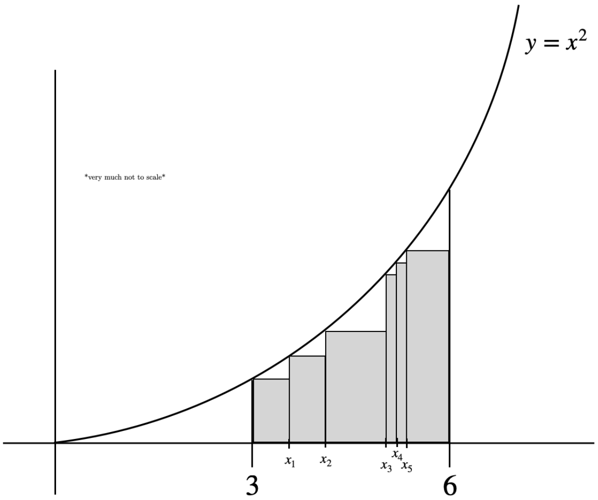


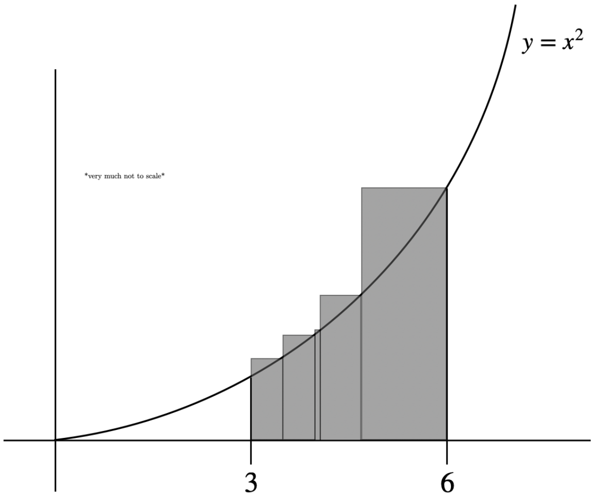
The light grey area is an underestimate of the area under the curve. Computing the sum of the areas of the light grey rectangles, we find:

$$3^2(4-3) + 4^2(5-4) + 5^2(6-5) = 50,$$

a reasonable result as $\displaystyle\int_3^6 x^2\ dx = 63.$

Of course, there is nothing special about these particular rectangles. We can split up our interval any way we want and still get an upper and lower bound. What's more, we don't need to use the same sub-intervals in constructing upper and lower bounds for the areas

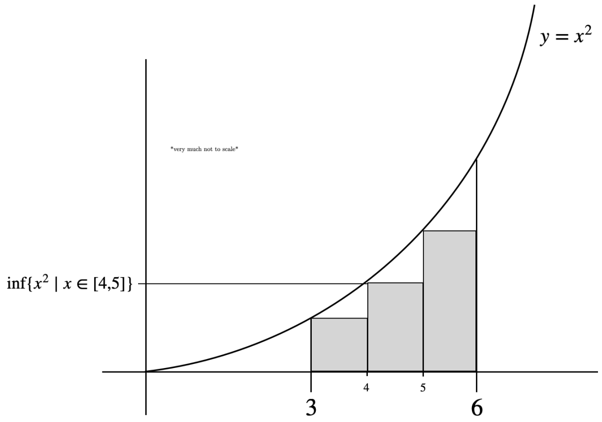




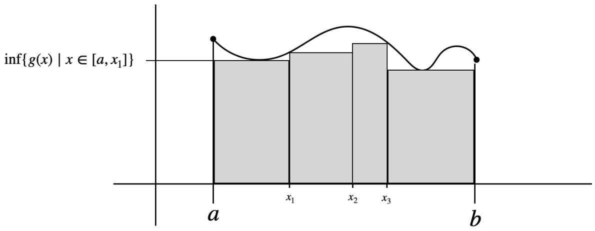
Computing the sum of the areas of the light grey rectangles, we would compute:

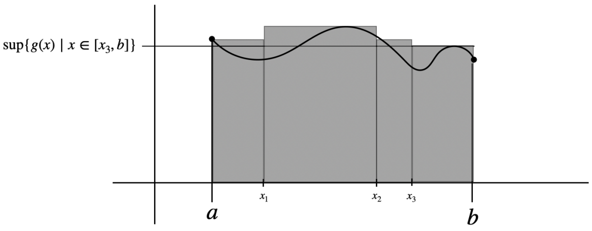
$$3^2(x_1 - 3) + x_1^2(x_2 - x_1) + x_2^2(x_3 - x_2) + x_3^2(x_4 - x_3) + x_4^2(x_5 - x_4) + x_5^2(6 - x_5)$$

Looking at our pictures above, the height of our lower bound rectangles is given by the minimum value of $f(x)$ on each sub-interval.



Similarly, we are finding the height of our upper bound rectangles by taking the maximum value of $f(x)$ on the sub-intervals. In this case these values occur at the end points of the sub-interval. But for an arbitrary curve, there is no reason for this to be the case.





Returning to our $f(x) = x^2$ example, consider the sets of all possible lower and upper bounds we can get for the area by using this method. Let L be set of all possible lower bounds. From our computation above we have $50 \in L$. Similarly, let U be the set of all possible upper bounds.

Since elements of L under estimate the area and elements of U over estimate the area, then, in general, elements of L cannot be bigger than elements of U. If we could find an under estimate and an over estimate that were the same value, then it is reasonable to conclude this value would be exactly equal to the area under the curve.

Let $L(f) = \sup L$ and $U(f) = \inf U$. The number $L(f)$ is the best possible upper bound for the set of lower bounds of the area[7]. Similarly, the number $U(f)$ is the best possible

[7] This sentence might make your head hurt. But it is an important conceptual step in the work we are doing in this section. Take the time to be sure you understand the meaning of this sentence

lower bound for the set of upper bounds of the area. One can prove:

$$L(f) = U(f) = 63 \ \left(= \int_3^6 x^2 \, dx\right)$$

This is all well and good for $f(x) = x^2$, a function we well understand. But, what can be say about an arbitrary $f : [a, b] \to \mathbb{R}$. Defining $L(f)$ and $U(f)$ as above, it would certainly be something special to have

$$L(f) = U(f)$$

Without knowing anything about our arbitrary function f, there is no reason to suspect that the lower and upper bounds ought to *meet in the middle.*

The line of reasoning here is a broad overview of the work we will undertake in 6.2 and 6.3. We will say a function f is $\underline{\text{integrable}}$ when $L(f) = U(f)$. Integrable functions are those for which we can compute $\int_a^b f(x)\, dx$. And so it will be this line of reasoning and the subsequent definition of integrable that will lead us to our proof of the Fundamental Theorem of Calculus.

Our line of reasoning above depended on us first dividing the interval $[3, 6]$ into smaller intervals and then constructing a rectangle in each sub-interval. And so we begin with a definition that brings precise meaning the the idea of dividing an interval into smaller intervals.

Definition 6.8. *A $\underline{\textit{partition of } [a,b]}$ is a set*

$$P = \{x_0, x_1, x_2, \ldots, x_n\}$$

where $x_0 = a, x_n = b$ and $x_0 < x_1 < x_2 < \cdots < x_n$. We refer to each interval of the form $[x_{i-1}, x_i]$ as a $\underline{\textit{sub-interval of } [a,b]}$

Returning to our example from the previous reading, we first partitioned $[3, 6]$ into the following three sub-intervals:

$$[3, 4], [4, 5], [5, 6]$$

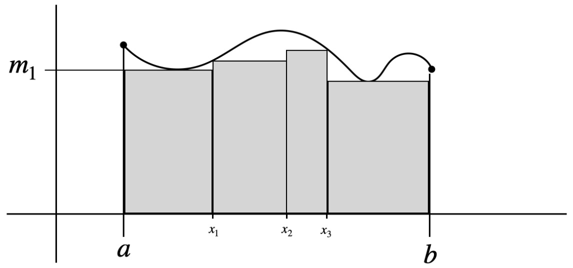
These sub-intervals arose from the partition $P = \{3, 4, 5, 6\}$ of $[3, 6]$.

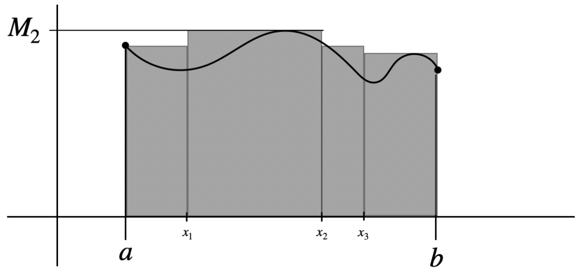
Of course, there is no reason why must choose *nice* values for our sub-intervals. For example, the partition $\left\{3, \pi, \pi + 1, 4.5, 6 - \frac{1}{1000}, 6\right\}$ divides the interval $[3, 6]$ into five subintervals:

$$[3, \pi], [\pi, \pi + 1], [\pi + 1, 4.5], \left[4.5, 6 - \frac{1}{1000}\right], \left[6 - \frac{1}{1000}, 6\right]$$

By estimating the area in each sub-interval we can estimate the area under the curve. On each sub-interval we use the minimum and maximum of f on the sub-interval to give the height of the estimating rectangle. (The width of the rectangle is given by the width of the interval). When f is continuous, the Extreme Value Theorem implies these values are respectively given by the infimum and supremum of f on each sub-interval.

For an interval $[x_{k-1}, x_k]$, let m_k denote the infimum of $\{f(x) \mid x \in [x_{k-1}, x_k]\}$. Let M_k denote the supremum of $\{f(x) \mid x \in [x_{k-1}, x_k]\}$.





For a partition $P = \{x_0, x_1, x_2, \dots, x_n\}$, the corresponding lower bound for the area is given by the sum of the areas of the rectangles:

$$m_1(x_1 - x_0) + m_2(x_2 - x_1) + \cdots + m_n(x_n - x_{n-1}) = \sum_{k=1}^{n} m_k(x_k - x_{k-1})$$

Similarly, the upper bound is given by:

$$\sum_{k=1}^{n} M_k(x_k - x_{k-1})$$

This technique is our key idea in defining integration. And so we give names for these two sums.

Definition 6.9. *Let $f : [a,b] \to \mathbb{R}$ be bounded on $[a,b]$ and let $P = \{x_0, x_1, \dots, x_n\}$ be a partition of $[a,b]$. The lower Riemann sum of f with respect to P is given by*

$$L(f,P) = \sum_{k=1}^{n} m_k(x_k - x_{k-1})$$

The upper Riemann sum of f with respect to P is given by

$$U(f,P) = \sum_{k=1}^{n} M_k(x_k - x_{k-1})$$

Aside. *Riemann sums are named for Bernhard Riemann. In the grand tradition of mathematics naming conventions, they should probably have been named for someone else. More on this later.*

From our example above, when $P = \{3, 4, 5, 6\}$ we have

$$\begin{aligned} L(f,P) &= \sum_{k=1}^{3} m_k(x_k - x_{k-1}) \\ &= m_1(x_1 - x_0) + m_2(x_2 - x_1) + m_3(x_3 - x_2) \\ &= m_1(4-3) + m_2(5-4) + m_3(6-5) \end{aligned}$$

As we can interpret $L(f, P)$ as a lower bound for the area under the curve, and $U(f, P)$ as an upper bound, we expect

$$L(f, P) \leq U(f, P)$$

On any particular sub-interval $[x_{k-1}, x_k]$ we have $m_k \leq M_k$ (m_k the infimum of the values of $f(x)$ on the interval, and M_i is the supremum). Therefore

$$m_k(x_k - x_{k-1}) \leq M_k(x_k - x_{k-1})$$

Summing over all sub-intervals in the partition yields

$$L(f, P) = \sum_{k=1}^{n} m_k(x_k - x_{k-1}) \leq \sum_{k=1}^{n} M_k(x_k - x_{k-1}) = U(f, P)$$

Lemma 6.10. *Let $f : [a, b] \to \mathbb{R}$ be bounded on $[a, b]$. For every partition P of $[a, b]$ we have*

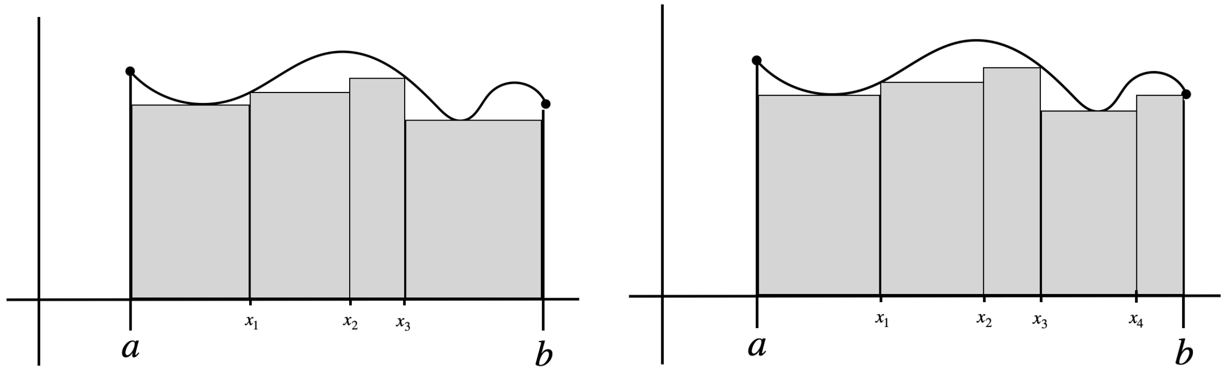
$$L(f, P) \leq U(f, P)$$

Lemma 6.10 permits us to compare the upper and lower sum of the same partition. However, there is no guarantee that the *best* partition for the lower sum is the same as the *best* partition for the upper sum. Ideally, we would want a result that lets us compare upper and lower sums of different partitions.

If P and P' are partitions of $[a, b]$, then

$$L(f, P) \leq U(f, P')$$

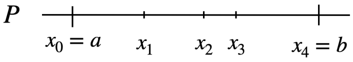
To get to this result we first consider the process of adding points to a partition. Consider the lower sum below and the result of adding an additional point between x_3 and b.

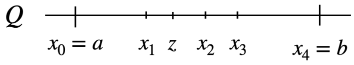


From the picture we see that adding a point to the partition doesn't make our estimate worse. In this case, it makes it better. Since a partition is a set, adding points to the partition results in a larger set. We call this larger set a refinement.

Definition 6.11. *Let P and Q be partitions of $[a,b]$. We say Q is a refinement of P when $P \subseteq Q$.*

Let $P = \{x_0, x_1, x_2, x_3, x_4\}$ be a partition of $[a,b]$ and consider the effect of adding a single extra point z to P between x_1 and x_2. Let Q be the resulting refinement.





Computing the lower sum of P gives:

$$L(f,P) = \sum_{k=1}^{4} m_k(x_k - x_{k-1}) = m_1(x_1 - x_0) + m_2(x_2 - x_1) + m_3(x_3 - x_2) + m_4(x_4 - x_3)$$

And for the lower sum of Q we have:

$$L(f,Q) = m_1(x_1 - x_0) + m_z(z - x_1) + m_z'(x_2 - z) + m_3(x_3 - x_2) + m_4(x_4 - x_3)$$

where

$$m_z = \inf\{f(x) \mid x \in [x_1, z]\} \text{ and } m_z' = \inf\{f(x) \mid x \in [z, x_2]\}$$

And so to compare the values of $L(f,P)$ and $L(f,Q)$ we need only compare the value of $m_z(z - x_1) + m_z'(x_2 - z)$ with $m_2(x_2 - x_1)$. In particular

If $m_z(z - x_1) + m_z'(x_2 - z) \geq m_2(x_2 - x_1)$, then $L(f,Q) \geq L(f,P)$.

If $m_z(z - x_1) + m_z'(x_2 - z) \leq m_2(x_2 - x_1)$, then $L(f,Q) \leq L(f,P)$.

Since m_2 is the infimum of f on $[x_1, x_2]$ and $[x_1, x_2] = [x_1, z] \cup [z, x_2]$, then $m_2 \leq m_z$ and $m_2 \leq m_z'$. Notice

$$m_2(x_2 - x_1) = m_2(x_2 - z + z - x_1) = m_2(x_2 - z) + m_2(z - x_1) \leq m_z(z - x_1) + m_z'(x_2 - z)$$

Therefore

$$L(f,P) \leq L(f,Q)$$

In other words $L(f,Q)$ is a not a worse lower bound for the area under the curve than $L(f,P)$.

In this argument we formed Q from P by adding in a single new point to our partition between x_1 and x_2. We could have inserted this point anywhere and the argument would have been the same, except with some different indices floating around. Inductively, we can add as many points to P as we want without making the resulting upper sum a worse upper bound for the area under the curve than $L(f,P)$.

Similarly, we can make the same argument for upper sums. And so we have the following result.

Lemma 6.12. *Let $f : [a,b] \to \mathbb{R}$ be bounded on $[a,b]$ and let P be a partition of $[a,b]$. If Q is a refinement of P, then $L(f,P) \leq L(f,Q)$ and $U(f,Q) \leq U(f,P)$.*

Consider a pair of partitions P and P' of $[a, b]$. Using refinements we can compare $L(f, P)$ and $U(f, P')$ as follows: Notice that $P \cup P'$ is a refinement of P and also a refinement of P'. And so by Lemma 6.12 we have

$$L(f, P) \leq L(f, P \cup P') \text{ and } U(f, P \cup P') \leq U(f, P')$$

By Lemma 6.10 we have

$$L(f, P \cup P') \leq U(f, P \cup P')$$

Therefore

$$L(f, P) \leq U(f, P')$$

Theorem 6.13. *Let $f : [a, b] \to \mathbb{R}$ be bounded on $[a, b]$. If P and P' are partitions of $[a, b]$, then*

$$L(f, P) \leq U(f, P')$$

Lemma 6.12 lets us take an upper or lower sum and potentially improve it by adding in extra points. Theorem 6.13 permits us to compare lower and upper sums of different partitions.

When $f \geq 0$, the lower sum gives a lower bound for the area under a curve. Similarly, each upper sum gives an upper bound for the area under a curve.

Let $\tilde{P}$ be the set of all partitions of $[a, b]$ and let L be the set of all lower sums we get over all possible partitions of $[a, b]$.

$$L = \{L(f, P) \mid P \in \tilde{P}\}$$

Similarly, let U be the set of all upper sums we get over all possible partitions of $[a, b]$.

$$U = \{U(f, P) \mid P \in \tilde{P}\}$$

By Theorem 6.13, If ℓ and u are respectively elements of L and U, then necessarily $\ell \leq u$. What would it mean for this inequality to hold with equality? That is, what can we deduce if there exists $\ell \in L$ and $u \in U$ so that $\ell = u$?

Since ℓ is a lower bound for the area and u is an upper bound for the area, if $\ell = u$, then the area under the curve must be equal to ℓ (and to u.).

Further, by Theorem 6.13, it must also mean that ℓ is the largest of all of the elements of L and u is the smallest of all of the elements of U. In other words, ℓ arose from the *best* partition for constructing a lower bound and u arose from from the *best* partition for constructing an upper bound But, how can we possibly hope to find the *best* partition, especially if we can use refinement to potentially find a better partition for an upper or lower sum.

By leveraging suprema and infima, we don't have to! We discuss this in the next section.

Aside. *There is something sneaky happening here. We are thinking about areas as build our intuition around upper and lower Riemann sums. However, our definitions and results do not explicitly refer to areas. In fact, there is no guarantee that $L(f, P)$ and $U(f, P)$*

are even positive values! The geometry is helping us to build our intuition, but our results will not limited to functions for which the geometric interpretation makes sense. Abstract mathematics need not have a physical interpretation to be meaningful.

Test Your Understanding

1. Let $f(x) = 2x + 1$. Compute the lower Riemann sum of f for the partition $P = \{0, 1.5, 2.5, 3\}$.

2. Let

$$\mathcal{I}(x) = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & x \notin \mathbb{Q} \end{cases}$$

   Compute the upper and lower Riemann sums of $\mathcal{I}$ for the partition $P = \{0, 1.5, 2.5, 3\}$.

3. Each of the results in this section have the hypothesis: *Let f be bounded on $[a, b]$.* Why is this hypothesis necessary in our discussion of Riemann sums?

Test Your Understanding - Answers

1. To compute the lower Riemann sum we must find the minimum value of f on each sub-interval. Since f is increasing, this will occur at the left endpoints

$$\begin{aligned} m_1 &= 2(0) + 1 = 1 \\ m_2 &= 2(1.5) + 1 = 4 \\ m_3 &= 2(2.5) + 1 = 6 \end{aligned}$$

   Therefore

$$\begin{aligned} L(f, P) &= \sum_{k=1}^{4} m_k(x_k - x_{k-1}) \\ &= m_1(x_1 - x_0) + m_2(x_2 - x_1) + m_3(x_3 - x_2) \\ &= 1(1.5) + 2(1) + 6(0.5) \\ &= 1.5 + 2 + 3 \\ &= 6.5 \end{aligned}$$

2. In each sub-interval we can find both a rational and an irrational number. Therefore $m_k = 0$ and $M_k = 1$ for each sub-interval of the partition. Therefore

$$\begin{aligned} &L(f, P) = 0 \\ &U(f, P) = (x_1 - x_0) + (x_2 - x_1) + (x_3 - x_2) = x_3 - x_0 = 3 \end{aligned}$$

3. To compute a lower sum we need to find the infimum of f on each sub-interval. The infimum of f will only exist if f is bounded below on each sub-interval. Thus we require f to be bounded below on the entire interval. Similarly, to compute an upper sum we require f to be bounded above on the entire interval. Thus we require f to be bounded on the entire interval.

6.3 The Riemann Integral

Our work in the previous section defined the upper and lower Riemann sum of a partition. In estimating the area under a curve, our goal is to choose the best partition of the lower sum and the best partition for the upper sum, in the hopes that they are equal.

However, actually figuring out which partition to choose sounds difficult. What's worse, our work on refinements tells us that once a partition is in place, we can do no worse by refining our partition.

Rather than try to *find* the best possible partition of the upper and lower sum, instead let us take advantage of our knowledge of supremum and infimum. That is, let us consider the following two values:

$$L(f) = \sup\{L(f,P) \mid P \in \tilde{P}\} \text{ and } U(f) = \inf\{U(f,P) \mid P \in \tilde{P}\}$$

where $\tilde{P}$ be the set of all partitions of $[a,b]$. By definition, for any partition P of $[a,b]$ we have

$$L(f,P) \leq L(f) \text{ and } \; U(f) \leq U(f,P)$$

Aside. *We take the supremum of the lower bounds because we want the best (i.e., biggest) possible lower bound. Similarly, we take the infimum of the upper bounds because we want the best (i.e., smallest) possible upper bound.*

In the ideal scenario, we will have $U(f) = L(f)$ and thus we have found the area under the curve. However, this may not actually occur. Consider the following example, the *the rational indicator function* on the interval $[1,2]$

$$\mathcal{I}(x) = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & x \notin \mathbb{Q} \end{cases}$$

Regardless of our choice of partition, in every sub-interval $[x_{k-1}, x_k]$ we can find both a rational and an irrational number. Therefore $m_k = 0$ and $M_k = 1$ for all sub-intervals. And so, no matter which partition $P = \{x_0, x_1, \ldots, x_n\}$ we choose we will have

$$\begin{aligned} L(\mathcal{I}, P) &= \sum_{k=1}^{n} m_k(x_k - x_{k-1}) \\ &= \sum_{k=1}^{n} 0(x_k - x_{k-1}) \\ &= 0 \end{aligned}$$

and

$$\begin{aligned}
U(\mathcal{I}, P) &= \sum_{k=1}^{n} M_k(x_k - x_{k-1}) \\
&= \sum_{k=1}^{n} 1(x_k - x_{k-1}) \\
&= (x_1 - x_0) + (x_2 - x_1) + (x_3 - x_2) + \cdots + (x_n - x_{n-1}) \\
&= (\cancel{x_1} - x_0) + (\cancel{x_2} - \cancel{x_1}) + (\cancel{x_3} - \cancel{x_2}) + \cdots + (x_n - \cancel{x_{n-1}}) \\
&= x_n - x_0 \\
&= 2 - 1 \\
&= 1
\end{aligned}$$

Therefore

$$\begin{aligned}
L(\mathcal{I}) &= \sup\{L(f, P) \mid P \in \tilde{P}\} \\
&= \sup\{0\} \\
&= 0
\end{aligned}$$

and

$$\begin{aligned}
U(\mathcal{I}) &= \sup\{U(f, P) \mid P \in \tilde{P}\} \\
&= \sup\{1\} \\
&= 1
\end{aligned}$$

And so $L(\mathcal{I}) \neq U(\mathcal{I})$.

From this example we see that not every function will have the property $U(f) = L(f)$. It is only those functions for which we have $U(f) = L(f)$ we define the notation

$$\int_a^b f(x)\, dx$$

Definition 6.14. *Let $f : [a, b] \to \mathbb{R}$ be bounded. Let $\tilde{P}$ be the set of partitions of $[a, b]$. The lower Riemann integral of f on $[a, b]$ is denoted as $L(f)$ and given by*

$$L(f) = \sup\{L(f, P) \mid P \in \tilde{P}\}$$

The upper Riemann integral of f on $[a, b]$ is denoted as $U(f)$ and given by

$$U(f) = \inf\{U(f, P) \mid P \in \tilde{P}\}$$

Definition 6.15. *Let $f : [a, b] \to \mathbb{R}$ be bounded. We say f is integrable on $[a, b]$ when $L(f) = U(f)$. When f is integrable on $[a, b]$ we denote $L(f)$ and $U(f)$ as*

$$\int_a^b f(x)\, dx$$

In other words we define the notation $\int_a^b f(x)\,dx$ *so that*

$$L(f) = \int_a^b f(x)\,dx = U(f)$$

With this new terminology in mind, let us return back to our familiar example: the area under the curve $f(x) = x^2$ on the interval $[3, 6]$. To prove $f(x) = x^2$ is integrable on $[3, 6]$ we would have to prove

$$\sup\{L(f, P) \mid P \in \tilde{P}\} = \inf\{U(f, P) \mid P \in \tilde{P}\}$$

This is an entirely non-obvious fact! It only once we confirm $L(f) = U(f)$ that we can be certain the notation

$$\int_3^6 x^2\,dx$$

has meaning.

From the definitions above there are a few things to notice:

Firstly there is nothing in this definition that refers specifically to *area*. We have stripped away all of the geometric intuition for the meaning of integration and are left only with statements about sets, suprema and infima. Understandably, this might not seem like a particularly helpful thing for us to have done. However, there are a few good reasons for us to have done so.

Chiefly is the fact that it is difficult to draw pictures in full generality. If we define integration using the geometry of area, then we need to pay special attention to cases where our function may not be restricted to the upper right quadrant of the plane. We would need have different pictures depending on the different possible behaviours of f. What's more, our ability to draw helpful pictures doesn't extend beyond drawing pictures of functions that are continuous, a property that does not appear on the definitions above. It is very possible for a non-continuous function to be integrable.

Another good reason to strip away all of the geometric intuition for integration is that this definition permits us to define integration in other contexts. For example, all of the work we have done follows (almost) the same if we replace $\mathbb{R}$ with $\mathbb{C}$ or $\mathbb{R}^n$ or $\mathbb{C}^n$, places where we have no intuition at all for the geometry of areas.

Secondly *an integral* isn't something *to do*. An integral is a supremum/infimum of a set. *Computing* an integral amounts to finding the infimum/supremum of sets.

Thirdly, from our work above, the rational indicator function is not integrable on any domain $[a, b]$. This is because, no matter which partition we choose, the lower Riemann sum is always equal to 0 and the upper Riemann sum is always equal to 1. And so the notation

$$\int_a^b \mathcal{I}(x)\,dx$$

is meaningless!

The argument that the rational indicator function is not integrable required us to make an argument about partitions. Given that there are many partitions to consider, we can imagine that it will be quite difficult to use the definition of integrable to prove that a function is indeed integrable. Indeed, the Completeness Axiom only tells us that $L(f)$ exists, but tells us nothing about what value it takes. Fortunately we have a few results that help us with this task.

Theorem 6.16. *Let $f : [a, b] \to \mathbb{R}$. If f is continuous on $[a, b]$, then f is integrable on $[a, b]$.*

Theorem 6.17. *Let $f : [a, b] \to \mathbb{R}$ and let f be integrable on $[a, b]$.*

1. $\int_a^b f(x)\ dx = \int_a^c f(x)\ dx + \int_c^b f(x)\ dx$ *for all* $c \in (a, b)$
2. $\int_a^b f(x) + g(x)\ dx = \int_a^b f(x)\ dx + \int_a^b g(x)\ dx$
3. $\int_a^b \lambda f(x)\ dx = \lambda \int_a^b f(x)\ dx$ *for all* $\lambda \in \mathbb{R}$

The latter two parts of this theorem imply, just like the derivative, the integral is linear.

We omit the proofs of these theorems. The proof of the former theorem requires some tools about continuity we did not have the time to develop. The proof of the latter theorem follows from a theorem we will discuss on Assignment 5.

Notably, the converse of Theorem 6.16 is false. In tutorial we examine an example of a function that is not continuous on its domain but is integrable on its domain.

As we will soon come to experience, working directly with the definition of integrability is not for the feint of heart. To actually prove that a function is integrable using the definition we need to have enough information about the function to be able to describe the set of all upper and lower Riemann sums, no easy feat as there there are infinitely many possible partitions to consider.

Recall that a function is integrable when $L(f) = U(f)$. Let P be a partition, from our work above we have

$$L(f, P) \le L(f) \le U(f) \le U(f, P)$$

If $L(f, P) = U(f, P)$, then necessarily $L(f) = U(f)$, and thus the function is integrable.

Back in Section 2 of the course, we had saw the following theorem for establishing a supremum:

Theorem (Theorem 2.20)**.** *Let $B \subseteq \mathbb{R}$ be bounded above and non-empty. Let $\gamma \in \mathbb{R}$ be an upper bound of B in $\mathbb{R}$. We have* $\sup B = \gamma$ *if and only if for every $\epsilon \in \mathbb{R}$ with $\epsilon > 0$ there is an element of A greater than $\gamma - \epsilon$.*

Surprisingly, this theorem is useful in determining if a discontinuous function is Riemann integrable. Consider the function:

$$f(x) = \begin{cases} 1 & x \neq 2 \\ 0 & x = 2 \end{cases}$$

on the interval $[1,3]$ We claim $L(f) = U(f) = 2$. To prove $L(f) = 2$ we must prove that for every $\epsilon > 0$ there exists $a \in \{L(f,P) \mid; P \in \tilde{P}\}$ so that $a > 2 - \epsilon$. For every element $b \in \{L(f,P) \mid; P \in \tilde{P}\}$ there is a partition P so that $b = L(f,P)$. Thus, to establish $L(f) = 2$ we must prove that for every $\epsilon > 0$ there exists $P \in \tilde{P}$ so that $L(f,P) < 2 - \epsilon$.

For $\epsilon > 2$, the partition $P = \{1,3\}$ suffices. We can compute $L(f,P) = 0$ and note for $\epsilon > 2$ we have $0 > 2 - \epsilon$.

Consider now $\epsilon < 2$ and the partition $P = \left\{1, 2 - \frac{\epsilon}{4}, 2 + \frac{\epsilon}{4}, 3\right\}$. We have

$$\begin{aligned} m_1 &= 1 \\ m_2 &= 0 \\ m_3 &= 1 \end{aligned}$$

Therefore

$$L(f,P) = (2 - \frac{\epsilon}{4} - 1) + (3 - 2 - \frac{\epsilon}{4}) = 2 - \frac{\epsilon}{2} > 2 - \epsilon$$

Therefore for every $\epsilon > 0$ there exists $a \in \{L(f,P) \mid; P \in \tilde{P}\}$ so that $a > 2 - \epsilon$. By Theorem 1.28 we conclude $L(f) = 2$.

A similar argument shows $U(f) = 2$. And thus we conclude f is integrable and $\displaystyle\int_1^3 f(x)dx = 2$.

Using Theorem 2.20, one can prove the following theorem, which fully classifies integrable functions.

Theorem 6.18. *Let $f : [a,b] \to \mathbb{R}$ be bounded. Then f is integrable on $[a,b]$ if and only if for every $\epsilon > 0$ there exists a partition P_ϵ so that*

$$U(f,P_\epsilon) - L(f,P_\epsilon) < \epsilon$$

We consider the this theorem and its proof in the Test Your Understanding questions for this section.

Darboux, Riemann and Integration Notation

The work we have done so far in precisely defining integration does little to help us understand where our standard notation for integration comes from:

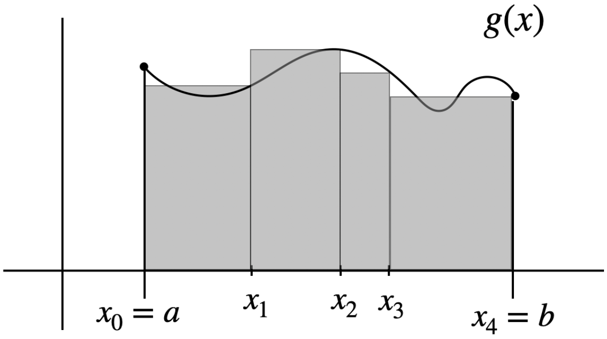
$$\int_a^b f(x)\, dx$$

Let us take a moment to do so. And, in doing so, give credit where it is due for the work we have done in this section.

Riemann integration is named for Bernhard Riemann (1826-1866). However the ideas that we introduced in the previous readings are not at all Riemann's. Instead they are due to Gaston Darboux (1842-1917). Though we define integration using Darboux's ideas, the notation we use is due to Riemann.

Riemann's approach to integration was similar to Darboux's, but differed in some key ways. Let $f:[a,b]\to\mathbb{R}$ be bounded. Let P be a partition of $[a,b]$.

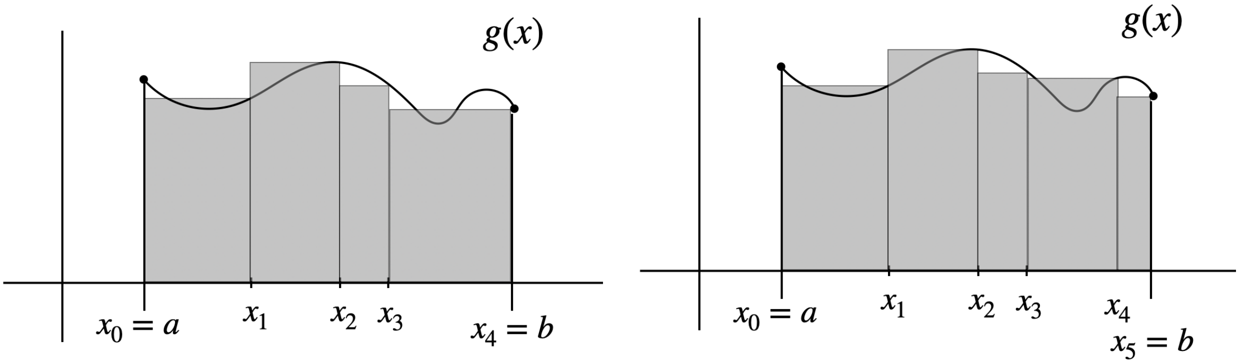
Following Riemann's approach, rather than take the maximum and minimum value of f on each interval, let us instead take the right endpoint[8] in each interval as the height of our rectangle.



The area under the curve is estimated by the sum

$$\sum_{k=1}^{4} g(x_k)(x_k - x_{k-1}) = g(x_1)(x_1-x_0)+g(x_2)(x_2-x_1)+g(x_3)(x_3-x_2)+g(x_4)(x_4-x_3).$$

Depending on the *shape* of the function, the resulting estimation may be an over estimate or an under estimate. Riemann's observation was that regardless of whether the resulting estimation was an over or under estimate, we could improve the estimate by refining the partition.



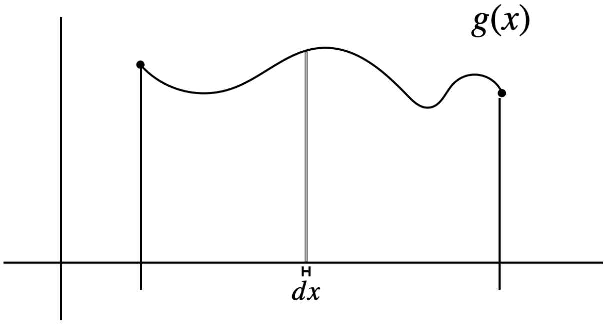
But without squeezing the area between upper and lower sums, how can we hone in on the actual area under the curve? Riemann's idea was to split the partition into infinitely

[8] This is a little bit misleading. In fact we can take any point we want in the interval to find the height of our rectangle. But choosing the right endpoint is a convenient simplification.

many parts. But to do so, one must reckon with an infinite sum! However, we can avoid this by repeatedly refining the partition, taking a sequential limit and hoping it converges.

$$\lim_{n\to\infty} \sum_{k=1}^{n} g(x_k)(x_k - x_{k-1}) = \text{ area under the curve}$$

As n gets very large, the width of each partition $(x_k - x_{k-1})$ is going to get very small. Let dx denote with width of an *infinitesimally small* partition.



As each partition gets very small, eventually every point on the curve will be the right endpoint of a sub-interval. And so rather than summing over the right end points of the sub-intervals, instead we can take $g(x)$ as the height of each rectangle, over all $x \in [a, b]$

With these notations in place, is it perhaps reasonable for us to write:

$$\text{area under the curve } = \lim_{n\to\infty} \sum_{k=1}^{n} g(x_i)(x_i - x_{i-1}) = \sum_{a\le x\le b} g(x)\ dx$$

Replacing the notation $\sum\limits_{a\le x\le b}$ with the notation $\displaystyle\int_a^b$, give us the following familiar expression for the area under a curve:

$$\int_a^b g(x)\ dx$$

Aside. *The work we have done discussing Riemann's approach is very sketchy! To fully reckon with Riemann's approach we would need to consider infinite sums. We return to this topic in Section 7 of the course.*

Test Your Understanding

1. Let $f(x) = 1$. Using Definition 6.15, prove f is integrable on $[2, 5]$.

2. Compute
$$\int_2^5 1\ dx$$
by appealing the the definition of this notation given in Definition 6.15

3. Let $k \in \mathbb{R}$ and let $g(x) = k$. Using Theorem 6.17 and the result from the previous question, prove
$$\int_2^5 g(x)\ dx = 3k$$

4. Consider the following proof of Theorem 6.18:

1 *Proof.* Assume f is integrable on $[a, b]$. Let $\epsilon > 0$. Let $\epsilon' = \frac{\epsilon}{2}$.

2 There exists $a \in \{L(f, P) \mid P \in \tilde{P}\}$ so that $a > L(f) - \epsilon'$

3 Similarly, there exists $b \in \{U(f, P) \mid P \in \tilde{P}\}$ so that $b < U(f) + \epsilon'$

4 Since $a \in \{L(f, P) \mid P \in \tilde{P}\}$ there exists $P \in \tilde{P}$ so that $L(f, P) = a$.

5 Similarly, there exists $Q \in \tilde{P}$ so that $U(f, Q) = b$.

6 Therefore $L(f, P) + \epsilon' > L(f)$ and $U(f) > U(f, Q) - \epsilon'$.

7 Since f is integrable on $[a, b]$ we have $L(f) = U(f)$.

8 And so $L(f, P) + \epsilon' > U(f, Q) - \epsilon'$.

9 Rearranging, we have $U(f, Q) - L(f, P) < \epsilon' + \epsilon' = \epsilon$.

10 Consider the partition $P \cup Q$. Notice $P \cup Q$ is a refinement of both P and Q.

11 We have $L(f, P) \leq L(f, P \cup Q)$ and $U(f, Q) \geq U(f, P \cup Q)$.

12 Therefore $U(f, P \cup Q) - L(f, P \cup Q) < \epsilon$.

13 Therefore for every $\epsilon > 0$ there exists a partition P_ϵ so that $U(f, P_\epsilon) - L(f, P_\epsilon) < \epsilon$

14 To prove the converse, we proceed by contradiction.

15 Assume for every $\epsilon > 0$ there exists a partition P so that $U(f, P) - L(f, P) < \epsilon$,
16 but f is not integrable on $[a, b]$.

17 Since f is not integrable on $[a, b]$ we have $L(f) < U(f)$.

18 Let $r = U(f) - L(f)$. Notice $r > 0$. Let $\epsilon = r$.

19 By hypothesis, there exists a partition P_ϵ so that $U(f, P_\epsilon) - L(f, P_\epsilon) < \epsilon$.

20 Notice $U(f, P_\epsilon) - L(f, P_\epsilon) = (U(f, P_\epsilon) - U(f)) + (L(f) - L(f, P_\epsilon)) + (U(f) - L(f))$

21 By definition $U(f, P_\epsilon) - U(f) \geq 0$ and $L(f) - L(f, P_\epsilon) \geq 0$.

22 Therefore $U(f, P_\epsilon) - L(f, P_\epsilon) \geq 0 + 0 + r = \epsilon$.

23 This is a contradiction.

24 Therefore if for every $\epsilon > 0$ there exists a partition P_ϵ so that $U(f, P_\epsilon) - L(f, P_\epsilon) < \epsilon$,
25 then f is integrable on $[a, b]$.

□

(a) How do you know the statement on line 2 is true. Cite a result or definition from the subject material to justify your response.

(b) How do you know the second sentence on line 10 is true. Cite a result or definition from the subject material to justify your response.

(c) How do you the inequalities on line 11 are true. Cite a result or definition from the subject material to justify your response.

(d) Which definition(s) is/are being referenced on line 21?

(e) What is the contradiction on line 23?

Test Your Understanding - Answers

1. To prove f is integrable on $[2, 5]$ we must prove $L(f) = U(f)$. For every subinterval of every partition P of $[2, 5]$ we have $m_k = M_k = 1$. Therefore for every partition P we have $L(f, P) = U(f, P)$. Therefore $L(f) = U(f)$. Thus f is integrable on $[2, 5]$

2. Since f is integrable, we have

$$L(f) = \int_2^5 1\, dx$$

Let $P = \{x_0, x_1, \ldots, x_n\}$ be a partition of $[2, 5]$.

$$L(f, P) = \sum_{i=1}^{n} m_i(x_i - x_{i-1}) = \sum_{i=1}^{n} 1(x_i - x_{i-1}) = 1 \sum_{i=1}^{n}(x_i - x_{i-1}) = (x_n - x_0) = (5 - 2) = 3$$

Therefore $L(f, P) = 3$ for all partitions P. Therefore $L(f) = 3$. And so

$$\int_2^5 1\, dx = 3$$

3. Since integration is linear we have

$$\int_2^5 g(x)\, dx = \int_2^5 k\, dx = k \int_2^5 1\, dx$$

From our work in the previous part,

$$k \int_2^5 1\, dx = k(3) = 3k$$

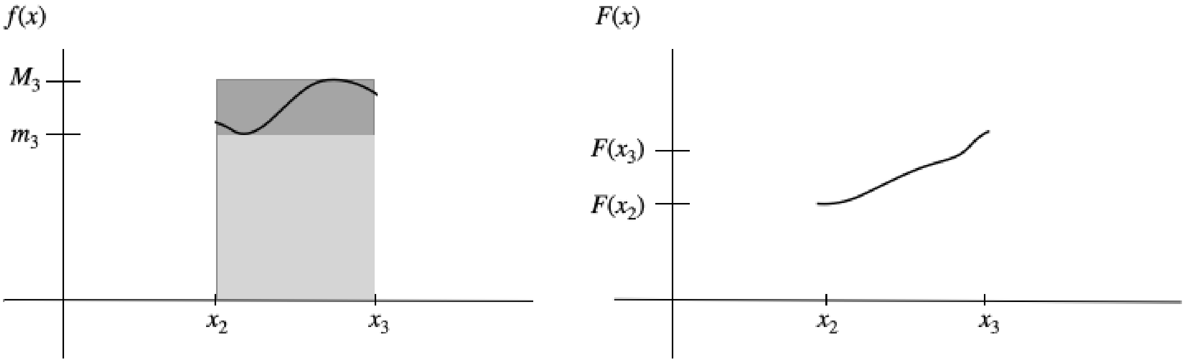
6.3.1 The Fundamental Theorem of Calculus Part I

Finally we reach our discussion on the Fundamental Theorem of Calculus.

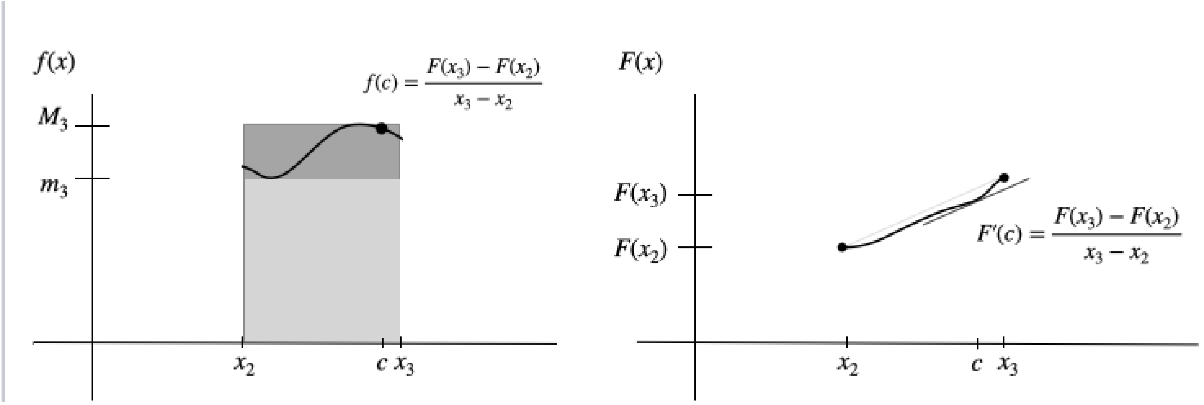
Theorem 6.19 (The Fundamental Theorem of Calculus Part I)**.** *If $f : [a,b] \to \mathbb{R}$ is integrable and $F : [a,b] \to \mathbb{R}$ satisfies $F'(x) = f(x)$ for all $x \in [a,b]$, then*

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

Before we dive in to our proof of Part I, let us take some time to develop some ideas. Let $f : [a,b] \to \mathbb{R}$ be integrable so that $F : [a,b] \to \mathbb{R}$ satisfies $F'(x) = f(x)$ for all $x \in [a,b]$. Let P be a partition of $[a,b]$ and let us take a look at F and f in the subinterval $[x_2, x_3]$



By the Mean Value Theorem there exists $c \in (x_2, x_3)$ so that $F'(c) = \frac{F(x_3)-F(x_2)}{x_3-x_2}$. Since $F'(x) = f(x)$, there exists $c \in (x_2, x_3)$ so that $f(c) = \frac{F(x_3)-F(x_2)}{x_3-x_2}$.


$f(x)$
$f(c) = \frac{F(x_3)-F(x_2)}{x_3-x_2}$
M_3
m_3
x_2
$c\ x_3$
$F(x)$
$F(x_3)$
$F(x_2)$
$F'(c) = \frac{F(x_3)-F(x_2)}{x_3-x_2}$
x_2
$c\ x_3$

Since m_3 is the infimum of f on $[x_2, x_3]$ we have $f(c) \geq m_3$. Similarly, $f(c) \leq M_3$. Therefore

$$m_3 \leq f(c) \leq M_3$$

Substituting in the value for $f(c)$ we have

$$m_3 \leq \frac{F(x_3) - F(x_2)}{x_3 - x_2} \leq M_3$$

Rearranging, we find

$$m_3(x_3 - x_2) \leq F(x_3) - F(x_2) \leq M_3(x_3 - x_2)$$

This is an interesting expression. When m_3 is positive, we can interpret the term on the left as the area of a rectangle. Similarly we can interpret the term on the right as the area of a rectangle. The inequality relates the areas of the two rectangles to the value of $F(x_3) - F(x_2)$, where F is anti-derivative of f.

This is quite similar to what we see in the statement of the Fundamental Theorem of Calculus Part I. An integral is related the value of anti-derivative at the end points of the an interval. With this in mind, we proceed with our proof of Part I of the Fundamental Theorem of Calculus.

Proof. Let $P = \{x_0, x_1, \ldots, x_n\}$ be a partition of $[a, b]$. Since F is differentiable on $[a, b]$, we can apply the Mean Value Theorem to each interval of the form $[x_{i-1}, x_i]$. Doing so, implies that for every $i \in \{1, 2, \ldots, n\}$ there exists $c_i \in (x_{i-1}, x_i)$ so that

$$F'(c_i) = \frac{F(x_i) - F(x_{i-1})}{x_i - x_{i-1}}$$

Since $F'(x) = f(x)$ for all $x \in [a, b]$ we have

$$f(c_i) = \frac{F(x_i) - F(x_{i-1})}{x_i - x_{i-1}}$$

Therefore

$$f(c_i)(x_i - x_{i-1}) = F(x_i) - F(x_{i-1})$$

Recall

$$L(f, P) = \sum_{i=1}^{n} m_k(x_i - x_{i-1})$$
$$U(f, P) = \sum_{i=1}^{n} M_k(x_i - x_{i-1})$$

Since m_i is the infimum of f on $[x_{i-1}, x_i]$ we have $f(c_i) \geq m_i$. Similarly, $f(c_i) \leq M_i$. Therefore

$$m_i \leq f(c_i) \leq M_i$$

Therefore

$$m_i(x_i - x_{i-1}) \leq f(c_i)(x_i - x_{i-1}) \leq M_i(x_i - x_{i-1})$$

From our work above, we can write this inequality as:

$$m_i(x_i - x_{i-1}) \leq F(x_i) - F(x_{i-1}) \leq M_i(x_i - x_{i-1})$$

Summing over all parts of the partition yields

$$\sum_{i=1}^{n} m_i(x_i - x_{i-1}) \leq \sum_{i=1}^{n} F(x_i) - F(x_{i-1}) \leq \sum_{i=1}^{n} M_i(x_i - x_{i-1})$$

Notice the sum on the left is exactly $L(f, P)$ and on the right is $U(f, P)$.

$$L(f,P) \leq \sum_{i=1}^{n}(F(x_i) - F(x_{i-1})) \leq U(f,P)$$

Consider now the centre sum:

$$F(x_1) - F(x_0) + F(x_2) - F(x_1) + F(x_3) - F(x_2) + F(x_4) - F(x_3) + \cdots + F(x_n) - F(x_{n-1})$$

Cancelling yields

$$\cancel{F(x_1)} - F(x_0) + \cancel{F(x_2)} - \cancel{F(x_1)} + \cancel{F(x_3)} - \cancel{F(x_2)} + \cancel{F(x_4)} - \cancel{F(x_3)} + \cdots + F(x_n) - \cancel{F(x_{n-1})}$$

Therefore

$$\sum_{i=1}^{n}(F(x_i) - F(x_{i-1})) = F(x_n) - F(x_0) = F(b) - F(a)$$

Therefore for every $P \in \tilde{P}$ we have

$$L(f,P) \leq F(b) - F(a) \leq U(f,P)$$

Since this inequality holds for every $P \in \tilde{P}$, $F(b) - F(a)$ is an upper bound for the set $\{L(f,P) \mid P \in \tilde{P}\}$. Since $L(f)$ is the supremum of this set we have

$$L(f) \leq F(b) - F(a)$$

Similarly

$$F(b) - F(a) \leq U(f)$$

In other words,

$$L(f) \leq F(b) - F(a) \leq U(f)$$

Since f is integrable on $[a, b]$, we have $L(f) = U(f)$. Therefore

$$L(f) = F(b) - F(a) = U(f)$$

And so by definition we have

$$\int_a^b f(x)\,dx = L(f) = U(f) = F(b) - F(a)$$

□

Looking at this proof, Part I of the Fundamental Theorem of Calculus seems like an incredible coincidence. The conclusion that we can compute an area by reversing the process of differentiation seems to fall out of seeing what happens when we apply the Mean Value Theorem on each interval. There are lots of examples of proofs we have seen in the course where the proof gives us some intuition on why a certain mathematical fact ought to be true. This isn't one of them.

The second part of the Fundamental Theorem of Calculus makes explicit the connection between differentiation and integration. Let us proceed to discuss the proof of this theorem.

6.3.2 The Fundamental Theorem of Calculus Part II

Theorem 6.20 (The Fundamental Theorem of Calculus Part II)**.** *Let $g : [a, b] \to \mathbb{R}$ be integrable and let $G : [a, b] \to \mathbb{R}$ so that*

$$G(x) = \int_a^x g(u)\, du$$

The following statements are true:

i The function G is continuous on $[a, b]$.

ii If g is continuous at $c \in [a, b]$, then G is differentiable at c and

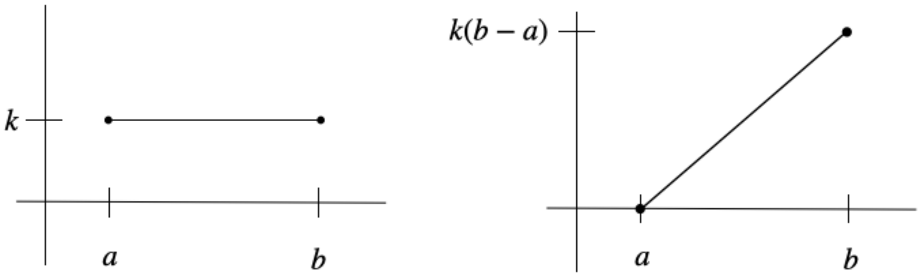
$$G'(c) = \frac{d}{dx}\left(\int_a^c g(u)\, du\right) = g(c)$$

In the interest of time, we will ignore *i* and focus only on *ii*.

To build some intuition, we return to our geometric interpretation of the integral as an area. To simplify matters, let us take $g(x)$ to be a constant function: $g(x) = k$

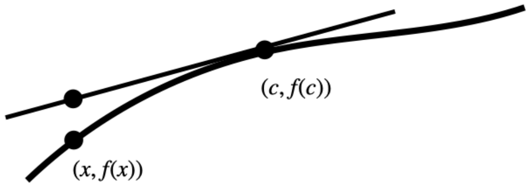
The function $G(x)$ is a measurement of the area under the curve of $g(x)$ in the interval $[a, x]$. As x increases, the area increases and so we expect $G(x)$ to be increasing. At $x = a$, there is no area under the curve, and so we expect $G(a) = 0$. At $x = b$ the area under the curve is given by $(b - a)k$. Therefore $G(b) = (b - a)k$. Finally, since g is constant, the rate at which the area is increasing is the same everywhere. And so G is linear.

In the picture below, g is on the left and G is on the right.



The slope of a tangent line of G at any point $x \in [a, b]$ is k. And so, $G'(x) = k = g(x)$, as expected. Therefore *ii* is true when G is linear.

Before we build up some intuition for non-constant functions let us take a moment to look more closely at a nice application of the derivative: estimation. Consider the curve drawn below and the tangent drawn at a point.



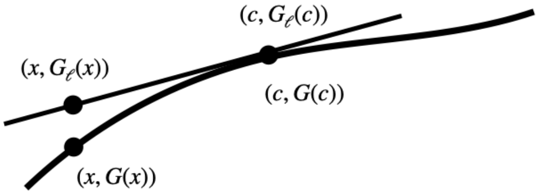
For values of x that are *close* to c, the line tangent to f at c is a pretty good approximation for $f(x)$.

We could formalize this observation by defining carefully the meaning of *approximation*, using ϵ's and δ's, but that won't get us any closer to understanding, and so we'll resist the urge. Instead, we will just remember:

When x is *close* to c, the line with slope $f'(c)$ through $(c, f(c))$ is a good approximation for $f(x)$

But what does this have to do with our theorem?

Consider G as drawn below



Let G_ℓ be the line tangent to the curve at c. By definition $G'(c)$ is the slope of the tangent line at c. From our observation above, this tangent line, G_ℓ, is a good approximation for $G(x)$ for points x that are close to c. Since G_ℓ is linear, we have $G'_\ell(x) = g_\ell(x)$. Since G_ℓ and G behave the same near c, we conclude $G'(c) = g(c)$.

Our work above is meant to give some insight in to why *ii* ought to be true, but it does not constitute a proof of this fact. We proceed to sketch a proof of *ii*. Unfortunately our proof sketch below is not based on the visual ideas above. Instead it works directly with the limit definition of differentiation and the partition definition of integration.

What follows is not a proof the second part of our theorem. Throughout we lean on some results we remember from previous studies and we make some leaps of logic that seem *reasonable*.

Given $G(x) = \int_a^x g(u)\, du$ we want to show $G'(c) = g(c)$ for all $c \in [a, b]$. To do this, we show that for every $\epsilon > 0$ we have

$$g(c) - \epsilon < G'(c) < g(c) + \epsilon$$

If this statement is true, then we can find no value between $G'(c)$ and $g(c)$. From this we conclude $G'(c) = g(c)$.

To consider $G'(c)$ we appeal back to our definition of differentiation:

$$G'(c) = \lim_{x \to c} \frac{G(x) - G(c)}{x - c}$$

Substituting in $G(x) = \int_a^x g(u)\,du$, the right side of this equation becomes

$$G'(c) = \lim_{x \to c} \frac{1}{x - c} \left(\int_a^x g(u)\,du - \int_a^c g(u)\,du \right)$$

Assuming that we can negate and sum integrals the way we expect[9], we have

$$\int_a^x g(u)\,du - \int_a^c g(u)\,du = \int_a^x g(u)\,du + \int_c^a g(u)\,du = \int_c^x g(u)\,du$$

Therefore

$$G'(c) = \lim_{x \to c} \frac{1}{x - c} \int_c^x g(u)\,du$$

By hypothesis, g is continuous on $[a, b]$. In particular, g is continuous at c.

Therefore for every $\epsilon > 0$ there exists $\delta > 0$ so that $|g(x) - g(c)| < \epsilon$ whenever $|x - c| < \delta$.

Consider some fixed value of $\epsilon > 0$. As x approaches c, eventually $|x - c| < \delta$ will be satisfied. And so when u is between x and c we have

$$|g(u) - g(c)| < \epsilon$$

Therefore as x approaches c we have

$$g(c) - \epsilon < g(u) < g(c) + \epsilon$$

for all u between x and c

Therefore as x approaches c we have

$$\int_c^x g(c) - \epsilon\,du < \int_c^x g(u)\,du < \int_c^x g(c) + \epsilon\,du$$

Since c is fixed, $g(c)$ is a constant. Since integration is linear we have

$$\begin{aligned}\int_c^x g(c) - \epsilon\,du &= \int_c^x g(c)\,du - \int_c^x \epsilon\,du \\ &= g(c)\int_c^x 1\,du - \epsilon \int_c^x 1\,du \\ &= g(c)(x - c) - \epsilon(x - c) \\ &= (x - c)(g(c) - \epsilon)\end{aligned}$$

[9] This is the first place this proof starts getting sketchy. We have not proven that our definition of integration based on partitions lets us negate integrals the way we expect.

Therefore

$$(x-c)(g(c)-\epsilon) < \int_c^x g(u)\ du$$

Similarly

$$\int_c^x g(u)\ du < (x-c)(g(c)+\epsilon)$$

With these two inequalities we can study the behaviour of

$$G'(c) = \lim_{x\to c} \frac{1}{x-c}\int_c^x g(u)\ du$$

Since

$$(x-c)(g(c)-\epsilon) < \int_c^x g(u)\ du < (x-c)(g(c)+\epsilon)$$

dividing each term by $(x-c)$ gives

$$\frac{1}{x-c}(x-c)(g(c)-\epsilon) < \frac{1}{x-c}\int_c^x g(u)\ du < \frac{1}{x-c}(x-c)(g(c)+\epsilon)$$

Therefore

$$g(c)-\epsilon < \frac{1}{x-c}\int_c^x g(u)\ du < g(c)+\epsilon$$

And so

$$\lim_{x\to c}(g(c)-\epsilon) < \lim_{x\to c}\frac{1}{x-c}\int_c^x g(u)\ du < \lim_{x\to c}(g(c)+\epsilon)$$

Notice

$$\lim_{x\to c}(g(c)-\epsilon) = g(c)-\epsilon$$

and

$$\lim_{x\to c}(g(c)+\epsilon) = g(c)+\epsilon$$

Therefore

$$g(c)-\epsilon < \lim_{x\to c}\frac{1}{x-c}\int_c^x g(u)\ du < g(c)+\epsilon$$

From our work above, we have

$$G'(c) = \lim_{x\to c}\frac{1}{x-c}\int_c^x g(u)\ du$$

Therefore

$$g(c)-\epsilon < G'(c) < g(c)+\epsilon$$

This statement is true no matter which value of ϵ we choose. And so we conclude

$$G'(c) = g(c)$$

There is lots of sketchy mathematics happening here. For example, at one point we said

Assuming we can negate and sum integrals the way we expect.

Further, our use of the phrase

As x approaches c...

is highly suspect. This phrase has no definition beyond our intuition for limits. Making our work above into a full proof would require some care in places where we have used this phase. What's more, there is some very sketchy algebra above. We have multiplied in inequalities without thinking carefully about whether or not our values are positive or negative. Turning this proof sketch in to a proof would require us to consider left side and right side limits to deal.canv

All of this said, however, this approach does work. There are lots of little details and technicalities to overcome. But the outline we have given above is the hardest part.

6.3.3 Improper Integrals

The second part of the Fundamental Theorem of Calculus is a statement about the behaviour of a function whose value depends on the upper limit of integration of a definite integral:

$$G(x) = \int_a^x g(u)\, du$$

If the domain of g is unbounded above (i.e., if it has a limit point at ∞), then the following statement is meaningful

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

Interpreting an integral as an area under a curve, considering the limit of G as x goes to infinity amounts to computing what seems to be an infinite area. We proceed with an example.

For every $b > 1$, the function u^2 is integrable on $[0, b]$. Define $G : (0, \infty) \to \mathbb{R}$ so that $G(x)$ is the integral of u^2 on the domain $[0, x]$. That is, let

$$G(x) = \int_0^x u^2\, du$$

We can use the first part of the Fundamental Theorem of Calculus to compute this integral directly:

$$G(x) = \int_0^x u^2\, du = \frac{1}{3}x^3$$

PICTURE

Using techniques from Calculus 2, we can verify

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

which seems to be a reasonable result.

Consider now the function

$$H(x) = \int_0^x \frac{1}{1+u^2} du = \arctan(x)$$

Here we have

$$\lim_{x \to \infty} \arctan(x) = \frac{\pi}{2},$$

a considerably less reasonable result that the one we got above for G.

PICTURE.

Definition 6.21. *Let $g : [a, \infty)$ be integrable on $[a, b]$ for all $b > a$. We define the following notation*

$$\int_a^\infty g(u)\, du = \lim_{x \to \infty} \int_a^x g(u)\, du$$

We call $\int_a^\infty g(u)\, du$ an improper integral on an infinite interval of integration

Using this notation we can denote our two examples above as:

$$\int_0^\infty u^2\, du = \infty \qquad \int_0^\infty \frac{1}{1+u^2}\, du = \frac{\pi}{2}$$

Of these two examples, it is the latter one that is most interesting. In each case the length of the curve is infinite. However in the latter case, even though the length of the curve is infinite, it bounds finite area.

In our work in limits we introduced limits to infinity after first taking a careful look at the limit of functions as it approaches a point. Using the same line of reasoning as above, we can consider integration when one of the bounds of integration approaches a point.

Consider the function $g(u) = \frac{1}{\sqrt{u}}$ on the interval $(0, 1]$. Since $g(u)$ is not bounded on $(0, 1]$, this function is not integrable on the interval $(0, 1]$. However, notice that $\frac{1}{\sqrt{u}}$ is integrable on the interval $[c, 1]$ for all $0 < c < 1$. Define $G : (0, 1) \to \mathbb{R}$ so that

$$G(x) = \int_x^1 \frac{1}{\sqrt{u}}\, du$$

PICTURE

Though 0 is not the domain of g, it is a limit point of the domain of g. And so we can consider the behaviour of G as x approaches 0 from the right

$$\lim_{x \to 0^+} G(x) = \lim_{x \to 0^+} =??$$

Just as we saw above, interpreting this integral as an area under a curve leads to a surprising result.

PICTURE.

Despite the fact that $\frac{1}{\sqrt{u}}$ is unbounded on the interval $(0, 1]$, somehow the area is encloses is finite. We can apply this same technique to study functions of other unbounded integrands.

Definition 6.22. *Let $g : (a, b] \to \mathbb{R}$ be integrable on $[c, b]$ for all $c \in (a, b)$ We define the following notation*

$$\int_a^b g(u)\, du = \lim_{x \to a^+} \int_x^b g(u)\, du$$

We call $\displaystyle\int_a^b g(u)\, du$ an <u>*improper integral of an unbounded integrand*</u>

EXAMPLE.

Though our definition above only applies to a function ... we can extend this notation to study integrands that are discontinuous on the bounds of integration. For example, consider

$$\int_{-1}^1 \frac{1}{x^2}\, dx$$

The function $\frac{1}{x^2}$ is not defined over the entire domain $[-1, 1]$. If we tried to apply the definition of Riemann integration, we would run in to trouble finding supremums on all subintervals. As x approaches 0 (from either direction) the function is unbounded above.

However, by interpreting this notation analogously to our definition of
<u>improper integral of an unbounded integrand</u>, we have:

$$\begin{aligned}\int_{-1}^1 \frac{1}{u^2}\, du &= \int_{-1}^0 \frac{1}{u^2}\, du + \int_0^1 \frac{1}{u^2}\, du \\ &= \lim_{x \to 0^-} \int_{-1}^x \frac{1}{u^2}\, du + \lim_{x \to 0^+} \int_x^1 \frac{1}{u^2}\, du \\ &= ANSWER\end{aligned}$$

Test Your Understanding

1. For each integral below determine, without proof, whether the improper integral on an infinite integral of integration converges or diverges. Feel welcome to use a computational algebra system rather than relying on your, perhaps rusty, integration skills.

   (a) $\displaystyle\int_0^\infty u^2\, du$

   (b) $\displaystyle\int_1^\infty \frac{1}{u^2}\, du$

   (c) $\displaystyle\int_1^\infty \frac{\log u}{u^2}\, du$

2. For each integral below, determine, without proof, whether the improper integral of each unbounded integrand converges. Feel welcome to use a computational algebra system rather than relying on your, perhaps rusty, integration skills

   (a) $\displaystyle\int_0^1 \frac{1}{u^2}\,du$

   (b) $\displaystyle\int_0^1 \frac{1}{\sqrt{1-u^2}}\,du$

   (c) $\displaystyle\int_{\frac{1}{3}}^1 \frac{1}{3u-1}\,du$

Test Your Understanding - Answers

1. (a) $\lim_{x\to\infty} \displaystyle\int_0^x u^2\, du = \lim_{x\to\infty} \frac{x^3}{3} = \infty$

   (b) $\lim_{x\to\infty} \int_1^x \frac{1}{u^2}\, du = \lim_{x\to\infty} -\frac{1}{x} + 1 = 1$

   (c) $\lim_{x\to\infty} \int_1^\infty \frac{\log u}{u^2}\, du = \lim_{x\to\infty} -\frac{\log(x)+1}{x} + \frac{\log(1)+1}{1} = 1$

2. (a) Converges to $\frac{\pi}{2}$

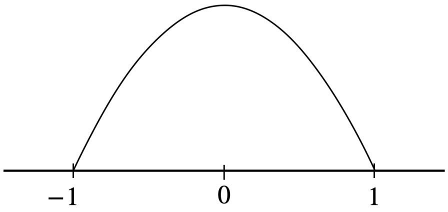
   (b) Does not converge

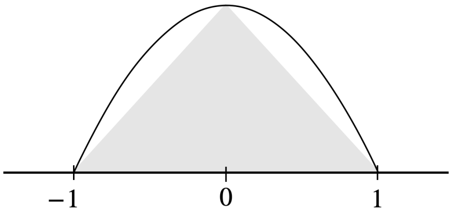
   (c) Does not converge

7 Series

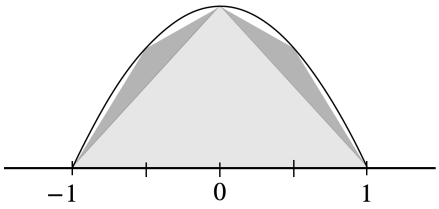
In the second half of Section 6 we looked at integration, a topic we were well familiar with before the start of our work this semester. And, perhaps, one that seems slightly less familiar as a result of our work in this subject.

Our familiarity with integration comes from our previous studies, which teach integration because it is just so very useful in many scientific and industrial contexts. However, integration is not the only means we have to compute the area under a curve. Consider $f(x) = 1 - x^2$ on the interval $[-1, 1]$. Rather than computing an antiderivative, instead let us estimate the area with a triangle.





The area of the triangle under estimates the area under the curve This triangle has area 1 and so the area under the curve is more than 1. To fill some of the remaining area, we again use some triangles.



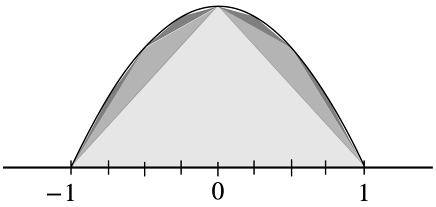
It takes geometric finesse to prove this, but these two new triangles each have area $\frac{1}{8}$ and so the area under the curve is bounded below by

$$1 + \frac{1}{4}.$$

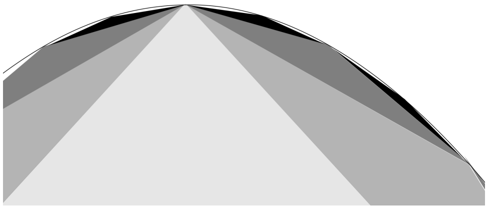
We can continue with this process, adding new triangles at every step to further fill the area. Adding a third iteration of triangles, we can compute[1] the total area added is $\frac{1}{4^2}$. And so at this point the total shaded area is given by

$$1 + \frac{1}{4} + \frac{1}{4^2}$$

[1] Again, this takes some geometric finesse we are not concerned with here. Do an internet search for `Archmedean Method` for more information



We can continue with this pattern, adding an area of $\frac{1}{4^k}$ in the kth-iteration[2].



After n applications of this process, the shaded area is given by

$$A_n = 1 + \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \cdots + \frac{1}{4^n}$$

The value A_n gives the area of a polygon that under estimates the area under the curve. Let us consider the behaviour of A_n as n approaches infinity.

Since A_n is a geometric series[3] with $a = 1$ and ratio $r = 4^{-1}$ we have

$$A_n = 1 + \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \cdots + \frac{1}{4^n} = \left(\frac{1 - 4^{-(n+1)}}{3/4}\right)$$

Notice $4^{-(n+1)} \to 0$.

And so

$$\lim_{n\to\infty} A_n = \frac{4}{3}$$

Coincidentally,

$$\int_{-1}^{1} 1 - x^2 \, dx = \frac{4}{3}$$

Somehow, our knowledge of sequences has helped us to compute the area under a curve. Though this time our sequence

$$(A_0, A_1, A_2, A_3, \dots)$$

was a sequence of sums:

[2] Again, it is not obvious why this should be the case. But we'll just go with it.

[3] $1 + r + r^2 + \cdots + r^n = \left(\frac{1-r^{n+1}}{1-r}\right)$

$$\left(1,\ 1+\frac{1}{4},\ 1+\frac{1}{4}+\frac{1}{4^2},\ 1+\frac{1}{4}+\frac{1}{4^2}+\frac{1}{4^3},\ \dots\right)$$

Let $a_n = \frac{1}{4^n}$. We have

$$A_n = a_0 + a_1 + a_2 + \cdots + a_n$$

With this work in mind, perhaps it is reasonable to write:

$$\lim_{n\to\infty} \sum_{k=0}^{n} \frac{1}{4^k} = \frac{4}{3}$$

Or even

$$1 + \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \cdots = \frac{4}{3}$$

This last statement is troubling: How can we add together infinitely many things and how is it possible for the result to be finite?[4] Hold on to this thought for the moment, as we set the stage for an even more interesting application of *infinite sums*.

Where ever you happen to be sitting right now, undoubtedly there is a calculator near by. Compute the following sum:

$$1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \frac{1}{5!} + \frac{1}{6!}$$

It is impossible for this number to be equal to e, the number we have computed is rational and e is irrational. However, by adding more terms of the same form, this sum gets closer and closer e.

$$\begin{aligned}
1 &= 1 \\
2 &= 1 + \frac{1}{1!} \\
2.5 &= 1 + \frac{1}{1!} + \frac{1}{2!} \\
2.\overline{6} &= 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} \\
2.708\overline{3} &= 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} \\
2.71\overline{6} &= 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \frac{1}{5!}
\end{aligned}$$

Let $a_0 = 1$ and $a_n = \frac{1}{n!}$. Perhaps then it is reasonable for us to write:

$$e = \sum_{n=0}^{\infty} a_n = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \dots$$

[4] And, if we try to common factor, how can we possibly get a common factor result of 3 in the denominator?

Okay, but why is this interesting? Other than being a neat party trick, why would we ever need to estimate e with a sum? Well, whatever device you used to compute the sum above is also capable of computing the following value:

$$e^{1.3}$$

Except, the device you are using does not have infinite memory. It cannot fully store all of the digits of e, and so how can it possibly compute $e^{1.3}$? It doesn't; it estimates $e^{1.3}$ using a finite sum similar to the one above.

In fact, your calculator does this every time it does a computation with e^x, $\sin x$, $\cos x$, $\log x$, etc... Your calculator is very good at working with (finite) polynomials and so estimates these functions using finite polynomials.

For example[5]

$$\sin x \approx x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!}$$

Using this, we can compute:

$$\sin \frac{\pi}{11} \approx \pi/11 - \frac{(\pi/11)^3}{3!} + \frac{(\pi/11)^5}{5!} - \frac{(\pi/11)^7}{7!} = 0.2817\ldots$$

Our goal in Section 7 of the course is to understand how all of this works. We will make sense of *infinite sums* and derive a method to use them to estimate non-polynomial functions.

[5] We'll talk about where this approximation comes from in the subsequent section

A Caveat About Sequence Indexing During our work in Section 4 on sequences we indexed sequences starting with $n = 1$.

$$a_1, a_2, a_3, \dots$$

At times in this section it will be convenient for us to index our sequences starting at 0

$$a_0, a_1, a_2, \dots$$

7.1 Series Convergence

At the start of introductory reading we considered the sequence given by $a_n = \frac{1}{4^n}$ and a corresponding sequence of *partial sums*:

$$(A_n) = \left(1,\ 1 + \frac{1}{4},\ 1 + \frac{1}{4} + \frac{1}{4^2},\ 1 + \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3}, \dots\right)$$

In this sequence, each term A_n is of the form:

$$A_n = a_0 + a_1 + a_2 + \dots + a_n = \sum_{k=0}^{n} a_k$$

Since (A_n) is a sequence, we can consider the behaviour of (A_n) as n goes to infinity. From our work in the introduction we found

$$\lim_{n\to\infty} A_n = \frac{4}{3}$$

The sequence $a_n = \frac{1}{4^n}$ is one where the corresponding sequence of *partial sums* converges. However, this need not be the case.

For example, consider $b_n = n$. Let (B_n) be the corresponding sequence of *partial sums*:

$$B_n = b_0 + b_1 + b_2 + \dots + b_n$$

For example, $B_3 = b_0 + b_1 + b_2 + b_3 = 0 + 1 + 2 + 3 = 6$.

We can see $B_n \geq n$ for all $n \in \mathbb{N}$. Since $B_{n+1} \geq B_n$ for all $n \in \mathbb{N}$, we notice (B_n) is monotone increasing. Recall the following theorem from Section 4:

Theorem (Theorem 4.24)**.** *Let (f_n) be a monotone sequence. The sequence (f_n) converges if and only if (f_n) is bounded.*

Since (B_n) is not bounded, we conclude (B_n) diverges as n goes to infinity. What's more, after a little more work we can confirm

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

To make our discussion on sequences of sums easier, we introduce the following terminology.

Definition 7.1. *Let (a_n) be a sequence and let (A_n) be a sequence so that $A_n = \sum_{k=0}^{n} a_k$. We say (A_n) is a* series. *We refer to the terms of (A_n) as* partial sums.

As a notational convenience, we use sigma notation to refer to a series:

$$(A_n) = \sum_{n=0}^{\infty} a_n$$

For example, when $a_n = 1$ for all $n \in \mathbb{N}$, the corresponding series can be denoted as

$$\sum_{n=0}^{\infty} 1 = 1 + 1 + 1 + \dots$$

This notation, though convenient, comes with some caveats.

Series are weird. They look and feel like sums as we are used to. And so we expect that all of the algebraic *stuff* we can do with addition should also work for series. Unfortunately this isn't the case. For example, the following algebra is not valid:

$$\begin{aligned} 0 &= 0 + 0 + 0 + 0 + \dots \\ &= (1-1) + (1-1) + (1-1) + (1-1) + \dots \\ &= 1 + (-1+1) + (-1+1) + (-1+1) + \dots \\ &= 1 + 0 + 0 + 0 + \dots \\ &= 1 \end{aligned}$$

If you find yourself doing algebraic manipulation within a series, turn back and try another route; the algebra you are doing may not be valid.

Aside. *There are cases for which our usual algebraic techniques are valid for series. We won't discuss them much here. Though they may come up in an aside or two over the rest of this section.*

The notation $\sum_{n=0}^{\infty} a_n$ refers to a sequence. Each term in this sequence is a partial sum. Since $\sum_{n=0}^{\infty} a_n$ is a sequence we may consider the behaviour of this sequence as n approaches infinity. We do so using our terminology and notation from Section 4.

Definition 7.2. *Let (a_n) be a sequence. We say the series* (A_n) converges to L *when there exists $L \in \mathbb{R}$ so that $\lim_{n\to\infty} A_n = L$. When $\sum_{n=0}^{\infty} a_n$ converges to L we write*

$$\sum_{n=0}^{\infty} a_n = L$$

Definition 7.3. *Let (a_n) be a sequence. We say the series $\sum_{n=0}^{\infty} a_n$ <u>diverges</u> when (A_n) diverges as n approaches infinity. When (A_n) diverges to ∞ we write $\sum_{n=0}^{\infty} a_n = \infty$.*

From our examples above

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

$$\sum_{n=0}^{\infty} n = 0 + 1 + 2 + 3 + \cdots = \infty$$

The notation $\sum_{n=0}^{\infty} a_n = L$ is a statement about a convergent sequence. Stating $\sum_{n=0}^{\infty} a_n = L$ means $A_n \to L$. And so we have an Algebra of Series Theorem that is analogous to our Algebra of Limits theorem for sequences.

Theorem 7.4 (Algebra of Series Theorem)**.** *If $\sum_{n=0}^{\infty} a_n = \alpha$ and $\sum_{n=0}^{\infty} b_n = \beta$, then*

1. $\sum_{n=0}^{\infty} a_n + b_n = \alpha + \beta$*; and*
2. $\sum_{n=0}^{\infty} \lambda a_n = \lambda\alpha$ *for all* $\lambda \in \mathbb{R}$.

Just as we did in Section 4 of the subject, our goal in Section 7 is to classify those series that are convergent. To do so we will lean heavily on material we have seen throughout the subject.

To begin with, let us see what we can learn about convergent series. Let $(A_n) = \sum_{n=1}^{\infty} a_n$ be a convergent series. In other words there exists $L \in \mathbb{R}$ so that

$$\lim_{n\to\infty} A_n = L$$

Consider now the sequence (a_n). Let us consider what possible behaviour (a_n) can exhibit. In particular, let us consider the behaviour of a_n as n approaches infinity.

Consider, for example, a sequence so that $a_n \to 1$. Since $A_n - A_{n-1} = a_n$ we have

$$\lim_{n\to\infty} A_n - A_{n-1} = a_n = 1$$

In other words, as n gets large, the difference between successive terms approaches 1. And so as n gets large, we have[6]

$$A_n \approx A_{n-1} + 1$$

[6] This is not a proof. We do not have a precise defined meaning for the notation $\approx$.

This means that our sum increases by approximately 1 when we add the next term. And so it seems quite unlikely that the sequence of partial sums is bounded above. If the sequence of partial sums is not bounded above, then the sequence of partial sums must diverge. (see Theorem 4.11)

In this argument, there is nothing special about $a_n \to 1$. Similar reasoning would apply for any $a_n \to T$ with $T \neq 0$.

Theorem 7.5. *If $a_n \not\to 0$, then the series $\sum_{n=0}^{\infty} a_n$ diverges.*

Rather than prove this theorem, we instead prove the contrapositive.

Theorem 7.6. *If $\sum_{n=0}^{\infty} a_n$ converges, then $a_n \to 0$.*

To prove this theorem, we recall the characterisation of convergent sequences as Cauchy sequences.

Proof. Let $(A_n) = \sum_{n=0}^{\infty} a_n$ be a convergent series.

If (A_n) converges, then, by the Cauchy Convergence Criterion, (A_n) is Cauchy. Therefore for every $\epsilon > 0$ there exists $M \in \mathbb{N}$ so that $|A_j - A_k| < \epsilon$ whenever $j, k > M$.

To show $a_n \to 0$ we must show that for every $\epsilon' > 0$ there exists $M' \in \mathbb{N}$ so that $|a_n - 0| < \epsilon'$ whenever $n > M'$

Let $\epsilon' > 0$. And let $\epsilon = \epsilon'$.

Since (A_n) is Cauchy, there exists $M \in \mathbb{N}$ so that $|A_j - A_k| < \epsilon$ whenever $j, k > M$. Consider the statement we get when $j = k + 1$

$$|A_{k+1} - A_k| < \epsilon \text{ whenever } k > M.$$

Notice $A_{k+1} - A_k = a_{k+1}$. Therefore

$$|a_{k+1}| < \epsilon \text{ whenever } k > M.$$

Manipulating the indices, we may write

$$|a_k| < \epsilon \text{ whenever } k - 1 > M.$$

Let $M' = M + 1$. Recall $\epsilon = \epsilon'$. Therefore $|a_n - 0| < \epsilon'$ whenever $n > M'$. In other words, $a_n \to 0$.

□

Theorem 7.6 does not help us determine if a series converges or diverges; the hypothesis of the theorem requires us to already be considering a convergent series. However, Theorem 7.5 can help us confirm that a series diverges, as in the hypothesis of the theorem we have an arbitrary series.

For example the series

$$\sum_{k=0}^{\infty} \frac{k^3}{k^2+4}$$

must diverge as

$$\lim_{n\to\infty} \frac{n^3}{n^2+4} = \infty$$

Similarly, by Theorem 7.5, the series

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

must diverge, as

$$\lim_{n\to\infty} \frac{2n^2}{n^2+1} = 2$$

As a last example, consider the sequence given by $h_n = \frac{1}{n}$. We have

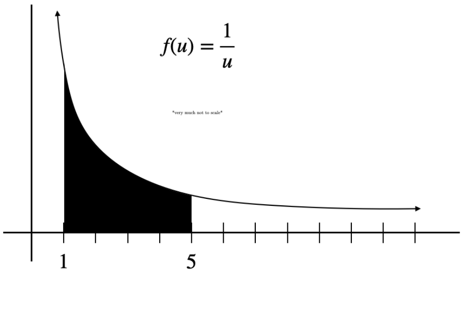
$$\lim_{n\to\infty} h_n = 0$$

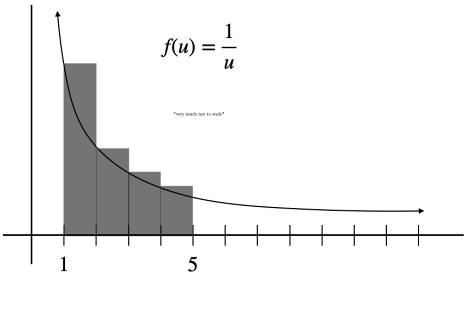
Neither the hypotheses of Theorem 7.6 nor Theorem 7.5 apply to (H_n). These theorems tell us nothing about the behaviour of

$$\sum_{n=1}^{\infty} h_n = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots$$

But, come on. This must converge, right? The terms of this sum get very small very quickly. It would be a completely unintuitive result if this infinite sum diverged.

Consider the picture below of $f(u) = \frac{1}{u}$ on the interval $[1, \infty)$ and the corresponding upper Riemann sum for the partition $P = \{1, 2, 3, 4, 5\}$.



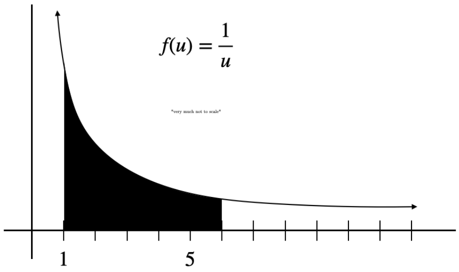


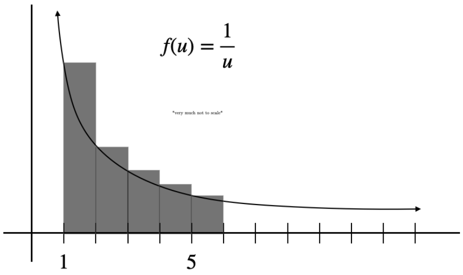
From our work in the last section we have

$$\int_1^5 \frac{1}{u}\,du \leq U(f,P) = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = H_4$$

Changing our partition to $\{1,2,3,4,5,6\}$ gives the corresponding inequality:

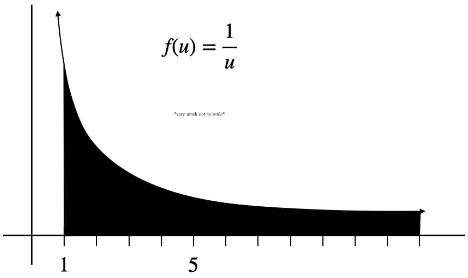
$$\int_1^6 \frac{1}{u}\,du \leq U(f,P) = \sum_{k=1}^{5} \frac{1}{k} = H_5$$

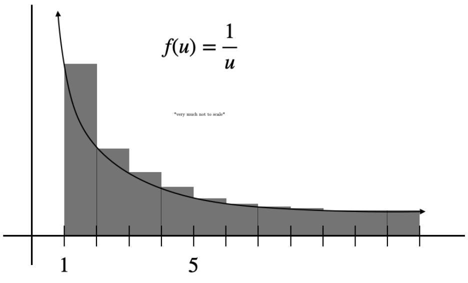




In general, we have the following inequality:

$$\int_1^{n+1} \frac{1}{u}\,du \leq \sum_{k=1}^{n} \frac{1}{k} = H_n$$





Computing the definite integral we have

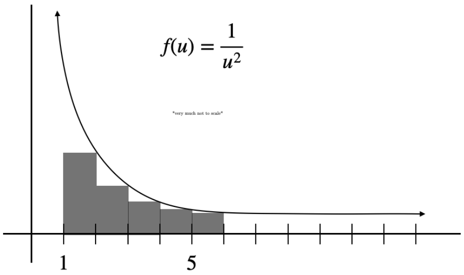
$$\log(n+1) - \log(1) \leq H_n$$

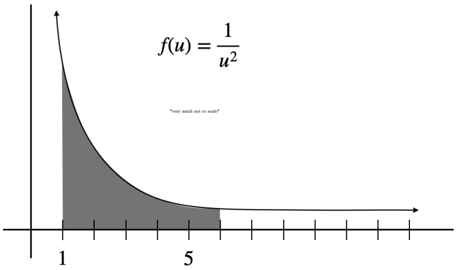
for all $n > 1$. Therefore the set

$$\{H_n \mid n \in \mathbb{N}^+\}$$

is not bounded above. And so by Theorem 4.13, the series $(H_n) = \displaystyle\sum_{n=0}^{\infty} \frac{1}{k}$ diverges.

Consider now undertaking the same process as above, replacing $h_n = \frac{1}{n}$ with $g_n = \frac{1}{n^2}$ and the upper Riemann Sum with the lower Riemann sum.





The corresponding lower Riemann sum for the partition $P = \{1, 2, 3, 4, 5, 6\}$ is

$$\frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} = G_6 - 1$$

As above, we conclude

$$G_n - 1 \leq \int_1^n \frac{1}{u^2}\, du$$

Taking the definite integral on the right and re-arranging results in the following expression

$$G_n \leq 2 - \frac{1}{n} < 2$$

Since G_n is monotone increasing and bounded, by Theorem 4.24 we conclude

$$1 + \frac{1}{2^2} + \frac{1}{3^2} + \dots$$

converges.

This technique works anytime we can compare our sequence with an appropriate integral.

Theorem 7.7 (Integral Test)**.** *Let f be integrable on $[1, \infty)$ so that f is positive and decreasing. The series $\sum_{n=1}^{\infty} f(n)$ converges if and only if there exists $L \in \mathbb{R}$ so that*

$$\lim_{x \to \infty} \int_1^x f(u)\, du = L$$

Using the integral test, we see $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges since $\lim_{x \to \infty} \int_1^x \frac{1}{u}\, du = \infty$.

Following the line of reasoning above comparing a series with upper/lower Riemann sums, we could study the convergence of series of the form

$$1 + \frac{1}{2^p} + \frac{1}{3^p} + \frac{1}{4^p} + \dots$$

Rather than do this, however, we can directly apply the integral test. The function

$$f(u) = \frac{1}{u^p}$$

is integrable on $[1, \infty]$ for all $p > 0$

If $p \neq 1$, then

$$\int u^{-p}\, du = -\frac{1}{p-1} u^{-p+1} + C$$

Therefore

$$\begin{aligned} \int_1^x u^{-p}\, du &= -\frac{1}{p-1} x^{-p+1} + \frac{1}{p-1} 1^{-p+1} \\ &= \frac{1}{p-1}\left(1 - \frac{1}{x^{p-1}}\right) \end{aligned}$$

When $0 < p < 1$, $\lim_{x\to\infty} \int_1^x u^{-p}\, du$ does not converge.

However, when $p > 1$, we have

$$\lim_{x\to\infty} \int_1^x u^{-p}\, du = \frac{1}{p-1}$$

In the former case, the Integral Test implies $\sum_{n=1}^{\infty} \frac{1}{n^p}$ diverges for $0 < p < 1$. Whereas in the latter case the Integral Test implies $\sum_{n=1}^{\infty} \frac{1}{n^p}$ converges for $p > 1$.

We refer to series of the form $\sum_{n=1}^{\infty} \frac{1}{n^p}$ as p-series and we have the following result.

Corollary 7.8 (p-series Test)**.** *The series*

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

- *converges when* $p > 1$*; and*
- *diverges when* $0 < p \leq 1$

When $p = 1$ the resulting series is called the Harmonic series [7]

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots$$

[7] The Harmonic series is named as such because of the importance of the fraction $\frac{1}{n}$ in music. Given a string vibrating with wavelength λ, overtones (i.e., harmonics) occur at wavelengths $\frac{\lambda}{2}, \frac{\lambda}{3}, \frac{\lambda}{4}$, etc.

Though Theorems 7.5 and 7.6 tell us nothing about the convergence of the Harmonic series, by the p-Series Test we confirm the Harmonic Series diverges.

The rest of our work in this section proceeds as follows. We examine a number of *tests* we can use to determine the behaviour of a sequence of partial sums. Throughout we make use of material we developed in Section 4 of the course. In the interest of time, we will not give full proofs of all results, but give some intuition on why each result is reasonable.

7.1.1 Geometric Series and Ratios

Recall our series from the start of this section:

$$1 + \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \dots$$

From our previous work, perhaps we recognize this series as being called a *geometric* series. Each term in the sum arises by multiplying the previous term by $\frac{1}{4}$.

More generally, a geometric series is one of the form:

$$a + ar + ar^2 + ar^3 +$$

Definition 7.9. *Let $a, r \in \mathbb{R}$ with $a \neq 0$. We refer to the series*

$$a + ar + ar^2 + ar^3 + \dots$$

as a geometric series

By recalling some school math and some work from above, we fully classify the convergence of geometric series.

When $r \geq 1$ or $r \leq -1$, the sequence

$$a, ar, ar^2, ar^3, \dots$$

does not converge to 0. And so by Theorem 7.5 the geometric series

$$a + ar + ar^2 + ar^3 + \dots$$

must diverge.

Assume now $-1 < r < 1$. From our previous work in mathematics, we perhaps recall partial sums of the form

$$a + ar + ar^2 + ar^3 + \dots + ar^n$$

can be computed directly:

$$\sum_{k=0}^{n} ar^k = a + ar + ar^2 + ar^3 + \dots + ar^n = a\left(\frac{1 - r^{n+1}}{1 - r}\right)$$

And so

$$\sum_{k=0}^{\infty} ar^k = \lim_{n\to\infty} \sum_{k=0}^{n} ar^k = \lim_{n\to\infty} a\left(\frac{1 - r^{n+1}}{1 - r}\right)$$

When $|r| < 1$, we have $r^{n+1} \to 0$. And so

$$\lim_{n\to\infty} a\left(\frac{1-r^{n+1}}{1-r}\right) = \frac{a}{1-r}$$

Therefore when $|r| < 1$, we have

$$a + ar + ar^2 + ar^3 + \cdots = \frac{a}{1-r}$$

Theorem 7.10 (Geometric Series Test)**.** *The geometric series*

$$a + ar + ar^2 + ar^3 + \ldots$$

converges if and only if $|r| < 1$.

Geometric series are classified as arising from sequences for which subsequent terms differ by a common ratio, r:

$$a, ar, ar^2, ar^3, \ldots$$

For any pair of subsequent terms in the sequence we have

$$\frac{a_{n+1}}{a_n} = r$$

The Geometric Series test tells us that the corresponding series converges if and only if

$$\left|\frac{a_{n+1}}{a_n}\right| = |r| < 1$$

In a geometric series, the ratio between successive terms is always equal. However, a similar result applies when the ratio approaches a common value.

Let $a_n = \frac{n^2}{2^n}$. Consider the series

$$\sum_{k=0}^{\infty} a_k = 0 + \frac{1}{2} + \frac{4}{4} + \frac{9}{8} + \frac{16}{16} + \frac{25}{32} + \frac{36}{64} + \ldots$$

Since $a_n \to 0$, it is possible this series converges.

This is not a geometric series as the ratio between successive terms, $\frac{a_{n+1}}{a_n}$, is not a constant.

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

Though $\frac{1}{2}\frac{(n+1)^2}{n^2}$ is not a constant, it almost feels like one. As n gets very large, the ratio between successive terms in the series approaches $1/2$.

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \frac{1}{2}$$

In other words, as n gets very large, the series behaves as a geometric series with ratio $r = 1/2$. Such a series converges, and so it is reasonable to expect that the series

$$\sum_{k=0}^{\infty} a_k = 0 + \frac{1}{2} + \frac{4}{4} + \frac{9}{8} + \frac{16}{16} + \frac{25}{32} + \frac{36}{64} + \dots$$

also converges.

This intuition turns out to be just about correct.

Theorem 7.11 (Limit Ratio Test)**.** *Let (a_n) be a sequence so that $\frac{a_{n+1}}{a_n} \to r$.*

- *If $|r| < 1$, then $\sum\limits_{n=0}^{\infty} a_n$ converges.*
- *If $|r| > 1$, then $\sum\limits_{n=0}^{\infty} a_n$ diverges.*

One caveat to notice is that the Limit Ratio Test tells us nothing when $\frac{a_{n+1}}{a_n} \to 1$ or $\frac{a_{n+1}}{a_n} \to -1$. For example, when $a_n = \frac{1}{n}$ we have

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \lim_{n\to\infty} \frac{n}{n+1} = 1$$

From our work above, the Harmonic series diverges. However, when $a_n = \frac{1}{n^2}$, again we have

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

In this case the corresponding series $\sum\limits_{k=1}^{\infty} \frac{1}{k^2}$ converges.

7.1.2 Bounded and Monotone Sequences

Moving on to our next convergence test, let us take a moment to examine series arising from sequences where each term is non-negative. Let (a_n) be such a sequence. That is, let (a_n) be a sequence so that $a_n \geq 0$ for all $n \in \mathbb{N}$.

Since each term of the sequence is positive, the sequence of partial sums must be monotone increasing:

$$0 \leq a_0 \leq a_0 + a_1 \leq a_0 + a_1 + a_2 \leq \dots$$

That is

$$0 \leq A_0 \leq A_1 \leq A_2 \leq \dots$$

Recalling the statement of Theorem **??**, to prove such a sequence converges it suffices to prove that it is bounded. In other words:

Theorem 7.12. *Let (a_n) be a sequence so that $a_n \geq 0$ for all $n \in \mathbb{N}$. The series $\sum\limits_{n=0}^{\infty} a_n$ converges if and only if there exists $K \in \mathbb{R}$ so that $A_n \leq K$ for all $n \in \mathbb{N}$.*

7.1.3 Alternating Series

In our discussion of Theorem 7.12 we restricted ourselves to sequences with non-negative terms. However, recalling our time from Section 4, sequences of the form

$$a_n = (-1)^n b_n$$

gave us interesting examples of converging and diverging sequences. Let us take a look at such a sequence and its resulting series. Consider the series arising from the sequence given by $(-1)^n \frac{1}{n}$.

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots$$

This series is quite similar to the Harmonic series, a series that defied our intuition about infinite sums. The Alternating Harmonic series also has a surprise in store for us.

The following statement is true

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots = \log 2$$

Yes. Really. As we mentioned above, series are weird. We'll return to this discussion of this fact at the end of the next reading. For now let us be satisfied with classifying series with alternating signs.

Let (b_n) be a sequence. And consider the corresponding *alternating* series:

$$b_0 - b_1 + b_2 - b_3 + b_4 - \dots$$

By Theorem 7.5, if the terms of the sequence $(-1)^n b_n$ do not converge to 0, then the sequence necessarily diverges. Notice

$$(-1)^n b_n \to 0 \text{ if and only if } b_n \to 0$$

Unlike for non-alternating series, it turns out that having terms converge to 0 implies the series converges provided the sequence is monotone decreasing.

Theorem 7.13. *[Alternating Series Test] Let (b_n) be a monotone decreasing sequence so that $b_n \to 0$. The alternating series*

$$b_0 - b_1 + b_2 - b_3 + b_4 - b_5 + \dots$$

converges.

This is a surprising result! For non-alternating series, having the terms go to 0 in the limit is not enough to imply the series converges. Let is take a moment to discuss the proof of Theorem 7.13.

Let (b_n) be a monotone decreasing sequence and consider a partial sum of an even number of terms

$$B_5 = b_0 - b_1 + b_2 - b_3 + b_4 - b_5$$

Since (b_n) is monotonically deceasing we have

$$\begin{aligned} b_0 - b_1 &\geq 0 \\ b_2 - b_3 &\geq 0 \\ b_4 - b_5 &\geq 0 \end{aligned}$$

Therefore

$$(b_0 - b_1) + (b_2 - b_3) + (b_4 - b_5) \geq 0$$

Grouping another way we see:

$$b_0 \geq b_0 - (b_1 - b_2) - (b_3 - b_4) - b_5$$

Therefore

$$0 \leq b_0 - b_1 + b_2 - b_3 + b_4 - b_5 \leq b_0$$

Let us compare the sum above with the sum

$$B_7 = b_0 - b_1 + b_2 - b_3 + b_4 - b_5 + b_6 - b_7$$

Since $b_6 - b_7 \geq 0$ we have

$$B_5 = (b_0 - b_1 + b_2 - b_3 + b_4 - b_5) \leq (b_0 - b_1 + b_2 - b_3 + b_4 - b_5) + b_6 - b_7 = B_7$$

Of course there is nothing special about only taking the sum of the first six terms. The reasoning above will apply for the sum of any even number of terms:

$$0 \leq b_0 - b_1 + \cdots + b_{2k} - b_{2k+1} \leq b_0$$

and

$$B_{2k+1} = (b_0 - b_1 + \cdots + b_{2k} - b_{2k+1}) \leq (b_0 - b_1 + \cdots + b_{2k} - b_{2k+1}) + b_{2k+2} - b_{2k+3} = B_{2k+3}$$

Recall (B_n) is the sequence of partial sums:

$$b_0, b_0 - b_1, b_0 - b_1 + b_2, \ldots$$

Consider the subsequence (B_{2n+1})

$$(B_1, B_3, B_5, \ldots)$$

From our work above, (B_{2n+1}) is monotonically increasing and bounded above. Therefore (B_{2n+1}) converges.

In other words, there exists $L \in \mathbb{R}$ so that $B_{2n+1} \to L$.

A similar argument gives that there exists $U \in \mathbb{R}$ so that $B_{2n} \to U$.

If indeed (B_n) converges, then since (B_{2n+1}) and (B_{2n}) are subsequences of (B_n), these subsequences must converge to the same value.

By construction we have

$$L - U = \lim_{n\to\infty} B_{2n+1} - B_{2n} = \lim_{n\to\infty} b_{2n+1}$$

Since $b_n \to 0$, we have

$$L - U = \lim_{n\to\infty} b_{2n+1} = 0$$

and so, $L = U$.

We have shown than the odd terms of (B_n) and the even terms of (B_n) converge to the same value. It remains to show, however, that (B_n) converges to this same value. Doing so requires us to construct an ϵ-M argument in which we choose M as a maximum of values of M obtained from B_{2n+1} and B_{2n}, a task we are well familiar with. And so for this reason we omit the remainder of the argument.

7.1.4 When can we treat a series like a sum?

Throughout our work so far we have been careful to make the distinction between a series and a sum of real numbers. In our example at the start of this section we saw the pitfalls of assuming that we can manipulate a series in the same way we can manipulate a sum:

$$\begin{aligned}0 &= 0 + 0 + 0 + 0 + \dots\\ &= (1-1) + (1-1) + (1-1) + (1-1) + \dots\\ &= 1 + (-1+1) + (-1+1) + (-1+1) + \dots\\ &= 1 + 0 + 0 + 0 + \dots\\ &= 1\end{aligned}$$

We end our discussion in this section by finding some criteria for when we can re-group and re-arrange in a series in the same manner we do so for sums. We begin with grouping.

Consider the series

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots$$

Recalling that a series is defined to be a sequence of partial sums, the series above is the sequence

$$\left(1,\ 1 + \frac{1}{2},\ 1 + \frac{1}{2} + \frac{1}{4},\ 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8},\ \dots\right)$$

Consider now the series

$$1 + \left(\frac{1}{2} + \frac{1}{4}\right) + \frac{1}{8} + \dots$$

Grouping together these two terms in the series corresponds to a slightly different sequence of partial sums:

$$\left(1,\ 1 + \left(\frac{1}{2} + \frac{1}{4}\right),\ 1 + \left(\frac{1}{2} + \frac{1}{4}\right) + \frac{1}{8},\ \dots\right)$$

This is not the same sequence of partial sums as above. In the previous sequence of partial sums the second term is $1+\frac{1}{2}$, whereas in this sequence the second term is $1+\left(\frac{1}{2}+\frac{1}{4}\right)$.

$$\left(1,\ 1+\frac{1}{2},\ 1+\frac{1}{2}+\frac{1}{4},\ 1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8},\ \dots\right) \neq \left(1,\ 1+\left(\frac{1}{2}+\frac{1}{4}\right),\ 1+\left(\frac{1}{2}+\frac{1}{4}\right)+\frac{1}{8},\ \dots\right)$$

However, these two sequences of partial sums converge to the same limit:

$$\left(1,\ 1+\frac{1}{2},\ 1+\frac{1}{2}+\frac{1}{4},\ 1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8},\ \dots\right) \to 2$$

$$\left(1,\ 1+\left(\frac{1}{2}+\frac{1}{4}\right),\ 1+\left(\frac{1}{2}+\frac{1}{4}\right)+\frac{1}{8},\ \dots\right) \to 2$$

Therefore

$$1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\dots = 1+\left(\frac{1}{2}+\frac{1}{4}\right)+\frac{1}{8}+\dots$$

In this argument, there is nothing particularly special about the series $\sum_{n=1}^{\infty}\frac{1}{2^n}$, other than the fact that it converges. And so if $a_0+a_1+a_2+a_3+\dots$ converges, then so must $a_0+(a_1+a_2)+a_3+\dots$ converge to the same value. Further, there is nothing special in the discussion above about grouping $\frac{1}{2}+\frac{1}{4}$ in to $\left(\frac{1}{2}+\frac{1}{4}\right)$. And so we have the following result.

Theorem 7.14. *If the series* $\sum_{n=0}^{\infty} a_n$ *converges, then any series obtained by grouping the terms without changing their order converges to the same value.*

Aside. *Given our desire for preciseness throughout the rest of this subject, it is reasonable to find the statement of this theorem lacking. To give a proof of this theorem we would need to define more carefully what we mean to say* any series obtained by grouping the terms without changing their order. *This would require introducing some definitions and futzing around with indices. We'll avoid doing this as it doesn't aid our understanding any further.*

Theorem 7.14 permits us to treat a convergent series as a sum with respect to regrouping. In other words, Theorem 7.14 gives us a notion of associativity for convergent series.

Theorem 7.14 tells us nothing about regrouping in divergent series. Our alternating series example above does not converge And so our re-grouping to *prove* $0 = 1$ is nonsense.

Similar to grouping, the other manipulation that might feel natural in manipulating series is re-ordering. In considering sums we re-order terms sometimes without noticing. That operations in $\mathbb{R}$ are commutative is so deeply embedded in our psyche that seeing, for example, a pair of matrices for which $AB \neq BA$ violates our sensibilities. We violate the same sense of decency in the following example.

Consider the alternating harmonic series:

$$\sum_{n=1}^{\infty}\frac{(-1)^n}{n} = 1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\dots$$

By the Alternating Series test, this series converges. In other words, there exists $L \in \mathbb{R}$ so that

$$L = \sum_{n=1}^{\infty} \frac{(-1)^n}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots$$

Since this series converges, the Algebra of Series theorem applies; multiplying by a constant gives the expected result:

$$2L = \sum_{n=1}^{\infty} 2\frac{(-1)^n}{n} = (2)1 - (2)\frac{1}{2} + (2)\frac{1}{3} - (2)\frac{1}{4} + \dots$$

Simplifying each term on the right, we have

$$2L = 2 - 1 + \frac{2}{3} - \frac{1}{2} + \frac{2}{5} - \frac{1}{3} + \frac{2}{7} - \frac{1}{4} + \frac{2}{9} - \frac{1}{5} + \frac{2}{11} - \frac{1}{6} \dots$$

Naively *collecting like terms*, we have

$$\begin{aligned} 2L &= (2-1) - \frac{1}{2} + \left(\frac{2}{3} - \frac{1}{3}\right) - \frac{1}{4} + \left(\frac{2}{5} - \frac{1}{5}\right) - \frac{1}{6} + \dots \\ &= 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots \\ &= L \end{aligned}$$

Therefore $2L = L$ and so if $L \neq 0$, then $2 = 1$. Above we noted

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots = \log 2$$

And so this argument proves $2 = 1$, which is absurd. The error in our work above is naively *collecting like terms* (i..e, reordering). This example shows that, in general, that additions in series are not commutative. To think about why such a conclusion might be reasonable, recall that a series is a sequence of partial sums. Re-ordering the terms of the series changes the corresponding sequence of partial sums.

We spend the remainder of this section discussing criteria for when reordering the terms of a series does not change the outcome. Albeit, in a roundabout way. In our example above, we saw that permitting re-ordering within the Alternating Harmonic series led us to a nonsense result.

The Harmonic series and the corresponding Alternating Harmonic series are interesting examples of diverging and converging series. Though the Alternating Harmonic series converges, substituting each terms for its absolute value, the result series does not converge.

This is not the case for every series with positive and negative terms. Consider a geometric series with common ratio $r = -\frac{1}{2}$. Such a series converges by the Geometric Series test. Substituting each term of the series with its absolute value results in a geometric series with common ratio $r = \frac{1}{2}$, which also converges.

For geometric series, the converse of this property is also true: If $\sum_{n=1}^{\infty} |ar^n|$ converges, then $|r| \leq 1$, and thus $\sum_{n=1}^{\infty} ar^n$ also converges. Such a statement turns out to be true for every series:

Theorem 7.15. *If* $\sum_{n=1}^{\infty} |a_n|$ *converges, then* $\sum_{n=1}^{\infty} a_n$ *converges.*

Definition 7.16. *Let* $\sum_{n=1}^{\infty} a_n$ *be a series. We say* $\sum_{n=1}^{\infty} a_n$ *is* <u>*absolutely convergent*</u> *when* $\sum_{n=1}^{\infty} |a_n|$ *converges. When* $\sum_{n=1}^{\infty} a_n$*, but* $\sum_{k=1}^{\infty} |a_n|$ *does not converge, we say* $\sum_{n=1}^{\infty} a_n$ *is* <u>*conditionally convergent*</u>

Returning to our examples above, the Alternating Harmonic series is conditionally convergent. Any convergent geometric series is necessarily conditionally convergent.

Our discussion seems to have strayed a long way from talking about re-ordering terms of a series. In fact, absolute convergence .... WRITE MORE HERE.

Theorem 7.17. *If* $\sum_{k=1}^{\infty} a_n$ *is absolutely convergent, then every rearrangement of* $\sum_{k=1}^{\infty} a_n$ *absolutely converges to the same value.*

Test Your Understanding

1. If possible, apply the Geometric Series Test to each series below

   (a) $\displaystyle\sum_{n=0}^{\infty} \left(\frac{5}{4}\right)^n$

   (b) $\displaystyle\sum_{n=0}^{\infty} \frac{2^{n+2}}{6^{2n+1}}$

   (c) $\displaystyle\sum_{n=0}^{\infty} \frac{n^3}{\pi^n}$

2. If possible, apply the Limit Ratio Test to each series below.

   (a) $\displaystyle\sum_{n=0}^{\infty} \frac{n^3}{\pi^n}$

   (b) $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n!}$

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

3. Determine whether each of the following alternating series converge or diverge.

   (a) $\displaystyle\sum_{n=0}^{\infty} (-1)^n$

   (b) $\displaystyle\sum_{n=1}^{\infty} (-1)^n \frac{n^2}{n^3+1}$

(c) $\sum_{n=1}^{\infty}(-1)^n \frac{n^3+1}{n^2}$

4. (a) Find all values $r \in \mathbb{R}$ so that the following series converges

$$1 + 2r + 4r^2 + 8r^3 + \dots$$

(b) Give an expression for the limit of the series for each value $r \in \mathbb{R}$ you found in the previous part.

5. Let $a_n = 9 \times 10^{-n}$ and consider the series

$$(A_n) = a_1 + a_2 + \dots$$

Prove

$$A_n \to 1$$

Test Your Understanding - Answers

1. Apply the Geometric Series Test to each series below
   (a) diverges
   (b) converges, $r = \frac{1}{18}$
   (c) does not apply
2. Apply the Limit Ratio Test to each series below.
   (a) converges
   (b) converges
   (c) diverges $r = 2$
3. Determine whether each of the following alternating series converge or diverge.
   (a) diverges, terms do not go to 0
   (b) converges
   (c) diverges, terms do not go to 0
4. The common ratio of this geometric series is $2r$. By the Geometric Series Test, the series converges when $r \in (-\frac{1}{2}, \frac{1}{2})$. When it converges, the series converges to $\frac{1}{1+2r}$
5. This is a geometric series with $a = 0.9$ and $r = 0.1$. And so

$$0.9999999\cdots = \sum_{n=1}^{\infty} 9 \times 10^{-n} = \frac{0.9}{1-0.1} = 1$$

7.2 Power Series

Let us take a moment to return to geometric series. Let $a \neq 0$ and consider

$$\sum_{n=0}^{\infty} ar^n = a + ar + ar^2 + ar^3 + \dots$$

As we saw above, the convergence of this series depends on r. By the Geometric Series Test, the series converges if and only if $-1 < r < 1$.

Recalling one of our goals in Section 7, we want to use polynomials to approximate functions. Replacing r, the common ratio, with x, a variable, in a geometric series yields something that looks a lot like a polynomial:

$$G(x) = \sum_{n=0}^{\infty} ax^n = a + ax + ax^2 + ax^3 + \dots$$

This *polynomial* converges if and only if $-1 < x < 1$. When $-1 < x < 1$ we have

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

Consider the *polynomial*

$$1 + x + x^2 + x^3 + \dots$$

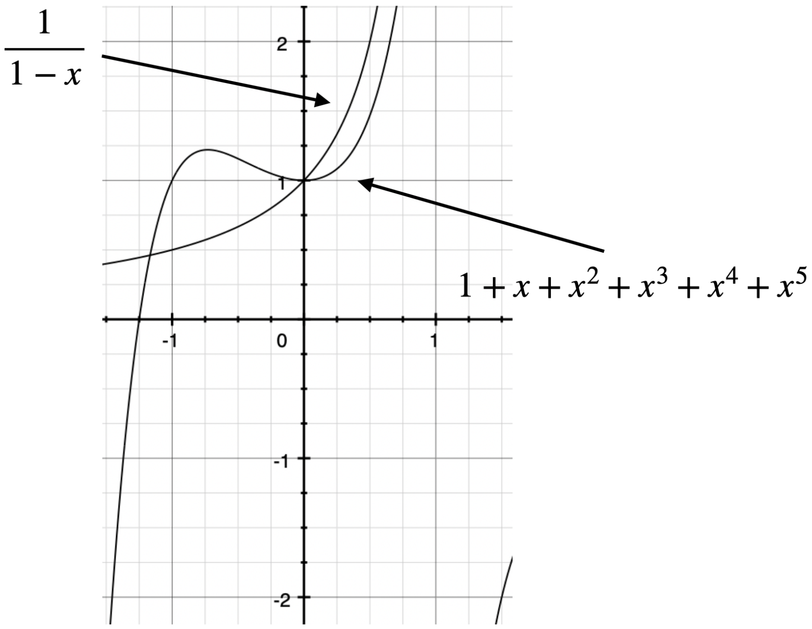
From our work above, when $|x| < 1$ we can write

$$1 + x + x^2 + x^3 + \dots = \frac{1}{1-x}$$

Consider truncating this sum after 6 terms:

$$f(x) = 1 + x + x^2 + x^3 + x^4 + x^5$$

It turns out that when we choose x so that $-1 < x < 1$, $f(x)$ is a pretty good approximation for $\frac{1}{1-x}$.



For example:

$$f(0.5) = 1 + (0.5) + (0.5)^2 + (0.5)^3 + (0.5)^4 + (0.5)^5 = 1.96875$$

and

$$\frac{1}{1-0.5} = 2$$

In the *polynomial*[8]

$$1 + x + x^2 + x^3 + \dots$$

each term has coefficient 1. However, using these same ideas, we can consider infinite *polynomials* whose coefficients take the value of any sequence. We refer to such series as power series

Definition 7.18. *Let (a_n) be a sequence. The power series of (a_n) is given by*

$$\sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + \dots$$

For example, the power series of $(-2)^n$ is given by

$$1 - 2x + 4x^2 - 8x^3 + 16x^4 - \dots$$

Though we recognize this series as an alternating series, it can also be considered as a geometric series with common ratio $r = -2x$. And so we can use the Geometric Series test to determine for which values of x the power series

$$G(x) = 1 - 2x + 4x^2 - 8x^3 + 16x^4 - \dots$$

[8] This isn't really a polynomial, but it certainly looks like one.

converges.

By the Geometric Series test, $G(x)$ converges for $-2x \in (-1, 1)$ and diverges otherwise. Therefore, $G(x)$ converges if and only if $x \in \left(-\frac{1}{2}, \frac{1}{2}\right)$.

Though the Geometric Series test will be useful for us in determining values for x for which a power series converges, our other tests will be useful as well. For example, consider the power series corresponding to our estimate for e in the introductory reading.

$$G(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

This is not a geometric series. The ratio between successive terms is not a constant. Here we have

$$\frac{g_{n+1}}{g_n} = \frac{x}{n+1}$$

Notice that when $x \in \mathbb{R}$ is fixed, we have

$$\lim_{n\to\infty} \frac{x}{n+1} = 0$$

And so by the Ratio Limit Test,

$$1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

converges for every $x \in \mathbb{R}$.

In each of our three examples above, the set of values of x for which the power series converged was non-empty. In fact, it is impossible to give an example of a power series for which the power series does not converge for any $x \in \mathbb{R}$. To see this, note that when $x = 0$ we have

$$\sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + \dots = a_0$$

Definition 7.19. *Let S be the set of values x for which the power series*

$$\sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + \dots$$

converges. We say S is the <u>*interval of convergence*</u> *of $\sum_{n=0}^{\infty} a_n x^n$. When S is not bounded we say $\sum_{n=0}^{\infty} a_n x^n$ has a* <u>*infinite interval of convergence*</u>

Returning to our examples above

- $\sum_{n=0}^{\infty} x^n$ has interval of convergence $(-1, 1)$.

- $\sum_{n=0}^{\infty}(-2x)^n$ has interval of convergence $\left(-\frac{1}{2}, \frac{1}{2}\right)$.
- $\sum_{n=0}^{\infty}\frac{x^n}{n!}$ has infinite interval of convergence

We interpret the interval of convergence of a power series as the set of values x for which the power series converges. That we use the word *interval* in our description of the set of values for which a power series converges suggests something more interesting is afoot. Based on the definition, there is no reason why the interval of convergence is necessarily an *interval.* In each of the examples we have seen, however, the interval of convergence has been an interval.

In the interest of brevity, we state the following result without proof or intuition.

Theorem 7.20. [9] *Let $G(x) = \sum_{n=0}^{\infty} a_n x^n$. The interval of convergence of $G(x)$ is one of the following:*

- $\{0\}$
- $\mathbb{R}$*; or*
- *an interval of the form* $(-\rho, \rho)$,$(-\rho, \rho]$, $[-\rho, \rho]$,*or* $[-\rho, \rho)$.

In the theorem above we refer to ρ as the <u>radius of convergence</u> of $G(x)$.

The radius of convergence marks line between values of x for which a power series converges and for which it diverges. We will return to our discussion on radius of convergence in at the end of Section ??.

Aside. *The symbol ρ is the greek letter* `rho`*. Since the word radius starts with r, our inclination is to refer to the radius of convergence with the letter r. However, r already has some other meanings in this reading and so we instead use a Greek letter that when pronounced in English begins with an* `r` *sound.*

7.2.1 When can we treat a power series as polynomial?

At the end of the previous section we took some time to think about manipulating a series in the same way we manipulate a sum. We saw that in certain circumstances, regrouping and rearranging within a series were reasonable things to do.

For any power series, choosing value for x within the interval of convergence results in a convergent sequence. And so we obtain the following from the Algebra of Power Series Theorem.

Theorem 7.21 (Algebra of Power Series Theorem)**.** *Let $G(x) = \sum_{n=0}^{\infty} a_n x^n$ and $H(x) =$*

[9] The presentation of this theorem is taken from *Real Analysis: A Long-Form Mathematics Textbook* by Jay Cummings

$\displaystyle\sum_{n=0}^{\infty} b_n x^n$ *be power series. Let* S_G *be the interval of convergence of* $G(x)$ *and* S_H *be the interval of convergence of* $H(x)$*.*

1. $G(c) + H(c) = \displaystyle\sum_{n=0}^{\infty}(a_n + b_n)c^n$ *for all* $c \in S_G \cap S_H$*; and*

2. $\lambda G(c) = \displaystyle\sum_{n=0}^{\infty} \lambda a_n c^n$ *for all* $\lambda \in \mathbb{R}$ *and* $c \in S_G$*.*

Other than algebraic manipulation, a natural question to ask is whether differentiation and integration are *valid* for power series.

Consider the equality

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

This equality is a statement about convergence of a series. In particular, this statement tells is that for any value x in the interval of convergence of $1 + x + x^2 + \dots$ the sequence $(1,\ 1 + x,\ 1 + x + x^2, \dots)$ converges to $\frac{1}{1-x}$. From this observation we are then perhaps convinced that a polynomial of the form

$$1 + x + x^2 + \cdots + x^n$$

is a good approximation for $\frac{1}{1-x}$ for all $x \in (-1, 1)$.

If $1 + x + x^2 + \cdots + x^n$ approximates the behaviour of $\frac{1}{1-x}$ on the interval $(-1, 1)$, then it not unreasonable to expect that the slopes of tangent lines to $\frac{1}{1-x}$ are approximated by the slopes of tangent lines to $1 + x + x^2 + \cdots + x^n$.

For $\frac{1}{1-x}$, the slope of a tangent line at any point is given by the derivative, $\frac{1}{1-2x+x^2}$. For $1 + x + x^2 + \cdots + x^n$, the slope of a tangent line at any point is given by the derivative $0 + 1 + 2x + \cdots + nx^{n-1}$. And so perhaps we can expect the following to be true for all $x \in (-1, 1)$.

In other words, perhaps:

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

Indeed our expectations hold!

Theorem 7.22. *Let* $G(x) = \sum_{n=0} a_n x^n$ *be a power series and let* f *be a integrable function so that for all* x *in the interval of convergence of* $G(x)$ *we have* $G(x) \to f(x)$*. For all* x *in the interval of convergence of* $G(x)$ *we have* $f'(x) = \sum_{n=1} n a_n x^{n-1}$*.*

Applying a similar line of reasoning for *area under the curve*, in place of *slope of a tangent line*, we have the following result.

Theorem 7.23. *Let* $G(x) = \sum_{n=0} a_n x^n$ *be a power series and let* f *be a integrable function so that for all* x *in the interval of convergence of* $G(x)$ *we have* $G(x) \to f(x)$*. For all* c *in the interval of convergence of* $G(x)$ *we have* $\int f(c)\, dx = \sum_{n=1} \frac{a_n}{n+1} c^{n+1}$*.*

The proofs of these theorems are not difficult in the sense that they dnot require any flicker of creativity. Instead they are an exercise in interpreting the various notations. For example, writing

$$f'(x) = \sum_{n=1} na_n x^{n-1}$$

means that for each x in the interval of convergence of $G(x)$, the sequence

$$\big(0,\ 0+a_1,\ 0+a_1+2a_2x,\ 0+a_1+2a_2x+3a_3x^2,\ \ldots\big)$$

converges to $f'(x)$. We omit the proofs of these theorems, as our interest in these two theorems is more in their usefulness than their proofs.

Returning to our example above, consider

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \ldots$$

By Theorem 7.22 we have

$$\frac{1}{1-2x-x^2} = 0 + 1 + 2x + 3x^2 + \ldots$$

for all $x \in (-1, 1)$

Part of the way through 7.1 we noted, with amazement, the following fact.

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots = \log 2$$

Using the Alternating Series Test we can confirm the Alternating Harmonic series converges. However, the test does not tell us what the series converges to. Consider the following power series:

$$1 - x + x^2 - x^3 + x^4 - x^5 + \ldots$$

The common ratio between successive terms is $r = -x$ and so when this power series converges, it converges to $\frac{1}{1-r} = \frac{1}{1+x}$.

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + x^4 - x^5 + \ldots$$

Applying Theorem 7.23, we have

$$log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} + \ldots$$

Evaluating both sides as $x = 1$, we get

$$\log 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \ldots$$

If we truncate the sum

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \ldots$$

after finitely many terms we get a reasonable approximation for $\log 2$.

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \frac{1}{9} - \frac{1}{10} + \frac{1}{11} = 0.7365\ldots$$

$$\log 2 = 0.6913$$

One caveat here, however is that 1 is not the interval of convergence of $1 - x + x^2 - x^3 + x^4 - x^5 + \ldots$. The interval of convergence of this power series is $(-1, 1)$. And so our result here doesn't follow directly from Theorem 7.22, but requires some further investigation. Do an internet search for `Abel's Theorem` for more information.

Theorems 7.22 and 7.23 are powerful tools in the study of *Combinatorial Enumeration.* Do an internet search for `generating function` for more information.

Test Your Understanding

1. Find the interval of convergence of each power series

   (a) $G(x) = \sum (-3)^n x^n$

   (b) $G(x) = \sum n^3 x^n$

   (c) $G(x) = \sum \frac{1}{n!} x^n$

   (d) $G(x) = \sum \frac{1}{n^2} x^n$ (You will need to apply more than one test here. The limit ratio test does give any information about $x = 1$ or $x = -1$)

Test Your Understanding - Answers

1. (a) $x \in \left(-\frac{1}{3}, \frac{1}{3}\right)$

   (b) For any fixed $x \neq 0$ the terms do not go to 0. And so $G(x)$ converges if and only if $x = 0$

   (c) The series converges for all $x \in \mathbb{R}$

   (d) By the limit ratio test, the series converges for $x \in (-1, 1)$. Further, when $x = -1$, the series converges by the Alternating series test. When $x = 1$, the series converges by the p-series test. Therefore the interval of convergence is $[-1, 1]$

7.2.2 Taylor Polynomials

As we have likely noticed by now, one of the major hurdles in moving from computational/procedural-based mathematics to conceptual/proof-based mathematics is a paucity of online resources for students. For example, there are many wonderful and well-crafted videos on topics from Calculus 2, but almost none of topics in Real Analysis. However, Taylor polynomials are a notable exception. Before going further, do an internet search for

```
3 blue 1 brown Taylor Polynomials
```

The video you will find, created by Dr Grant Sanderson, is a wonderful introduction to our final topic of the course. Take a moment to watch it now.

In the video, we saw that a polynomial of the form

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

might be a good choice to estimate the function e^x near $x = 0$. The function $T_3(x)$ has the following in common with e^x.

- $T_3(0) = 1 + 0 + \frac{0^2}{2!} + \frac{0^3}{3!} = 1 = e^0$
- $T_3'(0) = 0 + 1 + 2\frac{0}{2!} + 3\frac{0^2}{3!} = 1 = e^0 = [\frac{d}{dx}e^x]_{x=0}$
- $T_3''(0) = 0 + 0 + 1 + 3 \cdot 2\frac{0}{3!} = 1 = e^0 = [\frac{d^2}{dx^2}e^x]_{x=0}$.
- $T_3'''(0) = 0 + 0 + 0 + 1 = 1 = e^0 = [\frac{d^3}{dx^3}e^x]_{x=0}$.

Not only does the polynomial $T_3(x)$ agree with e^x at $x = 0$, but it also agrees with the first, second and third derivative of e^x at $x = 0$.

Adding on a similar term of the same form, the resulting polynomial

$$T_4(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!}$$

agrees with the fourth derivative of e^x at $x = 0$.

For any fixed value of $x \in \mathbb{R}$ we can consider the sequence

$$(T_n(x)) = (T_0(x), T_1(x), T_2(x), \dots)$$

Each term in this sequence is a partial sum, and so this sequence can be denoted as a series:

$$1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

For any fixed $x \in \mathbb{R}$, the terms of the sequence $(T_n(x))$ are increasingly better approximations of e^x.

For example

$$e^{1.3} = 3.669296667619244220457489916011486251431518884556551 46\dots$$

and

$$\begin{aligned}
T_0(1.3) &= 1\\
T_1(1.3) &= 1 + 1.3 = 2.3\\
T_2(1.3) &= 1 + 1.3 + \frac{1.3^2}{2!} = 3.145\\
T_3(1.3) &= 1 + 1.3 + \frac{1.3^2}{2!} + \frac{1.3^3}{3!} = 3.5111\overline{6}\\
T_4(1.3) &= 1 + 1.3 + \frac{1.3^2}{2!} + \frac{1.3^3}{3!} + \frac{1.3^4}{4!} = 3.630178\overline{3}\\
T_5(1.3) &= 1 + 1.3 + \frac{1.3^2}{2!} + \frac{1.3^3}{3!} + \frac{1.3^4}{4!} + \frac{1.3^5}{5!} = 3.66111191\overline{6}
\end{aligned}$$

Given this observation, perhaps

$$(T_n(1.3)) = (T_0(1.3), T_1(1.3), T_2(1.3), \dots) \to e^{1.3}$$

If this were the case, we could write

$$e^{1.3} = 1 + 1.3 + \frac{1.3^2}{2!} + \frac{1.3^3}{3!} + \dots$$

If $(T_n(x)) = (T_0(x), T_1(x), T_2(x), \dots) \to e^x$ for every $x \in \mathbb{R}$ we could write:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

Defining $0! = 1$ we can write this equality more compactly as

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

Recall the following two facts about e^x:

1. $e^0 = 1$
2. $\frac{d}{dx}e^x = e^x$

Let $G(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$ Ignoring all warnings about treating series as actual sums, we notice

$$G(0) = 1 + 0 + \frac{0^2}{2!} + \frac{0^3}{3!} + \dots = 1.$$

and

$$\begin{aligned}
G'(x) &= \frac{d}{dx}\left(1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots\right)\\
&= \frac{d}{dx}1 + \frac{d}{dx}x + \frac{d}{dx}\frac{x^2}{2!} + \frac{d}{dx}\frac{x^3}{3!} + \dots\\
&= 0 + 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots\\
&= G(x)
\end{aligned}$$

The power series $G(x)$ seems to share these two defining features of the function e^x. However, we should be cautious here. Fixing any value of x turns the power series into a series. As we have not verified any conditions for which algebraic manipulation of series is reasonable, we cannot be confident that any of this work makes any sense at all.

To prove

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

we must prove that for every x in the interval of convergence of $G(x)$, the sequence

$$(T_0(x), T_1(x), T_2(x), \dots)$$

converges to e^x.

Aside. *Before you go further, take a moment to make sense of this last sentence. The notation* $e^x = 1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\dots$ *is defined to mean that the sequence* $(T_0(x), T_1(x), T_2(x), \dots)$ *converges to* e^x*. If this doesn't make any sense, continuing to read on is not a good use of your time. Go back and review the definition of series and related notation. Remember:*

a series a sequence where each of the terms is a partial sum.

Using techniques from the previous section, we compute the interval of convergence of the power series

$$G(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

Let $e_n = \frac{x^k}{n!}$. Proceeding with the Limit Ratio test, we compute

$$\frac{e_{n+1}}{e_n} = \frac{1}{n+1}x$$

For fixed $x \in \mathbb{R}$, we have

$$\lim_{n\to\infty} \frac{e_{n+1}}{e_n} = 0$$

And so by the Limit Ratio test, the power series $G(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$ converges for all $x \in \mathbb{R}$.

To prove $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$ for all $x \in \mathbb{R}$, we must prove that for all $x \in \mathbb{R}$ the sequence

$$(T_0(x), T_1(x), T_2(x), \dots)$$

converges to e^x.

Recalling our definition of convergence for sequences, we want to prove that for every $\epsilon > 0$, there exists $M \in \mathbb{N}$ so that $|T_n(x) - e^x| < \epsilon$ whenever $n > M$.

From our construction of $T_n(x)$ above we have

$$|T_n(x) - e^x| = \left| \sum_{k=0}^{n} \frac{x^k}{k!} - e^x \right|$$

If we believe that polynomials of the form

$$1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!}$$

approximate e^x, then for any particular $x \in \mathbb{R}$, the expression

$$\left| \sum_{k=0}^{n} \frac{x^k}{k!} - e^x \right|$$

gives the difference between the approximation and the true value of e^x. In other words, this expression is the error in our approximation.

In proving

for every $\epsilon > 0$ there exists $M \in \mathbb{N}$ so that $|T_n(x) - e^x| < \epsilon$ whenever $n > M$

we are proving the error in the approximation can be made smaller than any ϵ. In other words, as n gets *very large*, the approximation for e^x approaches the true value for e^x. Proving this fact will be one of the main goals for this section.

Our main result, *Taylor's Theorem*, gives us a method to compute the error in our approximation. Further, Taylor's Theorem will apply to other functions that are difficult to compute, like $\sin x$ and $\cos x$, giving us a method to create approximations for these functions that are easy to compute.

Throughout the remainder of this section we denote by $f^{(n)}(x)$ the nth derivative of a function f.

Above we considered polynomials of the form

$$T_n(x) = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \cdots + \frac{1}{n!}x^n$$

The coefficients of this polynomials were chosen so that

- $T_n(x)$ agreed with e^x at $x = 0$; and
- each of the first n derivatives of $T_n(x)$ agreed with the corresponding derivative of e^x at $x = 0$.

We can use a similar construction to build a polynomial that approximates any differentiable function.

Let f be a function that is differentiable at 0. Consider the polynomial

$$T_n(x) = f(0) + f^{(1)}(0)x + \frac{f^{(2)}(0)}{2!}x^2 + \cdots + \frac{f^{(n)}(0)}{n!}x^n$$

This polynomial will agree with $f(x)$ at $x = 0$ and also agree with the first n derivatives of $f(x)$ at $x = 0$.

- $T_n(0) = f(0)$

- $T_n^{(1)}(0) = 0 + f^{(1)}(0) + 2\frac{f^{(2)}(0)}{2!}(0) + \cdots + n\frac{f^{(n)}(0)}{n!}0^{n-1} = f^{(1)}(0)$
- $T_n^{(2)}(0) = 0 + 0 + 2\frac{f^{(2)}(0)}{2!} + \cdots + n(n-1)\frac{f^{(n)}(0)}{n!}(0)^{n-2} = f^{(2)}(0)$
- $\dots$
- $T_n^{(n)}(0) = 0 + 0 + 0 + \cdots + 0 + n(n-1)(n-2)\cdots(3)(2)(1)\frac{f^{(n)}(0)}{n!} = f^{(n)}(0)$

Since $T_n(x)$ approximates the behaviour of $f(x)$ near $x = 0$, it is reasonable to expect $T_n(x)$ is a good approximation for $f(x)$ near $x = 0$.

Polynomials of this form will come in handy for us during this section, and so let us give them a name. We refer to $T_n(x)$ as the order n Taylor polynomial of f at 0.

Consider the function $f(x) = (1+x)^{-1/2}$. The order 3 Taylor polynomial of f at 0 is given by

$$T_3(x) = f(0) + f^{(1)}(0)x + \frac{f^{(2)}(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3$$

We compute

- $f(0) = 1$
- $f^{(1)}(0) = -\frac{1}{2}$
- $f^{(2)}(0) = \frac{3}{4}$
- $f^{(3)}(0) = -\frac{15}{8}$.

And so

$$T_3(x) = 1 - \frac{1}{2}x + \frac{3}{8}x^2 - \frac{5}{16}x^3$$

For every $n \in \mathbb{N}$, $T_n(x)$ approximates $f(x)$ near $x = 0$. For any particular value of x the error in this approximation is given by

$$|T_n(x) - f(x)|$$

Ideally, as n gets large the approximation improves and this error term goes to 0. In other words, for every $\epsilon > 0$ there will exist $M \in \mathbb{N}$ so that

$$|T_n(x) - f(x)| < \epsilon.$$

Recalling our definition of sequential limits, this is equivalent to saying

$$(T_0(x), T_1(x), T_2(x), \dots) \to f(x)$$

Studying the behaviour of this sequence tells us about the error in our approximation as n gets large. To ease our discussion, we give this sequence a name.

Definition 7.24. *Let f be a infinitely differentiable*[10] *function that is defined at* 0. *Let x be an element of the domain of f. The sequence*

$$(T_0(x), T_1(x), T_2(x), \dots)$$

[10] We haven't defined infinitely differentiable. In this context we should take it to mean that all possible deratives of f exist: $f, f^{(1)}, f^{(2)}, f^{(3)}, \dots$. Not every differentiable function has this property, but every differentiable function we are likely to care about has this property.

is referred to as the *Taylor series of f at* 0*.*

The Taylor series of f at 0 is a sequence where each term is a partial sum of the form

$$f(0) + f^{(1)}(0)x + \frac{f^{(2)}(0)}{2!}x^2 + \cdots + \frac{f^{(n)}(0)}{n!}x^n$$

Therefore we may denote the Taylor series of f at 0 as:

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

Returning to our example above, the Taylor series of $(1+x)^{-1/2}$ at 0 is given by the following power series

$$1 - \frac{1}{2}x + \frac{3}{8}x^2 - \frac{5}{16}x^3 + \frac{35}{128}x^4 - \frac{63}{256}x^5 + \dots$$

When $x = 1$, the sequence

$$(T_n(1)) = (T_0(1), T_1(1), T_2(1), \dots)$$

is a sequence of real numbers. This sequence converges to

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

(We have not proven this.)

To prove that for every x within the interval of convergence of the Taylor series we have

$$1 - \frac{1}{2}x + \frac{3}{8}x^2 - \frac{5}{16}x^3 + \frac{35}{128}x^4 - \frac{63}{256}x^5 + \cdots = (1+x)^{-1/2}$$

we would have to prove that for each x in the interval of convergence of the Taylor series, the sequence

$$(T_0(x), T_1(x), T_2(x), \dots)$$

converges to $(1+x)^{-1/2}$.

Recalling the definition of convergence for sequences, we would prove that for every $\epsilon > 0$ there exists $M \in \mathbb{N}$ so that $|T_n(x) - (1+x)^{-1/2}| < \epsilon$ whenever $n > M$.

As discussed in the introduction, the term:

$$|T_n(x) - (1+x)^{-1/2}|$$

can be interpreted as the error in the approximation. And so to prove that a Taylor series indeed converges to the expected value[11], we require a theorem that lets us compute this error. Taylor's Theorem does exactly this.

[11] This is not always the case. See the problem sheet for more details

Theorem 7.25 (Taylor's Theorem at 0)**.** *Let $n \in \mathbb{N}$ and let $f : I \to \mathbb{R}$ so that $f, f^{(1)}, f^{(2)}, \dots f^{(n+1)}$ are defined for all $x \in \mathbb{R}$. For every $x \in I$ with $x \neq 0$ there exists c between 0 and x so that*

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

At this stage, this theorem has been bestowed upon us without any attempt to justify its truth. Let us take a moment to try and make some sense of it.

In the expression

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

x and n are fixed. They are not variables. Rearranging we can write this expression as

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

It is hard to see, but if we squint carefully, this almost looks like the an expression we saw in the Mean Value Theorem. The function $f^{(n+1)}(c)$ is the first-derivative of the function $f^{(n)}(x)$. The Mean Value Theorem asserts the existence of a value c where the slope of the tangent line has a specified value.

The proof of Taylor's Theorem requires us to craft a function so that applying the Mean Value Theorem to the function gives us c so that

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

This is similar to how we proved the Mean Value Theorem: we crafted the function $f(x) - s(x)$ so that applying Rolle's Theorem to $f(x) - s(x)$ guaranteed the existence of c as defined in the statement of the Mean Value Theorem.

Aside. *The proof of Taylor's Theorem is not straightforward. You will not be asked to do this on the final exam.*

As our interest in Taylor's Theorem is in its usefulness for proving a Taylor series converges, we'll pay the proof no more attention. Instead let us see an example of using Taylor's Theorem to prove a Taylor series converges.

Example 7.26. Prove $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$ for all x in the interval of convergence of $1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$.

To prove $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$ we must prove that for all x in the interval of convergence of $1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$ the sequence of Taylor polynomials evaluated at x converges to e^x.

The interval of convergence of this power series is $x \in \mathbb{R}$. And so we prove $(T_0(x), T_1(x), T_2(x), \dots) \to e^x$ for all $x \in \mathbb{R}$.

Since Taylor's theorem will not apply when $x = 0$, we first consider the resulting series when $x = 0$.

When $x = 0$ we have

$$(T_0(0), T_1(0), T_2(0), \dots) = (1, 1, 1, \dots)$$

This sequence converges to $1 = e^0$. And so we conclude

$$e^0 = 1 + 0 + \frac{0^2}{2!} + \frac{0^3}{3!} + \dots$$

Assume $x \neq 0$. We appeal to the definition of convergence for sequences. Let $\epsilon > 0$. Recall $\frac{d^n}{dx^n} e^x = e^x$.

By Taylor's Theorem, there exists c_n between 0 and x so that

$$|T_n(x) - e^x| = |e^x - T_n(x)| = \left| \frac{e^{c_n}}{(n+1)!} x^{n+1} \right|$$

Therefore

$$\lim_{n \to \infty} \frac{e^{c_n}}{(n+1)!} x^{n+1} = 0$$

(This limit actually takes some work to justify. But verifying this limit is not the point of this exercise and so let's just accept it. You wouldn't be expected to prove such a limit on the final exam.)

Therefore there exists $M \in \mathbb{N}$ so that

$$\left| \frac{e^{c_n}}{(n+1)!} x^{n+1} - 0 \right| < \epsilon$$

whenever $n > M$. Recall

$$|T_n(x) - e^x| = \left| \frac{e^{c_n}}{(n+1)!} x^{n+1} \right|$$

Therefore $|T_n(x) - e^x| < \epsilon$ whenever $n > M$. Therefore

$$\lim_{n \to \infty} T_n(x) = e^x$$

And so

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

In the work we have done so far, we have approximated our function with a polynomial by considering the behaviour of derivatives at $x = 0$. Subsequently, we arrive at Taylor polynomials that do a good job approximating the function near $x = 0$. It turns out that our technique can be extended to consider derivatives (and thus approximations) at $x = b$ for $b \neq 0$.

We begin by generalising our definition of power series.

Definition 7.27. *Let (a_n) be a sequence. The power series of (a_n) at b is given by*

$$\sum_{n=0}^{\infty} a_n(x-b)^n = a_0 + a_1(x-b) + a_2(x-b)^2 + \dots$$

For a power series at b for $b \neq 0$, an interval of convergence of the form $(-\rho, \rho), [-\rho, \rho), (-\rho, \rho]$, or $[-\rho, \rho]$ is replaced by one of the form $(-\rho + b, \rho + b), [-\rho + b, \rho + b), (-\rho + b, \rho + b]$, or $[-\rho + b, \rho + b]$. In effect, replacing x by $x - b$ in the power series shifts the interval of convergence by b units.

For example, consider the following power series at $b = 2$.

$$G(x) = 1 + \frac{4}{5}(x-2) + \left(\frac{4}{5}\right)^2 (x-2)^2 + \left(\frac{4}{5}\right)^3 (x-2)^3 + \dots$$

For any particular value $x \in \mathbb{R}$ this power series becomes a series. And so it is reasonable to consider for which values $x \in \mathbb{R}$ the resulting series converges.

The ratio between successive terms in this power series is $r = \frac{4}{5}(x-2)$. For any particular value of x, this power series is a geometric series. By the Geometric Series test, the power series converges if and only if

$$\left|\frac{4}{5}(x-2)\right| < 1$$

Simplifying, we find the power series converges if and only if $x \in \left(\frac{3}{4}, \frac{13}{4}\right)$.

Notice the power series

$$H(x) = 1 + \frac{4}{5}x + \left(\frac{4}{5}\right)^2 x^2 + \dots$$

has interval of convergence

$$\left(-\frac{5}{4}, \frac{5}{4}\right)$$

And our power series $G(x)$ has interval of convergence

$$\left(2 - \frac{5}{4}, 2 + \frac{5}{4}\right)$$

Just as the interval of convergence of a power series is *centred* at 0, an interval of convergence for a power series at b is centred at b. In this example, the interval of convergence for $G(x)$ arises from the interval of convergence for $H(x)$ by shifting $b = 2$ units right.

In our work above in developing Taylor series at 0, we considered power series at 0. We proceed similarly to define Taylor series at b.

The order n Taylor polynomial of f at b is given by

$$T_{n,b}(x) = f(b) + f^{(1)}(b)(x-b) + \frac{f^{(2)}(b)}{2!}(x-b)^2 + \frac{f^{(b)}(0)}{3!}(x-b)^3 + \dots + \frac{f^{(n)}(b)}{n!}(x-b)^n$$

An order n Taylor polynomial of f at 0 approximates f near 0 by ensuring the following things are true:

- $T_n(0) = f(0)$
- $T_n^{(1)}(0) = f^{(1)}(0)$
- $T_n^{(2)}(0) = f^{(2)}(0)$
- $\dots$
- $T_n^{(n-1)}(0) = f^{(n-1)}(0)$
- $T_n^{(n)}(0) = f^{(n)}(0)$

Similarly, an order n Taylor polynomial of f at b approximates f near b by ensuring the following things are true:

- $T_{n,b}(b) = f(b)$
- $T_{n,b}^{(1)}(b) = f^{(1)}(b)$
- $T_{n,b}^{(2)}(b) = f^{(2)}(b)$
- $\dots$
- $T_{n,b}^{(n-1)}(b) = f^{(n-1)}(b)$
- $T_{n,b}^{(n)}(b) = f^{(n)}(b)$

Just as we did for Taylor polynomials at 0, we define a sequence of Taylor polynomials at b to be a Taylor series.

Definition 7.28. *Let f be a infinitely differentiable function that is defined at b. And let x be an element of the domain of f. The sequence*

$$(T_{0,b}(x), T_{1,b}(x), T_{2,b}(x), \dots)$$

is referred to as the Taylor series of f at b.

Theorem 7.29 (Taylor's Theorem at b)**.** *Let $n \in \mathbb{N}$, $\rho \in \mathbb{R}$ and let $f : (-\rho+b, \rho+b) \to \mathbb{R}$ so that $f, f^{(1)}, f^{(2)}, \dots f^{(n+1)}$ are defined for all $x \in (-\rho + b, \rho + b)$. For every $x \in (-\rho + b, \rho + b)$ with $x \neq b$ there exists c between b and x so that*

$$f(x) - T_{n,b}(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x-b)^{n+1}$$

Aside. *Though unstated, here ρ is the radius of convergence of the Taylor series of f at 0.*

In proving that a sequence of Taylor polynomials of $f(x)$ at b converges to $f(x)$, Taylor's Theorem at b plays the same role as Taylor's Theorem at 0.

To prove sequence of Taylor polynomials of $f(x)$ at b, $(T_{0,b}(x), T_{1,b}(x), T_{1,b}(x), \dots)$, converges to f, we must prove that for every $\epsilon > 0$ there exists $M \in \mathbb{N}$ so that

$$|T_{n,b}(x) - f(x)| < \epsilon.$$

Taylor's Theorem at b permits us to replace $|T_{n,b}(x) - f(x)|$ with a term of the form

$$\left| \frac{f^{(n+1)(c)}}{(n+1)!}(x-b)^{n+1} \right| .$$

To prove the sequence of Taylor polynomials of f at b converges to $f(x)$, it suffices to prove

$$\lim_{n\to\infty} \frac{f^{(n+1)(c)}}{(n+1)!}(x-b)^{n+1} = 0$$

for every x in the interval of convergence of the Taylor series of f at b.

Test Your Understanding

1. Compute the order 6 Taylor polynomials of $\cos\theta$ and $\sin\theta$ at 0.
2. Find the interval of convergence of the following power series: $G(x) = \sum \frac{1}{n^2}(x-3)^n$
3. Let f be a function and let $T_{2,1}(x)$ be the order 2 Taylor polynomial of f at 1. Prove
   (a) $T_{2,1}(1) = f(1)$
   (b) $T_{2,1}^{(1)}(1) = f'(1)$
   (c) $T_{2,1}^{(2)}(1) = f^{(2)}(1)$
4. Let $S_6(x)$ be the order 6 Taylor polynomial of $\sin x$ at 0. Using Taylor's Theorem, find an upper bound on the error of the approximation:

   $$|S_6(0.1) - \sin(0.1)|$$

   Hint: $-1 \leq \cos c \leq 1$.

Test Your Understanding - Answers

1. Let $C_n(\theta)$ denote the order n Taylor polynomial of $\cos\theta$ at 0. Let $S_n(\theta)$ denote the order n Taylor polynomial of $\sin\theta$ at 0.

$$\begin{aligned} C_6(\theta) &= 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} \\ S_6(\theta) &= \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} \end{aligned}$$

2. Using Limit ratio we find the series converges for all $x \in (2,4)$. At $x = 2$ the series converges by Alternating Series test. At $x = 4$ the series converges by p-series test. Therefore the interval of convergence is $[2,4]$.

3. We compute

$$\begin{aligned} T_{2,1}(x) &= f(1) + f^{(1)}(1)(x-1) + \frac{f^{(2)}(1)}{2!}(x-1)^2 \\ T_{2,1}^{(1)}(x) &= 0 + f^{(1)}(1) + 2\frac{f^{(2)}(1)}{2!}(x-1) \\ T_{2,1}^{(2)}(x) &= 0 + 0 + 2\frac{f^{(2)}(1)}{2!} \end{aligned}$$

and so

$$\begin{aligned} T_{2,1}(1) &= f(1) + f^{(1)}(1)(1-1) + \frac{f^{(2)}(1)}{2!}(1-1)^2 + = f(1) \\ T_{2,1}^{(1)}(1) &= 0 + f^{(1)}(1) + 2\frac{f^{(2)}(1)}{2!}(1-1) + = f^{(1)}(1) \\ T_{2,1}^{(2)}(1) &= 0 + 0 + 2\frac{f^{(2)}(1)}{2!} = f^{(2)}(1) \end{aligned}$$

4. By Taylor's theorem there exists c so that

$$S_6(0.1) - \sin(0.1) = \frac{f^{(7)}(c)}{7!}(0.1)^7$$

The seventh derivative of $\sin\theta$ is $\cos\theta$. Therefore

$$S_6(0.1) - \sin(0.1) = \frac{\cos(c)}{7!}(0.1)^7$$

We have

$$-1 \le \cos(c) \le 1$$

and so

$$-\frac{0.1^7}{7!} < S_6(0.1) - \sin(0.1) < \frac{0.1^7}{7!}$$

Therefore

$$|S_6(0.1) - \sin(0.1)| \le 0.00000000\overline{1984126}$$

7.2.3 Euler's Formula

In any conversation about aesthetics in mathematics, Euler's Formula is usually mentioned:

$$e^{i\pi} + 1 = 0$$

It turns out that Taylor polynomials give us the tools we need to prove this fact. Rearranging this formula we have

$$e^{i\pi} = -1 + 0$$

Recall the following fact we have been told sometime in the past:

$$e^{i\theta} = \cos\theta + i\sin\theta$$

Taking $\theta = \pi$ here gives $e^{i\pi} = -1 + 0$. And so to prove $e^{i\pi} + 1 = 0$ it suffices to prove $e^{i\theta} = \cos\theta + i\sin\theta$ for all $\theta \in \mathbb{R}$. Let us begin by computing the Taylor polynomials for $\cos\theta$ and $\sin\theta$.

Let $C_n(\theta)$ denote the order n Taylor polynomial of $\cos\theta$ at 0. Let $S_n(\theta)$ denote the order n Taylor polynomial of $\sin\theta$ at 0.

Recall the following facts

$$\cos 0 = 1 \quad \sin 0 = 0 \qquad\qquad \frac{d}{d\theta}\cos\theta = -\sin\theta \quad \frac{d}{d\theta}\sin\theta = \cos\theta$$

Using these fact we can compute, for example,

$$\begin{aligned} C_6(\theta) =& 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} \\ S_6(\theta) =& \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} \end{aligned}$$

What does this have to do with $e^{i\theta} = \cos\theta + i\sin\theta$?

Consider the polynomial

$$C_n(\theta) + iS_n(\theta)$$

Let us compare this polynomial at $n = 6$ with the order 6 Taylor polynomial of e^x evaluated at $x = i\theta$.

$$\begin{aligned} T_6(i\theta) &= 1 + i\theta + \frac{i^2\theta^2}{2!} + \frac{i^3\theta^3}{3!} + \frac{i^4\theta^4}{4!} + \frac{i^5\theta^5}{5!} + \frac{i^6\theta^6}{6!} \\ &= 1 + i\theta - \frac{\theta^2}{2!} - \frac{i\theta^3}{3!} + \frac{\theta^4}{4!} + \frac{i\theta^5}{5!} - \frac{\theta^6}{6!} \\ &= \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!}\right) + \left(i\theta - \frac{i\theta^3}{3!} + \frac{i\theta^5}{5!}\right) \\ &= \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!}\right) + i\left(\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!}\right) \\ &= C_6(\theta) + iS_6(\theta) \end{aligned}$$

In general we have

$$T_n(i\theta) = C_n(\theta) + iS_n(\theta)$$

Therefore the sequence of Taylor polynomials of e^x at 0 evaluated at $x = i\theta$

$$(T_0(i\theta), T_1(i\theta), \dots)$$

must converge to the same value as the sequence

$$(C_0(\theta) + iS_0(\theta),\ C_1(\theta) + iS_1(\theta),\ C_2(\theta) + iS_2(\theta) \dots)$$

The former sequence converges to $e^{i\theta}$. And so if

$$1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \cdots = \cos\theta$$

and

$$\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \cdots = \sin\theta$$

then by the Algebra of Series theorem

$$(C_0(\theta) + iS_0(\theta),\ C_1(\theta) + iS_1(\theta),\ C_2(\theta) + iS_2(\theta) \dots) \to \cos\theta + i\sin\theta$$

And thus we conclude

$$e^{i\theta} = \cos\theta + i\sin\theta$$

There are a few gaps to fill in this argument, but nothing major. Throughout this argument we have completely avoided any discussion of interval of convergence. Using the Limit Ratio Test we can prove each of the Taylor series above converge for all $x \in \mathbb{R}$. However, this does not guarantee to us that the Taylor Series for e^x converges for $x = i\theta$. The value $i\theta$ is not an element of $\mathbb{R}$. It is an element of $\mathbb{C}$. To prove e^x converges for $x = i\theta$ we would first have to extend our definition of convergence to consider complex numbers. This wouldn't be so bad; *absolute value* in our definition of convergence can be interpreted as *modulus* for elements of $\mathbb{C}$.

Once we had done this, we would then need to develop convergence tests for complex power series. In $\mathbb{C}$ an interval of convergence will be a circle in $\mathbb{C}$ centred at the origin. Hence the phrase *radius* of convergence.

With these tests we could prove the Taylor series for e^x converges at $i\theta$ and thus conclude

$$e^{i\theta} = \cos\theta + i\sin\theta$$

for all $\theta \in \mathbb{R}$.

None of this is beyond our grasp, but is work better suited for the study of *complex analysis.*

Index

Table of Notation