Course notes for Math 10850/10860, fall 2019 & spring 2020

following Spivak's *Calculus*

David Galvin, University of Notre Dame

Last updated April 25, 2020

Abstract

This is a set of notes to accompany lectures for Math 10850 (Honors Calculus I) and Math 10860 (Honors Calculus II), University of Notre Dame, fall 2019 and spring 2020. Honors Calculus I & II is officially a rigorous course on limits, differentiation, integration, and the connections between them. But it is really a first course in mathematical reasoning for the strongest and most motivated mathematically inclined. It's the course where you learn how to reason and how to prove. A major goal is to develop your ability to write arguments clearly and correctly. This is done in the context of an epsilon-delta approach to limits and calculus.

The recommended text for the course is

M. Spivak, *Calculus* (4th ed.), Publish or Perish Press, Houston, 2008. [3]

This is the size of a typical college calculus book, but the similarities end there. The text comprises a thorough theoretical introduction to differential and integral calculus, with expansive discussions. The exercises eschew the standard "practice technique X forty times in a row" format, and instead are challenging and probing extensions of the text.

I'll be following Spivak closely, skipping just a few topics of his and adding just a few, so it will be very useful to have a copy. The 3rd edition, which may be easier to get hold of, should work just as well as the 4th. At various points I may also be referring to other sources, mainly various handouts that other people have prepared for courses similar to this one. In particular, since Spivak jumps straight into proofs using the axioms of real numbers, but doesn't have any preamble on more foundational issues of logic and proofs, we will take a little while — nearly two weeks — to get to Spivak Chapter 1. Once we do get to Spivak, my notes will often follow his text *very* closely.

This document is currently in draft form. Comments & corrections are welcome! Email `dgalvin1@nd.edu`.

Contents

1 A quick introduction to logic

Mathematics deals in *statements* — assertions that are either unambiguously true or unambiguously false — and *proofs* — watertight arguments, based on fundamental rules of logic, that establish irrefutably the truth or falsity of statements. It's a language, and as such has a vocabulary and a grammar. The vocabulary includes technical words, that we will learn as we need them, but it also includes many everyday words and phrases, such as "if", "and", "implies", "it follows that", et cetera, whose mathematical meanings sometimes differ slightly from their ordinary meanings[1]. The grammar consists of the basic rules of logic or inference that allow us to understand the truth or falsity of complex statements from our knowledge of the truth or falsity of simpler statements.

This section of the notes introduces the basic vocabulary and grammar of mathematics. In Section 2, we will begin to talk about how the vocabulary and grammar are used to prove statements. In writing these two section I have drawn heavily on notes by John Bryant & Penelope Kirby [1] (Florida State), and Tom Hutchings [2] (Berkeley).

Underlying many of the examples in these first two sections (and all of the rest of the notes) is the basic language of sets. A quick introduction to the language of sets is given in Appendix A.

1.1 Statements

A basic object in mathematics is the *statement*: an assertion that is either true or false:

- "3 is a prime number." A true statement — we say that it has *truth value True* or simply *T*.
- "November 26, 1971 was a Friday." This is also true (you could look it up).
- "If I go out in the rain without a coat on, I will get wet."
- "There is no whole number between 8 and 10". A fine statement, albeit a false one — we say that it has *truth value False* or simply *F*.
- "There is life on Mars." Even though we don't know (yet) whether this is a true statement or a false one, everyone would agree that it has a definite truth value — there either is or there isn't life on Mars. So this is a statement.
- "All positive whole numbers are even". A false statement.
- "At least one positive whole number is even". A true statement.

[1] For example: if I tell you that tonight I'll either see Lion King at the movie theater, or go to a concert at DeBartolo, you might be surprised when I end up going to the movie *and* the concert: in ordinary language, "or" is most usually exclusive. But when we say mathematically that "either p or q is true", that always leaves open the possibility that *both p and q* are true: in mathematical language, "or" is always inclusive.

- "If you draw a closed curve on a piece of paper, you can find four points on the curve that form the four corners of a square". This is a statement, but is it a true one or a false one? Surprisingly, we don't know. The assertion was conjectured to be true in 1911 (by Otto Toeplitz)[2], but it has resisted all efforts at either a proof or a disproof.

Here are examples of things which are *not* statements:

- "Do you like ice cream?" A question, not an assertion.
- "Turn in the first homework assignment by Friday." An imperative, or a command, not an assertion.
- "$3x^2 - 2x + 1 = 0$." This is an assertion, but it does not have a definite truth-value — there are some x's for which it is true, and some for which it is false. So this is not a statement.
- "This statement is false." This is certainly an assertion. Is it true? If so, then it must be false. But if it is false, then it must be true. We can't assign a definite truth value to this assertion, so it is not a statement. This, and many other assertions like it, are referred to as *paradoxes*, and we try to avoid them as much as possible!

Some of our statement examples were quite simple ("There is life on Mars"), while others were more complicated beasts built up from simpler statements ("If I go out in the rain without a coat on, I will get wet"). Here we review the ways in which we build more complicated statements from simpler ones[3].

- **Negation**: If p is a statement, then the negation of p, which we sometimes call "not p" and sometimes write symbolically as $\neg p$[4], is a statement that has the opposite truth value to p. So:
  - $\neg p$ is false when p is true, and
  - $\neg p$ is true when p is false.

  Here are two clarifying examples:
  - If p is "There is life on Mars" then the negation of p is "There is no life on Mars". (It could also be "It is not the case that there is life on Mars".)
  - If p is

[2] O. Toeplitz, Über einige Aufgaben der Analysis situs, *Verhandlungen der Schweizerischen Naturforschenden Gesellschaft in Solothurn* **94** (1911), p. 197; see also https://en.wikipedia.org/wiki/Inscribed_square_problem

[3] A good analogy for what we are about to do, is how we combine numbers to form more complicated expressions that are still numbers, via the operations of addition, subtraction, multiplication, division, taking square roots, et cetera.

[4] Read this symbol as "not p".

"For every whole number n, there is a field with n elements"

(it doesn't matter what "field" might mean), then the negation of p is

"There is *some* whole number n for which there is not a field with n elements".

The negation is **NOT**: "There is *no* whole number n for which there is a field on n elements". Between "For every ..." and "There is no ..." there is a huge gap that is not covered by either statement — what if there are some n for which there is a field with n elements, and others for which there isn't? Then both of "For every ..." and "There is no ..." are false. But there's no such gap between "For every ..." and "There is some that is not ..." — whatever the possible sizes of fields, either one or other statement is true and the other is false. We'll come back to this idea when we talk about quantifiers.

We can use a *truth table* to summarize the effect of negation on a statement:

p	not p
T	F
F	T

We read this as: if p is true then $\neg p$ is false (first row), and if p is false then $\neg p$ is true (second row).

- **Conjunction**: If p and q are statements, then the conjunction of p and q is a statement that is true whenever both p *and* q are true, and false otherwise. We will typically write "p and q" for the conjunction. The symbolic notation is $p \wedge q$[5].

  If p is "There is life on Mars" and q is "There is water on Mars", then the conjunction $p \wedge q$ is "There is both life and water on Mars", and would only be true if we found there to be *both* life *and* water on Mars; finding that there is only one of these, or none of them, would not be good enough to make the conjunction true.[6]

  Here is the truth table for conjunction:

p	q	p and q
T	T	T
T	F	F
F	T	F
F	F	F

  Notice that since there are two options for the truth values of each of p, q, the truth table needs $2 \times 2 = 4$ rows.

[5] Read this as "p and q"

[6] Instead of saying "the conjunction $p \wedge q$" we will sometimes say "the conjunction of p and q", or even "the 'and' of p and q", or simply "p and q".

- **Disjunction**: If p and q are statements, then the disjunction of p and q is the statement that is true whenever at least one of p, q is true, and false otherwise. We write "p or q" for this compound statement, and sometimes denote it symbolically by $p \vee q$[7]. It is very important to remember that, by universal convention among mathematicians, "or" is *inclusive*: $p \vee q$ is true if p is true, or if q is true, *or if both are true*. This is a different from the ordinary language use of "or", which tends to be exclusive (see the footnote in the introductory text to Section 1).

  If p is "There is life on Mars" and q is "There is water on Mars", then the disjunction of p and q is "There is either life or water on Mars", and would only be false if we found there to be *neither* life *nor* water on Mars; finding any one of these (or both) would be good enough to make the disjunction true.[8]

  Here is the truth table for disjunction:

p	q	p or q
T	T	T
T	F	T
F	T	T
F	F	F

  As a specific, important, example, here's the truth table of "p or (not p)":

p	not p	p or (not p)
T	F	T
F	T	T

  (notice only two lines were needed here, because only one basic statement — p — is involved in "p or (not p)", and it can either be True or False).

  Observe: regardless of the truth value of p, the truth value of "p or (not p)" is True. This makes "p or (not p)" something called a *tautology* — a complex statement, made up of a combination of simpler statements, that turns out to be True regardless of the truth values of the simpler statements. In other words, a tautology is an *absolutely*, *always*, *eternally* True statement. Much of mathematics is engaged in the hunt for tautologies (usually going by the name "theorems"). The particular tautology we have just seen is an especially simple, but important, one. It's called *the law of the excluded middle*, because it says that

  > every statement is either True, or False

[7] Read this as "p or q"

[8] As with conjunction, instead of saying "the disjunction $p \vee q$" we will sometimes say "the disjunction of p and q", or even "the 'or' of p and q", or simply "p or q".

(there is no middle).[9]

From these basic operations we can build up much more complicated compound statements. For example, if we have three statements p, q and r we can consider the compound statement

$$\text{not } (p \text{ and } q) \text{ or not } (p \text{ and } r) \text{ or not } (q \text{ and } r).$$

or

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

(Notice that the symbolic formulation is typographically a lot nicer in this case; I'll stick with that formulation throughout this example.) If p and q are true and if r is false, then $p \wedge q$ is true, so $\neg(p \wedge q)$ is false. By similar reasoning $\neg(p \wedge r)$ and $\neg(q \wedge r)$ are both true. So we are looking at the disjunction of three statements, one of which is false and the other two of which are true. We haven't defined the disjunction of three statements, but it's obvious what it must be: the disjunction is true as long as at least one of the three statements is true. That means that in the particular case under consideration (p, q true, r false), the compound statement $\neg(p \wedge q) \vee \neg(p \wedge r) \vee \neg(q \wedge r)$ is true.

We can do this for all $2 \times 2 \times 2 = 8$ possible assignments of truth values to p, q and r, to form a truth table for the compound statement:

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

It appears that the statement $\neg(p \wedge q) \vee \neg(p \wedge r) \vee \neg(q \wedge r)$ is false only when all three of p, q, r are true, so in words it is the statement "At least one of p, q, r is false".

As another example, consider $\neg(p \wedge q \wedge r)$ (where again we haven't defined the conjunction of three statements, but it's obvious what it must be: the conjunction is true only if all three

[9] Given how we defined a statement, the law of the excluded middle must be true. It is a "tautology" in the ordinary language sense.

of the three statements are True). Here's the truth table for this compound statement:

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

It's exactly the same as the truth table for $\neg(p \wedge q) \vee \neg(p \wedge r) \vee \neg(q \wedge r)$, which of course it should be: even without writing the full truth table, it should have been evident that the statement $\neg(p \wedge q \wedge r)$ is the same as "At least one of p, q, r is false". This illustrates that two apparently different compound statements may have the same truth tables, and so may be considered "the same" statement.

Formally, if A is one statement built from the simpler statements p, q and r, using combinations of $\neg$, $\wedge$ and $\vee$, and B is another one, then A and B are *equivalent* (though we will often somewhat sloppily say *the same*) if: for each possible assignment of truth values to p, q and r, the resulting truth value of A is the same as the resulting truth value of B. Effectively, this means that if you use a single truth table to figure out what A and B look like, then the column corresponding to A is the same as the column corresponding to B. Of course, this can be extended to pairs of statements built from any number of simpler statements.

Here are a few pairs of equivalent statements; the equivalence of each pair is quickly verified by comparing truth tables.

- $(p \wedge q) \wedge r$ and $p \wedge (q \wedge r)$
- $(p \vee q) \vee r$ and $p \vee (q \vee r)$
- $\neg(p \wedge q)$ and $(\neg p) \vee (\neg q)$
- $\neg(p \vee q)$ and $(\neg p) \wedge (\neg q)$

If you were uncomfortable with line "We haven't defined the conjunction of three statements, but it's obvious what it must be ...", then you will be happy with the equivalence of the first pair: it shows that whatever pair-by-pair order we choose to deal with the conjunction of three statements, the resulting truth table is the same (and is the same as the truth table of "All of p, q, r are true"), so it is really ok to slightly sloppily talk about the conjunction of three statements. The equivalence of the second pair does the same job for the disjunction of three statements. With some (actually a lot) more work we could show that if $p_1, p_2, \ldots, p_n$

are n statements, then whatever pair-by-pair order we choose to deal with the conjunction $p_1 \wedge p_2 \wedge \ldots \wedge p_n$ the resulting truth table is the same, and is the same as the truth table of "All of p_1, p_2, …, p_n are true"); and there is an analogous statement for $p_1 \vee p_2 \vee \ldots \vee p_n$. (We will return to this, in the slightly different but essentially equivalent context of "associativity of addition", when we come to discuss proofs by induction.)

The third and fourth pairs of equivalences above are called *De Morgan's laws*, which we will return to in more generality later.

To illustrate what was meant earlier by "use a single truth table to figure out what A and B look like", here's the single truth table that shows the validity of the second of De Morgan's laws, that $\neg(p \vee q)$ and $(\neg p) \wedge (\neg q)$ are equivalent:

p	q	$p \vee q$	$\neg(\mathbf{p \vee q})$	$\neg p$	$\neg q$	$(\neg\mathbf{p}) \wedge (\neg\mathbf{q})$
T	T	T	**F**	F	F	**F**
T	F	T	**F**	F	T	**F**
F	T	T	**F**	T	F	**F**
F	F	F	**T**	T	T	**T**

Some comments:

- We've introduced some auxiliary columns into the truth-table, mostly for bookkeeping purposes[10]; some people find this helpful, others don't, it is entirely a matter of personal taste.

- The columns corresponding to $\neg(p \vee q)$ and $(\neg p) \wedge (\neg q)$ are identical. Since for each particular row, the truth values assigned to p and q for the purposes of determining the truth value of $\neg(p \vee q)$, are the same as the truth values assigned to p and q for the purposes of determining the truth value of $(\neg p) \wedge (\neg q)$, this allows us to conclude from the truth table that $\neg(p \vee q)$ and $(\neg p) \wedge (\neg q)$ are equivalent statements.

1.2 An note on parentheses

There's an inherent ambiguity in any reasonably complicated combination of numbers via the basic arithmetic operations of $+$, $-$, et cetera. For example, if we perform the addition first and then the multiplication we get

$$1 + 2 \times 3 = 3 \times 3 = 9$$

while if we perform the multiplication first and then the addition we get

$$1 + 2 \times 3 = 1 + 6 = 7.$$

[10] I believe that "bookkeeping" (and variants "bookkeeper", "bookkeepers") is the only unhyphenated English word with three double letters in a row ("woolly" doesn't count). There's also exactly one with four doubles — subbookkeeper. This will not be on the exam.

Similarly, there's an inherent ambiguity in any reasonably complex statement, related to the order in which to perform operations such as $\neg$, $\wedge$ and $\vee$ mentioned in the statement. Different choices of order may lead to different truth tables. For example, consider the statement "*notporq*". This could mean "take the disjunction of the following two statements:

- q, and
- the negation of p".

Or it could mean "take the negation of the following statement:

- the disjunction of q and p".

This are unfortunately different statements: if p and q are both true, then the first is true while the second is false.

One way to avoid both these ambiguities (arithmetical and logical) is to decide, once and for all time, on an order of precedence among arithmetic operations, and an order of precedence among logical operations. There *is* a fairly standard such order[11] for arithmetic operations. There are also established orders of precedence among logical operations. But there Tare two problems with taking this approach to eliminate ambiguity:

- first, there are *competing* orders of precedence, none universal, so that order-of-precedence alone does not eliminates all ambiguity, and
- second, an order of precedence among logical operators is something that one must remember, and there is no obvious motivation behind it to act as a memory aid. Who wants that?

For these reasons, I prefer to avoid ambiguity by using parentheses to indicate order of operation, with the convention being that you work from the inside out. So for example, to indicate "take the disjunction of the two statements 'q' and 'the negation of p'" I would write

"(not p) or q"

(indicating that the negation should be performed before the disjunction), while to indicate "take the negation of the statement 'the disjunction of q and p'" I would write

"not (p or q)"

(indicating that the disjunction should be performed before the negation).

I encourage you to carefully consider any statement that you write for ambiguity, and make every effort to de-ambiguize!

[11] see, for example, https://en.wikipedia.org/wiki/Logical_connective#Order_of_precedence

1.3 Implication

Implication is by far the most important operation that builds more complicated statements from simpler ones — it lies at the heart or virtually all logical arguments — and so (unsurprisingly) is probably the most subtle.

If p and q are two statements, then we can form the *implication* assertion "p implies q", symbolically $p \Rightarrow q$[12]. We *define* "p implies q" to be (a statement that is logically equivalent to) the statement

"(not p) or (p and q)".

To illustrate: we have already seen the example "If I go out in the rain without a coat on, I will get wet." This is the implication "p implies q" where p is "I go out in the rain without a coat" and q is "I get wet". Viewed as an ordinary language sentence it conveys precisely the same meaning as "either

- I don't go out in the rain without a coat,

or

- I do go out in the rain without a coat, and (as a result), I get wet"

which is indeed "(not p) or (p and q)".

The implication "p implies q" can be rendered into ordinary language in many other ways, such as:

- If p (happens) then (so does) q
- (The occurrence of) p is (a) sufficient (condition) for q (to happen)
- q (happens) whenever p (happens).

Mirroring the list above, the implication "If I go out in the rain without a coat on, I will get wet" can be expressed in ordinary language as

- My going out in the rain without a coat leads to (implies) my getting wet
- If I go out in the rain without a coat on, then I get wet
- My going out in the rain without a coat is a sufficient condition for me to get wet
- I get wet whenever I go out in the rain without a coat.

Some more notation related to implication:

- p is referred to as the *premise* or *hypothesis* of the implication

[12] Read this as "p implies q".

- q is referred to as the *conclusion*.

We can easily form the truth table of "p implies q", by forming the truth table of "(not p) or (p and q)":

p	q	not p	p and q	(not p) or (p and q)	p implies q ($p \Rightarrow q$)	(not p) or q
T	T	F	T	T	T	T
T	F	F	F	F	F	F
F	T	T	F	T	T	T
F	F	T	F	T	T	T

For good measure, I've thrown in the truth table of "(not p) or q" above; notice that it is exactly the same as that of "(not p) or (p and q)", that is, the same as "p implies q", and many texts take this simpler statement to be the definition of "p implies q".

> **Convention**: From here on, we will take "p implies q" to mean simply "(not p) or q".

Two lines of the truth table for "p implies q" should be obvious: If p is true, and q is true, then surely we want "p implies q" to be true; while if p is true but q is false, then surely we want "p implies q" to be false. The other two lines of the truth table, corresponding to when p is false, are more subtle. To justify them, think of "p implies q" as a promise, a contract, that says that **IF** something (p) happens, **THEN** something else (q) happens. The contract is a good one in the case when p and q both happen, and a bad one when p happens but q doesn't (that justifies the first two lines). If p *doesn't* happen (the last two lines of the table) then the contract is never invoked, so there is no basis on which to declare it bad, so we declare it good.

In terms of our example ("If I go out in the rain without a coat on, I will get wet"): suppose the TV weather forecaster tells me that if I go out without my coat in today's rain, I will get wet. From this, I would expect to get wet if I did go out in the rain without my coat; if that happened I would say "true promise" about the forecaster's statement, whereas if I went out in the rain without my coat and didn't get wet, I would say "false promise". But what if I didn't go out in the rain without a coat? The forecaster said nothing about what happens then, so whether I stay dry (by going out with a coat, or by staying home), or get wet (by taking a bath, or because my house has a leaking roof), she would not be breaking any kind of promise. If either of these last two things occur, I should still say that the implication stated was true because she did not break her promise.

If this isn't convincing, another justification for the "implies" truth table is given in the first homework.

The negation of an implication

We have observed via a truth table that "p implies q" is equivalent to "(not p) or q" (and, to repeat a remark from earlier, in many texts this is the definition of "implies"). By De

Morgan's law (and the easy fact that we have not yet explicitly stated, that "not (not p)" always has the same truth value as p), the negation "not (p implies q)" of an implication is equivalent to "p and (not q)". Think of it in these terms: to "falsify" the weather forecasters contract, to demonstrate that it was not a sound contract, you would need to both

- go out in the rain without a coat on

and

- *not* get wet.

The contrapositive of an implication

Here is another statement that is equivalent to "p implies q": the statement

"(not q) implies (not p)".

This reformulation is called the *contrapositive* of the original implication; we will see many times as the year progresses.

There are a few ways to check that the contrapositive of an implication is equivalent to the implication:

Via a truth table The easiest way, but not too informative.

Via DE Morgan's law By definition, "(not q) implies (not p)" is equivalent to "(not (not q)) or (not p)", which is in turn equivalent to "q or (not p)", which is equivalent to "(not p) or q", which by definition is equivalent to "p implies q".[13]

Via "reasoning" (probably the best way to initially get an understand of what's going on) After thinking through some examples (both non-mathematical and mathematical) you should be fairly comfortable with the fact that "p implies q" carries exactly the same content and meaning as "(not q) implies (not p)". Some examples:

"If I go out in the rain without a coat, then I will get wet"

carries just the same meaning as

"If I didn't get wet, then I didn't go out in the rain without a coat",

and

"If $b^2 - 4ac < 0$ then the equation $ax^2 + bx + c = 0$ doesn't have real solutions"

carries just the same meaning as

"If $ax^2 + bx + c = 0$ has real solutions, then $b^2 - 4ac \geq 0$".

[13] This can be thought of as our first example of a "proof" in the way mathematics understands the word: we have carefully put together some simpler truths ("not (not p)" is the same as p; "a or b" is the same as "b or a") in just the right order to deduce a non-obvious truth, that two superficially different statements are in fact the same.

The converse of an implication

A statement related to "p implies q" that is **NOT** equivalent to it (**NOT**, **NOT**, **NOT**, really **NOT**), is the *converse* "q implies p", which has truth table

p	q	q implies p
T	T	T
T	F	T
F	T	F
F	F	T

Note that this is *not* the same as the truth table of "p implies q". One strategy we will **NEVER** employ to show that p implies q is to show q implies p and then say that that's enough to deduce that p implies q — as the truth table shows, it **ISN'T**!!! Convincing me that

if I get wet, then I go out in the rain without my coat on

does not help towards convincing me that

if I go out in the rain without my coat on, then I get wet.

Indeed, I don't think you could convince me of the former, since it's not a true statement: if on a warm, dry morning I take a bath, then I get wet (so invoke the contract in the statement above), but I don't go out in the rain without my coat on, so the contract fails. The latter statement, on the other hand ("if I go out in the rain without my coat on, then I get wet") is, I think, true.

One reason that the converse of an implication gets confused with the implication itself, is that sometimes when we use implicative language in ordinary-speak, we actual mean the converse. A classic example is: "If you don't eat your vegetables, you won't get dessert". Formally this is the implication "p implies q" where p:"you don't eat your vegetables" and q:"you don't get dessert". And if it really is this implication, then you would not be upset when, upon eating your vegetables, you were *not* given dessert: you didn't "not eat your vegetables", so the contract *wasn't* invoked, and in this case there is no promise regarding dessert!

But in fact, you would be justified in being very peeved if, after diligently eating your vegetables, you were denied dessert. This is because you, like most sensible people, would have interpreted "If you don't eat your vegetables, you won't get dessert" as a contract whose meaning is that if you eat your vegetables, then you get rewarded with dessert. In other words, "(not p) implies (not q)", or, contrapositively (and equivalently), "q implies p." So although the ordinary language wording of the statement formally parsed as an implication in one direction, its meaning was really an implication in the converse direction.

These kinds of ambiguities occur a lot in ordinary language, and for this reason I will try to keep my examples in the mathematical realm, where there should be no ambiguity.

The "if and only if" ("iff") statement

Related to implication is *bidirectional implication*, or the *if and only if* statement. The statement "p if and only if q" is shorthand for "(p implies q) and (q implies p)", and is denoted symbolically by $p \Leftrightarrow q$[14] (an explanation for the terminology "if and only of" is given in Section 1.5). The sense of this statement is that p and q sink or swim together; for the bidirectional implication to be true, it must be that either p and q are simultaneously true or simultaneously false. The truth table for bidirectional implication is:

p	q	p if and only if q
T	T	T
T	F	F
F	T	F
F	F	T

The phrase "if and only if" is often abbreviated to "iff".

The bidirectional implication statement can be rendered into ordinary english as:

- p (happens) if and only if q (happens)
- (The occurrence of) p is (a) necessary and sufficient (condition) for q (to happen)
- (The occurrence of) q is (a) necessary and sufficient (condition) for p (to happen).

Here's an example from the world of sports:

"A team wins the World Series
if and only if .
they win the last MLB game of the calendar year"

Indeed, to the win the World Series, it is *necessary* to win the last game of the year; it is also *sufficient*.

1.4 An note on symbols versus words

The field of logic is filled with precisely defined symbolic notations, such as $\neg$, $\wedge$, $\vee$, $\Rightarrow$, $\Leftrightarrow$, $\therefore$, $\because$, $\exists$, $\forall$ (we'll see these last two in a while), et cetera. Many of these appear in this section, usually to keep complex propositions manageable. But if you pick up a mathematical paper, you will notice an

almost complete absence

of these symbols. The convention that is almost universally adhered to by mathematicians today is to

[14] Read this as "p if and only if q".

write mathematics in prose.

For example you will frequently see things like "A implies B, which implies C, which in turn implies D" in a mathematical exposition (note that this is a complete English sentence), and almost never see the symbolically equivalent

$$\text{“}A \Rightarrow B \Rightarrow C \Rightarrow D\text{”}.$$

And often a proof will be presented in the following way in a paper: "Since A and B are true, and we have argued that A and B together imply C, we deduce that C is true", but you will (almost) never see the symbolically equivalent

$$\begin{array}{lll} & & \text{“}A \text{ is true} \\ & & B \text{ is true} \\ \therefore & & A \wedge B \\ & \text{also} & (A \wedge B) \Rightarrow C \text{ is true} \\ \therefore & & C \text{ is true”}. \end{array}$$

Although the symbolic notation is sometimes a nice shorthand, when we come to write proofs of propositions I will be **strongly encouraging** you to follow the standard convention, and present proofs in (essentially) complete English sentences, avoiding logical symbols as much as possible. There will be much more on this later, particularly in discussions of homework problems.

1.5 An note on language: "if and only if", and "necessary and sufficient"

The logical implication "p implies q" is often rendered into ordinary language as "p is sufficient for q" (this should make sense: "p is sufficient for q" says that if p happens, then q happens, but allows for the possibility that q happens even when p doesn't, and this is exactly what "p implies q" means). Another phrase that you will encounter a lot is

"p is necessary for q".

"p is necessary for q" means the same thing as "if p doesn't happen, then q doesn't happen", which is the same as "(not p) implies (not q)", which is the contrapositive of (and so equivalent to) "q implies p".

It is for this reason that the bidirectional equivalence "p if and only if q" is often rendered into ordinary language as

"p is necessary and sufficient for q".

The "sufficient" part captures "p implies q", and as we have just seen, the "necessary" part captures "q implies p".

What about the phrase "if and only if" itself? Remember that the bidirectional implication between p and q is shorthand for "(p implies q) *and* (q implies p)". The logical implication "q implies p" is often rendered into English as "p if q" (this should make sense: "p if q" says that if q happens, then p happens, and this is exactly what "q implies p" means). Another phrase that you will encounter a lot is

"p only if q".

"p only if q" clearly means the same thing as "q is necessary for p", and so (by the discussion earlier in this section) means the same as "p implies q".

It is for this reason that the bidirectional equivalence is often rendered into ordinary language as

"p if and only if q".

The "if" part captures "q implies p", and as we have just seen, the "only if" part captures "p implies q".

1.6 A collection of useful equivalences

We have seen that some pairs of superficially different looking statements are in fact the same, in the sense that their truth tables are identical. One example was

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

This says that whenever we encounter the negation of the conjunction of two things (p and q, in this case) in the middle of a proof, we can replace it with the disjunction of the negations (if that would be helpful).

This example might seem somewhat silly — once "$\neg(p \wedge q)$" and "$(\neg p) \vee (\neg q)$" are translated into ordinary language, it is hard to tell them apart! Indeed, "$\neg(p \wedge q)$" translates to "it is not the case that both p and q are true", while "$(\neg p) \vee (\neg q)$" translates to "it is the case that at least one of p and q is false". So it's unclear how much mileage we might get from replacing one with the other. (We will in fact see that this substitution is sometimes quite useful.)

A more substantial example is the pair, discussed earlier in the implication section,

"$\neg(p \Rightarrow q)$" and "$p \wedge (\neg q)$".

Suppose we are are in the middle of a proof, and we find ourselves working with the statement "p doesn't imply q". What can we do with this? Using the above equivalent pair, we can replace it with the statement "p and not q". This formally doesn't change anything, but the new statement is so different-*looking* from the old that we might well be able to use this new viewpoint to move the argument forward, in a way that we couldn't have done sticking with the old statement.

Much of the business of formulating proofs involves this kind of manipulation — substituting one expression for another, equivalent, one, and leveraging the change of viewpoint to make progress — and so it is useful to have an arsenal of pairs of equivalent statements at one's disposal. Here is a list of the most common pairs of equivalent statements, together with their common names. They can all be verified by constructing truth tables for each of the two pairs of statements, and checking that the truth tables have identical final columns. For the most part, it's not important to remember the specific names of these pairs.

Name	Pair of equivalent statements
Identity law	$p \wedge T$ and p $p \vee F$ and p
Domination law	$p \vee T$ and T $p \wedge F$ and F
Idempotent law	$p \vee p$ and p $p \wedge p$ and p
Double negation law	$\neg(\neg p)$ and p
Commutative law	$p \vee q$ and $q \vee p$ $p \wedge q$ and $q \wedge p$
Associative law	$(p \vee q) \vee r$ and $p \vee (q \vee r)$ $(p \wedge q) \wedge r$ and $(p \wedge q) \wedge r$
Distributive law	$p \vee (q \wedge r)$ and $(p \vee q) \wedge (p \vee r)$ $p \wedge (q \vee r)$ and $(p \wedge q) \vee (p \wedge r)$
De Morgan's law	$\neg(p \wedge q)$ and $(\neg p) \vee (\neg q)$ $\neg(p \vee q)$ and $(\neg p) \wedge (\neg q)$
De Morgan's law for n terms	$\neg(p_1 \wedge p_2 \wedge \cdots \wedge p_n)$ and $(\neg p_1) \vee (\neg p_2) \vee \cdots \vee (\neg p_n)$ $\neg(p_1 \vee p_2 \vee \cdots \vee p_n)$ and $(\neg p_1) \wedge (\neg p_2) \wedge \cdots \wedge (\neg p_n)$
Absorption law	$p \wedge (p \vee q)$ and p $p \vee (p \wedge q)$ and p
Tautology law	$p \vee (\neg p)$ and T
Contradiction law	$p \wedge (\neg p)$ and F
Equivalence law	$p \Leftrightarrow q$ and $(p \Rightarrow q) \wedge (q \Rightarrow p)$
Implication law	$p \Rightarrow q$ and $(\neg p) \vee q$
Implication Negation law	$\neg(p \Rightarrow q)$ and $p \wedge (\neg q)$
Contrapositive law	$p \Rightarrow q$ and $(\neg q) \Rightarrow (\neg p)$

Table 0: Common pairs of equivalent statements.

These pairs of equivalent statements can be used to form "proofs" that various other, more complex, pairs of statements are in fact equivalent. Indeed, one way to establish that statement X is equivalent to statement Y is to create a chain of statements, all of which are equivalent, with X at the start of the chain and Y at the end, each one obtained from the

previous one by substituting out one expression for another, equivalent expression (either using a pair from the table above, or a pair that has already been established as being equivalent, either via a truth table, or by the method described here).

Here's an example. Suppose you want to show that $(p \wedge (p \Rightarrow q)) \Rightarrow q$ is equivalent simply to T (so, in the language we will introduce shortly, is a *tautology*), without using a truth table. In other words, you want to "reason out" that modus ponens is a valid logical inference, rather than "brute force" it. Here would be one possible approach:

$$
\begin{array}{lll}
(p \wedge (p \Rightarrow q)) \Rightarrow q & \text{is equivalent to} & \neg(p \wedge (p \Rightarrow q)) \vee q \quad \text{(Implication law (or, definition of implication))} \\
& \text{which is equivalent to} & ((\neg p) \vee \neg(p \Rightarrow q)) \vee q \quad \text{(De Morgan's law)} \\
& \text{which is equivalent to} & ((\neg p) \vee (p \wedge (\neg q))) \vee q \quad \text{(Negation of implication)} \\
& \text{which is equivalent to} & (((\neg p) \vee p) \wedge ((\neg p) \vee (\neg q))) \vee q \quad \text{(Distributive law)} \\
& \text{which is equivalent to} & ((p \vee (\neg p)) \wedge ((\neg p) \vee (\neg q))) \vee q \quad \text{(Commutative law)} \\
& \text{which is equivalent to} & (T \wedge ((\neg p) \vee (\neg q))) \vee q \quad \text{(Tautology law)} \\
& \text{which is equivalent to} & (((\neg p) \vee (\neg q)) \wedge T) \vee q \quad \text{(Commutative law)} \\
& \text{which is equivalent to} & ((\neg p) \vee (\neg q)) \vee q \quad \text{(Identity law)} \\
& \text{which is equivalent to} & (\neg p) \vee ((\neg q) \vee q) \quad \text{(Associative law)} \\
& \text{which is equivalent to} & (\neg p) \vee (q \vee (\neg q)) \quad \text{(Commutative law)} \\
& \text{which is equivalent to} & (\neg p) \vee T \quad \text{(Tautology law)} \\
& \text{which is equivalent to} & T \quad \text{(Domination law).}
\end{array}
$$

This seems like overkill: it would have been much faster, *in this particular case*, to use a truth table. But as the number of simpler propositions involved grows, the truth table approach becomes less and less desirable. For example, with two propositions (as we have here) the truth table has only $2^2 = 4$ rows; but with 20 propositions, the truth table has $2^20 \approx 1,000,000$ rows! At that point, it is completely impractical to use a truth table to verify an equivalence, and it is absolutely necessary to the pure reasoning illustrated in the example above.

1.7 Predicates

We've defined a statement to be an assertion that is either unambiguously true or unambiguously false. So "A human has walked on the moon" is a (true) statement, while "A Martian has walked in South Bend" is a (false) statement.

Returning to the mathematical world, what about something like "$x^2 + y^2 = 4$"? This is neither true nor false, because no specification has been made as to the values of x and y. If (x, y) is on the circle of radius 2 centered at the origin — say, $x = \sqrt{3}$, $y = -1$ — then the assertion is true, otherwise it is false.

A *predicate* is an assertion involving a variable or variables, whose truth or falsity is not absolute, but instead depends on the particular values the variables take on. So "$x^2 + y^2 = 4$"

is a predicate. Predicates abound in mathematics; we frequently are studying objects that depend on some parameter, and want to know for which values of the parameter some various assertions are true.

There are three ways in which predicates might be built up to become statements. One is by asserting an implication between predicates involving the same variables, of the form "if the first predicate is true, then the second must be also". Here's an example:

$$\text{"if } x-2=1 \text{ then } x^2-9=0\text{"}.$$

This is "p implies q" where p: "$x-2=1$" and q: "$x^2-9=0$" are both predicates, not statements. It happens to be a true statement: either

- $x-2\neq 1$, in which case "not p" is true, so "(not p) or q" is true,

or

- $x-2=1$, in which case $x=3$, so $x^2-9=0$, so q is true, making "(not p) or q" true.

This way of turning a predicate into a statement is somewhat boring: "if $x-2=1$ then $x^2-9=0$" amounts to saying "consider the predicate '$x^2-9=0$' in the specific situation where $x-2=1$". Two (much) more interesting ways of turning a predicate into a statement involve the notion of quantifying over some universal set, which we now discuss.

Quantifiers

Predicates may also become statements by adding *quantifiers*. One quantifier, "for all", says that the predicate holds for all possible values. As an example consider the (false) statement

$$\text{for every number } n,\ n \text{ is a prime number.}$$

We notate this statement as

$$(\forall n)p(n)$$

(read: "For all n, $p(n)$ (holds)") where $p(n)$ is the predicate "n is a prime number" — the "(n)" is added to p to indicate that p depends on the variable n. The formal reason the statement is false lies in the precise meaning that we assign to it: for any predicate $p(n)$, the statement "$(\forall n)p(n)$" is declared to be

- true, if $p(n)$ is true for *every* possible choice of n, and
- false, if there is even a single n for which $p(n)$ is not true.

The quantifier $\forall$ is referred to as the *universal quantifier*.

The *existential quantifier*, symbolically $\exists$, says that the predicate holds for *some* choice of the variable (but not necessarily for all of them). So with $p(n)$ as above, the true statement

$$(\exists n)p(n)$$

(read: "There exists n such that $p(n)$ is true") asserts that *some* number is prime. Formally, for any predicate $p(n)$, the statement "$(\exists n)p(n)$" is declared to be:

- true, if there is *at least one* n for which $p(n)$ is true, and
- false, if $p(n)$ is false for *every* n.

The universe of discourse

When one hears things like "for every n", or "there is an n", one should immediately ask "Where is one looking for n?" — the truth or otherwise of the associated statement may depend crucially on what the pool of possible n is. For example, consider the statement "There exists an x such that $x^2 = 2$" (or: "$(\exists x)r(x)$" where $r(x)$ is the predicate "$x^2 = 2$"). This is a true statement, if one is searching among real numbers — the value $x = \sqrt{2}$ witnesses the truth. On the other hand, if one is searching only among positive integers, then the statement becomes false — there is clearly no positive integer x with $x^2 = 2$. (Later, we'll talk about what happens if one is searching among rational numbers).

For this reason it is imperative, when using quantifiers, to know exactly what is the universe of possible values for the variable (or variables) of the predicate (or predicates) involved in the quantification. This is referred to as the *universe of discourse* of the variable. Usually, it is abundantly clear, from the context, what the universe of discourse is; if it is not clear, it needs to be made explicit in the quantification.

One way to make the universe of discourse explicit is to simply say what it is:

"With x running over positive real numbers, there exists x such that $x^2 = 2$"

or

"With the universe of discourse for x being positive real numbers,
there exists x such that $x^2 = 2$".

Another, more common way, is to build the universe of discourse into the quantification: e.g.,

"There exists a positive real x such that $x^2 = 2$".

Symbolically, this last statement could be written

"$(\exists x \in \mathbb{R}^+)(x^2 = 2)$".

Here, "$\mathbb{R}^+$" is a standard notation for the positive real numbers (we'll see this later), and the symbol "$\in$" is the set theory notation for "is an element of" (so "$x \in \mathbb{R}^+$" conveys that x lives inside the set of positive real numbers). We will have more to say on basic set theory later.

This last method of building the universe of discourse into quantification is especially useful when a statement involves multiple variables, each with a different universe of discourse. Consider, for example, the following statement, which essentially says that one can find a rational number as close as one wants to any real number:

"for every real number x, for every positive real number ε, there is a rational number r that is within ε of x".

There are three variables — x, ε and r — each with a different universe of discourse. The above rendering of the statement is much cleaner than the (equivalent):

"With the universe of discourse for x being real numbers, for ε being positive real numbers, and for r being rational numbers, for every x and ε there is r such that r is within ε of x".

Symbolically, the statement we are discussing might be succinctly expressed as:

$$\text{"}(\forall x \in \mathbb{R})(\forall \varepsilon \in \mathbb{R}^+)(\exists r \in \mathbb{Q})(-\varepsilon < x - r < \varepsilon)\text{"}.$$

Here, "$\mathbb{R}$" is a standard notation for the real numbers, and "$\mathbb{Q}$" is a standard notation for the rational numbers. If it is absolutely clear from the context (as it will be throughout most of this course) that all variables are real numbers (that is, that all universes of discourse are either the set of real numbers, or subsets thereof), then we could also write

$$\text{"}(\forall x)(\forall \varepsilon > 0)(\exists r \in \mathbb{Q})(-\varepsilon < x - r < \varepsilon)\text{"}.$$

The statement we are discussing happens to be true, although proving it will involve a great deal of machinery. We will come to it towards the middle of the semester. Notice that we read quantifiers, as in ordinary reading, from left to right.

A predicate needs a quantifier for *every* variable to turn into a statement. Consider, for example,

$$\text{"for all } x,\ xy = 0\text{"}.$$

This is not a valid statement. It is true if y happens to be 0, and it is false if y is not 0, so its truth or falsity depends on the choice of y. On the other hand, both

$$\text{"for all } x, \text{ for all } y,\ xy = 0\text{"} \text{ and } \text{"for all } x, \text{ there is a } y \text{ such that } xy = 0\text{"}$$

are statements (the first is false, the second is true).

Order of quantifiers

When a predicate with multiple variables gets turned into a statement by the addition of variables, the order in which we list the quantifiers is very important — typically a statement will change its meaning quite dramatically if we flip the order. For example, consider the predicate $p(m, n)$: "m is greater than n", or, more succinctly, "$m > n$". With the universe of discourse for all variables being the set of real numbers, the (true) statement

$$(\forall n)(\exists m)p(m, n), \quad \text{or} \quad (\forall n)(\exists m)(m > n)$$

says "For every number n, there is a number m such that m is greater than n". Flipping the order of the quantifiers leads to the false statement

$$(\exists m)(\forall n)(m > n),$$

or, "there is some number m such that every number is smaller than m".

For another example, let the variable x range over Major League baseball players, and the variable y range over Major league baseball teams. Let $p(x, y)$ be the predicate "player x is a shortstop for team y". Consider the following four statements which formally are similar-looking, but that translate into four *very* different statements in ordinary language:

- $(\forall x)(\exists y)p(x, y)$: for every player x, there is a team y such that x is the shortstop for y; in other words, every baseball player is *some* team's shortstop — false.
- $(\exists y)(\forall x)p(x, y)$: there is a team y such that every player x is the shortstop for y; in other words, there is some particular team such that *every* baseball player is that team's shortstop — false.
- $(\forall y)(\exists x)p(x, y)$: for every team y, there is a player x such that x is the shortstop for y; in other words, every team has a shortstop — true.
- $(\exists x)(\forall y)p(x, y)$: there is a player x such that every team y has x as its shortstop; in other words, there is some particular player who is *every* team's shortstop — false.

You should keep these absurd examples in mind as you work with quantifiers, and remember that it is very important to keep careful track of the order in which you introduce them — at least if you care about the meaning of the statements that you get in the end!

For a slightly more complicated example, here is what is called the *Archimedean principle* of positive real numbers:

"If N and s are positive numbers, there's a positive number t with $ts > N$."

(This is true no matter how big N is or how small s is.) We can take the predicate $p(N, s, t)$ to be "$ts > N$", and then encode the statement as $(\forall N)(\forall s)(\exists t)p(N, s, t)$. Note that this is implicitly assuming that we have agreed that we are working in the world of positive real numbers; if we were instead working in the world of all real numbers, we could write something like:

$$(\forall N)(\forall s)\left[((N > 0) \wedge (s > 0)) \Rightarrow (\exists t)((t > 0) \wedge p(N, s, t))\right]$$

(which we might read as, "For every N and s, if N and s are positive, then there is a t such that both of the following hold: t is positive and $ts > N$").

Negation of predicates

The operation of negation has a very simple effect on quantifiers: it turns $\forall$ into $\exists$ and vice versa, while bringing the negation inside the predicate. Think about $(\forall x)p(x)$ and $(\exists x)(\neg p(x))$, for example. If the first is true then $p(x)$ holds for every x, so $\neg p(x)$ holds for no x, so the second is false, while if the first is false then there is an x for which $p(x)$ doesn't hold, and so the second is true. This argument shows that

$$\neg((\forall x)p(x)) \quad \text{is equivalent to} \quad (\exists x)(\neg p(x)),$$

and similarly we can argue that

$$\neg((\exists x)p(x)) \quad \text{is equivalent to} \quad (\forall x)(\neg p(x)).$$

If the universe of discourse for the variable x is finite, then the two equivalences above are just DeMorgan's laws, rewritten. Indeed, if the possible choices for x are $x_1, x_2, \ldots, x_n$, then $(\forall x)p(x)$ is the same as $p(x_1) \wedge p(x_2) \wedge \cdots \wedge p(x_n)$, and so by DeMorgan's law the statement $\neg((\forall x)p(x))$ is the same as $(\neg p(x_1)) \vee (\neg p(x_2)) \vee \cdots \vee (\neg p(x_n))$, which is just another way of saying $(\exists x)(\neg p(x))$. Similarly, one can argue that $\neg((\exists x)p(x))$ means the same as $(\forall x)(\neg p(x))$. So negation of quantifiers can be thought of as DeMorgan's law generalized to the situation where the number of predicates being and-ed or or-ed is not necessarily finite.

What about a statement with more quantifiers? Well, we can just repeatedly apply what we have just established, working through the quantifiers one by one. For example, what is the negation of the statement $(\exists x)(\exists y)(\forall z)p(x, y, z)$?

$$\begin{array}{lcl} \neg((\exists x)(\exists y)(\forall z)p(x,y,z)) & \text{is equivalent to} & (\forall x)(\neg((\exists y)(\forall z)p(x,y,z))) \\ & \text{which is equivalent to} & (\forall x)(\forall y)(\neg((\forall z)p(x,y,z))) \\ & \text{which is equivalent to} & (\forall x)(\forall y)(\exists z)(\neg p(x,y,z)). \end{array}$$

The homework will included some more involved examples, such as negating the Archimedean principle (and interpreting the result).

1.8 Tautologies

A *tautology* is a statement that is built up from various shorter statements or quantified predicates p, q, r, ..., that has the property that no matter what truth value is assigned to each of the shorter statements, the compound statement is true.

A simple example is "There either is life on Mars, or there is not", which can be expressed as $p \vee (\neg p)$ where p is the statement "There is life on Mars". If p is true then so is $p \vee (\neg p)$, while if p is false then $\neg p$ is true, so again $p \vee (\neg p)$ is true. In general the tautology $p \vee (\neg p)$ is referred to as the *law of the excluded middle* (there is no middle ground in logic: either a statement is true or it is false).

Looking at the truth table of the bidirectional implication $\Leftrightarrow$, it should be evident that if p and q are any two compound statements that are build up from the same collection of

shorter statements and that have the same truth tables, then $p \Leftrightarrow q$ is a tautology; so for example,

$$\neg(p \vee q \vee r) \Leftrightarrow (\neg p) \wedge (\neg q) \wedge (\neg r)$$

is a tautology. A tautology can be thought of as indicating that a certain statement is true, in an unqualified way; the above tautology indicates the truth of one of De Morgan's laws.

An important tautology is

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

(easy check: this could only be false if q is false and both p and $p \Rightarrow q$ are true; but if q is false and p is true, $p \Rightarrow q$ must be false; so there is no assignment of truth values to p and q that makes the compound statement false). This tautology indicates the truth of the most basic rule of logical deduction, that if a statement p is true, and it is also true that p implies another statement q, then it is correct to infer that q is true. This is called *modus ponens*.

For example, if you know (or believe) the truth of the implication

"If I go out in the rain without a coat on, I will get wet"

and I also give you the information that I go out in the rain without a coat on, then it is legitimate for you to reach the conclusion that I get wet.

The ideas raised in this short section are fundamental to mathematics. A major goal of mathematics is to discover *theorems*, or statements that are true. In other words, a theorem is essentially a tautology. Most complex theorems are obtained by

- starting with a collection of simpler statements that are already known to be true (maybe because they have already been shown to be true, or maybe because they are assumed to be true, because they are axioms of the particular mathematical system under discussion), and then
- deducing the truth of the more complex statement via a series of applications of rules of inference, such as modus ponens.

This process is referred to as *proving* the theorem. Almost every result that we use in this class, we will prove; this is something that sets Math 10850/60 apart from courses such as Math 10550/60 (Calculus 1/2), which explain and apply the techniques of calculus, without laying a rigorous foundation. The notion of proof will be explored in more detail in Section 2.

2 An introduction to proofs

Before we talk about proofs, we give a very brief guide to the basics of a mathematical theory.

2.1 The basics of a mathematical theory

We begin with a collection of

- **Axioms**: propositions that we agree in advance are true.

Axioms may be thought of as the fundamental building blocks of any mathematical theory. They are usually chosen to be simple, intuitive statements that capture the essential structure of the objects that we want in our theory. You may be a little dissatisfied by a supposedly "rigorous" mathematical course starting out by making unprovable "assumptions"; but remember, we can do *nothing* unless we have *something* to build from!

A famous example of a set of axioms is the set of five that Euclid used in his book *Elements* to lay down the ground rules of the mathematical system that we now call "Euclidean geometry"[15][16][17]:

1. A straight line segment can be drawn joining any two points.
2. Any straight line segment can be extended indefinitely in a straight line.
3. Given any straight line segment, a circle can be drawn having the segment as radius and one endpoint as center.
4. All right angles are congruent.
5. If two lines are drawn which intersect a third in such a way that the sum of the inner angles on one side is less than two right angles, then the two lines inevitably must intersect each other on that side if extended far enough. (This axiom is equivalent to what is known as the "parallel postulate": parallel lines don't meet.)

[15] Statements taken from http://mathworld.wolfram.com/EuclidsPostulates.html.

[16] Why not just "Geometry"? For over two millennia after Euclid proposed his five axioms, mathematicians struggled with the fifth of them. The first four were obvious, simple, necessary building blocks of geometry, but the fifth seemed overly complex. Generations of mathematicians attempted to reduce the complexity of the axioms by *proving* (in the sense that we are using in this section) the fifth axiom from the first four. All attempts were unsuccessful, and in 1823, Janos Bolyai and Nicolai Lobachevsky independently discovered why: there are systems of geometry that satisfy the first four axioms of Euclid, but not the fifth; that is, there are entirely consistent notions of geometry in which sometimes parallel lines *do* eventually meet. So to describe geometry in the plane, as Euclid was trying to do, something like the complex fifth axiom is needed. Systems of geometry that satisfy the first four axioms of Euclid, but not the fifth, are referred to as "non-Euclidean geometries".

[17] These are actually Euclid's five *postulates*; his *axioms* define the basic properties of equality. We will mention these later.

(As another example of a set of axioms, consider the fundamental building blocks of the United States laid down by the founding fathers in the Declaration of Independence:

> "We hold these truths to be self-evident, that all men are created equal, that they are endowed by their creator with certain unalienable rights, that among these are life, liberty and the pursuit of happiness.")

Along with axioms, we have

- **Definitions**: statements that specify what particular terms mean.

Think of the list of definitions as a dictionary, and the list of axioms as a rule-book. As an example, Euclid presents four definitions, explaining what words like "point" and "line" (used in the axioms) mean[18][19]:

1. A *point* is that which has no part.
2. A *line* is a breadthless length.
3. The extremities of lines are points.
4. A straight line lies equally with respect to the points on itself.

Once we have Axioms and Definitions, we move on to the meat of a mathematical theory, the

- **Theorems**: statements whose truth follows from the axioms and the definitions via rules of logical inference.

If the definitions are the dictionary, and the axioms are the rule-book, then the theorems are the structures that can be legitimately formed from the words, "legitimately" meaning following the rules of the rule-book. As an example, here is Euclid's famous Theorem I.47, the *Pythagorean theorem*[20]:

> **Theorem**: In right-angled triangles the square on the side opposite the right angle equals the sum of the squares on the sides containing the right angle.

How do we know that the Pythagorean theorem is indeed a theorem, that is, is indeed a statement that follows Euclid's rule-book? We know, because Euclid provided a *proof* of the theorem, and once we follow that proof we have no choice but to accept that if we agree with the axioms, we must agree with the Pythagorean theorem.

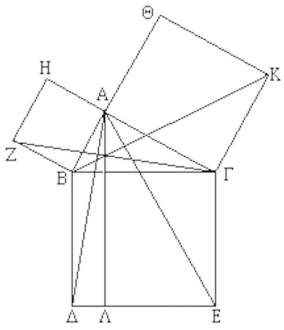
[18] Statements taken from http://www-history.mcs.st-and.ac.uk/HistTopics/Euclid_definitions.html.

[19] As pointed out at http://www-history.mcs.st-and.ac.uk/HistTopics/Euclid_definitions.html, these seem a little strange, as Euclid seems to be defining "point" twice (first and third definitions), and "line" twice (second and fourth). Fortunately we don't need to worry about this, as Math 10850 is not a course in Euclidean geometry!

[20] Statement from https://www.cut-the-knot.org/pythagoras/Proof1.shtml.

- **Proofs**: the truth of a statement p is established via a *proof*, a sequence of statements, ending with the statement p, each of which is either
  - an axiom,
  - an instance of a definition,
  - a theorem that has previously been proved, or
  - a statement that follows from some of the previous statements via a rule of inference.

For completeness, here's a treatment of Euclid's proof of the Pythagorean theorem, as presented in Wikipedia:



1. Let ACB be a right-angled triangle with right angle CAB.
2. On each of the sides BC, AB, and CA, squares are drawn, CBDE, BAGF, and ACIH, in that order. The construction of squares requires the immediately preceding theorems in Euclid, and depends upon the parallel postulate.[14]
3. From A, draw a line parallel to BD and CE. It will perpendicularly intersect BC and DE at K and L, respectively.
4. Join CF and AD, to form the triangles BCF and BDA.
5. Angles CAB and BAG are both right angles; therefore C, A, and G are collinear. Similarly for B, A, and H.
6. Angles CBD and FBA are both right angles; therefore angle ABD equals angle FBC, since both are the sum of a right angle and angle ABC.
7. Since AB is equal to FB and BD is equal to BC, triangle ABD must be congruent to triangle FBC.
8. Since A-K-L is a straight line, parallel to BD, then rectangle BDLK has twice the area of triangle ABD because they share the base BD and have the same altitude BK, i.e., a line normal to their common base, connecting the parallel lines BD and AL. (lemma 2)
9. Since C is collinear with A and G, square BAGF must be twice in area to triangle FBC.
10. Therefore, rectangle BDLK must have the same area as square BAGF = AB^2.
11. Similarly, it can be shown that rectangle CKLE must have the same area as square ACIH = AC^2.
12. Adding these two results, $AB^2 + AC^2 = BD \times BK + KL \times KC$
13. Since BD = KL, $BD \times BK + KL \times KC = BD(BK + KC) = BD \times BC$
14. Therefore, $AB^2 + AC^2 = BC^2$, since CBDE is a square.

A proof of the Pythagorean, by Euclid

2.2 Proofs, more slowly

In the last section we said what a proof is, slightly formally. Now we'll look much more closely at the notion of proof, starting more informally. Quoting Hutchings [2]:

> "A mathematical proof is an argument which convinces other people that something is true. Math isn't a court of law, so a 'preponderance of the evidence' or 'beyond any reasonable doubt' isn't good enough. In principle we try to prove things beyond any doubt at all."

We've discussed in the last section the grammar and vocabulary that will be used to present proofs; nor we turn to the rather more exciting and challenging process of actually production proofs. Let's begin with an example, a dialogue between Alice and Bob (two frequently occurring characters in mathematics) taken verbatim from Hutchings [2] (the footnotes are mine).

Alice I've just discovered a new mathematical truth!

Bob Oh really? What's that?

Alice For every integer x, if x is even, then x^2 is even.

Bob Hmm... are you sure that this is true?

Alice Well, isn't it obvious?

Bob No, not to me.

Alice OK, I'll tell you what. You give me any integer x, and I'll show you that the sentence 'if x is even, then x^2 is even' is true. Challenge me.

Bob (eyes narrowing to slits): All right, how about $x = 17$.

Alice That's easy. 17 is not even, so the statement 'if 17 is even, then 17^2 is even' is vacuously true[21]. Give me a harder one.

Bob OK, try $x = 62$.

Alice Since 62 is even, I guess I have to show you that 62^2 is even.

Bob That's right.

Alice (counting on her fingers furiously): According to my calculations, $62^2 = 3844$, and 3844 is clearly even...

Bob Hold on. It's not so clear to me that 3844 is even. The definition says that 3844 is even if there exists an integer y such that $3844 = 2y$. If you want to go around saying that 3844 is even, you have to produce an integer y that works.

Alice How about $y = 1922$.

Bob Yes, you have a point there. So you've shown that the sentence 'if x is even, then x^2 is even' is true when $x = 17$ and when $x = 62$. But there are billions[22] of integers that x could be. How do you know you can do this for every one?

Alice Let x be any integer.

Bob Which integer?

[21] Think about the implication "p implies q". It's the same as "(not p) or q". If p is false (as in this case: p is the statement "17 is an even number") then "not p" is true, so the disjunction "(not p) or q" is indeed true. And this is going to be the case regardless of what p and q are. This is what Alice means when she says that 'if 17 is even, then 17^2 is even' is vacuously true; and more generally, the statement 'p implies q' is vacuously true whenever p is false. This idea is so universally understood, that we (almost) never refer to it when proving implications.

On the other hand, if p is true (as will happen in a moment, when x is set to 62), then "not p" is false, so the only way we are going to show that the disjunction "(not p) or q" is true is by showing that q is true. Of course, we get to use the fact that p is true in the argument, which will probably be a big help!

[22] Billions? Actually infinitely many!

Alice Any integer at all. It doesn't matter which one. I'm going to show you, using only the fact that x is an integer and nothing else, that if x is even then x^2 is even.

Bob All right ... go on.

Alice So suppose x is even.

Bob But what if it isn't?

Alice If x isn't even, then the statement 'if x is even, then x^2 is even' is vacuously true. The only time I have anything to worry about is when x is even.

Bob OK, so what do you do when x is even?

Alice By the definition of 'even', we know that there exists at least one integer y such that $x = 2y$.

Bob Only one, actually.

Alice I think so[23]. Anyway, let y be an integer such that $x = 2y$. Squaring both sides of this equation, we get $x^2 = 4y^2$. Now to prove that x^2 is even, I have to exhibit an integer, twice which is x^2.

Bob Doesn't $2y^2$ work?

Alice Yes, it does. So we're done.

Bob And since you haven't said anything about what x is, except that it's an integer, you know that this will work for any integer at all.

Alice Right.

Bob OK, I understand now.

Alice So here's another mathematical truth. For every integer x, if x is odd, then x^2 is...

Two comments are in order here (the first paraphrased from [2]):

- A proof is an explanation which convinces someone that a statement is true. A good proof also helps them understand *why* it is true.
- The proof we have just given is long (though we will soon start presenting proofs much more compactly) but, importantly, it has only *finite* length. And yet, it is verifying *infinitely many* things: 2^2 is even, 4^2 is even, 6^2 is even, 8^2 is even, and so on. This is a first illustration of the remarkable power of mathematical logic — we have, in a finite amount of space, verified infinitely many truths!

[23] But it doesn't really matter — we just need *at least one* such y.

In this course, we will prove (almost) everything that we discuss. It might make sense, then, that we start the semester off with a thorough introduction to proofs and logic. We won't, though. We will instead learn about proofs mostly by *doing* them. For the rest of this section, we'll just discuss fairly briefly some of the common techniques of proofs, with simple examples to illustrate some of them. Then we will move on to the *real* examples, and develop a familiarity with proofs by exposure and immersion.

2.3 A summary of basic proof techniques

Here, taken from [2], is a table ("Table 1: Logic in a nutshell") that summarizes, informally, "just about everything you will need to know about logic. It lists the basic ways to prove, use, and negate every type of statement. In boxes with multiple items, the first item listed is the one most commonly used. Don't worry if some of the entries in the table appear cryptic at first; they will make sense after you have seen some examples."

Statement	Ways to Prove it	Ways to Use it	How to Negate it
p	• Prove that p is true. • Assume p is false, and derive a contradiction.	• p is true. • If p is false, you have a contradiction.	not p
p and q	• Prove p, and then prove q.	• p is true. • q is true.	(not p) or (not q)
p or q	• Assume p is false, and deduce that q is true. • Assume q is false, and deduce that p is true. • Prove that p is true. • Prove that q is true.	• If $p \Rightarrow r$ and $q \Rightarrow r$ then r is true. • If p is false, then q is true. • If q is false, then p is true.	(not p) and (not q)
$p \Rightarrow q$	• Assume p is true, and deduce that q is true. • Assume q is false, and deduce that p is false.	• If p is true, then q is true. • If q is false, then p is false.	p and (not q)
$p \iff q$	• Prove $p \Rightarrow q$, and then prove $q \Rightarrow p$. • Prove p and q. • Prove (not p) and (not q).	• Statements p and q are interchangeable.	(p and (not q)) or ((not p) and q)
$(\exists x \in S)\ P(x)$	• Find an x in S for which $P(x)$ is true.	• Say "let x be an element of S such that $P(x)$ is true."	$(\forall x \in S)$ not $P(x)$
$(\forall x \in S)\ P(x)$	• Say "let x be any element of S." Prove that $P(x)$ is true.	• If $x \in S$, then $P(x)$ is true. • If $P(x)$ is false, then $x \notin S$.	$(\exists x \in S)$ not $P(x)$

Table 1: Logic in a nutshell.

Note that in the fifth row (Statement $p \iff q$), second column (Ways to Prove it), you should read "Prove p and q" to mean "prove p (is true), and, in a separate argument, prove q (is true)", rather than "prove the conjection 'p and q"', and the same for "Prove (not p) and (not q)". Also, for all statements involving quantifiers, the table takes "S" to be the universe of discourse.

2.4 Examples of proofs

Here we present a list of examples of using proof techniques. The statements we will prove will all be very simple ones. This is fine; the point of this section is not to derive deep truths, but rather to illustrate the techniques that we will be using repeatedly as the year goes on, on many rather more serious examples.

Most of the examples here are taken from [2], essentially verbatim.

An example of proving a "for every" statement

This first example involves proving a 'for every' statement, but can also be thought of as proving an implication — an 'if ... then' statement. Along the way we will see how to leverage knowing that a 'there exists' statement happens to be true. All of these ideas have been already introduced in Alice and Bob's dialogue.

Example: Give a proof that for every integer x, if x is odd, then $x+1$ is even.

This is a 'for every' statement, so the first thing we do is write

> Let x be any integer.

Think of x as a particular integer here ... 11, or -2, or 1729, just one that we are not explicitly naming. If we can show that 'if x is odd then $x+1$ is even', using *only* that fact that x is an integer, and not using any properties of a particular integer, then we will have shown that *for all integers* x, if x is odd, then $x+1$ is even, as required.

Again, we have to show, using only the fact that x is an integer, that if x is odd then $x+1$ is even. As discussed earlier, this implication is vacuously true if x happens to be even, so we get to assume from here on that x is odd. So we write

> Suppose x is odd.

We must somehow use this assumption to deduce that $x+1$ is even. Now the statement 'x is odd' means 'there exists an integer y such that $x = 2y+1$'. So we get to leverage the assumption 'x is odd', by writing x in the form $2y+1$, where y is *known for certain* to be an integer[24]. So we write

> Let y be an integer such that $x = 2y+1$ (such a y exists because x is odd).

[24] Which integer? We don't know; it depends on what x is, and we haven't said explicitly what x is. But that's ok; all we are going to use is that y is *some* integer.

Note that we don't just write "Let y be an integer such that $x = 2y + 1$" and leave it at that; we need to justify that such a y actually exists, hence the comment in parentheses.

Now we want to prove that $x + 1$ is even. In other words, we want to show that there exists an integer y such that $x + 1 = 2y$. **BUT**, we have to be careful! The name 'y' is already in use; it is a witness to the fact that x is odd. So in telling ourselves what we mean by the statement '$x + 1$ is even', we need to use a different name for the witness. That's fine, as there are lots of available names. Let's use w. We write

> To show that $x + 1$ is even, we need to find an integer w such that $x + 1 = 2w$.

How do we find such a w? Well, we know $x = 2y+1$, so $x+1 = (2y+1)+1 = 2y+2 = 2(y+1)$, so it looks like we have a candidate w, namely $y + 1$. We finish the proof by writing

> Adding 1 to both sides of $x = 2y + 1$, we get
>
> $$x + 1 = (2y + 1) + 1, \quad \text{or, equivalently,} \quad x + 1 = 2(y + 1).$$
>
> Since y is an integer, so is $y + 1$. Taking $w = y + 1$ we see that $x + 1$ is even.

Notice that we haven't just explained how $x + 1$ gets written in the forn $2w$ — we have also justified that the w in question *is an integer*, which is part of the definition of $x + 1$ being even.

The proof is now done. It's typical to insert a symbol to indicate that a proof has come to an end. Historically that symbol was often the string "Q.E.D." (*quod erat demonstrandum*, Latin for "what was to be shown"[25]). It is much more common these days to use the symbol "□".

Here's how the proof would look without the extraneous discussion:

Claim: For every integer x, if x is odd, then $x + 1$ is even.

Proof: Let x be any integer. Suppose x is odd. Let y be an integer such that $x = 2y + 1$ (such a y exists because x is odd). To show that $x + 1$ is even, we need to find an integer w such that $x + 1 = 2w$. Adding 1 to both sides of $x = 2y + 1$, we get

$$x + 1 = (2y + 1) + 1, \quad \text{or, equivalently,} \quad x + 1 = 2(y + 1).$$

Since y is an integer, so is $y + 1$. Taking $w = y + 1$ we see that $x + 1$ is even. □

Notice that

- the proof is written in *prose*. When read out loud (reading the symbols in the usual way we say them; "equals" for "=", et cetera) it makes perfect sense as a paragraph in ordinary language, and

[25] Or: "quite easily done".

- every step of the proof is justified: every line follows from previous ones, by an application of a definition, or by an appeal to something which has been established as true earlier in the prrof.

These two properties are hallmarks of a good proof, and you should strive towards them in your proof writing!

An example of a proof involving an "and" statement

Example: Write a proof that for every integer x and[26] for every integer y, if x is odd and y is odd then xy is odd.

As before we begin

> Let x and y both be integers. Suppose x and y are both odd.

Because we are assuming that *both x and y* are odd, we can leverage *both* of these statements to provide alternate ways of representing *both x and y*. And since the fact that both x and y are odd is the *only* thing that we know about these two numbers, using the definition of oddness to represent them in terms of other integers is basically the *only* thing that we can do to have any hope of proceeding with this particular proof. So we write

> Then there are integers w and v such that $x = 2w + 1$ and $y = 2v + 1$.

Two things are worth noting here. The first is that to prove the result in its full generality, we have to use different names for the witnesses that x and y are odd. It would not do to say that there is integer w such that $x = 2w + 1$ and integer w such that $y = 2w + 1$. If we did this, we would be implicitly forcing x and y to be equal for the rest of the proof, and would end up proving the much more restrictive statement that if x, y are both odd, and equal, then xy is odd — in other words, if x is odd so is x^2.

The second point worth noting is that we are already engaging in a very common practice in the writing of mathematical proofs, that of skipping steps that are obvious enough not to be spelled out. In this case, what we really should have written is something like:

> Since x is odd and y is odd, it follows that in particular x is odd. So there exists an integer w such that $x = 2w + 1$.
>
> Also, since x is odd and y is odd, it follows that in particular y is odd. So there exists an integer v such that $y = 2v + 1$.

[26] This 'and' is not a logical 'and'; it is just there to make the statement easier to read. Some people might write "for all integers x, y" here. Symbolically, what we are being asked to prove involves a double universal qualifier, and can be written

$$(\forall x \in \mathbb{Z})(\forall x \in \mathbb{Z})(((x \text{ odd}) \wedge (y \text{ odd})) \Rightarrow (xy \text{ odd})).$$

That is, we should have inferred "p" from "p and q", and drawn our first conclusion; and then inferred "q" from "p and q", and drawn our second conclusion. But this really is overkill![27]

We want to conclude that xy is odd, and there is really only one thing we can do: use the two expressions we have derived for x and y to derive a new expression for xy, and see if we can "massage" it into the form "twice an integer, plus one". That's easily done, so we go straight to the formal write-up:

> We now have
>
> $$xy = (2w+1)(2v+1) = 4wv + 2w + 2v + 1 = 2(2wv + w + v) + 1.$$
>
> Since v and w are integers, so is $2wv + w + v$, so we conclude that xy is odd. □

Here's the full proof in one pass:

> Let x and y both be integers. Suppose x and y are both odd. Then there are integers w and v such that $x = 2w + 1$ and $y = 2v + 1$.
>
> We now have
>
> $$xy = (2w+1)(2v+1) = 4wv + 2w + 2v + 1 = 2(2wv + w + v) + 1.$$
>
> Since v and w are integers, so is $2wv + w + v$, so we conclude that xy is odd. □

An example of an "if and only if" proof

Remember that "p if and only if q", sometimes abbreviated to "p iff q", is shorthand for the conjunction

(p implies q) and (q implies p).

As such, proofs of if-and-only-if statements usually require two distinct parts:

- a proof that p implies q,

and

- a proof that q implies p.

In the example we are about to give, the two parts are quite similar, but in general they might use quite different techniques.

Claim: For every integer x, x is even if and only if $x + 1$ is odd.

Proof: We begin by proving that for every integer x, if x is even then $x + 1$ is odd. Indeed, let x be any even integer. There is an integer y such that $x = 2y$, so $x + 1 = 2y + 1$, showing that $x + 1$ is odd.

[27] Knowing what steps to skip, because they are obvious enough not to be spelled out, is a witchy art, not a science, and depends a lot on your own mathematical maturity, and your understanding of the mathematical maturity of your audience. It is an art that we will feel our way into as the year goes on.

Next we prove that for every integer x, if $x+1$ is odd then x is even. Indeed, let x be any integer such that $x+1$ is odd. There is an integer y such that $x+1=2y+1$, so $x=2y$, showing that x is even. □

Notice that the proof was written quite compactly, with some of the justifications elided. For example, we didn't say that it is from the definition of evenness that "x is even" allows us to conclude "there is an integer y with $x=2y$". This is an example of me understanding my audience — we have seen a number of proofs now involving evenness and oddness, so I don't feel the need any more to spell out that I am using the definition. Nonetheless, the proof is still written *using full sentences*, with *no substantial step left unjustified*.

The value of this if-and-only-if proof is that now we can conclude that for any integer x, the statements 'x is even' and '$x+1$ is odd' are interchangeable; this means that we can take any true statement and replace some occurrences of the phrase 'x is even' with the phrase '$x+1$ is odd' to get another true statement; and, more importantly, we can do this *without having to provide a proof to justify the exchange* — once the equivalence has been proven once, it remains true for all time, and can be used as part of our arsenal of basic true statements. For example, in their dialogue Alice and Bob proved that

> For every integer x, if x is even then x^2 is even.

Since "x is even" is now known to be interchangeable with "$x+1$ is odd", we can conclude, without need for further proof, that the following is also a true statement:

> For every integer x, if $x+1$ is odd then x^2 is even.

Proof by cases

Sometimes it is very helpful to break a proof into cases, with the different possible cases being treated differently (sometimes slightly differently, sometimes substantially so). Let's start with an example (from which the general method should be very apparent), before discussion the matter more formally.

Example: Prove that for every integer x, the number $x(x+1)$ is even.

Some integers are even, some are odd, none are both, and every integer is one of the two.[28] So it makes a great deal of sense to consider first what happens when x is even, and then what happens when x is odd. If we can establish that

- if x is even, then $x(x+1)$ is even

[28] This is obvious, no? Formally, it requires a proof: x being even means that there is an integer y with $x=2y$, while x being odd means that there is an integer y' with $x=2y'+1$. It's not unreasonable that there might be some number z that can be expressed both as $2y$ and $2y'+1$, for some integers y, y'; and/or that there might be some number z that *cannot* be expressed as either $2y$ or $2y'+1$, for any integers y, y'. Of course, neither such z exists, but it actually takes some effort to prove. We'll defer it until we have seen prove by induction.

and

- if x is odd, then $x(x+1)$ is even

then we will have covered all possibilities, and shown that no matter what sort of integer x is, $x(x+1)$ is even.

The details of the proof are quite easy, so what follows is mainly introduced as a template for the presentation of proofs by cases.

Proof: We consider two cases.

Case 1, x even In this case there is an integer y such that $x = 2y$. We have $x+1 = 2y+1$, so

$$x(x+1) = (2y)(2y+1) = 4y^2 + 2y = 2(2y^2 + y).$$

Since $2y^2 + y$ is an integer, this shows that in this case $x(x+1)$ is even.

Case 2, x odd In this case there is an integer y such that $x = 2y+1$. We have $x+1 = 2y+2$, so

$$x(x+1) = (2y+1)(2y+2) = 4y^2 + 6y + 2 = 2(2y^2 + 3y + 1).$$

Since $2y^2 + 3y + 1$ is an integer, this shows that in this case $x(x+1)$ is even.

The two cases cover all possibilities, so we conclude that $x(x+1)$ is always even. □

Formally, here's what's going on in a proof by cases: we are trying to prove the implication "P implies q", and we realize that the predicate P can be expressed as "p_1 or p_2 or $\cdots$ or p_n", for some simpler predicates $p_1, \ldots, p_n$. So really what we are trying to prove is that

$$(p_1 \text{ or } p_2 \text{ or } \cdots \text{ or } p_n) \text{ implies } q. \tag{1}$$

(In the example just given, P is "x is an integer", p_1 is "x is an even integer" and p_2 is "x is an odd integer"). Instead of proving this single implication, we proceed by cases, proving each of the implications "p_1 implies q", "p_2 implies q", et cetera. From this we immediately conclude that the following statement is true:

$$(p_1 \text{ implies } q) \text{ and } (p_2 \text{ implies } q) \text{ and } \cdots \text{ and } (p_n \text{ implies } q). \tag{2}$$

Is this enough to conclude (1)? It's left as an exercise for the reader[29] to argue that in fact (1) and (2) are equivalent.

Proofs by cases are often quite tedious![30]. For example:

Claim: If $1 \leq n \leq 40$ then $n^2 - n + 41$ is a prime number.

Proof: The hypothesis here is the disjunction (the "or") of 40 separate cases: $n = 1$, $n = 2$, et cetera. So to complete the proof, it's enough to check each of those cases in turn:

[29] This is often code for: the author is too lazy to write down the details. But in these notes, "exercise for the reader" will usually be code for: look out for this on a homework/quiz/exam.

[30] I've seen plenty of research papers where at one point the authors have to deal with Case 7, Subcase C, Subsubcase viii, or some such!

- Case 1 ($n = 1$): Here $n^2 - n + 41 = 41$, which is indeed a prime number.
- Case 2 ($n = 2$): Here $n^2 - n + 41 = 43$, which is again a prime number.
- Case 3 ($n = 3$): Here $n^2 - n + 41 = 47$, again a prime number.
- Cases 4 through 39 ($n = 4$ through 39): Left to reader[31]
- Case 40 ($n = 40$): Here $n^2 - n + 41 = 1601$, which is a prime number[32].

□

Indirect proofs (contradiction and contrapositive)

All of the proof examples we have presented so far have been what are usually called *direct* proofs: we verified various implications of the form "p implies q" by assuming that p is true, and then *directly* arguing, via definitions, rules of logic, and earlier established truths, that q inevitably must be true.

Sometimes it is easier to argue indirectly. The most common form of an indirect argument is *proof by contradiction*: suppose that we want to prove that the statement p is true. We begin by assuming that p is false. We then try to deduce a contradiction, that is, some statement q which we know is false. If we succeed, then our assumption that p is false must be wrong! So p must be true, and our proof is finished.

To be a little more specific: often, when arguing that an implication "p implies q" is true, we begin by assuming that p is true and q is false (the one situation in which the implication "p implies q" is false) and derive a contradiction, meaning: we deduce that some statement r is true, and also that its negation "not r" is true, so that "r and (not r)" is true. This can't be (the truth value of "r and (not r)" is always false). So the only possible conclusion is that it is *not* the case that p is true and q is false, which is the same as saying that it *is* the case that "p implies q" is true.

Example (Here the universe of discourse for all variables is the set of real numbers): Prove the statement "If $5x + 25y = 2019$, then at least one of x, y is not an integer".

Proof: We argue by contradiction. Let us assume both that $5x + 25y = 2019$ and that both of x, y are integers (this is the negation of "at least one of x, y is not an integer"). We have that

$$5x + 25y = 5(x + 5y),$$

so $5x + 25y$ is a multiple of 5; and since $5x + 25y = 2019$, this says that 2019 is a multiple of 5. But also, by a direct calculation, we see that 2019 is *not* a multiple of 5. We have arrived at a contradiction, and so conclude that the statement we are trying to prove is indeed true. □

Some comments are in order:

[31] Here really because of the author's laziness.

[32] How do I know? I asked https://www.isprimenumber.com/prime/1601.

- Why didn't we try a direct proof? Because assuming the truth of the hypothesis in this case (that $5x + 25y = 2019$) gives us very little to work with — it tells us nothing specifically about x and y.

- In this example of proof by contradiction, we had p:"$5x + 25y = 2019$", q:"at least one of x, y is not an integer", and r:"2019 is a multiple of 5".

Another indirect proof technique is *proof by contrapositive*: this involves proving "p implies q" by giving a *direct* proof of the statement "(not q) implies (not p)". This is the contrapositive of, and equivalent to, "p implies q). By "giving a direct proof" I mean assuming "not q" and then using axioms, definitions, logical equivalences and rules of inference to deduce "not p".

Example: Prove "if mn is even, then either m is even or n is even".

Proof: We give a direct proof of the contrapositive statement: if both m and n are odd, then mn is odd.

Let us assume that m and n are odd. Then there are whole numbers k and ℓ with $m = 2k + 1$ and $n = 2\ell + 1$. We have

$$\begin{aligned} mn &= (2k+1)(2\ell+1) \\ &= 2k\ell + 2k + 2\ell + 1 \\ &= 2(k\ell + k + \ell) + 1. \end{aligned}$$

Since $k\ell + k + \ell$ is a whole number, mn is odd. □

A few comments are in order on this example:

- Proof by contrapositive can be thought of as a special case of proof by contradiction: we assume p and "not q", and reach the contradiction "p and (not p).

- Why didn't we try a direct proof here? Because, as in the last example, assuming the truth of the hypothesis in this case (that mn is even) gives us very little to work with — it tells us nothing specifically about m and n.

- Most of the serious proofs that we will see in this course will either be proofs by contradiction, or proofs by contrapositive.

Various other examples of proving implications

Most of the theorems we will prove will have statements of the form "if X holds then Y holds", where X is a string of assumptions, which we will call the *hypotheses* of the theorem, and Y is the *conclusion*. This is an implication: "X implies Y". So in discussing proofs, we will mostly be concerned with ways of rigorously justifying implications.

Suppose we are faced with the implication "p implies q", and we want to prove that it is valid. Here are two more proof techniques we can use. Both of them are slightly degenerate, but important to know about.

- **Trivial proof**: If we know q is true then $p \Rightarrow q$ is true regardless of the truth value of p.

  Example: (For this example, and most subsequent examples in this section, the universe of discourse for all variables is the set of positive natural numbers. The exceptions to this convention will be noted as we come across them). "If n is a prime number, then $n - n = 0$". Here the conclusion "$n - n = 0$" is true, whether n is a prime number or not; so the implication is *trivially* true.

- **Vacuous proof**: If we know p is false then "p implies q" is true regardless of the truth value of q (we've discussed this earlier, in a footnote during Alice and Bob's dialogue).

  Example 1: "If $4n$ is a prime number, then $n - n = 0$". Here the hypothesis is "$4n$ is a prime number". But this is *false*, regardless of what n we pick ($4n$ will always be a multiple of 4). So, by the truth table of implication, the implication is *true*, and we can say this without even looking at the conclusion statement.

  Here's another way to look at this, which explains why we refer to this as a "vacuous" proof: to prove the implication, we are being asked to verify that *whenever it holds that* $4n$ *is prime, it also holds that* $n - n = 0$. It *never* holds that $4n$ is prime, so there are *no* cases to check, there is no possible witnesses to the *incorrectness* of the implication, so, having no evidence to the contrary, we must conclude that the implication is *correct*.

  Example 2: "If $4n$ is a prime number, then $n - n = 1$". In Example 1 we could have equally well argued that the implication is trivial(ly true), since the conclusion is true. Here, the conclusion is *false*. But again, the *implication* is true, vacuously. The premise is false, and so there are no witnesses to refute the implication.

  These examples should illustrate that implication is a subtle (you might say "slippery") logical operation, that takes some getting used to.

2.5 An note on equality

We haven't said it explicitly yet, but it has been implicit: in proving statements involving numbers, as well as using the rules of inference, axioms, definitions, and logical equivalences, we also use a few basic properties of the equality symbol "=", namely:

- **E1**: For all numbers a, $a = a$ ("Things which coincide with one another are equal to one another.")

- **E2**: For all a, b and c, if $a = c$ and $b = c$ then $a = b$ ("Things which are equal to the same thing are also equal to one another.")

- **E3**: For all a, b, c and d, if $a = b$ and $c = d$ then $a + c = b + d$ and $a - c = b - d$ ("If equals are added to equals, the whole are equal" and "If equals be subtracted from equals, the remainders are equal.")

These "axioms of equality" were first formulated by Euclid around 300BC; I've put his statements in parentheses above[33].

2.6 A (slightly) more formal look at logic

For those who might be interested, we say a little bit more here about the formal business of logic and proofs. In reality we won't need to use any of this language as the year progresses, so reading this section is *entirely* optional, and I won't mention any of this in class or quizzes.

Rules of inference

The rules of inference are the basic rules of logic, that allow us to infer, or deduce, the truth of new propositions from old. Each rule is a re-statement of a tautology. Take "Hypothetical syllogism" below as an example. Suppose I know that

If Notre Dame beats Michigan this year, I will celebrate with beer at Rohr's

(this is an axiom: how else would I celebrate?) and also that

If I drink beer at Rohr's, I will Uber home

(again an axiom: I don't drink and drive). Then I should legitimately be able to conclude

If Notre Dame beats Michigan this year, I will Uber home that night.

Why can I conclude this? Because the statement

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

is a *tautology*; it's a true statement, regardless of the truth values that p, q and r happen to take. (In this case p:"Notre Dame beats Michigan", q:"I celebrate" and r:"I Uber home".)

Once we've verified that the relevant propositions are tautologies, each of the rules of inference should be quite palatable. Here is the list of rules that we will most commonly use:

Name	If you know ...	you can infer ...	because ... is a tautology
Modus ponens	p and $p \Rightarrow q$	q	$(p \wedge (p \Rightarrow q)) \Rightarrow q$
Modus tollens	$\neg q$ and $p \Rightarrow q$	$\neg p$	$(\neg q \wedge (p \Rightarrow q)) \Rightarrow \neg p$
Disjunction introduction	p	$p \vee q$	$p \Rightarrow (p \vee q)$
Conjunction elimination	$p \wedge q$	p	$(p \wedge q) \Rightarrow p$
Hypothetical syllogism	$p \Rightarrow q$ and $q \Rightarrow r$	$p \Rightarrow r$	$((p \Rightarrow q) \wedge (q \Rightarrow r)) \Rightarrow (p \Rightarrow r)$
Conjunction introduction	p and q	$p \wedge q$	$(p \wedge q) \Rightarrow (p \wedge q)$
Disjunctive syllogism	$p \vee q$ and $\neg p$	q	$((p \vee q) \wedge (\neg p)) \Rightarrow q$
Constructive dilemma	$p \Rightarrow q, r \Rightarrow s$ and $p \vee r$	$q \vee s$	$((p \Rightarrow q) \wedge (r \Rightarrow s) \wedge (p \vee r)) \Rightarrow q \vee s$

[33] Wording taken from http://www.friesian.com/space.htm.

In the next few paragraphs, we'll make some remarks on the rules of inference. This section won't have many examples, because soon we will launch into the main topic of the first half of the course, working with the axioms of the real numbers, and we will get a chance there to see plenty of proofs that use these methods.

- **Modus ponens**: If you know p, and you know p implies q, you can deduce q.

  The name comes from the Latin phrase *modus ponendo ponens*, meaning *the way that affirms by affirming*, conveying the sense that modus ponens is quite a direct method of inference. It is by far the most used and most important method.

- **Modus tollens**: If you know that p implies q, and you know that q is *false*, you can deduce that p is false.

  The name comes from the Latin phrase *modus tollendo tollens*, meaning *the way that denies by denying*, conveying the sense that modus tollens is an *indirect* method of inference. It is sometimes called *proof by contrapositive*, because the contrapositive of "p implies q" is (the equivalent statement) "not q implies not p", and knowing this together with "not q" allows the immediate deduction of "not p", by modus ponens.

  The next few rules are quite obvious and require no discussion:

- **Disjunction introduction**: If you know that p is true, then regardless of the truth or otherwise of some other statement q, you can immediately deduce that at least one of p or q are true.

- **Conjunction elimination**: If you know that both p and q are true, then you can immediately deduce that p is true.

- **Hypothetical syllogism**: If you know that p implies q, and that q implies r, then (by following the obvious chain) you can deduce that p implies r. This says that "implies" is a *transitive* relation.

- **Conjunction introduction**: If you know that both p and q are true, then you can deduce that the compound statement "p and q" is true. This is a sort of converse to Conjunction elimination.

- **Disjunctive syllogism**: If you know that either p or q are true, and you know that p is false, then you can deduce that q is true.

- **Constructive dilemma**: If you know both that p implies q, and that r implies s, and you also know that at least one of the two premises p, r are true, then (since you can deduce that at least one of the conclusions q, s are true), you can deduce that the compound statement "r or s" is true.

This is a "constructive dilemma" because you deduce that one of two things (q or s) is true, but you have no way of knowing explicitly *which* is true; you can't "construct" a simple true statement out of the knowledge that the complex statement "q or s" is true.

An note on *invalid* inferences

Modus ponens says: from p and $p \Rightarrow q$ you can infer q, and modus tollens says: from $\neg q$ and $p \Rightarrow q$ you can infer $\neg p$.

There are two other tempting "rules of inference" that are both **INVALID**:

- "from q and $p \Rightarrow q$ you can infer p": this is called *affirming the consequent*, or the *converse error* (using the *conclusion* to say something about the *hypothesis*), and is invalid, because

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

  is not a tautology.
- "from $p \Rightarrow q$ you can infer $(\neg p) \Rightarrow (\neg q)$": this is called *denying the antecedent*, or the *inverse error* (confusing the direction of implication), and is invalid, because

  $$(p \Rightarrow q) \Rightarrow ((\neg p) \Rightarrow (\neg q))$$

  is not a tautology.

Let's illustrate all of this with the statement:

"If you fall of a wall, you break a bone".

This is an implication, with hypothesis "you fall off a wall" and conclusion "you break a bone".

- Suppose you know that the implication is true, and you also know that you fell off a wall. Then you conclude that you broke a bone. That is modus ponens in action.
- Suppose you know that the implication is true, and you also know that you *do not* have a broken bone. Then you conclude that you *did not* fall off a wall. That is modus tollens in action.
- Suppose you know that the implication is true, and you also know that you have a broken bone. Then you *cannot* conclude that you fell off a wall — there are other ways to break a bone. If you did make that inference, you would be making the converse error.

- Suppose you know that the implication is true. Then you *cannot* conclude that if you do not fall off a wall, then you do not have a broken — again, this implication is easily shown to be false by considering any non-falling-off-a-wall circumstance that leads to a broken bone. If you did make that inference, you would be making the inverse error.

There are also some rules of inference relating to quantification, all of which are quite evident:

- **Universal instantiation**: If you know $(\forall x)p(x)$, you can infer $p(c)$ for any particular c in the universe of discourse
- **Universal generalization**: If you know $p(c)$ for an arbitrary/generic element in the universe of discourse, you can infer $(\forall x)p(x)$
- **Existential instantiation**: If you know $(\exists x)p(x)$, you can infer $p(c)$ for some c in the universe of discourse (this allows you to define a variable c to stand for some fixed element of the universe of discourse, whose specific name may not be known, for which $p(c)$ is true)
- **Existential generalization**: If you know $p(c)$ for some fixed element of the universe of discourse you can infer $(\exists x)p(x)$.

Approaches to proving quantified statements

Often the statements of theorems are of the form $(\exists x)p(x)$ or $(\forall x)p(x)$. Here we discuss some general approaches to these types of theorems.

- **Constructive existential proofs**: To prove $(\exists x)p(x)$, one approach is to find ("construct") an explicit element c in the universe of discourse for x, such that $p(c)$ is true.

  Example: "There exist arithmetic progressions of length 4, all terms of which are prime numbers".

  Proof: The numbers 251, 257, 263 and 269 form an arithmetic progressions of length 4, and all of these numbers are prime.

- **Non-constructive existential proofs**: One can sometimes prove $(\exists x)p(x)$ by showing that there must be an element c in the universe of discourse for x, such that $p(c)$ is true, *without explicitly exhibiting such a c.*

  Example: "Among any 13 people, some two of them must have their birthdays in the same month".

Proof: Let k_i be the number of people, among the 13, who have their birthday in the ith month of the year. We want to show $k_i \geq 2$ for some i. Suppose, for a contradiction, that $k_i \leq 1$ for each i. Then

$$k_1 + k_2 + \cdots + k_{12} \leq 1 + 1 + \cdots + 1 = 12.$$

But since everybody has a birth-month, we also have

$$k_1 + k_2 + \cdots + k_{12} = 13$$

That $k_1 + \cdots + k_{12}$ is simultaneously at most 12 and exactly 13 is a contradiction, and this proves the statement.

Some remarks:

- The principle exposed in this proof is sometimes called the *pigeon-hole principle*: "If more than n pigeons distribute themselves among at most n pigeon holes, then there must be at least one pigeon-hole that has at least two pigeons in it". This simple-sounding principle turns out to be incredibly powerful ('though unfortunately quite hard to apply!) Some applications appear in the homework.
- There are quite a few major open problems in the field of combinatorics (a brach of mathematics closely related to theoretical computer science) that involve finding *constructive* existential proofs of statements that are very easy proven in a non-constructive way.

- **Non-existence proofs**: Suppose we wish to show that there is *no* element of the universe of discourse that satisfies a particular predicate; that is, we wish to prove $\neg(\exists x)p(x)$. This is equivalent to $(\forall x)(\neg p(x))$; so one approach is to choose a generic element c in the universe of discourse for x, assume that $p(c)$ holds, and derive a contradiction. This allows us to conclude that for generic c, $\neg p(c)$ is true, and so (by universal generalization) $(\forall x)(\neg p(x))$ is true.

  Example: (Here the universe of discourse for m is positive whole numbers). "There is no m for which $4m + 3$ is a perfect square".

  Proof: Let m be an arbitrary positive integer, and assume that $4m + 3$ is a perfect square, say $4m + 3 = k^2$ for some integer k. Because $4m + 3$ is odd, so too is k^2; and because the square of an even number is even, it must be that k is an odd number, say $k = 2a + 1$ for some integer a. So we have

$$4m + 3 = (2a + 1)^2.$$

  Rearranging terms, this is equivalent to

$$2 = 4(a^2 + a - m),$$

and this implies, dividing both sides by 2, that

$$1 = 2(a^2 + a - m).$$

This is a contradiction: the left-hand side above is odd, and the right-hand side is even (since $a^2 + a - m$ is a whole number). This contradiction shows that $4m + 3$ is never a perfect square.

- **Universal quantification proofs**: In general to establish $(\forall x)p(x)$ one starts with a generic element c of the universe of discourse, and argues the truth of $p(c)$. We will see plenty of examples going forward.

- **Counterexamples**: Sometimes we want to establish $\neg(\forall x)p(x)$ — that is is *not* the case that $p(c)$ is true for every element of the universe of discourse. What is required here is an example of a *single*, *specific* c in the universe of discourse for which $p(c)$ is *false*. This is often referred to as a *counterexample* to the statement $(\forall x)p(x)$.

  Example (actually, exercise): We've seen that $n^2 - n + 41$ is prime for $1 \leq n \leq 40$. Show that is *not* true that $n^2 - n + 41$ is prime for every positive integer n.

3 Axioms for the real number system

Calculus is concerned with differentiation and integration of functions of the real numbers. Before understanding differentiation and integration, we need to understand functions, and before understanding functions, we need to understand the real numbers.

3.1 Why the axiomatic approach?

We all have an intuitive idea of the various number systems of mathematics:

1. the *natural numbers*, $\mathbb{N} = \{1, 2, 3, 4, \ldots\}$ — the ordinary counting numbers, that can be added, multiplied, and sometimes divided and subtracted, and the slightly more sophisticated version of the natural numbers, that includes 0 — we'll denote this set as $\mathbb{N}^0$;

2. the *integers*, $\mathbb{Z} = \{\ldots, -4, -3, -2, -1, 0, 1, 2, 3, 4, \ldots\}$ — the natural numbers together with their negatives, which allow for subtraction in all cases;

3. the *rationals*, $\mathbb{Q}$, the set of all numbers of the form a/b where a and b are integers[34], and b is not 0, which allows for division in all cases; and

4. the *real numbers*, which we denote by $\mathbb{R}$, and which "fill in the gaps" in the rational numbers — π, for example, or $\sqrt{2}$, are not expressible as the ratio of two integers, but they are important numbers that we need to have in our number system. Mathematical life is much easier when we pass from

   - the rational numbers, which may be thought of as a large (actually infinite) collection of marks on the number line:

     ....,

     that's very dense, but doesn't cover the whole line,

   to

   - real numbers, which may be thought of as the *entire* number line:

     ———————————————————,

[34] There is a subtlety here. Any given fraction appears infinitely often among the set of all numbers of the form a/b; for example, two-thirds appears as 2/3, 4/6, 6/9, et cetera.To be proper, we should say that each rational number is an *infinite set* of expressions of the form a/b, where a and b are integers, $b \neq 0$, satisfying $a/b = a'/b'$ for all pairs a, b and a', b' in the set. The specific rational number represented by the infinite set is the common value of all these ratios.

It will be helpful for us to keep these intuitive ideas in mind as we go through the formal definition of the real numbers; we will use them to provide (hopefully helpful) illustrative examples as we go along. But it is very important to remember that in this section our goal is to

rigorously derive all important properties of the real numbers;

and so our intuitive ideas will only ever be used for illustration.

The approach we are going to take to understanding the real numbers will be *axiomatic*: we will write down a collection of axioms, that we should all agree that the real numbers should satisfy, and then we will *define* the real number system to be the set of objects that satisfy all the axioms.

This begs four questions:

- **Q1**: What axioms should we choose?
- **Q2**: How do we know that there is actually a set of objects that satisfies the axioms?
- **Q3**: Even if we know there is such a set, how do we know that it is the *only* such set (as the language we've used above — "*the* real number system [is] *the* set of objects that satisfy all the axioms" — strongly suggests?
- **Q4**: If there is a unique set of objects that satisfies the axioms, why don't we approach the real numbers by actually *constructing* that set? Why the more abstract axiomatic approach?

We'll briefly discuss possible answers now.

- **A1**: What axioms should we choose? We'll try to capture what we believe are the essential properties on out "intuitive" real numbers, with a collection of axioms that are as simple as possible. Most of the axioms will turn out to be very obvious uncontroversial; a few will be less obvious, motivated by considering our intuitive understanding of the reals, but still uncontroversial; and only one will be non-obvious. This last will be the axiom that separates the rational numbers from the reals, and will be the engine that drives almost every significant result of calculus.
- **A2**: Is there a set of objects that satisfies the axioms? Starting from the basic rules of set theory, it *is* possible to construct a set of sets, and to define addition and multiplication on that set, in such a way that the result behaves exactly as we would expect the real numbers to behave. We could therefore take a *constructive* approach to the real numbers, and indeed Spivak devotes the last few chapters of his text to explaining much of this construction.
- **A3**: Is there a *unique* such a set? Essentially, yes; again, this is discussed in the final chapters of Spivak's text.

- **A4**: Why the more abstract axiomatic approach, then? A good question: if we can *construct* the reals, why not do so?

  One reason not to is that the construction is quite involved, and might well take a whole semester to fully describe, particularly since it requires understanding the axioms of set theory before it can get started.

  Another reason is a more practical, pedagogical one: most mathematical systems don't have the luxury, that the reals have, of having an essentially *unique* model. In a little while in your mathematical studies, for example, you will see the incredibly important notion of a *vector space*. It will turn out that there are many, many (infinitely many) essentially different instances of a vector space. That means that it hopeless to try to study vector spaces constructively. Instead we have to approach axiomatically — we set down the basic properties that we want a "vector space" to satisfy, then derive all the further properties that it must satisfy, that follow, via rules of logic, from the basic properties, and then know that *every* one of the infinitely many instances of a vector space must satisfy all these further properties.

  The same goes for most of the other basic mathematical objects, such as groups, metric spaces, rings, fields, ... Since so many mathematical objects need to be studied axiomatically, it's good to get started on the axiomatic approach as early as possible!

3.2 The axioms of addition

From now on, whenever we talk about the *real numbers*, we are going to mean the following: the real numbers, which we denote by $\mathbb{R}$, is a set of objects (which we'll call *numbers*)

- including two special numbers, 0 and 1 ("zero" and "one")

together with

- an operation, $+$ ("addition"), which can combine any two numbers a, b to form another (not necessarily different) number, $a + b$, and
- another operation, $\cdot$ ("multiplication"), which can combine any two numbers a, b to form another number, $a \cdot b$,

and which satisfies a collection of 13 axioms, which we are going to label P1 through P13, and which we will introduce slowly over the course of the next few sections.

The first four axioms, P1 through P4, say that addition behaves in all the obvious ways, and that 0 plays a special role in terms of addition:

- **P1, Additive associativity**: For all a, b, c,

$$a + (b + c) = (a + b) + c.$$

- **P2, Additive identity**: For all a,

$$a + 0 = 0 + a = a.$$

- **P3, Additive inverse**: For all a there's a number $-a$ such that

$$a + (-a) = (-a) + a = 0.$$

- **P4, Additive commutativity**: For all a, b,

$$a + b = b + a.$$

Comments on axiom P1

Axiom P1 says that when we add together three numbers, the way in which we parenthesize the addition is irrelevant (this property is referred to as *associativity*); so whenever we work with real numbers, we can *unambiguously* write expressions like

$$a + b + c.$$

But what about adding together *four* numbers? There are five different ways in which we can add parentheses to a sum of the form $a + b + c + d$, to describe the order in which the addition should take place:

- $(a + (b + c)) + d$
- $((a + b) + c) + d$
- $a + (b + (c + d))$
- $a + ((b + c) + d)$
- $(a + b) + (c + d)$.

Of course, if the set of "real numbers" we are axiomatizing here is to behave as we expect the real numbers to behave, then we want all five of this expressions to be the same. Do we need to add an axiom, to declare that all five expressions are the same? And then, do we need to add another axiom to say that all 14 ways of parenthesizing $a + b + c + d + e$ are the same? And one that says that all 42[35] ways of parenthesizing $a + b + c + d + e + f$ are the

[35] How many different ways are there to parenthesize the expression $a_1 + a_2 + \ldots + a_n$? For $n = 2, 3, 4, 5, 6, \ldots$, the answer is $1, 2, 5, 14, 42, \ldots$, as can be verified by a brute-force search. The sequence of numbers that comes up in this problems is very famous (among mathematicians ...): it is the answer to literally hundreds of different problems, has been the subject of at least five books, and has been mentioned in at least 1400 mathematical papers. Although it's not obvious, the terms of the sequence follow a very simple rule.

same? And ... you see where I'm going — do we need to add infinitely many axioms, just to say that for every n,

$$a_1 + a_2 + \ldots + a_n$$

is an unambiguous expression, whose value doesn't depend on the way in which parenthesize?

Fortunately, no! Using the rules of inference, we can *deduce* that **if**

$$a + (b + c) = (a + b) + c$$

for all possible choices of a, b, c, **then**

$$(a + (b + c)) + d = ((a + b) + c) + d = a + (b + (c + d)) = a + ((b + c) + d) = (a + b) + (c + d)$$

for all possible choices of a, b, c, d. We formulate this as a claim; it will be the first proper proof of the course.

Claim 3.1. *If a, b, c and d are real numbers, then each of $(a + (b + c)) + d$, $((a + b) + c) + d$, $a + (b + (c + d))$, $a + ((b + c) + d)$ and $(a + b) + (c + d)$ are the same.*

Proof: We first show that $(a + (b + c)) + d = ((a + b) + c) + d$. By axiom P1, $(a + (b + c)) = ((a + b) + c)$, and so $(a + (b + c)) + d = ((a + b) + c) + d$ follows immediately from the rules of equality (specifically, from **E3**).

By virtually the same reasoning[36], $a + (b + (c + d)) = a + ((b + c) + d)$.

We now consider $(a + (b + c)) + d$ and $a + ((b + c) + d)$. We apply axiom P1, but with a twist: P1 says that for any A, B and C, $(A + B) + C = A + (B + C)$. We apply this with $A = a$, $B = b + c$ and $C = d$ to conclude that $(a + (b + c)) + d = a + ((b + c) + d)$.

All this shows that first four expressions are all equal to each other. So what is left to show is that fifth equals *any one* of the first four. We leave this as an exercise to the reader.[37] □[38]

Note that this was an example of a *direct* proof.

It would take much more work to show that all 14 ways of parenthesizing $a + b + c + d + e$ lead to the same answer, but this too can be shown to follow from P1. Sadly, using this approach it would take infinitely much work to show that for *all* n, and all $a_1, a_2, \ldots, a_n$, all ways of parenthesizing $a_1 + a_2 + \cdots + a_n$ lead to the same answer. We could get over this problem by adding infinitely many P1-like axioms to our set of axioms; fortunately, we will

[36] If two parts of a proof are basically identical, it's quite acceptable to describe one of them in detail, and then say that the other is essentially the same. **But**: you should *only* do this if the two arguments really are basically identical. You should *not* use this to worm your way out of writing a part of a proof that you haven't fully figured out!

[37] You'll see this expression — "we leave this as an exercise to the reader" — a lot throughout these notes. I *strongly* encourage you to do these exercise. They will help you understand to concepts that are being discussed, and they are good practice for the quizzes, homework and exams, where some of them will eventually appear.

[38] It's traditional to use this symbol — □ — to mark the end of a proof.

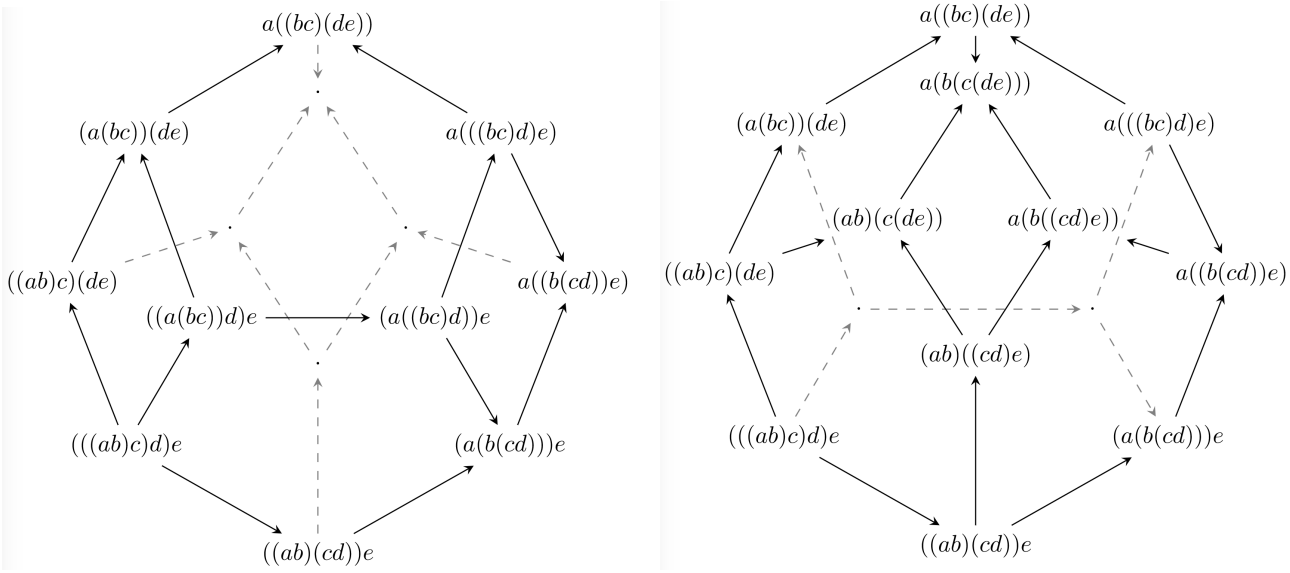
soon come to a method — proof by induction — that allows us to prove statements about *all* natural numbers in a finite amount of time, and we will use this method to show that it is indeed the case that the expression

$$a_1 + a_2 + \cdots + a_n$$

doesn't depend on the way in which it is parenthesized. So from here on, we will allow ourselves to assume this truth.

Aside: The *associahedron* K_n is an object that consists of points and edges. The points are all the different ways of parenthesizing the expression $a_1 + a_2 + \cdots + a_n$, and there is an edge between two points, if it is possible to show that the two associated expressions are equal, by applying the associativity axiom $a + (b + c) = (a + b) + c$ (with appropriate choices of a, b and c). So, for example, the associahedron K_4 has five points, namely $(a_1 + (a_2 + a_3)) + a_4$, $((a_1+a_2)+a_3)+a_4$, $a_1+(a_2+(a_3+a_4))$, $a_1+((a_2+a_3)+a_4)$ and $(a_1+a_2)+(a_3+a_4)$, and there is an edge joining $(a_1+(a_2+a_3))+a_4$ and $a_1+((a_2+a_3)+a_4)$, because one application of axiom P1 (with $a = a_1$, $b = a_2+a_3$ and $c = a_4$ to conclude that $a_1+((a_2+a_3)+a_4) = (a_1+a_2)+(a_3+a_4)$.

It's evident that the associahedron K_3 is just a line segment joining two points (and so is 1-dimensional), and after a little bit of work you should be able to convince yourself that K_4 is a pentagon (and so is 2-dimensional). The associahedron K_5, corresponding to the different ways of parenthesizing the sum of 5 terms, is a pretty 3-dimensional shape, built out of squares and pentagons, shown in the pictures below (front and back):



The associahedron K_5

The fact that the associahedron K_5 is connected — it's possible to move from any one point to any other, along edges — shows that all 14 ways of parenthesizing the sum of 5 things end up yielding the same final value.

In higher dimensions — when "5 things" is replaced with "6 things", "7 things", and so on — the associahedra continue to be connected, showing that the way that an arbitrary sum is parenthesized doesn't change the final answer; but this isn't (or at least shouldn't be) obvious!

I'm mentioning associahedra, because there is a Notre Dame connection: much of what we know about associahedra was discovered by Jim Stasheff, who worked on them while he was a professor here from 1962 to 1968. See https://en.wikipedia.org/wiki/Associahedron for more information, including a nice animation of the three-dimensional associahedron.

Comments on axioms P2 and P3

Axiom P2 says that the number 0 is special, in that when it is added to anything, or when anything is added to it, nothing changes. We of course know (from our intuitive understanding of real numbers) that 0 should be the *unique* number with these special properties. The axiom doesn't say that, but fortunately this extra property of 0 can be *deduced* (proven) from the axioms as presented. If fact, something a little stronger is true:

Claim 3.2. *If x and a are any numbers satisfying $a + x = a$, then $x = 0$.*

Proof: We simply "subtract a" from both sides of the equation $a + x = a$:

Since $a + x = a$, we know that

$$-a + (a + x) = (-a) + a.$$

Using P1 on the left and P3 on the right, this says that

$$((-a) + a) + x = 0.$$

Using P3 on the left, this says that $0 + x = 0$, and using P2 on the left we finally conclude that $x = 0$. □

Notice that this allows us to deduce the uniqueness of 0: if x is such that $a+x = x+a = a$ for all a, then in particular $a + x = a$ for some particular number a, so by the above claim, $x = 0$.

Notice also that we had to use all three of axioms P1, P2 and P3 to prove an "obvious" fact; this will be a fairly common feature of what follows. We're trying to produce as simple as possible a set of axioms that describe what we think of as the real numbers. So it makes sense that we're going to have to make wide use of this simple set of axioms to verify the more complex, non-axiomatic properties that we would like to verify.

The proof above proceeded by adding $-a$ to both sides of the equation, which we though of as "subtracting a". We formalize that idea here:

for any numbers a, b, we define "$a - b$" to mean $a + (-b)$.

Axiom P3 says that every number a has an *additive inverse* (which we denote by $-a$): a number which, when added to a, results in the answer 0. Of course, the additive inverse of each number should be *unique*. We leave it as an exercise to the reader to verify this: for any numbers a and b, if $a + b = 0$, or if $b + a = 0$, then $b = -a$.

Another property we would expect to be true of addition is the *cancellation* property. We leave it as an exercise to prove that if a, b, c are any numbers, and if $a + b = a + c$, then $b = c$.

Comments on axiom P4

Axiom P4 tells us the order in which we add two numbers doesn't affect the sum. This property of addition is referred to as *commutativity*. This is not true of all operations that we will perform on numbers — we don't expect $a - b$ to be equal to $b - a$ in general, for example — so for addition, it really needs to be explicitly said.

We know that in fact if we add n numbers, for any n, the order in which we add the numbers doesn't impact the sum. It should be fairly clear that we don't need to add any new axioms to encode this more general phenomenon. For example, while there are six different ways of ordering three numbers, a, b, c, to be added[39], namely

- $a + b + c$
- $a + c + b$
- $b + a + c$
- $b + c + a$
- $c + a + b$
- $c + b + a$,

it's easy to see that all six of them are equal to $a + b + c$. All we need to do is to repeatedly apply commutativity to neighboring pairs of summands, first to move the a all the way to the left, then to move the b to the middle position. For example,

$$c+b+a = c+(b+a) = c+(a+b) = (c+a)+b = (a+c)+b = a+(c+b) = a+(b+c) = a+b+c$$

(overkill: since we have already fully discussed associativity, I could have just written

$$c + b + a = c + a + b = a + c + b = a + b + c).$$

As with associativity, once we have proof by induction we will easily prove that when we add any n terms, $a_1, \ldots, a_n$, the sum doesn't depend on the pairwise order in which the pairs are added. So from here on, we will allow ourselves to assume this truth.

[39] In an earlier footnote we asked the question, "How many different ways are there to parenthesize the expression $a_1 + a_2 + \ldots + a_n$?". We can ask the analogous question here: "How many different ways are there to order the n summands $a_1, a_2, \ldots, a_n$?" This question is much easier than the one for parenthesizing. For $n = 2, 3, 4, 5, 6, \ldots$ the sequence of answers is $2, 6, 24, 120, 720, \ldots$, and you should quickly be able to see both the pattern, and the reason for the pattern.

3.3 The axioms of multiplication

The next four axioms, P5 through P8, say that multiplication behaves in all the obvious ways, and that 1 plays a special role in terms of multiplication:

- **P5, Multiplicative associativity**: For all a, b, c,

$$a \cdot (b \cdot c) = (a \cdot b) \cdot c.$$

- **P6, Multiplicative identity**: For all a,

$$a \cdot 1 = 1 \cdot a = a.$$

- **P7, Multiplicative inverse**: For all a, if $a \neq 0$ there's a number a^{-1} such that

$$a \cdot a^{-1} = a^{-1} \cdot a = 1.$$

- **P8, Multiplicative commutativity**: For all a, b,

$$a \cdot b = b \cdot a.$$

These axioms look almost identical to those for addition. Indeed, replace "$\cdot$" with "$+$" and "1" with "0" in P5 through P8, and we have almost exactly P1 through P4. *Almost* exactly: we know that *every* number should have a negative (an additive inverse), but it it only *non-zero* numbers that have a reciprocal (a multiplicative inverse), and so in P7 we explicitly rule out 0 having an inverse.

Looking at what we did for addition, you should be able to prove the following properties of multiplication:

- 1 is unique: if x is such that $a \cdot x = a$ for all a, then $x = 1$, and
- the multiplicative inverse is unique: if $a \neq 0$, and if x is such that $a \cdot x = 1$, then $x = a^{-1}$,

and you should also be able to convince yourself that associativity and commutativity of multiplication extend to the product of more than two terms, without the need for additional axioms

With regards the first bullet point above: it will not be possible to prove for multiplication, the analog of the most general statement we proved for addition: it is *false* that "if a and x are any numbers, and $a \cdot x = a$, then $x = 1$". Indeed, if $a = 0$ then a counterexample to this statement is provided by any $x \neq 0$. Instead, the most general statement one can possibly prove is "if a and x are any numbers, with $a \neq 0$, and if $a \cdot x = a$, then $x = 1$".

We have mentioned the cancellation property of addition. There is a similar property of multiplication, that says we can "cancel" a common factor on both sides of an equation, *as long as that factor is not* 0.

Claim 3.3. *If a, b, c are any numbers, and $a \cdot b = a \cdot c$, then* either $a = 0$ or $b = c$.

Proof: Suppose that $a \cdot b = a \cdot c$. We want to argue that either $a = 0$ or $b = c$ (or perhaps both).

There are two possibilities to consider. If $a = 0$, then we are done (since if $a = 0$ it is certainly the case that either $a = 0$ or $b = c$). If $a \neq 0$, then we must argue that $a \cdot b = a \cdot c$ implies $b = c$. We get to use that there is a number a^{-1} such that $a^{-1} \cdot a = 1$. Using this, the argument goes like:

$$\begin{array}{cc} a \cdot b = a \cdot c & \text{implies (using P7, valid since } a \neq 0) \\ a^{-1} \cdot (a \cdot b) = a^{-1} \cdot (a \cdot c) & \text{which implies (using P5)} \\ (a^{-1} \cdot a) \cdot b = (a^{-1} \cdot a) \cdot c & \text{which in turn implies (using P7)} \\ 1 \cdot b = 1 \cdot c & \text{which finally implies (using P6)} \\ b = c. & \end{array}$$

□

Notice that

- *the proof above is presented in full sentences.* We did not simply write

$$\begin{array}{c} \text{“}a \cdot b = a \cdot c \\ a^{-1} \cdot (a \cdot b) = a^{-1} \cdot (a \cdot c) \\ (a^{-1} \cdot a) \cdot b = (a^{-1} \cdot a) \cdot c \\ 1 \cdot b = 1 \cdot c \\ b = c, \text{”} \end{array}$$

  and notice also that

- *every step of the proof was justified* (in this case, by reference to a particular axiom).

My expectation is that you will always present your proofs in complete sentences, and initially with every step justified (this condition will get relaxed soon, but for now it is the expectation!)

The proof above proceeded by multiplying both sides of the equation by a^{-1}, which we think of as “dividing by a”. We formalize that idea here:

for any numbers a, b, with $b \neq 0$, we define “a/b” to mean $a \cdot (b^{-1})$.

We've mentioned two special properties of 0 — it is the additive inverse, and it is the unique number that we do not demand has a multiplicative inverse. There is a third special property of 0, that we know from our intuitive understanding of real numbers, namely that $a \cdot 0 = 0$ for any a. This hasn't been mentioned in the axioms so far, but we definitely want it to be true. There are two possibilities:

- *either* we can deduce $(\forall a)(a \cdot 0 = 0)$ from the axioms so far,
- *or* we can't, in which case we really need another axiom!

It turns out that we are in the second situation above — it is not possible to prove that for all a, $a \cdot 0 = 0$, just using Axioms P1 through P8. And in fact, we can *prove* that we can't prove this! We won't bother to make the digression and do that here, but I'll explain how it works. Suppose that we can find a set X of "numbers", that include numbers "0" and "1", and we can define operations "+" and "$\cdot$" on this set of numbers, in such as way that all of the axioms P1 through P8 hold, *but for which also there is some number a with $a \cdot 0 \neq 0$.* Then that set X would act as a witness to prove that P1 through P8 alone are not enough to prove that for all a, $a \cdot 0 = 0$. (It's actually quite simple to find such a set X. Consider it a challenge!)

An obvious choice of new axiom is simply the statement "for all a, $a \cdot 0 = 0$" — if we want this to be true, and it doesn't follow from the axioms so far, then let's force it to be true by adding it as a new axiom. The route we'll take is a little different. We'll add an axiom that talks in general about how addition and multiplication interact with each other.

3.4 The distributive axiom

The next axiom, that links addition and multiplication, lies at the heart of almost every algebraic manipulation that we will ever do.

- **P9, Distributivity of multiplication over addition**: For all a, b, c,

$$a \cdot (b + c) = (a \cdot b) + (a \cdot c).$$

As a first substantial consequence, let's use P9 to prove that $a \cdot 0 = 0$.

Claim 3.4. *If a is any number then $a \cdot 0 = 0$.*

Proof: By P2,

$$0 + 0 = 0.$$

Multiplying both sides by a, we get that

$$a \cdot (0 + 0) = a \cdot 0.$$

By P9, this implies

$$a \cdot 0 + a \cdot 0 = a \cdot 0. \quad (\star)$$

Adding $-(a \cdot 0)$ to the left-hand side of $(\star)$, applying P1, then P3, then P2, we get

$$a \cdot 0 + a \cdot 0 + (-a \cdot 0) = a \cdot 0 + (a \cdot 0 - a \cdot 0) = a \cdot 0 + 0 = a \cdot 0. \quad (\star\star)$$

Adding $-(a \cdot 0)$ to the right-hand side of $(\star)$, and applying P3, we get

$$a \cdot 0 + (-a \cdot 0) = 0. \quad (\star\star\star)$$

Since the left- and right-hand sides of $(\star)$ are equal, they remain equal on adding $-(a \cdot 0)$ to both sides, so combining $(\star\star)$ and $(\star\star\star)$ we get

$$a \cdot 0 = 0.$$

□

Another important consequence of P9 is the familiar property of real numbers, that if the product of two numbers if zero, then at least one of the two is zero.

Claim 3.5. *If $a \cdot b = 0$ then either $a = 0$ or $b = 0$.*

Proof: If $a = 0$, then there is no work to do, so from here on we assume that $a \neq 0$, and we argue that this forces $b = 0$.

Since $a \neq 0$ there is a^{-1} with $a^{-1} \cdot a = 1$. Multiplying both sides of $a \cdot b = 0$ by a^{-1}, we get

$$\begin{array}{rl} a^{-1} \cdot (a \cdot b) = a^{-1} \cdot 0, & \text{which implies (by P5, P7 and Claim 3.4) that} \\ 1 \cdot b = 0, & \text{which implies (by P6) that} \\ b = 0. & \end{array}$$

□

Notice that as well as using the axioms in this proof, we have also used a previously proven theorem, namely Claim 3.4. As the results we prove get more complicated, this will happen more and more. Notice also that we condensed three lines — applications of P5, P7 and Claim 3.4 — into one. This is also something that we will do more and more of as we build more proficiency at constructing proofs.

To illustrate the use of P9 in algebraic manipulations, consider the identity

$$x^2 - y^2 = (x - y) \cdot (x + y),$$

valid for all real x, y, where "x^2" is shorthand for "$x \cdot x$". (If you are not familiar with this identity, you should familiarize yourself with now; it will prove to be very useful.) To verify that it is a valid identity, note that, by P9, we have

$$\begin{array}{rcl} (x - y) \cdot (x + y) & = & (x - y) \cdot x + (x - y) \cdot y \\ & = & (x + (-y)) \cdot x + (x + (-y)) \cdot y \quad \text{(definition of } -y) \\ & = & x \cdot x + (-y) \cdot x + x \cdot y + (-y \cdot y) \quad \text{(P9)} \\ & = & x \cdot x - y \cdot x + y \cdot x - y \cdot y \quad \text{(P8)} \\ & = & x \cdot x - y \cdot y \quad \text{(various axioms, applied in obvious ways).} \end{array}$$

Now defining x^2 to mean $x \cdot x$, and $y^2 = y \cdot y$, we get the result. (Note: in the third line we are using a version of P9 that follows immediately from P9 using P8: for all a, b, c, $(b + c) \cdot a = (b \cdot a) + (c \cdot a)$.)

But wait! In the middle of the proof, we slipped in

$$(-y) \cdot x = -(y \cdot x)$$

(the product of (the additive inverse of y) and (x), is the same as the additive inverse of (the product of y and x)). This isn't an axiom, so needs to be proved! As a last example of the power of P9, we present a proof that suggests the rule "negative times negative equals positive", and along the way verifies the sneaky step in the above proof.

Claim 3.6. *For all numbers a, b, $(-a) \cdot (-b) = ab$*[40].

Proof: We begin by arguing that $(-a) \cdot (b) = -(a \cdot b)$. We know that $-(a \cdot b)$ is the additive inverse of $a \cdot b$, and that moreover it is the *unique* such inverse (a previous exercise for the reader). So if we could show $a \cdot b + (-a) \cdot (b) = 0$, then we could deduce $(-a) \cdot (b) = -(a \cdot b) = 0$. But by P9[41],

$$a \cdot b + (-a) \cdot (b) = (a + (-a)) \cdot b = 0 \cdot b = 0.$$

So indeed, $(-a) \cdot (b) = -(a \cdot b)$.

But now we have

$$\begin{aligned} (-a) \cdot (-b) + (-(a \cdot b)) &= (-a) \cdot (-b) + (-a) \cdot (b) \quad \text{(by what we just proved above)} \\ &= (-a)((-b) + b) \quad \text{(by distributivity)} \\ &= (-a) \cdot 0 = 0. \end{aligned}$$

But also, directly from P3,

$$a \cdot b + (-(a \cdot b)) = 0.$$

It follows, either by uniqueness of additive inverses, or by cancellation for addition, that $(-a) \cdot (-b) = a \cdot b$. $\square$

This is an example of a proof that is simple, in the sense that every step is easy to justify; but not *easy*, because to get the proof right, it is necessary to come up with just the right steps!

When we come (very shortly) to introduce positive and negative numbers, we will see that Claim 3.6 really can be interpreted to say that

[40] The seemingly similar statement that $-(-a) = a$ is much simpler, and follows from P1, P2 and P3. It's left as an exercise

[41] P9 says that $X \cdot (Y + Z) = (X \cdot Y) + (X \cdot Z)$. But using P8 (commutativity of multiplication), this is exactly the same as $(Y + Z) \cdot X = (Y \cdot X) + (Z \cdot X)$, and this is the form in which P9 is being used here. As we become more familiar with the concepts of commutativity, associativity, and distributivity, will we start to make this shortcuts more and more, without explicitly saying so.

in the real numbers, negative times negative is positive. ($\diamond$)

This is a rule of numbers that's hard to make intuitive sense of, but it is an unavoidable one. If we believe axioms P1 through P9 to be true statements about real numbers (and they all seem uncontroversial), then the proof of Claim 3.6 tells us that we must, inevitably, accept ($\diamond$) as a true fact.

Before moving on we make one more (attempt at a) proof. If it clearly true that in the real numbers, the only solution to the equation

$$a - b = b - a$$

is $a = b$. We "prove" this by starting from $a - b = b - a$, adding $a + b$ to both sides, reordering terms, and applying the additive inverse axiom to get $a + a = b + b$, using multiplicative identity and distributivity to deduce $(1 + 1) \cdot a = (1 + 1) \cdot b$, and then multiplying both sides by the multiplicative inverse of $1 + 1$ to deduce $a = b$.

What's wrong with this "proof"? What is wrong is that we multiplied by the inverse of $1 + 1$. But we can only do this if $1 + 1 \neq 0$. Of course, we know that in the real numbers, $1 + 1$ is *not* 0. But, how do we know this in our axiomatic approach?

It turns out that we *cannot* prove $1 + 1 \neq 0$ using axioms P1 through P9 only. There is set of "numbers", including "0" and "1", together with operations "+" and "$\cdot$", that satisfy all of axioms P1 through P9, but that also has $1 + 1 = 0$! To capture the real numbers, we therefore need more axioms. In the next section, we introduce the three axioms of *order*, that together rule out the possibility $1 + 1 = 0$.

3.5 The axioms of order

Nothing in the axioms so far have captured the notion that there is an *order* on the real numbers — that for any two distinct numbers a, b, one of a, b is bigger than the other. One way to rectify this is to introduce a relation "$<$" (with "$a < b$" meaning "a is less than b", or "b is greater than a"), and then add some axioms that describe how "$<$" should behave.

Another approach, the one we will take, is to declare a subset of the real numbers to be the "positive" numbers, add axioms that describe how positivity behaves with respect to addition and multiplication, and then *define* an order relation in terms if positivity.

The axioms of order that we will use say that there is a collection $\mathbb{P}$ (of *positive* numbers) satisfying

- **P10, Trichotomy law**: For every a exactly one of

  1. $a = 0$
  2. $a \in \mathbb{P}$
  3. $-a \in \mathbb{P}$.

  holds.

- **P11, Closure under addition**: If $a, b \in \mathbb{P}$ then

$$a + b \in \mathbb{P}.$$

- **P12, Closure under multiplication**: If $a, b \in \mathbb{P}$ then

$$ab \in \mathbb{P}.$$

Numbers which are neither positive nor zero are referred to as *negative*; there is no special notation for the set of negative numbers. This last definition immediately says that each number is exactly one of positive, negative or 0. The trichotomy axiom also fairly immediately implies the following natural facts:

Claim 3.7. *If a is positive then $-a$ is negative; and if a is negative then $-a$ is positive.*

Proof: We start with the second point. Suppose a is negative. Then, by definition, a is neither positive nor 0, so by Trichotomy $-a$ is positive.

Now for the first point. Suppose a is positive. Then $-a$ is not positive (that violates the Trichotomy axiom). Also, $-a$ is not 0, because then $-(-a)$ would be 0 too, and (exercise) $-(-a) = a$, so $a = 0$, violating trichotomy. So by definition of negativity, $-a$ is negative. $\square$.

Quickly following on from this, we get the familiar, fundamental, and quite non-intuitive statement that the product of two negative numbers is positive.

Claim 3.8. *If a and b are negative, then ab is positive.*

Proof: Since a and b are negative we have that $-a$ and $-b$ are positive, so by P12 $(-a)(-b)$ is positive. But by Claim 3.6 $(-a)(-b) = ab$, so ab is positive. $\square$

A fundamental and convenient property of the real numbers is that it is possible to put an *order* on them: there is a sensible notion of "greater than" ("$>$") and "less than" ("$<$") such that for any two numbers a, b with $a \neq b$, either a is greater than b or a is less than b. We now define one such notion of order, using the positive-negative-zero trichotomy.

Order definitions:

- "$a > b$" means $a - b \in \mathbb{P}$
- "$a < b$" means $b > a$
- "$a \geq b$" means that either $a > b$ or $a = b$ (i.e, "$a \geq b$" is the same as "$(a > b) \vee (a = b)$")
- "$a \leq b$" means that either $a < b$ or $a = b$

Note that "$a < b$" means the same as "$b - a \in \mathbb{P}$", so the same as "$-(b - a)$ is negative", which is easily seen to be the same as "$a - b$ is negative".

Applying the trichotomy law to $a - b$, and using the definitions of $<$ and $>$, we easily determine that for every a, b, exactly one of

- $a = b$
- $a < b$
- $a > b$

holds.

In Spivak's text (Chapter 1), many properties of $<$ are derived. You should look over these, and treat them as (excellent) exercises in working with axioms and definitions. There will also be some of them appearing on the homework. You shouldn't have to memorize them (and you *shouldn't* memorize them), because they are all obvious properties, that you are already very familiar with. Nor should you be memorizing proofs. Your goal should be to do enough of these types of proofs that they become instinctive.

We'll give two examples; there will be plenty more in the homework. In the sequel, we will freely use many properties of inequalities that we have not formally proven; but we will use nothing that you couldn't prove, if you chose to, using the ideas of this section.

Claim 3.9. *If $a < b$ and $b < c$ then $a < c$.*

Proof: Since $a < b$ we have $b - a \in \mathbb{P}$ and since $b < c$ we have $c - b \in \mathbb{P}$, so by P11, closure under addition, we get that $(b - a) + (c - b) = c - a$[42] $\in \mathbb{P}$, which says $a < c$. □

Claim 3.10. *If $a < b$ and $c > 0$ then $ac < bc$.*

Proof: Since $a < b$ we have $b - a \in \mathbb{P}$ and since $c > 0$ we have $c \in \mathbb{P}$, so by P12, closure under multiplication, we get that $(b - a)c = bc - ac \in \mathbb{P}$, which says $ac < bc$. □

Related to Claim 3.10 is the fact that when an inequality is multiplied by a *negative* number, the direction of the inequality is *reversed*:

$$\text{If } a < b \text{ and } c < 0 \text{ then } ac > bc.$$

We leave the proof of this as an exercise to the reader.

We now highlight an important consequence of Claim 3.8, and also use this to introduce for the first time the word "Corollary": a result that follows in a quite direct way as an application of a previous result.

Corollary 3.11. *(Corollary of Claim 3.8) If $a \neq 0$ then $a^2 > 0$.*

[42]Note that we're using associativity, commutativity, additive inverse and additive identity axioms here, without saying so. At this point these steps should be so obvious that they can go without saying.

Proof: If $a \neq 0$ then either a is positive, in which case $a^2 = a \cdot a$ is positive by P12, of a is negative, in which case $a^2 = a \cdot a$ is also positive, this time by Claim 3.8.[43] $\square$

A few more important corollaries tumble out now. The first is the (obvious?) fact that

- $1 > 0$.

Indeed, we have $1^2 = 1 \cdot 1 = 1$ by P6, so since $1 \neq 0$[44] Corollary 3.11 applies to conclude that $1^2 = 1$ is positive.

The second is that $1 + 1 \neq 0$; this follows from the facts that 1 is positive and that positivity is closed under addition. This is the fact that we need to go back and complete our earlier "proof" that $a - b = b - a$ only if[45] $a = b$.

Have we pinned down the real numbers with axioms P1 through P12? It seems not. Our intuitive notion of the rational numbers $\mathbb{Q}$ seems to satisfy all of P1 through P12, as does our intuitive notion of the reals $\mathbb{R}$; but we have a sense that the reals are "richer" than the rationals, containing "irrational" numbers like $\sqrt{2}$ and π. So it seems that more axioms are needed to precisely pin down the notion of real numbers — more on that in a short while.

For the moment, let us mention one more set of "numbers" that satisfy P1 through P9, but fail to satisfy the order axioms. This is the set $\mathbb{C}$ of *complex* numbers, numbers of the form $a + bi$ where a and b are real numbers and i is a newly introduced symbol that acts as a "square root" of -1 — it satisfies $i^2 = -1$ (since -1 is negative, and any non-zero number, when squared, is positive, there can be no real number whose square is -1).

The complex numbers are algebraically manipulated in all the obvious ways:

- $(a + bi) + (c + di) = (a + c) + (b + d)i$, and
- $(a + bi) \cdot (c + di) = ac + adi + bci + bdi^2 = ac + adi + bci - bd = (ac - bd) + (ad + bc)i$.

Their importance lies in the following fact: in the rationals, we can solve any equation of the form $ax + b = 0$ for $a \neq 0$, but we can't solve all quadratic equations, for example we can't solve $x^2 - 2 = 0$. Moving to the reals will allow us to solve that and many other quadratic equations, but not all of them, for example we can't solve $x^2 + 1 = 0$. We *can* solve this

[43] This is our first natural example of a *proof by cases*: the assertion to be proved is of the form $p \Rightarrow q$ where p:"$a \neq 0$". By trichotomy the premise p can be written as $p_1 \vee p_2$ where p_1:"$a > 0$" and p_2:"$a < 0$". Proof by cases says that to prove $(p_1 \vee p_2) \Rightarrow q$ it is necessary and sufficient to prove both $p_1 \Rightarrow q$ and $p_2 \Rightarrow q$, that is, to "break into cases", which is exactly what we have just done.

[44] Really? Is it true that $1 \neq 0$? You could try and prove this from the axioms, but you would fail, for the simple reason that there is a set of "numbers", together with special numbers 0 and 1, operations "+" and "$\cdot$" and a subset "$\mathbb{P}$" of positive numbers, that satisfies all of axioms P1 through P12, but for which $1 \neq 0$ *fails*, that is, for which $1 = 0$! The setup is simple: let 0 be the *only* number in the set (so $0 + 0 = 0$ and $0 \cdot 0 = 0$), let the special element 1 be that same number 0, and let $\mathbb{P}$ be empty. It's easy to check that all axioms are satisfied in this ridiculous setup. To rule out this giving a perfectly good model for real numbers, we actually have to build in to our definition of real numbers the fact that $0 \neq 1$. We'll say this explicitly when we summarize the axioms later.

[45] Our first "natural" use of "only if" for "implies".

quadratic in the complex numbers, via $x = i$ or $x = -i$. But presumably there are more complex polynomials that we can't even solve in complex numbers, no? No! Amazingly, once i is introduced to the number system, *every* polynomial

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = 0$$

has a solution![46]

So, the complex numbers form a set that satisfies P1 through P9. What about P10, P11 and P12?

Claim 3.12. *It is not possible to find a subset $\mathbb{P}$ of the complex numbers for which axioms P10 through P12 hold.*

Proof: Suppose it was possible to find such a subset $\mathbb{P}$ of "positive" complex numbers. Consider the number i. We certainly have $i \neq 0$ (if $i = 0$ then $i^2 = 0$, but also $i^2 = -1$ by definition, so $-1 = 0$; and adding 1 to both sides gives $0 = 1$, a contradiction).

Since $i \neq 0$ we have $-1 = i^2 > 0$ by Corollary 3.11, so -1 is positive, so 1 is negative; but this contradicts the fact that 1 is positive.

This contradiction proves that no such a subset $\mathbb{P}$ can exist. □

Axioms P1 through P12 describe a mathematical object called an *ordered field*; the above claim demonstrates that $\mathbb{C}$ is *not* an example of an ordered field.

3.6 The absolute value function

The *absolute value* of a number is a measure of "how far from 0" the number is, without regard for whether it is positive or negative.

- 0 itself has absolute value 0;
- if two positive numbers a and b satisfy $a < b$ (so b is bigger, "more positive" than a, "further from 0"), then the absolute value of a is smaller than the absolute value of b;
- if they are both negative and $a < b$ (so a is "more negative" than b) then the absolute value of a is bigger than the absolute value of b; and
- if a and b are negatives of each other ($a = -b$, $b = -a$) then they have the same absolute value.

The formal definition of the absolute value function is: for real a, the *absolute value* of a, denoted $|a|$, is given by

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

[46] This is the *fundamental theorem of algebra*.

So, for example, $|2| = 2$, $|-\pi| = \pi$, and $|1 - \sqrt{2}| = \sqrt{2} - 1$. We will frequently define functions using this "brace" notation, so you have to get used to reading and using it. The brace notation above says that there are three different, disjoint regimes for the answer to the question "what is $|a|$?" — the regime $a > 0$ (where the answer is "a"); the regime $a = 0$ (where the answer is "0"); and the regime $a < 0$ (where the answer is "$-a$"). The three regimes have no overlap, so there is no possible ambiguity in the definition[47], and by trichotomy they cover all reals, so there are no gaps in the definition.

Noting that when $a = 0$ the numbers "a" and "0" coincide, we could have been a little more efficient, and broken into the two regimes $a \geq$ [48]0 and $a < 0$, to get:

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

It's a matter of taste which approach to take.

We will use the absolute value to create a notion of "distance" between numbers: the distance between a and b is $|a - b|$. Representing a and b on a number line, $|a - b|$ can be thought of as the length of the line segment joining a and b (a quantity which is positive, whether $a < b$ or $a > b$).

A fundamental principle of the universe is that "the shortest distance between two points is a straight line". A mathematical interpretation of this principle says that for any sensible notion of "distance" in a space, for any three points x, y, z

$$\text{(the distance from } x \text{ to } y\text{)}$$

is no larger than

$$\text{(the distance from } x \text{ to } z\text{) plus (the distance from } z \text{ to } y\text{)},$$

that is, it can never be *quicker* to get from x to y, if you demand that you must pass through a particular point z on the way (though it might be just as quick, if z happens to lie on a shortest path between x and y). The mathematical study of *metric spaces* explores these ideas.

In the context of using absolute value as a notion of distance between two real numbers, consider $x = a$, $y = -b$ and $z = 0$. The distance from x to y is $|a - (-b)| = |a + b|$, the distance from x to 0 is $|a - 0| = |a|$, and the distance from 0 to $-b$ is $|0 - (-b)| = |b|$. If we believe that absolute value is sensible as a notion of distance, then we would expect that $|a + b| \leq |a| + |b|$. This is indeed the case. The following, called the *triangle inequality* is one of the most useful tools in calculus.

[47] Sometimes we will present braced definitions in which there *is* overlap between regimes. As long as the two potentially conflicting clauses agree at the points of overlap (and they usually are just points), this is fine, if a little sloppy. As an example, this is an unambiguous and complete definition of absolute value, with two overlapping regimes:

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

[48] Remember "$a \geq b$" is shorthand for "either $a > b$ or $a = b$"

Claim 3.13. *(Triangle inequality) For all reals a, b, $|a+b| \leq |a| + |b|$.*

Proof: Because the absolute value function is defined in cases, it makes sense to consider cases for a, b.

Case 1, $a, b \geq 0$ In this case, $|a| = a$, $|b| = b$, and (since $a + b \geq 0$), $|a + b| = a + b$, and so $|a + b| = |a| + |b|$.

Case 2, $a, b \leq 0$ In this case, $|a| = -a$, $|b| = -b$, and (since $a + b \leq 0$), $|a + b| = -a - b$, and so again $|a + b| = |a| + |b|$.

Case 3, $a \geq 0$**,** $b \leq 0$ Here we know $|a| = a$ and $|b| = -b$, but what about $|a + b|$?

If $a + b \geq 0$ then $|a + b| = a + b$, and to verify the triangle inequality in this case we need to establish

$$a + b \leq a - b,$$

or $b \leq -b$. Since $b \leq 0$, we have $-b \geq 0$ and $0 \leq -b$, so indeed $b \leq -b$.

If, on the other hand, $a + b < 0$ then $|a + b| = -a - b$, and to verify the triangle inequality in this case we need to establish

$$-a - b \leq a - b,$$

or $-a \leq a$. Since $a \geq 0$ (and so $0 \leq a$), we have $-a \leq 0$, so indeed $-a \leq a$.

Case 4, $a \leq 0$**,** $b \geq 0$ This is almost identical to Case 3, and we omit the details.

□

Another, more conceptual, proof of the triangle inequality appears in Spivak.

The absolute value function appears in two of the most important definitions of calculus — the definitions of limits and continuity — so it behooves us to get used to working with it. The standard approach to dealing with an expression involving absolute values is to break into cases, in such a way that within each case, all absolute value signs can be removed. As an example, let us try to find all real x such that

$$|x - 1| + |x - 2| > 1.$$

The clause in the absolute value definition that determines $|x - 1|$ changes at $x = 1$, and the clause that determines $|x - 2|$ changes at $x = 2$. It makes sense, then, to consider five cases: $x < 1$, $x = 1$, $1 < x < 2$[49], $x = 2$ and $x > 2$.

[49]This is shorthand for "$1 < x$ and $x < 2$."

Case 1: $x < 1$ Here $|x-1| = 1-x$ (since $x-1 < 0$ in this case), and $|x-2| = 2-x$, so $|x-1| + |x-2| = 3-2x$ and $|x-1| + |x-2| > 1$ is the same as $3-2x > 1$ or $1 > x$. So: in the regime $x < 1$, $|x-1| + |x-2| > 1$ is true exactly when $x < 1$, which it always is in this regime, and we conclude that the set of all $x < 1$ is one set of numbers satisfying the inequality.

Case 2: $x = 1$ Here $|x-1| + |x-2| = 1$, so the inequality is not satisfied.

Case 3: $1 < x < 2$ Here $|x-1| + |x-2| = x - 1 + 2 - x = 1$ and again the inequality is not satisfied.

Case 4: $x = 2$ Here $|x-1| + |x-2| = 1$, so again the inequality is not satisfied.

Case 5: $x > 2$ Here $|x-1| + |x-2| = 2x - 3$ and the inequality becomes $x > 2$, which is true always in this regime, and we conclude that the set of all $x > 2$ is another set of numbers satisfying the inequality.

Having finished the case analysis, we conclude that the inequality is satisfied when x is less than 1 and when x is greater than 2.[50]

3.7 The completeness axiom

This section introduces the completeness axiom, which allows us to give a complete (no pun intended) description of the real numbers. Almost immediately after we are done with this section, the complete axiom will fade into the background. But in a few weeks, when we come to the major theorems of continuity — the intermediate value theorem and the extreme value theorem — it will come blazing back to the foreground, spectacularly.

Our intuition about the real numbers suggests that it cannot be the case that axioms P1 through P12 are *not* enough to pin down the real precisely, or uniquely: both $\mathbb{Q}$ and $\mathbb{R}$ (as we understand them, informally) satisfy all the axioms so far; but surely $\mathbb{R}$ contains more numbers than $\mathbb{Q}$ — numbers like $\sqrt{2}$, π, and e — so in particular $\mathbb{Q}$ and $\mathbb{R}$ should be different sets that both satisfy P1 through P12. We formalize this now, by presenting the ancient[51] proof that $\sqrt{2}$ is not a rational number.

Claim 3.14. *There do not exists natural numbers a, b with $\frac{a^2}{b^2} = 2$.*

Proof: Suppose, for a contradiction, that there are natural numbers a, b with $\frac{a^2}{b^2} = 2$. If a and b are both even, say $a = 2m$ and $b = 2n$ for some natural numbers m, n, then we have

$$\frac{m^2}{n^2} = \frac{4m^2}{4n^2} = \frac{(2m)^2}{(2n)^2} = \frac{a^2}{b^2} = 2,$$

[50] We will soon see the standard way to represent sets like this.

[51] Literally ancient — this proof is hinted at in Aristotle's *Prior Analytics*, circa 350BC.

so we could just as well use the pair of numbers m, n. By repeating this process, of dividing each number by 2 if both are even, until we can no longer do this, we reach a point where we have two natural numbers a', b', *not both even*, with $\frac{(a')^2}{(b')^2} = 2$.

We have $(a')^2 = 2(b')^2$, so $(a')^2$ is even. But that implies that a' is even (an odd number — say one of the form $2k + 1$ for natural number k — can't square to an even number, since $(2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1$, which is odd since $2k^2 + 2k$ is a whole number).

So $a' = 2c$ for some whole number c. Plugging in to $(a')^2 = 2(b')^2$, this yields $4c^2 = 2(b')^2$ or $2c^2 = (b')^2$. So $(b')^2$ is even, implying that b' is even.

This is a contradiction — we have that a', b' are not both even, and simultaneously are both even. It follows that there are *no* natural numbers a, b with $\frac{a^2}{b^2} = 2$. □

Given that we would *like* there to be a square root of 2 in the real numbers, it makes sense to hunt for some more axioms. It turns out that we actually need just one more. Our intuitive sense is that though there are "gaps" in the rationals, the gaps are "small", and in fact the rationals are in some sense "dense" in the reals[52] — for every real r, there can be found a sequence of rational numbers that approaches arbitrarily close to r. Indeed, if we believe that every real r has a decimal expansion

$$r = a.d_1d_2d_3d_4\ldots,$$

then the sequence of rational numbers

$$a, a.d_1, a.d_1d_2, a.d_1d_2d_3, \ldots$$

always lies below r, but eventually gets as close to r as one wishes. So, if we make sure that not only do the real contain all the rational numbers, but also contain all the "numbers" that can be approached arbitrarily closely by sequences of rationals, then it doesn't seem beyond the bounds of possibility that we would have "filled in all the gaps" in the rationals.

We'll formalize this idea with a single extra axiom. To state it sensibly, we need to introduce the notions of upper bounds and least upper bounds.

Informally, b is an upper bound for a set S of numbers if b is at least as large as everything in S. Formally, say that b is an *upper bound* for S if

$$\text{for all } s, \text{ if } s \text{ is an element of } S \text{ then } s \leq b.$$

Some sets have upper bounds. For example,

- the set A of

  $$\text{all numbers that are strictly bigger than 0 and strictly less than 1}$$

  has 12 as an upper bound, and also 6, and 3, and 2.5, and 1; but not 1/2, or .99, or 0 or -10; and

[52] We will formalize this later.

- the set B of

  non-positive numbers (number that are 0 or less than 0)

  has 0 as an upper bound, and also *any* number bigger than 0, but not any number less than 0.

Other sets *don't* have upper bounds, such as

- the set of reals itself (there is no real number that is at least as large as all other real numbers, since for every real r the number $r + 1$ is a real that is bigger than r); and
- the natural numbers (for the same reason).

What about the empty set, the set that contains *no* numbers? We denote this set by $\emptyset$. Does $\emptyset$ have an upper bound? Yes! The number b is an upper bound for a set if

for all s, if s is an element of the set, then $s \leq b$.

Pick an arbitrary b, and then an arbitrary s. The premise of the implication "if s is an element of the empty set, then $s \leq b$" is "s is an element of the empty set". This premise is *false*. So the implication is true, for an arbitrary s and for all s. And so b is an upper bound for an arbitrary b, and so for all b.

This may seem counter-intuitive, but it is a consequence of the way we define implication. This is a situation where it might be very helpful to think of "p implies q" as meaning "either p is false, or q is true". Then the definition of b being an upper bound for S becomes

b is an upper bound for S if, for all s, either s is not a member of S, or $s \leq b$.

In this form, it is clear that *every* number is an upper bound for the empty set.

Having talked about upper bounds, we now introduce the concept of a least upper bound. A number b is a *least upper bound* for a set S if

- it is an upper bound for S (this is the "upper bound" part of the definition) and
- if b' is any other upper bound for S, then $b \leq b'$ (this is the "least" part of the definition).

The definition talks about "a" least upper bound; but it should be clear that if S has a least upper bound, then it has only one. Indeed, if b and b' are both least upper bounds then $b \leq b'$ (because b is a *least* upper bound), and also $b' \leq b$ (because b' is a *least* upper bound), so in fact $b = b'$.

Let's look at three examples: for the set A above, 1 and all numbers greater than 1 are upper bounds, and no other numbers are. So A has a least upper bound, and it is the number 1. Note that in this example, the least upper bound is *not* in the set A.

For the set B above, 0 and all numbers greater than 0 are upper bounds, and no other numbers are. So B has a least upper bound, and it is the number 0. Note that in this example, the least upper bound *is* in the set B.

It might seem like I'm working towards suggesting that *every* set that has an upper bound has a least upper bound. But our third example, the empty set, nixes that suggestion: every number is an upper bound for $\emptyset$, so it has upper bounds but no *least* upper bound.

But $\emptyset$ seems quite special. Maybe every *non-empty* set that has an upper bound, has a *least* upper bound? This is not a theorem that we can prove, using just axioms P1 through P12. Here's an informal reason: we sense that there is a "gap" in the rationals, where $\sqrt{2}$ should be. So, inside the rationals, consider the set C of all numbers x satisfying $x^2 < 2$. This set has an upper bound — 2, for example. But does it have a least upper bound? If b is any upper bound, then it must be that $b^2 > 2$ (we can't have $b^2 = 2$, since we are in the world of rationals; nor can we have $b^2 < 2$, because then we should be able to find another rational b', slightly larger than b, with $(b')^2$ still less than 2 — this using the idea that $\sqrt{2}$ can be approached arbitrarily closely by rationals). But (again using the idea that $\sqrt{2}$ can be approached arbitrarily closely by rationals) if $b^2 > 2$, then it can't be a *least* upper bound, because we should be able to find another rational b', slightly smaller than b, with $(b')^2$ still greater than 2, that acts as a lesser upper bound for the set C.

So, that every *non-empty* set that has an upper bound, has a *least* upper bound, is not a theorem we can prove in P1 through P12; but it seems like a good fact to have, because it seems to allow for the "filling in of the gaps" in the rationals — in the example discussed informally above, if C *had* a least upper bound b, it seems pretty clear that it should satisfy $b^2 = 2$, and so b should act as a good candidate for being the square root of 2.

This motivates the last, and most subtle, and most powerful, axiom of the real numbers, the *completeness axiom*:

- **P13, Completeness**: If A is a non-empty set of numbers that has an upper bound, then it has a least upper bound.

The notation that is traditionally used for the (remember, it's unique if it exists) least upper bound of a set A of numbers is "l.u.b. A" or (much more commonly)

$$\sup A$$

(sup here is short for "supremum"). So the completeness axiom says:

If A is a non-empty set of numbers that has an upper bound, then $\sup A$ exists.

There's an equivalent form of the Completeness axiom, that involves lower bounds. Say that b is a *lower bound* for a set S if it is no bigger than any element of S, that is, if

$$\text{for all } s, \text{ if } s \text{ is an element of } S \text{ then } b \leq s.$$

(As before with upper bounds, some sets have lower bounds, and some don't; and *every* number is a lower bound for the empty set.) A number b is a *greatest lower bound* for a set S if

- it is a lower bound for S and
- if b' is any other lower bound for S, then $b \geq b'$.

(As with least upper bounds, this number, if it exists, is easily seen to be unique). The notation that is traditionally used for the greatest lower bound of a set A of numbers is "g.l.b. A" or (much more commonly)

$$\inf A$$

(inf here is short for "infimum").

Following through our discussion about least upper bounds, but now thinking about greatest lower bounds, it seems reasonable clear that non-empty sets of numbers with lower bounds should have greatest lower bounds. This doesn't need a new axiom; it follows from (and in fact is equivalent to) the Completeness axiom.

Claim 3.15. *If A is a non-empty set of numbers that has a lower bound, then* $\inf A$ *exists.*

Proof: Consider the set $-A := \{-a : a \in A\}$. This is non-empty since A is non-empty, and it has an upper bound; if b is a lower bound for A, then $-b$ is an upper bound for $-A$. So $-A$ has a least upper bound, call it α.

We claim that $-\alpha$ is a greatest lower bound for A. Since α is an upper bound for $-A$, we have $\alpha \geq -a$ for all $a \in A$, so $-\alpha \leq a$, so $-\alpha$ is certainly a lower bound for A. Now suppose β is another lower bound for A. Then $-\beta$ is an upper bound for $-A$, so $-\beta \geq \alpha$, so $\beta \leq -\alpha$, so $-\alpha$ is the *greatest* lower bound for A. □

The key point in this proof is the following fact, which is worth isolating and remembering:

$$\inf A = -\sup(-A) \text{ where } -A = \{-x : x \in A\}.$$

3.8 Examples of the use of the completeness axiom

When we come to discuss continuity, we will see plenty of examples of the power of P13. For now, we give a few fairly quick examples. As a first example, we give a formalization of the informal discussion about the existence of $\sqrt{2}$ that came up earlier.

Claim 3.16. *There is a number x with $x^2 = 2$.*

Proof: Let C be the set of numbers a satisfying $a^2 < 2$.

C is non-empty — for example, 0 is in C. Also, C has an upper bound. For example, 2 is an upper bound. Indeed, we have $2^2 = 4 > 2$, and if $y \geq 2$ then $y^2 \geq 2^2 = 4 > 2$, so an element of C must be less than 2.

By the completeness axiom, C has a (unique) least upper bound. Call it x. We claim that $x^2 = 2$; we will prove this by ruling out the possibilities $x^2 < 2$ and $x^2 > 2$.

Suppose $x^2 < 2$. Consider the number

$$x' = \frac{3x+4}{2x+3}.$$

(It certainly is the case that $x > 0$, so x' really is a number — we are not guilty of accidentally dividing by zero.) Note that $x < x'$ is equivalent to $x < (3x+4)(2x+3)$, which is equivalent to $2x^3 + 3x < 3x + 4$, which is equivalent to $x^2 < 2$, which is true, so $x < x'$. And note also that $(x')^2 < 2$ is equivalent to

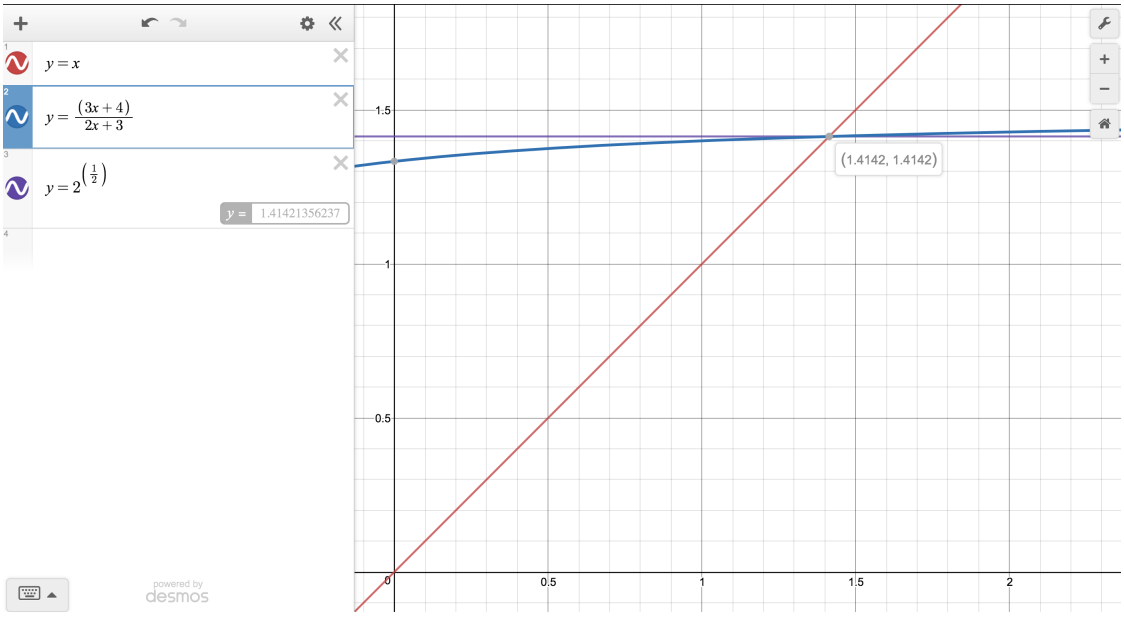
$$\left(\frac{3x+4}{2x+3}\right)^2 < 2,$$

which (after some algebra) is equivalent to $x^2 < 2$, which is true, so $(x')^2 < 2$. It follows that $x \in C$, but since $x < x'$ this contradicts that x is an upper bound for C. So we conclude that it is not the case that $x^2 < 2$.

Now suppose $x^2 > 2$. Again consider $x' = (3x+4)/(2x+3)$. Similar algebra to the last case shows that now $x' < x$ and $(x')^2 > 2$, so $y^2 > 2$ for any $y \geq x'$, so all elements of C are less than x', so x' is an upper bound for C, contradicting that x is the *least* upper bound for C. So we conclude that it is not the case that $x^2 > 2$.

We are left only with the possibility $x^2 = 2$, which is what we wanted to prove. □

The picture below illustrates what's going on in the proof above. The red line is $y = x$, the blue curve is $y = (3x+4)/(2x+3)$, and the purple line is $y = \sqrt{2}$. All three lines meet at $(\sqrt{2}, \sqrt{2})$. The blue curve is between the two lines: above the red & below the purple before $\sqrt{2}$, and below the red, above the purple after $\sqrt{2}$.



So: if we take any number x with $x^2 < 2$ (so in regime where red is below blue is below purple), the three graphs illustrate that $(3x+4)/(2x+3)$ is bigger than x, but its square is

still less than 2. So x can't be the l.u.b. for the set of numbers whose square is less than 2 — it's not even an upper bound, since it's smaller than $(3x+4)/(2x+3)$.

And: if we take any number x with $x^2 > 2$ (so in regime where purple is below blue is below red), the graphs illustrate that $(3x+4)/(2x+3)$ is smaller than x, but its square is still greater than 2. So x can't be the l.u.b. for the set of numbers whose square is less than 2 — it's an upper bound, but $(3x+4)/(2x+3)$ is a better (lesser) upper bound.

So the l.u.b. for the set of numbers whose square is less than 2 — which exists by the completeness axiom — has a square which is neither less than nor greater than 2. By trichotomy, it must be equal to 2.

The algebra in the proof is simply verifying that indeed red is below blue is below purple when $x^2 < 2$, and purple is below blue is below red when $x^2 > 2$.

Later we will prove the Intermediate Value Theorem (IVT), a powerful result that will make it essentially trivial to prove the existence of the square root, or cubed root, or any root, of 2, or any positive number. Of course, there is no such thing as a free lunch — we will need the completeness axiom to prove the IVT.

The next three examples are probably best looked at after reading the section on Natural numbers, as a few concepts from that section get used here.

The first of these is the use of completeness is to demonstrate the "obvious" fact that the natural numbers

$$\mathbb{N} = \{1, 1+1, 1+1+1, \ldots\} = \{1, 2, 3, \ldots\}$$

forms an *unbounded* set (a set with no upper bound). While this seems obvious, it should not actually be; there exist examples of sets of numbers satisfying P1-P12, in which $\mathbb{N}$ is *not* unbounded, meaning that P13 is absolutely necessary to prove this result.

The proof goes as follows. $\mathbb{N}$ is non-empty. Suppose it is bounded above. Then, by P13, it has a least upper bound, i.e., there's an $\alpha = \sup \mathbb{N}$. We have $\alpha \geq n$ for all $n \in \mathbb{N}$. Now if $n \in \mathbb{B}$, so is $n+1$, so this says that $\alpha \geq n+1$ for all $n \in \mathbb{N}$. Subtracting one from both sides, we get that $\alpha - 1 \geq n$ for all $n \in \mathbb{N}$. That makes $\alpha - 1$ an upper bound for $\mathbb{N}$, and one that is smaller than α, a contradiction! So $\mathbb{N}$ must not be bounded above.

Closely related to this is the *Archimedean property* of the real numbers:

> Let r be any positive real number (think of it as large), and let ε be any positive real number (think of it as small). Then there is a natural number n such that $n\varepsilon > r$.[53]

The proof is very quick: Suppose the property were false. Then there is some $r > 0$ and some $\varepsilon > 0$ such that $n\varepsilon \leq r$ for all $n \in \mathbb{N}$, so $n \leq r/\varepsilon$, so $\mathbb{N}$ is bounded above, and that's a contradiction.

[53] In his book *The sand-reckoner*, Archimedes put an upper bound on the number of grains of sand that could fit in the universe. Think of the Archimedean property as saying "no matter how small your grains of sand, or how large your universe, if you have enough grains of sand you will eventually fill the entire universe."

A simple and tremendously useful corollary of the Archimedean property is the special case $r = 1$:

> for all $\varepsilon > 0$ there is a natural number n such that $n\varepsilon > 1$, that is, so that $1/n < \varepsilon$.

The final application we give of the Completeness axiom we give in this quick introduction is to the notion of density.

> A set $S \subseteq \mathbb{R}$ is *dense in* $\mathbb{R}$ if for all $x < y$ in $\mathbb{R}$, there is an element of S in (x, y).[54] We also say that S is a *dense subset* of $\mathbb{R}$.

For example, the set of reals itself forms a dense subset of the reals, rather trivially, as does the set of reals minus one point. The set of positive numbers is *not* dense (there is no positive number between -2 and -1), and nor is the set of integers (there is no integer between 1.1 and 1.9).

Our intuition is that the rationals *are* dense in the reals. This is indeed the case.

Claim 3.17. *$\mathbb{Q}$ is dense in $\mathbb{R}$ — if x, y are reals with $x < y$, then there is a rational in the interval (x, y).*

Proof: We'll prove that for $0 \leq x < y$ there's a rational in (x, y). Then given $x < y \leq 0$, there's a rational r in $(-y, -x)$, so $-r$ is a rational in (x, y); and given $x < 0 < y$, any rational in $(0, y)$ is in (x, y)[55].

So, let $0 \leq x < y$ be given. By the Archimedean property, there's a natural number n with $1/n < y - x$. The informal idea behind the rest of the proof is that, because the gaps between consecutive elements in the "$1/n$" number line

$$\{\ldots, -3/n, -2/n, -1/n, 0, 1/n, 2/n, 3/n, \ldots\}$$

are all smaller than the distance between x and y, one (rational) number in this set must fall between x and y.

Formally: Because $\mathbb{N}$ is unbounded, there's $m \in \mathbb{N}$ with $m \geq ny$. Let m_1 be the least such (this is an application of the well-ordering principle). Note that $m_1 > 1$, because $y - x > 1/n$, so $y > 1/n$, so $1 < ny$. Consider $(m_1 - 1)/n$. We have $(m_1 - 1) < ny$ (or else m_1 would not have been the least integer at least as large as ny) and so $(m_1 - 1)/n < y$. If $(m_1 - 1)/n \leq x < y \leq m_1/n$ then $1/n \geq y - x$, a contradiction. So $(m_1 - 1)/n > x$, and thus $(m_1 - 1)/n \in (x, y)$. □

We also should believe that the *irrational* numbers are dense in $\mathbb{R}$. There's a quite ridiculous proof of this fact, that used the irrationality of $\sqrt{2}$.

Claim 3.18. *$\mathbb{R} \setminus \mathbb{Q}$ is dense in $\mathbb{R}$.*

Proof: Let $x < y$ be given. There's a rational r in $(x/\sqrt{2}, y/\sqrt{2})$, by density of the rationals. But then $\sqrt{2}r \in (x, y)$, and $\sqrt{2}r$ is irrational! □

[54] S stands for *Starbucks* — between any two points in New York City, there is a Starbucks.

[55] Or even more simply (as was pointed out by someone in class one day) $0 \in (x, y)$

3.9 A summary of the axioms of real numbers

To summarize this chapter: the real numbers, denoted $\mathbb{R}$, is a set of objects, which we call *numbers*, with

- two special numbers, 0 and 1, that are distinct from each other,[56]
- an operation $+$, *addition*, that combines numbers a, b to form the number $a + b$,
- an operation $\cdot$, *multiplication*, that combines a, b to form $a \cdot b$, and
- a set $\mathbb{P}$ of *positive* numbers,

that satisfies the following 13 axioms:

P1, Additive associativity For all a, b, c, $a + (b + c) = (a + b) + c$.

P2, Additive identity For all a, $a + 0 = 0 + a = a$.

P3, Additive inverse For all a there's a number $-a$ with $a + (-a) = (-a) + a = 0$.

P4, Additive commutativity For all a, b, $a + b = b + a$.

P5, Multiplicative associativity For all a, b, c, $a \cdot (b \cdot c) = (a \cdot b) \cdot c$.

P6, Multiplicative identity For all a, $a \cdot 1 = 1 \cdot a = a$.

P7, Multiplicative inverse For all a, if $a \neq 0$ there's a number a^{-1} such that $a \cdot a^{-1} = a^{-1} \cdot a = 1$.

P8, Multiplicative commutativity For all a, b, $a \cdot b = b \cdot a$.

P9, Distributivity of multiplication over addition For all a, b, c, $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$.

P10, Trichotomy law For every a exactly one of

- $a = 0$
- $a \in \mathbb{P}$
- $-a \in \mathbb{P}$.

holds.

P11, Closure under addition If $a, b \in \mathbb{P}$ then $a + b \in \mathbb{P}$.

P12, Closure under multiplication If $a, b \in \mathbb{P}$ then $ab \in \mathbb{P}$.

[56] See the footnote just after the proof of Corollary 3.11.

P13, Completeness If A is a non-empty set of numbers that has an upper bound, then it has a least upper bound.

We can legitimately talk about *the* real numbers: there is (essentially) a *unique* structure that satisfies all the above axioms, and it can be explicitly constructed. We will say no more about this; but note that Spivak discusses this topic in Chapters 28 through 30 of his book.

4 Induction

Let X be any set of numbers that satisfies each of the axioms P1 through P12 (X might be the rational numbers, or the real numbers, or any number of other possibilities). Inside X there is a copy of what we will think of as the "natural numbers", namely

$$\mathbb{N} = \{1, 1+1, 1+1+1, \ldots\} \quad \text{or} \quad \mathbb{N} = \{1, 2, 3, \ldots\}.$$

(I'm going to assume that everyone is familiar with the standard naming convention of Arabic numbers!) Notice that we have

$$1 < 1+1 < 1+1+1 < \cdots \quad \text{or} \quad 1 < 2 < 3 < \cdots,$$

since $1 > 0$, so adding one more "1" to a sum of a collection of 1's increases the sum.

This definition of the natural numbers is somewhat informal (what exactly does that "..." mean?), but it will work perfectly well for us while we introduce the most important property of the natural numbers, the principle of mathematical induction. In this section we'll discuss induction first in this informal setting. We'll then present a more formal definition of $\mathbb{N}$, and indicate how (in principle at least) we could establish all of $\mathbb{N}$'s expected properties in this formal setting.

4.1 The principle of mathematical induction (informally)

We have already encountered a number of situations in which we would like to be able to prove that some predicate, that depends on a natural number n, is true for *every* $n \in \mathbb{N}$. Examples include:

- if $a_1, \ldots, a_n$ are n arbitrary reals, then the sum $a_1 + a_2 + \cdots + a_n$ does not depend on the order in which parentheses are put around the a_i's, and
- if $a_1, \ldots, a_n$ are n arbitrary reals, then the sum of the a_i's does not depend on the order in which the a_i's are arranged in the sum.

We know that we can, in principle, use the axioms of the real numbers to prove each of these statements *for any particular* n, but it seems like this case-by-case approach would require *infinite* time to prove either of the statements for *every* n.

There's a fix. Let's pick one of these predicates, call it $p(n)$. Suppose we can prove

A that $p(1)$ is true

and we can also give an argument that shows that

B for any arbitrary natural number n, $p(n)$ implies $p(n+1)$.

Armed with these two weapons, we have a convincing argument that $p(n)$ is true for *every* n. Indeed, if a friend were to challenge us to provide them with a proof of $p(7)$, we would tell them:

- well, $p(1)$ is true (that's **A**), so
- since $p(1)$ is true, and $p(1)$ implies $p(2)$ (that's **B**, in the specific case $n = 1$), we conclude via modus ponens that $p(2)$ is true, so
- since $p(2)$ is true, and $p(2)$ implies $p(3)$ (**B** for $n = 2$), modus ponens again tells us that $p(3)$ is true, so
- since $p(3)$ is true, and $p(3)$ implies $p(4)$, $p(4)$ is true, so
- since $p(4)$ is true, and $p(4)$ implies $p(5)$, $p(5)$ is true, so
- since $p(5)$ is true, and $p(5)$ implies $p(6)$, $p(6)$ is true, so
- since $p(6)$ is true, and $p(6)$ implies $p(7)$, $p(7)$ is true.

And if instead they challenged us to prove $p(77)$, we would do the same thing, just with many more lines. There's a *uniform* proof of $p(n)$ for *any* n — one that doesn't require a specific examination of $p(n)$, but simply one appeal to **A** followed by $n - 1$ identical appeals to **B** and modus ponens. Because of this uniformity, we can simply present **A** and **B** as a proof of $p(n)$ for *all* n. If our friend wants a specific proof of $p(777)$ from this, they are free to supply the 777 required steps themselves!

As long as **A** and **B** can both be given finite length proofs, this gives a finite length proof of $p(n)$ for infinitely many n. We summarize this:

> **The principle of mathematical induction**: Let $p(n)$ be a predicate, with the universe of discourse for n being natural numbers. If $p(1)$ is true, and if, for arbitrary n, $p(n)$ implies $p(n + 1)$, then $p(n)$ is true for all n.

Some notation:

- a proof using the principle of mathematical induction is commonly called a *proof by induction*;
- the step in which $p(1)$ is verified is called the *base case* of the induction; and
- the step in which it is established that for arbitrary n, $p(n)$ implies $p(n + 1)$ (a step which will almost always involve symbolic manipulations of expressions involving n, where no *specific* properties of n are used), is called the *induction step*.

Here is a very tangible illustration of what's going on with induction:

> **The principle of mathematical induction, ladder version**: If you have a way of getting on a ladder, and if you have a way of going from any rung of the ladder to the next rung up, then you can get as high up the ladder as you wish.

Proving identities via induction

Let's have an example. What's the sum of the first n natural numbers? Well:

- $1 = 1$,
- $1 + 2 = 3$,
- $1 + 2 + 3 = 6$,
- $1 + 2 + 3 + 4 = 10$,
- $1 + 2 + 3 + 4 + 5 = 15$,
- $1 + 2 + 3 + 4 + 5 + 6 = 21$,
- $1 + 2 + 3 + 4 + 5 + 6 + 7 = 28$,
- $1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 = 36$,
- $1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 = 45$,
- $1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 = 55$.

A pattern seems to be emerging: it appears that $1 + 2 + \ldots + n = n(n + 1)/2$.

Claim 4.1. *For every natural number n,*

$$1 + 2 + 3 + \ldots + n = \frac{n(n+1)}{2}.$$

Proof: Let $p(n)$ be the predicate

$$p(n) : \text{“}1 + 2 + 3 + \ldots + n = \frac{n(n+1)}{2}\text{”}$$

(where the universe of discourse for n is natural numbers). We prove that $p(n)$ is true for all n, by induction.

Base case: $p(1)$ is the assertion that $1 = 1(2)/2$, or $1 = 1$, which is true.

Induction step: Let n be an arbitrary natural number. We want to establish the implication

$$p(n) \text{ implies } p(n + 1),$$

that is to say, we want the establish that this statement has truth value T. By definition of implication, this is the same as showing that the statement

$$\text{either (not } p(n)) \text{ or } p(n + 1)(\star)$$

has truth value T.

If $p(n)$ is false, then (not $p(n)$) is true, so $(\star)$ is indeed true. If $p(n)$ is true, then (not $p(n)$) is false, so to establish that $(\star)$ is true we need to show that $p(n+1)$ is true. *But*, we don't have to start an argument establishing $p(n+1)$ from scratch — we are in the case where $p(n)$ is true, so *we get to assume* $p(n)$ *as part of our proof of* $p(n+1)$.

$p(n+1)$ is the assertion

$$1+2+3+\cdots+n+(n+1)=\frac{(n+1)((n+1)+1)}{2}$$

or

$$(1+2+3+\cdots+n)+(n+1)=\frac{(n+1)((n+2)}{2}.(\star\star)$$

Since $p(n)$ is being assumed to be true, we get the assume that

$$1+2+3+\cdots+n=\frac{n(n+1)}{2},$$

and so $(\star\star)$ (the statement whose truth we are trying to establish) becomes

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

Multiplying both sides by 2, and expanding out the terms, this becomes

$$n^2+n+2n+2=n^2+3n+2,$$

which is true.

We have established the truth of the implication "$p(n)$ implies $p(n+1)$", for arbitrary n, and so we have showing that the induction step is valid.

Conclusion: By the principle of mathematical induction, $p(n)$ is true for all natural numbers n, that is,

$$1+2+3+\ldots+n=\frac{n(n+1)}{2}.$$

□

Of course, this write-up was filled with overkill. In particular, in proving the truth of the implication $p \Rightarrow q$ we almost never explicitly write that if the premise p is false then the implication is true; so it is very typical to start the induction step with "Assume $p(n)$. We try to deduce $p(n+1)$ from this." Also, we very often don't even explicitly introduce the predicate $p(n)$. Here is a more condensed write-up of the proof, that should act as template for other proofs by induction.

Claim 4.2. *For every natural number n,*

$$1+2+3+\ldots+n=\frac{n(n+1)}{2}.$$

Proof: We proceed by induction on n.

Base case: The base case $n = 1$ is obvious.

Induction step: Let n be an arbitrary natural number. Assume

$$1 + 2 + 3 + \ldots + n = \frac{n(n+1)}{2}.$$

From this we get

$$\begin{aligned} 1 + 2 + 3 + \ldots + n + (n+1) &= (1 + 2 + 3 + \ldots + n) + (n+1) \\ &= \frac{n(n+1)}{2} + n + 1 \\ &= \frac{n^2 + n + 2n + 2}{2} \\ &= \frac{n^2 + 3n + 2}{2} \\ &= \frac{(n+1)(n+2)}{2} \\ &= \frac{(n+1)((n+1)+1)}{2}. \end{aligned}$$

The equality of the first and last expressions in this chain is the case $n + 1$ of the assertion, so we have verified the induction step.[57]

By induction the assertion is true for all n. □

In proving an identity — an equality between two expressions, both depending on some variable(s) — by induction, it is often very helpful to start with one side of the $n + 1$ case of the identity, and manipulate it via a sequence of equalities in a way that introduces one side of the n case of the identity into the mix; this can then be replaced with the *other* side of the n case, and then the whole thing might be massage-able into the other side of the $n + 1$ identity. That's exactly how we proceeded above.

Now is a good time to introduce *summation notation*. We write

$$\sum_{k=1}^{n} a_k$$

as shorthand for

$$a_1 + a_2 + a_3 + \cdots + a_{k-1} + a_k.$$

k is called the *index of summation*, and the a_k's are the *summands*. For example, we have

$$\sum_{k=1}^{7} k^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 + 7^2,$$

[57] Notice that the induction step is presented here as a complete english-language paragraph, even though it involves a lot of mathematics. Read it aloud!

$$\sum_{k=1}^{2} f(k) = f(1) + f(2)$$

and

$$\sum_{k=1}^{n} 1 = 1 + 1 + \ldots + 1 = n,$$

where there are n 1's in the sum (so the summand doesn't actually have to change as k changes).

More generally $\sum_{k=\ell}^{u} a_k$ means $a_\ell + a_{\ell+1} + \cdots + a_{u-1} + a_u$, so

$$\sum_{j=-3}^{2} 2^j = \frac{1}{8} + \frac{1}{4} + \frac{1}{2} + 1 + 2 + 4.$$

If there happen to be no numbers in the range between ℓ and u inclusive, then the sum is called *empty*, and by convention is declared to be 0, so, for example,

$$\sum_{k=3}^{1} a_k = 0$$

(starting from 3 and working upwards along the number line, no numbers between 3 and 1 are encountered).

If "$\sum$" is replaced with "$\prod$", then we replace addition with multiplication, so

$$\prod_{i=1}^{5} i = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120.$$

The empty product is by convention declared to be equal to 1.

In summation notation, the statement of Claim 4.2 is

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

There are similar formulas for the sums of the first n squares, cubes, et cetera. The following are good exercises in proof by induction:

- $\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$,
- $\sum_{k=1}^{n} k^3 = \frac{n^2(n+1)^2}{4}$.

Recursively defined sequences

Hand-in-glove with proof by induction goes definition by recursion. A sequence of numbers $(a_1, a_2, a_3, \ldots)$ is defined *recursively* if

- the values of the a_i for some small indices are specified, and

- for all other indices i, a procedure is given for calculating a_i, in terms of a_{i-1}, a_{i-2}, et cetera.

Properties of sequences defined recursively are often proved by induction, as we will now see.

The most famous example of a recursively defined sequence is the *Fibonacci numbers*. Define a sequence $(f_0, f_1, f_2, \ldots)$ by[58]

- $f_0 = 0$, $f_1 = 1$ and
- for $n \geq 2$, $f_n = f_{n-1} + f_{n-2}$.

The sequence begins $(0, 1, 1, 2, 3, 5, 8, 13, 21, 34, \ldots)$. Fibonacci numbers count many different things, for example:

- f_{n+1} is the number of ways of tiling a 1 by n strip with 1 by 1 and 1 by 2 tiles;
- f_{n+1} is the number of hopscotch boards that can be made using n squares[59];
- f_{n+1} in the number of ways of covering 2 by n strip with 2 by 1 dominoes;
- f_{n+2} is the number of words of length n that can be formed from the letters a and b, if two a's are not ever allowed to appear consecutively; and
- the Fibonacci numbers count the number of pairs of rabbit on an island after a certain amount of time has passed, under some very contrived conditions.[60]

The Fibonacci numbers exhibit many nice patterns. For example, define s_n to be the sum of all the Fibonacci numbers up to and including f_n, that is, $s_n = f_0 + f_1 + \cdots + f_n$, or $s_n = \sum_{k=0}^{n} f_k$. Here is a table of some values of s_n, compared to f_n:

n	0	1	2	3	3	5	6	7	8
f_n	0	1	1	2	3	5	8	13	21
s_n	0	1	2	4	7	12	20	33	54.

There seems to be a pattern: $s_n = f_{n+2} - 1$. We can prove this by induction on n. The base case $n = 0$ is clear, since $s_0 = 0 = 1 - 1 = f_2 - 1$. For the induction step, suppose that for some $n \geq 0$ we have $s_n = f_{n+2} - 1$. Then

$$\begin{aligned} s_{n+1} &= s_n + f_{n+1} \\ &= (f_{n+2} - 1) + f_{n+1} \quad \text{(inductive hypothesis)} \\ &= (f_{n+2} + f_{n+1}) - 1 \\ &= f_{n+3} - 1 \quad \text{(recursive definition of Fibonacci numbers)} \\ &= f_{(n+1)+2} - 1, \end{aligned}$$

[58] Notice that here I'm starting indexing at 0, rather than 1.

[59] See https://en.wikipedia.org/wiki/Hopscotch.

[60] The Fibonacci numbers are named for Leonardo of Pisa, nicknamed "Fibonacci", who discussed them in his book *Liber Abaci* in 1202, in the context of rabbits on an island. They had already been around for a while, though, having been studied by the Indian mathematician Pingala as early as 200BC.

so, by induction, the claimed identity is proven.

Other sum identities satisfied by the Fibonacci numbers include the following, that you can try to prove by induction:

- (Sum of odd-indexed Fibonacci numbers) $\sum_{k=0}^{n} f_{2k+1} = f_{2n+2}$;
- (Sum of even-indexed Fibonacci numbers) $\sum_{k=0}^{n} f_{2k} = f_{2n+1} - 1$; and
- (Sum of squares of Fibonacci numbers) $\sum_{k=0}^{n} f_k^2 = f_n f_{n+1}$ (hard!).

Many important mathematical operations are defined recursively. For example, although it is tempting simply to define a^n, for real a and natural number n, by

$$\text{“}a^n = a \cdot a \cdot \cdots \cdot a\text{”}$$

where there are n a's in the product on the right, this somewhat informal definition is an awkward one to use when trying to establish basic properties of powers. If instead (as we do) we define a^n recursively, via:

$$a^n = \begin{cases} a & \text{if } n = 1 \\ a \cdot a^{n-1} & \text{if } n \geq 2 \end{cases}$$

then proving all the expected properties becomes a fairly straightforward exercise in induction. For example, on the homework you will be asked to prove that for all natural numbers n, m, it holds that $a^{n+m} = (a^n)(a^m)$, and this should be done via induction.

We can also define $a^0 = 1$ for all non-zero a. We do not define 0^0[61].

Now that we've defined powers, it's possible to present another application of induction, the *Bernoulli inequality*. In the future (not this year) the content of the inequality will be quite useful; right now, it's just an example of an *inequality* proved inductively.

Claim 4.3. *For all $x \geq -1$ and all $n \in \mathbb{N}$, $(1+x)^n \geq 1 + nx$.*

Proof: We proceed by induction on n. We could if we wished start the induction at $n = 0$, where the assertion is that for all $x \geq -1$, $(1+x)^0 \geq 1 + 0 \cdot x$. This *seems* true enough: it's “$1 \geq 1$”. But, it's not always that, because at $x = -1$ we are required to interpret 0^0, which we have chosen not to do. So we'll start our induction (as the claim suggests) at $n = 1$, where the assertion is that for all $x \geq -1$, $(1+x)^1 \geq 1 + 1 \cdot x$, or $1 + x \geq 1 + x$, which is true not only for $x \geq -1$ but for all x.

[61] A very strong case can be made for $0^0 = 1$, because for natural numbers a and b, a^b counts the number of functions from a set of size b to a set of size a. When $b = 0$ and $a \neq 0$, there should be one function from the empty set to a set of size a, namely the “empty function” that does nothing, and this agrees with $a^0 = 1$ for $a \neq 0$; and when both a and b are 0, there is again one function from the empty set to itself, again the empty function, justifying setting 0^0 to be 1. If none of this makes sense, that's fine, as we haven't yet said what a function is. It might make more sense after we do.

We now move on to the induction step. Assuming $(1+x)^n \geq 1+nx$ holds for all $x \geq -1$, we consider how $(1+x)^{n+1}$ compares with $1+(n+1)x$ for $x \geq -1$. We have

$$\begin{aligned}
(1+x)^{n+1} &= (1+x)(1+x)^n \text{ by definition of powers} \\
&\geq (1+x)(1+nx) \text{ (by induction hypothesis)} \\
&= 1+(n+1)x+nx^2 \text{ (by some algebra)} \\
&\geq 1+(n+1)x \text{ (since } nx^2 \geq 0).
\end{aligned}$$

This proves the validity of the induction step, and so the claim is proved by induction. □

But wait ... where did we use $x \geq -1$ in the proof? The result is *false* without this assumption — for example, if $x = -4$ and $n = 3$, then $(1+x)^n = -27$ while $1+nx = -11$, so $(1+x)^n < 1+nx$. I'll leave it as an exercise to identify where the hypothesis got used.

4.2 A note on variants of induction

The principle of induction says that for $p(n)$ a predicate, with the universe of discourse for n being natural numbers, if $p(1)$ is true, and if, for arbitrary n, $p(n)$ implies $p(n+1)$, then $p(n)$ is true for all n. There are numerous natural variants, too numerous to possibly mention, and too similar to the basic principle for use to need to mention. I'll say a few here, so you can get the idea; looking at these examples you should realize that induction can be quite flexible. In all cases, $p(n)$ is a predicate with universe of discourse for n being natural numbers.

- If, for some natural number k, $p(k)$ is true, and if, for arbitrary $n \geq k$, $p(n)$ implies $p(n+1)$, then $p(n)$ is true for all $n \geq k$.
- If $p(0)$ is true, and if, for arbitrary $n \geq 0$, $p(n)$ implies $p(n+1)$, then $p(n)$ is true for all $n \geq 0$.
- If $p(-5)$ is true, and if, for arbitrary $n \geq -5$, $p(n)$ implies $p(n+1)$, then $p(n)$ is true for all $n \geq -5$.
- If $p(2)$ is true, and if, for arbitrary $n \geq 2$, $p(n)$ implies $p(n+2)$, then $p(n)$ is true for all positive even numbers.
- $\cdots$.

4.3 Binomial coefficients and the binomial theorem

We all know that $(x+y)^2$ expands out to $x^2+2xy+y^2$. What about $(x+y)^3$, $(x+y)^4$, et cetera? Here is a table showing the various expansions of $(x+y)^n$ for some small values of n.

For completeness I've included $n = 0$ ($(x+y)^0 = 1$ as long as $x + y \neq 0$) and $(x+y)^1 = x + y$.

$$\begin{array}{rcc}
(x+y)^0 & = & 1 \\
(x+y)^1 & = & x+y \\
(x+y)^2 & = & x^2 + 2xy + y^2 \\
(x+y)^3 & = & x^3 + 3x^2y + 3xy^2 + y^3 \\
(x+y)^4 & = & x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4 \\
(x+y)^5 & = & x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + y^5 \\
(x+y)^6 & = & x^6 + 6x^5y + 15x^4y^2 + 20x^3y^3 + 15x^2y^4 + 6xy^5 + y^6 \\
\cdots & & \cdots .
\end{array}$$

There is a pattern here. Of course, $(x+y)^n$, when expanded out, has some x^n terms, some $x^{n-1}y$ terms, some $x^{n-2}y^2$ terms, and so on, down to some xy^{n-1} and some y^n terms, but by "a pattern" I mean a pattern in the coefficients of each of these terms. That pattern is best spotted when the table is re-written as a triangle of coefficients, without the terms $x^n y^{n-k}$ that are just cluttering up the picture:

$$\begin{array}{ccccccccccccc}
 & & & & & & 1 & & & & & & \\
 & & & & & 1 & & 1 & & & & & \\
 & & & & 1 & & 2 & & 1 & & & & \\
 & & & 1 & & 3 & & 3 & & 1 & & & \\
 & & 1 & & 4 & & 6 & & 4 & & 1 & & \\
 & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1.
\end{array}$$

Notice that there are 1's down the outside, and all other numbers are obtained by summing their neighbor above and to right and above and to the left. If this pattern were to continue, the next row would be

$$1 \quad 7 \quad 21 \quad 35 \quad 35 \quad 21 \quad 7 \quad 1,$$

and sure enough, after a lot of computation, we discover that

$$(x+y)^7 = x^7 + 7x^6y + 21x^5y^2 + 35x^4y^3 + 35x^3y^4 + 21x^2y^5 + 7xy^6 + y^7.$$

We'll prove, by induction, that this pattern holds for all n. But first we have to be a little more precise about what the pattern is. Define the number "$n!$" (read this as "n factorial") recursively, via:

$$n! = \begin{cases} 1 & \text{if } n = 1 \\ n \cdot (n-1)! & \text{if } n \geq 2 \\ 1 & \text{if } n = 0. \end{cases}$$

So, informally, $n!$ is the product of all the whole numbers from n down to 1:

- $3! = 3 \cdot 2! = 3 \cdot 2 \cdot 1! = 3 \cdot 2 \cdot 1 = 6$,

- $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$, et cetera.

We can interpret $n!$ as the number of ways of arranging n objects in order — there are n options for the first object; for each of those n options there are a further $n-1$ options for the second, so $n \cdot (n-1)$ options for the first two objects; for each of those $n(n-1)$ options there are a further $n-2$ options for the third, so $n(n-1)(n-2)$ options for the first three objects, and so on down to $n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1 = n!$ options for ordering all n objects.

That 0! is defined to be "1" may seem a little strange, as it is hard to imagine counting the number of ways of arranging 0 objects in order; but as we will see there are sound reasons for this, and also it adheres to our convention that an empty product is 1 (the product of all the numbers, starting at 0, and going *down* to 1, is empty).

Next define the expression "$\binom{n}{k}$", (read this as "n choose k") via

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Notice that because of the way we have defined the factorial function (never 0, and making sense exactly for the natural numbers including 0) this expression is well-defined for all $n \geq 0$, and for all k satisfying $0 \leq k \leq n$. That means we can form a triangle from the numbers $\binom{n}{k}$:

$$\begin{array}{ccccccccccccc}
 & & & & & & \binom{0}{0} & & & & & & \\
 & & & & & \binom{1}{0} & & \binom{1}{1} & & & & & \\
 & & & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & & & \\
 & & & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & & & \\
 & & \binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} & & \\
 & \binom{5}{0} & & \binom{5}{1} & & \binom{5}{2} & & \binom{5}{3} & & \binom{5}{4} & & \binom{5}{5} & \\
\binom{6}{0} & & \binom{6}{1} & & \binom{6}{2} & & \binom{6}{3} & & \binom{6}{4} & & \binom{6}{5} & & \binom{6}{6}.
\end{array}$$

Calculating the numerical values, this triangle becomes:

$$\begin{array}{ccccccccccccc}
 & & & & & & 1 & & & & & & \\
 & & & & & 1 & & 1 & & & & & \\
 & & & & 1 & & 2 & & 1 & & & & \\
 & & & 1 & & 3 & & 3 & & 1 & & & \\
 & & 1 & & 4 & & 6 & & 4 & & 1 & & \\
 & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1,
\end{array}$$

which looks just like the triangle of coefficients we got when we expanded out $(x+y)^n$.

In that case (coefficients in expansion of $(x+y)^n$) we *suspected* that each number in the triangle, other than the 1's on the outside, was obtained by summing its two neighbors above it. In this case (triangle of $\binom{n}{k}$), we can easily *prove* this relation.

Claim 4.4. *For $n \geq 0$, and for $0 \leq k \leq n$:*

- *if* $k = 0$ *or if* $k = n$ *then* $\binom{n}{k} = 1$*;*
- *otherwise,*

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Proof: If $k = 0$ then since $0! = 1$, we have

$$\binom{n}{k} = \frac{n!}{0!(n-0)!} = \frac{n!}{1 \cdot n!} = 1,$$

and if $n = k$ for a similar reason we have $\binom{n}{k} = 1$.[62]

Otherwise, we must have $n \geq 2$ and $1 < k < n$. We have

$$\begin{aligned}
\binom{n-1}{k-1} + \binom{n-1}{k} &= \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} + \frac{(n-1)!}{k!((n-1)-k)!} \\
&= \frac{(n-1)!}{(k-1)!(n-k)!} + \frac{(n-1)!}{k!(n-k-1)!} \\
&= \frac{(n-1)!}{(k-1)!(n-k-1)!}\left(\frac{1}{n-k} + \frac{1}{k}\right) \\
&= \frac{(n-1)!}{(k-1)!(n-k-1)!}\left(\frac{k+(n-k)}{(n-k)k}\right) \\
&= \frac{n!}{k!(n-k)!} \\
&= \binom{n}{k}.
\end{aligned}$$

(Notice that all steps above involve expressions that make sense, because $n \geq 2$ and $1 < k < n$). □

Just as there was a counting interpretation of $n!$, there's a counting interpretation of $\binom{n}{k}$. How many subsets of size k does a set of size n have? Well, we can select such a subset by choosing a first element, then a second, et cetera, leading to a count of $n \cdot (n-1) \cdots (n-k+1) = n!/(n-k)!$; but each particular subset has been counted many times. In fact, a particular subset has been counted $k!$ times, once for each of the $k!$ ways in ways in which its k elements can be arranged in order. So our count of $n!/(n-k)!$ was off by a multiplicative factor of $k!$, and the correct count is $(n!/(n-k)!)/k!$, which is exactly $\binom{n}{k}$. So:

$\binom{n}{k}$ is the number of subsets of size k of a set of size n.

This allows a alternate proof of Claim 4.4. When $k = n$, $\binom{n}{k}$ is the number of subsets of size n of a set of size n, and this is clearly 1 (the set itself). When $k = 0$, $\binom{n}{k}$ is the number of subsets of size 0 of a set of size n, and this is also 1 (the empty set is a subset of any set, and it is the only set with 0 elements). For $n \geq 2$, and $1 < k < n$, subsets of size k of a set X of size n fall into two classes:

[62] This is a strong justification for declaring $0! = 1$.

- those that include a particular fixed element x — there are $\binom{n-1}{k-1}$ of these, one for each subset of $X - \{n\}$ of size $k-1$, and
- those that *don't* include x — there are $\binom{n-1}{k}$ of these, one for each subset of $X - \{n\}$ of size k.

So X has $\binom{n-1}{k-1} + \binom{n-1}{k}$ subsets of size k; but it also (directly) has $\binom{n}{k}$ subsets of size k; so

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

This identity is called *Pascal's identity*[63]

We're now ready to formalize a theorem that captures the pattern we were noticing with $(x+y)^n$. It's called the *binomial theorem* (because the expansion of $(x+y)^n$ is a *binomial expansion* — an expansion of an expression involving two (*bi*) named (*nomial*) things, x and y), and the numbers $\binom{n}{k}$ that come up in it are often called *binomial coefficients*.

Theorem 4.5. *Except in the case when $n = 0$ and at least one of $x, y, x+y = 0$, for all $n \in \mathbb{N}^0$ and for all real x, y,*

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

or, more succinctly,

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k}y^k.$$

Proof: When $n = 0$, as long as all of $x, y, x+y$ are non-zero both sides of the identity are 1, so they are equal.

For $n \geq 1$ we proceed by induction on n (with predicate:

$$p(n) : \text{“for all real } x, y,\ (x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k}y^k\text{”}).$$

The base case $p(1)$ asserts $(x+y)^1 = \binom{1}{0}x + \binom{1}{1}y$, or $x + y = x + y$, which is true for all real x, y.

For the induction step, we assume that for some $n \geq 1$ we have

$$(x+y)^n = x^n + \binom{n}{1}x^{n-1}y + \binom{n}{2}x^{n-2}y^2 + \cdots + \binom{n}{n-2}x^2y^{n-2} + \binom{n}{n-1}xy^{n-1} + y^n$$

for all real x, y. Multiplying both sides by $x+y$, this yields

$$(x+y)^{n+1} = (x+y)\left(x^n + \binom{n}{1}x^{n-1}y + \binom{n}{2}x^{n-2}y^2 + \cdots + \binom{n}{n-2}x^2y^{n-2} + \binom{n}{n-1}xy^{n-1} + y^n\right).$$

[63] It's named for the French polymath Blaise Pascal. The triangle of values of $\binom{n}{k}$ is called *Pascal's triangle*, and has many lovely properties. It is easily googled.

Now the right-hand side above is

$$\begin{gathered} x^{n+1}+ \\ \tbinom{n}{1}x^n y + \tbinom{n}{2}x^{n-1}y^2 + \cdots + \tbinom{n}{n-2}x^3 y^{n-2} + \tbinom{n}{n-1}x^2 y^{n-1} + xy^n + \\ x^n y + \tbinom{n}{1}x^{n-1}y^2 + \cdots + \tbinom{n}{n-3}x^3 y^{n-2} + \tbinom{n}{n-2}x^2 y^{n-1} + \tbinom{n}{n-1}xy^n + \\ y^{n+1}. \end{gathered}$$

or

$$\begin{gathered} x^{n+1}+ \\ \tbinom{n}{1}x^n y + \tbinom{n}{2}x^{n-1}y^2 + \cdots + \tbinom{n}{n-2}x^3 y^{n-2} + \tbinom{n}{n-1}x^2 y^{n-1} + \tbinom{n}{n}xy^n + \\ \tbinom{n}{0}x^n y + \tbinom{n}{1}x^{n-1}y^2 + \cdots + \tbinom{n}{n-3}x^3 y^{n-2} + \tbinom{n}{n-2}x^2 y^{n-1} + \tbinom{n}{n-1}xy^n + \\ y^{n+1}. \end{gathered}$$

Applying Claim 4.4 to each pair of terms in matching columns in the second and third rows, this becomes

$$\begin{gathered} x^{n+1}+ \\ \tbinom{n+1}{1}x^n y + \tbinom{n+1}{2}x^{n-1}y^2 + \cdots + \tbinom{n}{n-2}x^3 y^{n-2} + \tbinom{n}{n-1}x^2 y^{n-1} + \tbinom{n+1}{n}xy^n + \\ y^{n+1} \end{gathered}$$

(for example,

$$\binom{n}{2}x^{n-1}y^2 + \binom{n}{1}x^{n-1}y^2 = \left(\binom{n}{2} + \binom{n}{1}\right)x^{n-1}y^2 = \binom{n+1}{2}x^{n-1}y^2)$$

Using $\binom{n+1}{0} = \binom{n+1}{n+1} = 1$, this last expression is exactly

$$\sum_{k=0}^{n+1} \binom{n+1}{k} x^{(n+1)-k}y^k.$$

So we have shown that $(1+x)^{n+1} = \sum_{k=0}^{n+1}\binom{n+1}{k}x^{(n+1)-k}y^k$ for all real x, y, which is $p(n+1)$. The induction step is complete, as is the proof of the theorem. □

At the end of Spivak Chapter 2, there are plenty of exercises that explore the many properties of the numbers $\binom{n}{k}$.

4.4 Complete, or strong, induction (informally)

Sometimes induction is not enough to verify a proposition that at first glance seems tailor-made for induction. For example, consider the recursively defined sequence

$$a_n = \begin{cases} 2 & \text{if } n = 0 \\ 3 & \text{if } n = 1 \\ 3a_{n-1} - 2a_{n-2} & \text{if } n \geq 2. \end{cases}$$

This table shows the first few values of a_n:

n	0	1	2	3	4	5	6	7
a_n	2	3	5	9	17	33	65	129.

There seems to be a pattern: it seems that $a_n = 2^n + 1$ for each n. If we try to prove this by induction, though, we run into a number of problems. The base case $n = 0$ is evident. The first problem arises when we think about the induction step: we assume, for some arbitrary $n \geq 0$, that $a_n = 2^n + 1$, and try to deduce that $a_{n+1} = 2^{n+1} + 1$.

Our inclination is to use the recursive definition $a_{n+1} = 3a_n - 2a_{n-1}$. But already in the very first instance of the induction step, we are stuck, because at $n = 0$ the recursive definition we would like to use is $a_1 = 3a_0 - 2a_{-1}$. This makes no sense (there is no a_{-1}). And indeed, it shouldn't make sense, because the clause $a_n = 3a_{n-1} - 2a_{n-2}$ of the definition of a_n kicks in only when $n \geq 2$. To say anything about a_1, we have to appeal to a different clause in the definition, namely $a_1 = 3$. Since $3 = 2^1 + 1$, this is still consistent with the general pattern we are trying to prove.

One way to think of this is that we are verifying *two* base cases ($n = 0$ *and* $a = 1$) before going on to the induction step; another way to think of it is that we are treating the induction step "$p(0) \Rightarrow p(1)$" as a special case, and showing that it is a true implication by showing that both $p(0)$ and $p(1)$ are simply true, always, so the implication is true; the remainder of the induction step, "$p(n) \Rightarrow p(n + 1)$ for every $n \geq 1$" will be dealt with in a different, more general, way. However we choose to think of it, this issue arises frequently in proofs by induction, especially when dealing with recursively defined sequences.

Having dealt with the first instance of the induction step, lets move on to the general inductive step, $p(n) \Rightarrow p(n+1)$ for $n \geq 1$. Here we can legitimately write $a_{n+1} = 3a_n - 2a_{n-1}$, because for $n \geq 1$, this is the correct clause for defining a_{n+1}. We would like to say that

$$3a_n - 2a_{n-1} = 2^{n+1} + 1,$$

using that $a_n = 2^n + 1$. But we can't: the best we can say is

$$3a_n - 2a_{n-1} = 3(2^n + 1) - 2a_{n-1},$$

because in trying to verify $p(n) \Rightarrow p(n + 1)$ we can assume nothing about $p(n - 1)$.

There's a fix: presumable, in getting this far in the induction, we have already established not just $p(n)$, but also $p(n - 1), p(n - 2), p(n - 3)$, et cetera. If we have, then we can, as well as using $a_n = 2^n + 1$, use $a_{n-1} = 2^{n-1} + 1$. Then we get

$$a_{n+1} = 3a_n - 2a_{n-1} = 3(2^n + 1) - 2(2^{n-1} + 1) = 3 \cdot 2^n + 3 - 2^n - 2 = 2 \cdot 2^n + 1 = 2^{n+1} - 1,$$

as we need to show to establish $p(n + 1)$.

We can formalize this idea in the *principle of complete induction*, also called the *principle of strong induction*:

> **The principle of complete mathematical induction**: Let $p(n)$ be a predicate, with the universe of discourse for n being natural numbers. If $p(1)$ is true, and if, for arbitrary n, the conjunction of $p(1), p(2), \ldots, p(n)$ implies $p(n + 1)$, then $p(n)$ is true for all n.

Going back through the discussion that we gave to justify the principle of induction, it should be clear that complete or strong induction is an equally valid proof technique. We can in fact argue that strong induction is *exactly* as strong as regular induction:

- Suppose that we have access to the principle of strong induction. Suppose that $p(n)$ is a predicate (with n a natural number) and that we know
  - $p(1)$ and
  - for arbitrary $n \geq 1$, $p(n)$ implies $p(n+1)$.

  Then we *also* know $p(1) \wedge p(2) \wedge \cdots \wedge p(n)$ implies $p(n+1)$ (if we can infer $p(n+1)$ from $p(n)$, we can certainly infer it from $p(1), p(2), \ldots, p(n)$!). So by strong induction, we can conclude that $p(n)$ is true for all n. In other words, if we have access to the principle of strong induction, we also have access to the principle of induction.

- Suppose that we have access to the principle of induction. Suppose that $p(n)$ is a predicate (with n a natural number) and that we know
  - $p(1)$ and
  - for arbitrary $n \geq 1$, $p(1) \wedge p(2) \wedge \cdots \wedge p(n)$ implies $p(n+1)$.

  We would like to conclude that $p(n)$ is true for all n; but we can't simple say that $p(n)$ implies $p(n+1)$, and use induction; we don't know whether $p(n)$ (on its own) implies $p(n+1)$. Here's a fix: consider the predicate $Q(n)$ define by

  $$Q(n) : \text{“}p(1) \wedge p(2) \wedge \cdots \wedge p(n).\text{”}$$

  We know $Q(1)$ (it's just $p(1)$. Suppose, for some arbitrary n, we know $Q(n)$. Then we know $p(1) \wedge p(2) \wedge \cdots \wedge p(n)$, and we can deduce $p(n+1)$. But, again since we know $p(1) \wedge p(2) \wedge \cdots \wedge p(n)$, we can now deduce $p(1) \wedge p(2) \wedge \cdots \wedge p(n) \wedge p(n+1)$, that is, we can deduce $Q(n+1)$. So we can apply induction to Q to conclude $Q(n)$ for all n. But a consequence of this is that $p(n)$ holds for all n (remember, $Q(n)$ is $p(1) \wedge p(2) \wedge \cdots \wedge p(n)$). In other words, if we have access to the principle of induction, we also have access to the principle of strong induction.

Here's an important application of complete induction, from elementary number theory. A natural number $n \geq 2$ is said to be *composite* if there are natural numbers a and b, both at least 2, such that $ab = n$. It is said to be *prime* if it is not composite. We can use strong (complete) induction to show that every natural number $n \geq 2$ can be written as a product of prime numbers.[64]

[64]The *fundamental theorem of arithmetic* states that the prime factorization of any number is *unique* up to the order in which the primes in the factorization are listed (note that this would not be true if 1 was considered a prime number, for then 3.2.1 and 3.2.1.1.1 would be different prime factorizations of 6). The fundamental theorem of arithmetic is also proven by induction, but takes a lot more work than the result we are about to prove, establishing the existence of *a* prime factorization.

Indeed, let $p(n)$ be the predicate "n can be written as the product of prime numbers". We prove that $p(n)$ is true for all $n \geq 2$ by complete induction.

Base case $n = 2$: This is trivial since 2 is a prime number.

Inductive step: Suppose that for some $n \geq 3$, we know that $p(m)$ is True for all m in the range $2 \leq m \leq n-1$[65]. We consider two cases.

- Case 1: n is prime. In this case $p(n)$ is trivial.
- Case 2: n is composite. In this case $n = ab$ for some natural numbers a and b with $2 \leq a \leq n-1$ and $2 \leq b \leq n-1$. Since $p(a)$ and $p(b)$ are both true (by the complete induction hypothesis) we have

  $$a = p_1 p_2 \cdots p_k$$

  and

  $$b = q_1 q_2 \cdots q_\ell$$

  where $p_1, p_2, \ldots, p_k, q_1, q_2, \ldots, q_\ell$ are all prime numbers. But that implies that n can be written as a product of prime numbers, via

  $$n = ab = p_1 p_2 \cdots p_k q_1 q_2 \cdots q_\ell.$$

  This shows that $p(n)$ follows from $p(2), p(3), \ldots, p(n-1)$.

By complete induction, we conclude that $p(n)$ is true for all $n \geq 2$.

Note that this proof would have gone exactly *nowhere* if all we were able to assume, when trying to factorize n, was the existence of a factorization of $n-1$.

We now give a more substantial example of complete induction. The associativity axiom for multiplication says that for all reals a, b, c, we have $a(bc) = (ab)c$ (note that I'm using juxtaposition for multiplication here, as is conventional, rather than the "$\cdot$" that I've been using up to now). Presumably, there is an "associativity axiom" for the product of n things, too, for all $n \geq 3$ (we've already seen the version for $n = 4$). Let GAA(n) be the predicate "for any set of n real numbers $a_1, \ldots, a_n$ the order in which the product $a_1 \cdots a_n$ is parenthesized does not affect the final answer", and let GAA be the generalized associativity axiom, that is, the statement that GAA(n) holds for all $n \geq 1$[66].

Claim 4.6. GAA *is true.*

Proof: Among all the ways of parenthesizing the product $a_1 \cdots a_n$ we identify one special one, the *right-multiply*:

$$R(a_1, \ldots, a_n) = (\cdots(((a_1 a_2)a_3)a_4)\cdots)a_n.$$

[65] Note that when proving things by induction, you can either deduce $p(n+1)$ from $p(n)$, or deduce $p(n)$ from $p(n-1)$; similarly, when proving things by strong induction you can either deduce $p(n+1)$ from $p(1) \wedge \cdots \wedge p(n)$, or deduce $p(n)$ from $p(1) \wedge \cdots \wedge p(n-1)$; it's a matter of taste or convenience

[66] Really the result is only interesting for $n \geq 3$, but it makes sense, and is true, for $n = 1, 2$ as well, so we'll throw those into the mix, too.

We will prove, by strong induction on n, that for all $n \geq 1$, the predicate "for any set of n real numbers $a_1, \ldots, a_n$, all the ways of parenthesizing the product $a_1 \cdots a_n$ lead to the answer $R(a_1, \ldots, a_n)$." This will show that GAA is true.

The base case $n = 1$ is trivial — with only one number, there is one one possible product. The same goes for the base case $n = 2$. The base case $n = 3$ is axiom P5.

For the inductive step, let $n \geq 4$ be arbitrary, and suppose that the predicate we are trying to prove (GAA(k)) is true for all values k of the variable between 1 and $n - 1$. Let P be an arbitrary parenthesizing of the product $a_1 \cdots a_n$. P has a final, outer, product, the last pair of numbers multiplied together before P is fully evaluated. We consider cases.

Case 1 The final product is of the form Aa_n. By induction (variable value $n - 1$) we have $A = R(a_1, \ldots, a_{n-1})$, so

$$P = Aa_n = R(a_1, \ldots, a_{n-1})a_n = R(a_1, \ldots, a_n).$$

Case 2 The final product is of the form AB where A is a parenthesizing of $a_1, \ldots, a_k$ and B is a parenthesizing of $a_{k+1}, \ldots, a_n$, where $1 \leq k \leq n - 2$. If $k = n - 2$ then we have

$$P = A(a_{n-1}a_n) = (Aa_{n-1})a_n$$

(by P5), and we are back in case 1, so $P = R(a_1, \ldots, a_n)$. If $k \leq n-3$ then by induction (variable value $n - k$) we have

$$B = R(a_{k+1} \cdots a_n) = R(a_{k+1} \cdots a_{n-1})a_n$$

and so, once again by P5,

$$P = AB = A(R(a_{k+1} \cdots a_{n-1})a_n) = (AR(a_{k+1} \cdots a_{n-1}))a_n,$$

and we are back in case 1, so $P = R(a_1, \ldots, a_n)$.[67]

[67]Here's an alternate way of presenting the two cases:
Case 1: $P = (\text{SOMETHING})a_n$. By induction (GAA($n-1$)), this is the same as

$$P = R(a_1, \ldots, a_{n-1})a_n = R(a_1, \ldots, a_n).$$

Case 2:

$$\begin{aligned}
P &= \underbrace{(\text{SOMETHING})}_{\text{involving } x_1, \ldots, x_k,\ 1 \leq k \leq n-2} \cdot \underbrace{(\text{SOMETHING ELSE})}_{\text{involving } x_{k+1}, \ldots, x_n} \\
&= \underbrace{R(a_1, \ldots, a_k)}_{\text{GAA}(k)} \underbrace{R(a_{k+1}, \ldots, a_n)}_{\text{GAA}(n-k)} \\
&= R(a_1, \ldots, a_k) \cdot (R(a_{k+1}, \ldots, a_{n-1}) \cdot a_n) \\
&= (R(a_1, \ldots, a_k)R(a_{k+1}, \ldots, a_{n-1}) \cdot) \cdot a_n \quad \text{(GAA(3))} \\
&= R(a_1, \ldots, a_{n-1})a_n \quad \text{(GAA}(n-1)\text{)} \\
&= R(a_1, \ldots, a_n).
\end{aligned}$$

In either case, $P = R(a_1, \ldots, a_n)$, and so the claim is proven by (strong) induction. $\square$

Notice that we needed the induction hypothesis for *all* values of the variable below n, so we really needed strong induction.

Of course, there is also an analogous generalized associativity for addition. Strong induction is in general a good way to extend arithmetic identities from a few terms to arbitrarily many terms. You should do some of the following as exercises:

Generalized commutativity For $n \geq 2$, for any set of n reals, the result of adding the n reals does not depend on the order in which the numbers are written down; and the same for multiplication.

Generalized distributivity For $n \geq 2$, and for any set of real numbers $a, b_1, b_2, \ldots, b_n$,

$$a(b_1 + \cdots b_n) = ab_1 + \cdots ab_n.$$

Generalized triangle inequality For $n \geq 2$, and for any set of real numbers $b_1, b_2, \ldots, b_n$,

$$|b_1 + \ldots b_n| \leq |b_1| + \cdots + |b_n|.$$

Generalized Euclid's rule For $n \geq 2$, and for any set of real numbers $b_1, b_2, \ldots, b_n$, if $b_1 b_2 \cdots b_n = 0$ then at least one of $b_1, b_2, \ldots, b_n$ must be 0.

4.5 The well-ordering principle (informal)

A set S has a *least element* if there is an element s *in the set* S with $s \leq s'$ for every $s' \in S$. Not every set has a least element: there is no least positive number (for every positive number p, $p/2$ is a smaller positive number), and there is no least negative number (for every negative number q, $q - 1$ is a smaller negative number).

The set of natural numbers, on the other hand (at least as we have informally defined it), has a least element element, namely 1. Moreover, it seems intuitively clear that every subset of $\mathbb{N}$ has a least element; or rather, every *non-empty* subset of $\mathbb{N}$ has a least element (the empty set has no least element). We formulate this as the *well-ordering principle* of the natural numbers:

Claim 4.7. *(The well-ordering principle of the natural numbers) If E is a non-empty subset of the natural numbers, then E has a least element.*

Proof: Suppose E is a subset of the natural numbers with no least element. We will show that E is empty; this is the contrapositive of, and equivalent to, the claimed statement.

Let $p(n)$ be the predicate "$n \notin E$". We will show, by strong induction, that $p(n)$ is true for all n, which will show that E is empty.

The base case $p(1)$ asserts $1 \notin E$, which is true; if $1 \in E$ then 1 would be the least element in E.

For the induction step, assume that $p(1), \ldots p(n-1)$ are all true, for some arbitrary natural number $n \geq 2$. Then none of $1, 2, \ldots, n-1$ are in E, so neither is n, since if $n \in E$ then would be the least element in E. So $p(n)$ is true, assuming $p(1), \ldots p(n-1)$ are all true, and by strong induction $p(n)$ is true for all n. $\square$

As an application of well-ordering, we give an alternate proof of the irrationality of $\sqrt{2}$.

Suppose (for a contradiction) $\sqrt{2}$ is rational. Let E be set of all natural numbers x such that $x^2 = 2y^2$ for some natural number y. Under the assumption that $\sqrt{2}$ is rational, E is non-empty, and so by well-ordering it has a least element, a say, with $a^2 = 2b^2$ for some natural number b.

Now it is an easy check that $b < a < 2b$ (indeed, since $a^2 = 2b^2$ it follows that $b^2 < a^2 < 4b^2$, from which $b < a < 2b$, via a homework problem).

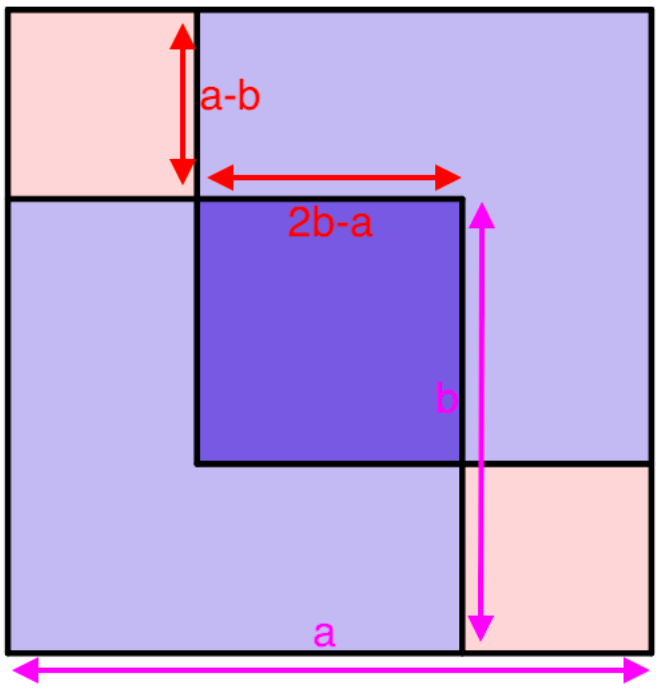
Set $a' = 2b - a$ and $b' = a - b$. By the relations $b < a < 2b$, both natural numbers, and since $b < a$ we have $a' < a$. But now note that

$$2(b')^2 = 2(a-b)^2 = 2a^2 - 4ab + 2b^2 = a^2 - 4ab + 4b^2 = (2b-a)^2 = (a')^2,$$

so $a' \in E$, contradicting that a is smallest element of E.

We conclude that E' is empty, so $\sqrt{2}$ is irrational.

This lovely proof was discovered by Stanley Tennenbaum; here is a visual illustration of it, that I have taken (and augmented) from https://divisbyzero.com/2009/10/06/tennenbaums-proof-of-the-irrationality-of-the-square-root-of-2/.



A visual illustration of Tennenbaum's proof of the irrationality of $\sqrt{2}$

4.6 Inductive sets

The purpose of the rest of this section is to make the "..." in

$$\mathbb{N} = \{1, 1+1, 1+1+1, \ldots\},$$

and the principle of mathematical induction, a little more formal.

Say that a set $S \subseteq X$ is *inductive* if it satisfies both of these properties:

1. 1 is in S and
2. $k+1$ is in S whenever k is in S.

So, for example:

- X is inductive.
- The set of positive numbers in X is inductive.
- The set of positive numbers excluding 5 is *not* inductive; it fails the second condition, since 4 is in S but not 5.
- The set of positive numbers, excluding $3/2$ is *not* inductive; it fails the second condition, since $1/2$ is in S but not $3/2$.
- The set of positive numbers that are at least 1, excluding $3/2$ is inductive; the absence of $3/2$ is not an obstacle, since $1/2$ is not in S, so the implication "If $1/2$ is in S then $3/2$ is in S" is true.
- The set of positive numbers that are greater than 1 is *not* inductive; it fails the first condition.
- If S_1 and S_2 are two inductive sets, then the set of elements that are in both S_1 and S_2 is also inductive.

It feels like the set $\{1, 1+1, 1+1+1, \ldots\}$ should be in *every* inductive set, because 1 is in every inductive set, so $1+1$ is also, and so on. To formalize that "and so on", we make the following definition.

Definition 4.8. A number n is a *natural number* if it is in every inductive set. We denote by $\mathbb{N}$ the set of all natural numbers.

So, for example, 1 is a natural number (because it is in every inductive set), and so is $1+1$, and so is $1+1+\ldots+1$ where there are 1876 1's in the sum. More generally if k is in $\mathbb{N}$ then k is in every inductive set, so (by definition of inductive sets) $k+1$ is in every inductive set, so $k+1$ is in $\mathbb{N}$. In other words, $\mathbb{N}$ is an inductive set itself.

By its definition, $\mathbb{N}$ is contained in every inductive set. Moreover, it is the only inductive set that is contained in every inductive set. To see this, consider an inductive set E that is contained in every inductive set. Since $\mathbb{N}$ is inductive, we have that E is contained in $\mathbb{N}$. Suppose that E is not equal to $\mathbb{N}$. Then there is some number k with k in $\mathbb{N}$ but k not in E. But if k is in $\mathbb{N}$ then by the definition of $\mathbb{N}$ we have that k is E, since being in $\mathbb{N}$ means being in *every* inductive set, including E. The contradiction — k is not in E and k is in E — shows that E is not equal to $\mathbb{N}$ is False, and so we conclude $E = \mathbb{N}$. We summarize what we have just proven in a claim.

Claim 4.9. *The natural numbers form an inductive set, and* $\mathbb{N}$ *is the* unique *minimal inductive set — it is contained in every inductive set, and no other inductive set has this property. In particular if* E *is a subset of* $\mathbb{N}$ *and* E *is inductive then* $E = \mathbb{N}$.

4.7 The principle of mathematical induction

Re-phrasing the last sentence of Claim 4.9 we obtain the important *principle of mathematical induction.*

Theorem 4.10. *Suppose that* E *is a set of natural numbers satisfying*

1. 1 *is in* E *and*
2. $k+1$ *is in* E *whenever* k *is.*

Then $E = \mathbb{N}$.

There is no need for a proof of this — it really is just a direct re-phrasing of the last sentence of Claim 4.9. To get a first hint of the power of Theorem 4.10 we use it to derive the following result, which is precisely the form of induction that we are by now familiar with.

Theorem 4.11. *Suppose that* $p(n)$ *is a predicate (a statement that is either True or False, depending on the value of* n*), where the universe of discourse for the variable* n *is all natural numbers. If*

- $p(1)$ *is true and*
- $p(k+1)$ *is true whenever* $p(k)$ *is true*

then $p(n)$ *is true for all* n *in* $\mathbb{N}$.

Proof: Let E be the set of all n for which $p(n)$ is True. We immediate have that 1 is in E and that $k+1$ is in E whenever k is. That $E = \mathbb{N}$, that is that $p(n)$ is True for all n in $\mathbb{N}$, now follows from Theorem 4.10. □

Slightly informally Theorem 4.11 says that if $p(n)$ is some proposition about natural numbers, and if we can show that

Base case $p(1)$ is True and

Induction step for all n the truth of $p(n)$ (the **induction hypothesis**) implies the truth of $p(n+1)$

then we can conclude that $p(n)$ is True for all natural numbers. The power here, that you should see from some examples, is that the principle of mathematical induction allows us to prove *infinitely many things* ($p(1)$, $p(2)$, $p(3)$, et cetera), with only a *finite amount of work* (proving $p(1)$ and proving the single implication $p(n) \Rightarrow p(n+1)$, involving a variable).

More informally still, induction says (repeating a previous observation) that if you can get onto the first rung of a ladder ($p(1)$), and you know how to climb from any one rung to any other ($p(n) \Rightarrow p(n+1)$), then you can climb as high up the ladder as you wish, by first getting on the ladder and then moving up as many rungs as you wish, one rung at a time.

We've already seen many examples of induction at work, in the informal setting, and of course all of those examples go through perfectly in the more formal setting we've given here. We give a few more examples of induction at work now, mostly to establish some very fundamental properties of the natural numbers, that will be useful later. You should get the sense that every property of numbers that you are already familiar with can be established formally in the context of the definition of natural numbers that we have given.

Claim 4.12. *For all natural numbers n, $n \geq 1$.*

Proof: Let $p(n)$ be the predicate "$n \geq 1$", where the universe of discourse for the variable n is all natural numbers. We prove that $p(n)$ is true for all n by induction.

Base case: $p(1)$ is the assertion $1 \geq 1$, which is true.

Induction step: Assume that for some n, $n \geq 1$. Then $n + 1 \geq 1 + 1 \geq 1 + 0 = 1$. So the truth of $p(n)$ implies the truth of $p(n+1)$.

By induction, $p(n)$ is true for all n, that is, for all natural numbers n, $n \geq 1$. □

Corollary 4.13. *There is no natural number x with $0 < x < 1$.*

Proof: Such an x would be a natural number that does not satisfy $x \geq 1$, contradicting Claim 4.12. □

Claim 4.14. *For every natural number n other than 1, $n - 1$ is a natural number.*

Proof: Let $p(n)$ be the predicate "$(n \neq 1) \implies (n - 1 \in \mathbb{N})$". We prove $(\forall n)p(n)$ by induction (with, as usual, the universe of discourse being $\mathbb{N}$).

Base case: $p(1)$ is the assertion $(1 \neq 1) \implies (1 - 1 \in \mathbb{N})$, which is true, since the premise $1 \neq 1$ is false.

Induction step: Assume that for some n, $(n \neq 1) \implies (n - 1 \in \mathbb{N})$. Then $n + 1 \neq 1$, for if $n + 1 = 1$ then $n = 0$, which is not a natural number. Also, $(n + 1) - 1 = n$, which is a

natural number. So both the premise and the hypothesis of $p(n+1)$ is are true, so $p(n+1)$ is true

By induction, $p(n)$ is true for all n, that is, for all natural numbers n, if $n \neq 1$ then $n-1 \in \mathbb{N}$. $\square$

Corollary 4.15. *There is no natural number x with $1 < x < 2$.*

Proof: Such an x would be a natural number other than 1, so $x - 1 \in \mathbb{N}$ by Claim 4.14. But $0 < x - 1 < 1$, contradicting Corollary 4.13. $\square$

All of these results have been leading up to the following, an "obvious" statement that requires a (somewhat sophisticated) proof. It captures in very concrete way that the natural numbers are indeed a set of the form $\{1, 2, 3, \ldots\}$.

Claim 4.16. *For every natural number n, there is no natural number x with $n < x < n + 1$.*

Proof: We proceed by induction on n, with the base case $n = 1$ being Claim 4.15. For the induction step, suppose that for some n there is no natural number x with $n < x < n+1$, but there is a natural number y with $n + 1 < y < n + 2$. Since $n \neq 0$ we have $y \neq 1$ so $y - 1 \in \mathbb{N}$, and since $n < y - 1 < n + 1$ this contradicts the induction hypothesis. We conclude that there is no such y, and so by induction the claim is true. $\square$

4.8 The principle of complete, or strong, induction

Sometimes it is helpful in an induction argument to be able to assume not just $p(n)$ when trying to prove $p(n + 1)$, but instead to assume $p(k)$ for all $k \leq n$. Here are the two forms of the method of *strong* or *complete* induction that this leads to.

Theorem 4.17. *Suppose that E is a set of natural numbers satisfying*

1. *1 is in E and*
2. *$k + 1$ is in E whenever every j with $j \leq k$ is.*

Then $E = \mathbb{N}$.

Theorem 4.18. *Suppose that $p(n)$ is a predicate with universe of discourse for n being all natural numbers. If*

- *$p(1)$ is True and*
- *$p(k + 1)$ is True whenever $p(j)$ is True for all $j \leq k$*

then $p(n)$ is True for all n in $\mathbb{N}$.

As we have observed earlier, complete induction (Theorem 4.18) and ordinary induction (Theorem 4.11) are equivalent, in the sense that any proof that can be carried out using one can be transformed into a proof that use the other. We repeat the justification of this claim here, in slightly different language.

Suppose we have a proof of the truth of some predicate $p(n)$ for all natural numbers n, that uses ordinary induction. Then the argument used to deduce the truth of $p(k+1)$ from that of $p(k)$, is exactly an argument that deduces the truth of $p(k+1)$ from the truth $p(j)$ for all $j \leq k$ (just one that never needs to use any of the hypotheses of the implication except $p(k)$). So any prove using ordinary induction can be transformed into one using complete induction, somewhat trivially.

On the other hand, suppose we have a proof of the truth of some predicate $p(n)$ for all natural numbers n, that uses complete induction. Let $q(n)$ be the predicate "$p(m)$ holds for all $m \leq n$". If $q(n)$ is True for all n then $p(n)$ is True for all n, and vice-versa, so to prove that $p(n)$ is True for all n it is enough to show that $q(n)$ is true for all n. This can be proved by ordinary induction: $q(1)$ is True because $p(1)$ is True, and if we assume that $q(k)$ is True for some $k \geq 1$ then we know $p(j)$ for all $j \leq k$, so we know $p(k+1)$ (by our complete induction proof of $p(n)$ for all n), so we know $p(j)$ for all $j \leq k+1$ (here we need that there are no natural numbers strictly between k and $k+1$, which is Claim 4.16) so we know $q(k+1)$, and now ordinary induction can be used to infer that $q(n)$ is true for all n.

4.9 The well-ordering principle

A *least element* of a set S of numbers is an element x_0 of S such that for all $x \in S$ we have $x_0 \leq x$. None of the set of all real numbers, or all rational numbers, or all positive numbers, or all integers, has a least element. But it seems "obvious" that the set of natural numbers has a least element, namely 1, and indeed it can be proven by induction that $n \geq 1$ for every natural number n. More generally, it should be equally obvious that every *non-empty* subset of the natural numbers has a least element (the empty set does not have any elements, so in particular does not have a least element). This "obvious" fact is hard to pin down precisely, because there are so many subsets to consider. However, it is a true fact, called the well-ordering principle.

Theorem 4.19. *Every non-empty subset of the natural numbers has a least element.*

Proof: We use the principle of complete induction. Let S be a subset of the natural numbers with no least element, and let T be the complement of S (the set of all natural numbers not in S).

We have that $1 \in T$, because if $1 \in S$ then S would have a least element, namely 1.

Suppose, for some $k \geq 1$, that for all $j \leq k$ we have $j \in T$. Then $k+1$ is in T. Indeed, suppose $k+1$ is in S. Then $k+1$ would be a least element of S, since no natural number j with $j \leq k$ is in S, so if n is in S then $n > k$, so $n \geq k+1$ (this last by Claim 4.16).

By the principle of complete induction $T = \mathbb{N}$ and so S is empty.

We have proven that a subset of $\mathbb{N}$ with no least element is empty, which is the contrapositive of the assertion we wanted to prove. $\square$

In the other direction, one can also prove the principle of complete induction using the well-ordering principle, and so, remembering that ordinary and complete induction are equivalent, we conclude that the three principles

the principle of mathematical induction
the principle of complete induction
the well-ordering principle

are equivalent (and all follow from the axioms of real numbers). We will use the three interchangeably.

5 Functions

5.1 An informal definition of a function

Informally a function is a rule that assigns, to each of a set of possible inputs, an unambiguous output. Two running examples we'll use are:

Example 1 Given a real number, square it and subtract 1, and

Example 2 Add 1 to the input, subtract 1 from the input, multiply the two answers to get the output.

In **Example 1**, the input 7 leads unambiguously to the output 48, as does the input -7 (there's no rule that says that different inputs must lead to *different* outputs). In **Example 2**, the input 3 leads unambiguously to the output $(3+1)(3-1)$ or 8.

Functions can be much more complex than this; for example we might input a natural number n, and output the nth digit after the decimal point of π, *if* that digit happens to be odd; and output the n digit after the decimal point of $\sqrt{n}$ otherwise. It's not easy to calculate specific values of this function, but you will agree that it is unambiguous[68].

As an[69] example of an ambiguous function, consider the rule "for input a positive number x, output that number y such that $y^2 = x$". What is the output associated with input 4? We have no way of knowing from the rule whether it is intended to be $+2$ or -2, so this rule doesn't define a function.

Every function has a

- **Domain**: the set of all possible inputs,

and a

- **Range**: the set of all outputs, as the inputs run over the whole domain.

For **Example** 1 the domain is the set of all real numbers. The range is less obvious, but it shouldn't be too surprising to learn that it is the set of all reals that are at least -1, or $\{x : x \geq -1\}$.

For **Example** 2 the domain is unclear. But we have the following universally agreed upon

> **Convention**: If the domain of a function of real numbers is not specified, then the domain is taken to be the largest set of reals for which the rule makes sense (i.e., does not involve dividing by zero, taking the square roof of a negative number, or evaluating 0^0); this set is called the *natural domain* of the function.

[68] Or is it?

[69] Possible "another"; see footnote above!

Based on this convention, the domain for **Example 2** is the set of all real numbers. The range is again $\{x : x \geq -1\}$.

In general it is pretty easy to determine the natural domain of a function — just throw out from the reals all values where the rule define the function leads to problems — but usually the range is far from obvious. For example, it's pretty clear that the rule that sends x to $(x^2+1)/(x^2-1)$ (call this **Example 3**) has domain $\mathbb{R} - \{-1, 1\}$, but there is no clear reason why its range is $(-\infty, -1] \cup (1, \infty)$.

This last example, by the way, makes it clear that we need some better notation for functions than "the rule that ...". If we have a compact, easily expressible rule that determines a function, and we know the domain X and range Y of the function, there is a standard convention for expressing the function, namely

$$f : X \to Y$$
$$x \mapsto \text{whatever expression describes the rule in question.}$$

For **Example 1** we might write

$$f : \mathbb{R} \to [-1, \infty)$$
$$x \mapsto x^2 - 1.$$

When using this notation, we will also use "$f(x)$" to indicate the output associated with input x, so $f(7) = 48$ and $f(-1) = 0$. But of course we can also do this for generic input x, and write $f(x) = x^2 - 1$; and since this is enough to completely specify what the function does on every possible input, we will often present an expression like this as the definition of the particular function f.

This convention is particularly convenient when we are not specifying the domain of the function we are working with, but instead taking it to have its natural domain. So we might completely specify **Example 3** by writing

$$\text{"the function } \tilde{r}_7(x) = (x^2+1)/(x^2-1)\text{"}.$$

That fully pins down the function, since we can (easily) compute the domain and (with difficulty) compute the range. (I'm deliberately using a wacky name here, $\tilde{r}_7$ rather than the more conventional f, or g, or h, to highlight that the name of a function can be *anything*).

A problem with the above notation is that it involves knowing the range, which is often very difficult to compute. We get over this by introducing the notion of

- **Codomain**: any set that *includes* the set of all possible outputs, but is not necessarily equal to the set of all possible outputs.

We then extend the notation above: if we know the domain X of a function, and also know *a* codomain Y, we can write

$$f : X \to Y$$
$$x \mapsto \text{whatever expression describes the rule in question.}$$

So **Example 2** could be written as

$$f : \mathbb{R} \to \mathbb{R}$$
$$x \mapsto (x+1)(x-1).$$

(Notice that when working with real numbers, we can *always* resort to a worst-case scenario and take all of $\mathbb{R}$ as a codomain).

Often the rule that defines a function is best expressed in pieces, as in

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x^2 & \text{if } x \geq 0. \end{cases}$$

We've seen this before, for example with the absolute value function.

5.2 The formal definition of a function

Going back to **Example 1** and **Example 2**, we might ask, are they different functions? Given our informal definition, the answer has to be "yes". The two rules — "square and subtract 1", and "add 1, subtract 1, multiply results" — are *different* rules. But we really would like the two examples to lead to the *same* function — they both have the same domains, and, because $x^2 - 1 = (x+1)(x-1)$ for all reals, for any given input each function has the same output.

This highlights one shortcoming of the informal definition we've given of a function. Another shortcoming is that it is simply too vague; what exactly do we mean by a "rule"? And without a precise formation of what is and isn't a rule, can we do any mathematics with functions?

We now give the *formal* definition of a function, which is motivated by the fact that all that's needed to specify a function is the information of what the possible inputs are, and what output is (unambiguously) associated with each input.

- A **function** is a set of *ordered pairs* (pairs of the form (a, b), where the order in which a and b are written down matters), with the property that each a which appears as the first co-ordinate of a pair in the set, appears as the first co-ordinate of exactly *one* pair.

Think of a as a possible input, and b as the associated output. The last part of the definition is what specifies that to each possible input there is an *unambiguous* assignment of output.

As an example, the function whose domain is all integers between -2 and 2 inclusive, and which is informally described by the rule "square the input", would formally be

$$f = \{(-2, 4), (-1, 1), (0, 0), (1, 1), (2, 4)\}.$$

We write "$f(-1) = 1$" as shorthand for $(-1, 1) \in f$ ($(-1, 1)$ is one of the pairs that makes up f). With this formal definition, the functions in **Example 1** and **Example 2** become the same function, because the sets of pairs (a, b) in both functions is the same.

In the context of this formal definition, we can now formally define domain, range and codomain.

- The **domain** of a function f, written Domain(f), is the set of all first co-ordinates of pair in the function;
- the **range** of f, written Range(f), is the set of all second co-ordinates of pair in the function; and
- a (not "the" — it's not unique) **codomain** of f is any set that contains the range as a subset.

Notice that although you have probably long been used to using the notation "$f(x)$" as the name for a generic function, with this formal definition it ok (and in fact more correct) to just use "f". A function is a set of pairs, and the name we use for the set (a.k.a. the name for the function) doesn't need to, and indeed shouldn't, use a variable. The expression "$f(x)$" should be understood not as a stand-in for the function, but (informally) as the output of the function when the input is x and (formally) the second co-ordinate of that pair in f whose first co-ordinate is x, if there is such a pair.

Having said that, in the future we will frequently use informality like "the function $f(x) = 3x - 2$" to specify a function, rather than the formal but more cumbersome

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

5.3 Combining functions

If f, g, h, et cetera, are all real functions (meaning: functions whose domains and ranges are all subsets of the real numbers), we can combine them to form other functions.

Addition and subtraction Informally the function $f+g$ is specified by the rule $(f+g)(x) = f(x) + g(x)$. Of course, this only makes sense for those x for which both $f(x)$ and $g(x)$ make sense; that is, for those x which are in both the domain of f and the domain of g. Formally we define

$$f + g = \{(a, b + c) : (a, b) \in f, (a, c) \in g\},$$

and observe that

$$\text{Domain}(f + g) = \text{Domain}(f) \cap \text{Domain}(g).$$

Notice that $f + g$ really is a function. It's a set of ordered pairs certainly. And suppose that a is a first co-ordinate of some pair (a, d) in $f + g$. It's in the set because there a b with $(a, b) \in f$ and a c with (a, c) in g; but by the definition of function (applied to f and g) we know that b and c are unique, so d can only be $b + c$.

Informally $f - g$ is defined by $(f - g)(x) = f(x) - g(x)$. You should furnish the formal definition for yourself as an exercise, verify that $f - g$ is indeed a function, and verify that $\text{Domain}(f - g) = \text{Domain}(f) \cap \text{Domain}(g)$ (so is the same as $\text{Domain}(f + g)$).

Notice that just like ordinary addition, addition of functions is commutative. This follows quickly from the commutativity of ordinary addition. We give the proof of this fact here; take it as a template for other, similarly straightforward facts that will be left as exercises.

Claim 5.1. *For any two real functions f and g, $f + g = g + f$.*

Proof: Suppose $(a, d) \in f + g$. That means there is a unique real b and a unique real c such that $(a, b) \in f$, $(a, c) \in g$, and $d = b + c$. But by commutativity of addition, we have $d = c + b$. This says that $(a, d) \in g + f$.

By the same reasoning, if $(a, d) \in g + f$ then $(a, d) \in f + g$. So as sets of ordered pairs, $f + g = g + f$. □

As a first exercise in similar manipulations, you should verify also that addition of real functions is associative.

Multiplication and division The product of two functions f, g is defined informally by $(fg)(x) = f(x)g(x)$, and formally by

$$fg = \{(a, bc) : (a, b) \in f, (a, c) \in g\}.$$

As with addition, $\text{Domain}(fg) = \text{Domain}(f) \cap \text{Domain}(g)$, and multiplication is commutative and associative. Moreover multiplication distributes across addition: $f(g + h) = fg + fh$.

We can also define the product of a function with a real number. If f is a function and c is a real number then cf is defined informally by $(cf)(x) = c(f(x))$, and formally by

$$cf = \{(a, cb) : (a, b) \in f\}.$$

(Notice that we never write fc — it's conventional to put the constant *in front* of the function name).

We can define $-f$ to mean $(-1)f$ (and easily check that this creates no clash with the previous use of "$-$" in the context of functions — $f + (-g) = f - g$).

Division of a function f by a function g is defined informally by $(f/g)(x) = f(x)/g(x)$, and formally by

$$f/g = \{(a, b/c) : (a, b) \in f, (a, c) \in g\}.$$

We have to be a little careful about the domain of f/g, as we not only have to consider whether f and g make sense at possible input x, but also whether the expression f/g makes sense (i.e., we have to make sure that we are not dividing by 0). We have

$$\text{Domain}(f/g) = (\text{Domain}(f) \cap \text{Domain}(g)) - \{x : (x, 0) \in g\},$$

that is, the domain of f/g is all things in the domain of both f and g, other than those things which get send to 0 by g.

Rational functions Two very important special functions are the

- **constant function**: $f(x) = 1$ for all x, formally $\{(x, 1) : x \in \mathbb{R}\}$,

and the

- **linear function**: $f(x) = x$ for all x, formally $\{(x, x) : x \in \mathbb{R}\}$,

both with domains all of $\mathbb{R}$.

Combining these two functions with repeated applications of addition, multiplication and multiplication by constants, we can form the family of

- **polynomial functions**: functions of the form $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a + 0$, where $a_0, \ldots, a_{n-1}$ are all real constants, and a_n is a non-zero constant.

Such a polynomial is said to have *degree* n, and the numbers $a_0, a_1, \ldots, a_n$ are said to be the *coefficients* of the polynomial. We will see a lot more of polynomials as the course progresses; for now we will just say that the domain of any polynomial all of $\mathbb{R}$.

Combining polynomial functions with division, we can form the family of

- **rational functions**: functions of the form $f(x) = P(x)/Q(x)$, where P and Q are polynomials, and Q is not the identically (or constantly) 0 function, $\{(x, 0) : x \in \mathbb{R}\}$.

The domain of a rational function P/Q is $\mathbb{R}$ minus all those places where Q is 0.

In the discussion above we've talked about the domains of the functions we have been building. In general it is *very difficult* to pin down ranges of functions. In fact, it's a theorem[70] that if the degree of a polynomial is an even number six or greater, and the coefficients of the polynomial are rational numbers, it is in general not possible to express the range of the polynomial using rational numbers together with addition, subtraction, multiplication, division and taking nth roots; and for a rational function P/Q, if the degree of Q is five or greater, it is in general not even possible to express the *domain* of the function succinctly!

We know that there are many more functions beyond polynomials and rational functions. Familiar examples include $\sqrt{\cdot}$, sin, log, and exp. These will be introduced formally as we go on. For now, we'll use them for examples, but won't use them in the proofs of any theorems.

5.4 Composition of functions

There's one more very important way of building new functions from old: *composition*. Informally, giving two functions f and g, the composition $f(g(x))$ means exactly what it says:

[70] A quite difficult one, using something called Galois theory.

first apply g to x, and then apply f to the result. As an example, suppose $f(x) = \sin x$ and $g(x) = x^2 + 1$. Then the composition would be $f(g(x)) = \sin(x^2 + 1)$.

Notice that unlike previous ways of combining functions,

composition is not commutative!!!.

Indeed, if you are familiar with the sin function then you will know that in the example above, since $g(f(x)) = (\sin x)^2 + 1$ and this is definitely a different function from $f(g(x)) = \sin(x^2 + 1)$, we have an example already of a pair of function f, g for which $f(g(x)) \neq g(f(x))$, in general. For a more prosaic example, consider $a(x) = x^2$ and $b(x) = x + 1$; we have

$$a(b(x)) = (x + 1)^2 = x^2 + 2x + 1 \neq x^2 + 1 = b(a(x)),$$

then inequality in the middle being witnessed by any x other than $x = 0$.

Because composition is not commutative, we have to be very careful with the informal language we use to describe composition. By convention, "f composed with g (applied to x)" means "$f(g(x))$". Notice that in this convention there is an inherent order among the functions: "f composed with g" means something quite different from "g composed with f". It is sometimes tempting to use language like "the composition of f and g", but *this is ambiguous, and should be avoided!*.

Along with the language "f composed with g", it's also common to see "f of g" and "f after g". Both of these last two an inherent order, and the last is particularly suggestive: if f is *after* g, then the action of g gets performed *first*.

What is the domain of the composition of f with g? The composed function makes sense exactly for those elements of the domain of g, for which the outputs of g are themselves in the domain of f. Consider, for example, the function given by the rule that x maps to $\sqrt{(x+1)/(x-1)}$. This is the composition of the square root function (call it sq), with the function (call it f) that maps x to $(x+1)/(x-1)$. Now the domain of f is all reals except 1; but since the domain of sq is non-negative numbers, the domain of the composition is exactly those real x that are not 1, and that have $(x + 1)/(x - 1) \geq 0$. It's an easy exercise that $(x + 1)/(x - 1) \geq 0$ precisely when either $x \leq -1$ or $x > 1$; so the domain of the composition is $\{x : x \leq -1 \text{ or } x > 1\}$, which we can also write as $(-\infty, -1] \cup (1, \infty)$.

Formally, we use the notation "$\circ$" (read "composed with", "after") to indicate composition:

$$f \circ g = \{(a, c) : (a, b) \in g \text{ for some } b, (b, c) \in f\}$$

Although composition is not commutative, it is associative; proving this is just a matter of unpacking the definition:

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

while

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

so indeed $(f \circ (g \circ h))(x) = ((f \circ g) \circ h)(x)$ for every x. To finish we just need to check that the domains of $f \circ (g \circ h)$ and $(f \circ g) \circ h$ are the same; but it's easy to check that x is in the domain of $f \circ (g \circ h)$ exactly when

- x is in the domain of h,
- $h(x)$ is in the domain of g, and
- $g(h(x))$ is in the domain of f,

and these are also exactly the conditions under which x is in the domain of $(f \circ g) \circ h$.

5.5 Graphs

Note: I haven't included any pictures in my first pass through this section. I *strongly* encourage you to read this section with desmos open on a browser, so that you can create pictures as you go along. Spivak (Chapter 4) covers the same material, and has plenty of pictures.

In this section we talk about representing functions as graphs. It's important to point out from the start, though, that a graphical representation of a function should only ever be used as an aid to thinking about a function, and to provide intuition; considerations of graphs should *never* serve as part of a proof. The example of $f(x) = \sin(1/x)$ below gives an illustration of why not, as does the graph of Dirichlet's function (again, see below).

To start thinking about graphs, first recall the real number line, a graphical illustration of the real numbers. The line is usually drawn horizontally, with an arbitrary spot marked in the center representing 0, and an arbitrary spot marked to the right of 0, representing 1. This two marks define a unit distance — the length of the line segment joining them. Relative to this unit distance, the positive number x is represented by the spot a distance x to the right of 0, while the negative number x' is represented by the spot a distance x' to the left of 0. In this way all real numbers are represented by exactly on point on the number line (assuming the line is extended arbitrarily far in each direction), and the relation "$a < b$" translates to "a is to the left of b" on the line.

Recall that after introducing the absolute value function, we commented that the (positive) number $|a - b|$ encodes a notion of the "distance" between a and b. This interpretation of absolute value makes it quite easy to represent on the number line solutions to inequalities involving absolute value. For example:

- the set of x satisfying $|x - 7| < 3$ is the set of x whose distance from 7 is at most 3; that is, the set of x which on the number line are no more than (and not exactly) 3 units above 7 and no less than (and not exactly) 3 units below 7; that is, the *open* interval of numbers between $7 - 3$ and $7 + 3$ ("open" meaning that the end-points are not in the interval); that is, the interval $(4, 10)$; and, more generally

- for fixed real x_0 and fixed $\delta > 0$, $\{x : |x - x_0| < \delta\} = (x_0 - \delta, x_0 + \delta)$.

This general example will play a major role in the most important definition of the semester, the definition of a *limit* (coming up soon).

Now we move on to graphing functions. The *coordinate plane* consists of two copies of the number line, called *axes* (singular: *axis*), perpendicular to each other, with the point of intersection of the lines (the *origin* of the plane) being the 0 point for both axes. Traditionally one of the axes is horizontal (the "x-axis"), with the right-hand direction being positive, and the other is vertical (the "y-axis"), with the upward direction being positive. It's also traditional for the location of 1 on the x-axis to be the same distance from the origin as the distance from 1 to the origin along the y-axis.

A point on the co-ordinate plane represents an *ordered pair* of numbers (a, b), with a (the "x-coordinate") being the perpendicular distance from the point to the y-axis, and b (the "y-coordinate") being the perpendicular distance from the point to the x-axis. In the other direction, each ordered pair (a, b) has associated with a unique point in the coordinate plane: to get to that point from the origin, travel a units along the x-axis (so to the right if a is positive, and to the left if a is negative), and then travel b units in a direction parallel to the y-axis (so up if b is positive, and down if b is negative).

(Some notation:

- the *first quadrant* of the coordinate plane is the top right sector consisting of points (a, b) with a, b positive;
- the *second quadrant* is the top left sector consisting of points (a, b) with a negative, b positive;
- the *third quadrant* is the bottom left sector consisting of points (a, b) with a, b negative; and
- the *fourth quadrant* is the bottom right sector consisting of points (a, b) with a positive, b negative.)

Since functions are (formally) nothing more or less than ordered pairs of numbers, the coordinate plane should be an ideal tool for representing them. Formally, the *graph* of a function is precisely the set of points on the coordinate plane that represent the pairs that make up the function. Informally, we think of the graph as encode the output for every input — to see the output associated with input x, travel x units along the x-axis, then move parallel to the y-axis until the graph is hit. Notice:

- if the graph is not hit by the line parallel to the y-axis, that passes through the point at distance x from the origin along the x-axis, then we can conclude that x is not in the domain of the function;

- the line parallel to the y-axis may need to be scanned in both directions (up and down) to find the graph; if one has to scan up, then the function is positive at x, and if one has to scan down, then it's negative at x; and
- if the line parallel to the y-axis hits the graph, it must hit it at a *single* point; otherwise the output of the function at input x is ambiguous. This leads to the
  - **Vertical line test**: A collection of points in the coordinate plane is the graph of a function, if and only if every vertical line in the plane (line parallel to the y-axis) meets the collection of points *at most once*.

A graph can only provide an imperfect representation of a function of the reals, at least if the function has infinitely many points in its domain, because we can only every plot finitely many points. Even the slickest computer, that renders lovely smooth images of graphs, is only actually displaying finitely many points — after all, there are only finitely many pixels on a screen. Except for the very simplest of graphs (e.g. straight line graphs) the best we can ever do is to plot a bunch of points, and make our best guess as to how to interpolate between the points. We can never be *certain*, just from looking at the graph, that weird things don't in fact happen in the places where we have interpolated. This is the main reason why we won't use graphs to reason about functions (but as we'll see in a while, there are other reasons).

Nonetheless, it behooves us to be familiar with the graphs of at least some of the very basic functions, and how these graphs change as the function changes slightly. The best way to become familiar with the shapes of graphs, is to draw lots of them.

The tool that I recommend for drawing graphs is desmos.com. After you hit "Start Graphing", you can enter a function in the box on the left, in the form

$$\text{“}f(x) = \text{something to do with } x\text{”}$$

(e.g., $f(x) = x^2 - 3\sqrt{x}$). The graph of the function (or at least, a good approximation to it) will appear on the right, where you can zoom in or out, and/or move to different parts of the graph. You can enter multiple functions (just give then different names), and they will helpfully appear in different colors (the color of the graph on the right matching the color of the text specifying the function on the left). This allows you to compare the graphs of different functions.

You can enter variables into the specification of a function, and you be able to create a "slider" that lets you change the specific value assigned to the variable. For example, entering "$f(x) = ax^2 + bx + c$" and creating sliders for each of a, b, c, allows you to explore how the graph of the general quadratic equation changes as the coefficients change.

In these notes, I won't go over all the graphs that might be of interest to us, and laboriously describe their properties. That would be pointless, mainly because (at the risk of beating a dead horse) *we will never use our understanding of a graph to prove something; we will only use it to aid intuition.* Instead, I invite you to go to desmos.com, and explore these families,

discovering their properties for yourself. I'll provide a list of suggested functions, and leading questions (with some answers):

- The constant function, $f(x) = c$ for real c. What happens to the graph as c changes?
- The linear function through the origin, $f(x) = mx$ for real m. What happens to the graph as m changes? What's the difference between positive m and negative m? Do all straight lines through the origin occur, as m varies over the reals? (The answer to this last question is "no". Think about the vertical line test).
- The linear function, $f(x) = mx + c$ for real m, c. What happens to the graph as c changes?

  Evidently, the graphs of the linear functions are straight lines. The number m is the *slope* of the line. It measures the ratio of the change in the y-coordinate brought about by change in the x-coordinate: if x is changed to $x + \Delta x$ (change Δx) then the output changes from mx to $m(x+\Delta x)$ for change $m\Delta x$, leading to ratio $m\Delta x/\Delta x = m$. Notice that this is *independent* of x — the linear functions are the only functions with this property, that the ration of the change in the y-coordinate to change in x-coordinate is independent of the particular x that one is at.

  This leads to an easy way to calculate the slope of a line, given two points (x_0, y_0) and (x_1, y_1) on the line: just calculate the ratio of change in y-coordinate to change in x-coordinate as one moves between these points, to get

$$m = \frac{y_1 - y_0}{x_1 - x_0}.$$

  And it also gives an easy way to calculate the precise equation of a line, given two (different) points (x_0, y_0) and (x_1, y_1): since the slope is independent of the x-value, consider a generic point $(x, f(x))$ on the line, and equate the calculations of the slope using the pair $(x, f(x))$, (x_0, y_0) and the pair $(x, f(x))$, (x_1, y_1), to get

$$\frac{f(x) - y_0}{x - x_0} = \frac{f(x) - y_1}{x - x_1},$$

  then solve for $f(x)$.
- The quadratic function, $f(x) = ax^2 + bx + c$ for real a, b, c. What is the general shape of the graph? What happens to the graph as a, b, c change? In particular, how does the *sign* of a (whether it is positive or negative) affect the shape?

  The shape of the graph of a quadratic function is referred to as a parabola. Parabolas have very clean geometric interpretations:

  – a *parabola* is the set of points in the coordinate plane that are equidistant from a fixed line and a fixed point.

We illustrate by considering the horizontal line $f(x) = r$, and the point (s, t), where we'll assume $t \neq r$. The (perpendicular, shortest) distance from a point (x, y) to the line $f(x) = r$ is $|y - r|$, and the (straight line) distance from (x, y) to (s, t) is, by the Pythagorean theorem

$$\sqrt{(x-s)^2 + (y-t)^2}.$$[71]

So the points (x, y) that are equidistant from the line and the point are exactly those that satisfy

$$|y - r| = \sqrt{(x-s)^2 + (y-t)^2}$$

which, because both sides are positive, is equivalent to

$$(y - r)^2 = (x - s)^2 + (y - t)^2$$

or

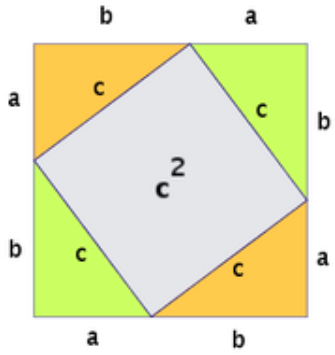
$$y = \frac{x^2}{2(t-r)} - \frac{sx}{(t-r)} + \frac{s^2 + t^2 - r^2}{2(t-r)},$$

so the graph of the set of points is the graph of a specific quadratic equation.

Of course, there are far more parabolas than graphs of quadratic equations: by drawing some lines and points in the plane, and roughly sketching the associated parabolas, you will quickly see that a parabola is only the graph of a quadratic (that is, only passes the vertical line test) if the line happens to be parallel to the x-axis.

- The general polynomial, $f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ for real a_i's, $a_n \neq 0$. What is the general shape? In particular, what happens to the graph for very large x

[71] Why is this a reasonable formula for the straight-line distance between (x, y) to (s, t)? This distance formula comes quickly from the Pythagorean theorem, which says that if the hypotenuse of a right-angled triangle has length c, and the other two side lengths are a and b, then $a^2 + b^2 = c^2$. But why is *this* true? There are *many* proofs, going back to Euclid, about 300BC. My favorite proof is conveyed succinctly in the following picture (taken from https://math.stackexchange.com/questions/563359/is-there-a-dissection-proof-of-the-pythagorean-theorem-for-tetrahedra):



The area of the big square is $(a + b)^2$; but it is also $c^2 + 4((1/2)ab)$. So $(a + b)^2 = c^2 + 4((1/2)ab)$, or $a^2 + b^2 = c^2$.

and very small x (i.e., very large negative x), and how does that depend on n and a_n? How many "turns" does the graph have, and how does change as n changes?

We will be able to answer these questions fairly precisely, once we have developed the notion of the derivative.

More generally one may ask,

- How does the graph of $f(cx)$ related to the graph of $f(x)$, for constant c? What's the different between positive and negative c here?
- What about the graph of $cf(x)$?
- and $f(x+c)$?
- and $f(x)+c$?

and one may explore the answers to these questions by plotting various graphs, and seeing what happens as the various changes are made.

One important graph that is not the graph of a function is that of a circle. Geometrically, a *circle* is the set of all points at a fixed distance r (the *radius*) from a given point (a, b) (the *center*), and algebraically the circle is set of all points (x, y) in the coordinate plane satisfying

$$(x-a)^2 + (y-b)^2 = r^2$$

(using the Pythagorean theorem to compute distance between two points). A circle that will be of special interest to us is the *unit circle centered at the origin*, given algebraically by

$$x^2 + y^2 = 1.$$

The circle is not the graph of a function, because it fails the vertical line test. A circle can be represented as the union of *two* functions, namely

$$f(x) = \sqrt{r^2 - (x-a)^2} + b, \quad x \in [a-r, a+r]$$

and

$$f(x) = -\sqrt{r^2 - (x-a)^2} + b, \quad x \in [a-r, a+r].$$

Related to the circle is the *ellipse*, a "squashed" circle, which geometrically is the set of all points, the sum of whose distances to two fixed points is a given fixed constant (so when the two points coincide, the ellipse becomes a circle). One also sometimes encounters the *hyperbola*, the set of all points the difference of whose distance from two points is the same. Circles, ellipses, parabolas and hyperbola are all examples of *conic sections*, shapes beloved of ancient mathematicians. In a modern calculus course like the present one, we will not have any need for conic sections, but if you interested there is a chapter in Spivak on the topic.

Two important functions that we will use for examples are the trigonometric functions sin and cos. We'll give a provisional definition here; it won't be until the spring semester, when we have studied the derivative, that we will give a precise definition.

Provisional definition of sin **and** cos The points reached on unit circle centered at the origin, starting from $(1, 0)$, after traveling a distance θ, measured counter-clockwise, is $(\cos\theta, \sin\theta)$.

The domain of point cos and sin is all of $\mathbb{R}$, since one can travel any distance along the circle. Negative distances are interpreted to mean clockwise travel, and distance greater than 2π (the circumference of the circle) simply traverse the circle many times.

Let's watch the trajectory of cos, as a point travels around the circle:

- at $\theta = 0$, we are at $(1, 0)$, and so $\cos 0 = 1$;
- as θ increases from 0 to $\pi/2$ (a quarter of the circle), we go from $(1, 0)$ to $(0, 1)$, with decreasing x-coordinate, and so $\cos\theta$ decreases from 1 to 0 as θ increases from 0 to $\pi/2$, and $\cos \pi/2 = 0$;
- as θ increases from $\pi/2$ to π, we go from $(0, 1)$ to $(-1, 0)$, with decreasing x-coordinate, and so $\cos\theta$ decreases from 0 to -1 as θ increases from $\pi/2$ to π, and $\cos\pi = -1$;
- as θ increases from π to $3\pi/2$, we go from $(-1, 0)$ to $(0, -1)$, with increasing x-coordinate, and so $\cos\theta$ increases from -1 to 0 as θ increases from π to $3\pi/2$, and $\cos 3\pi/2 = 0$;
- as θ increases from $3\pi/2$ to 2π, we go from $(0, -1)$ to $(1, 0)$, with increasing x-coordinate, and so $\cos\theta$ increases from 0 to 1 as θ increases from $3\pi/2$ to 2π, and $\cos 2\pi = 1$.

This gives the familiar graph of cos on the interval $[0, 2\pi]$, and of course, since we are back where we started after traveling fully around the circle, the graph just periodically repeats itself from here on.

Going the other direction, as θ decreases from 0 to $-\pi/2$ (a quarter of the circle, clockwise), we go from $(1, 0)$ to $(0, -1)$, with decreasing x-coordinate, and so $\cos\theta$ decreases from 1 to 0 as θ decreases from 0 to $-\pi/2$, and $\cos -\pi/2 = 0$, and continuing in this manner we see the graph also extends periodically on the negative side of the y-axis.

We can play the same game with sin, and discover that this provisional definition[72] yields the expected periodic graph there, too.

The sin function, suitably modified, gives us a ready example of a function whose behavior cannot be understand fully using a graph. Consider $f(x) = \sin(1/x)$ (on domain $\mathbb{R} - \{0\}$) (formally, the composition of sin with the function that takes reciprocal). Just like sin, this is a function that oscillates, but unlike sin the oscillations are not of length (2π in the case of sin. As x comes from infinity to $1/(2\pi)$, $1/x$ goes from 0 to 2π, so f has one oscillation in that (infinite) interval. Then, as x moves down from $1/(2\pi)$ to $1/(4\pi)$, $1/x$ goes from 2π to 4π, so f has another oscillation in that (finite) interval. The next oscillation happens in the shorter finite interval as x moves down from $1/(4\pi)$ to $1/(6\pi)$; the next in the even

[72] Why is this a *provisional* definition? Because it requires understanding length along the curved arc of a circle. To make the notion of length along a curve precise, we need to first study the integral.

shorter interval as x moves down from $1/(6\pi)$ to $1/(8\pi)$. As x gets closer to 0, the oscillations happen faster and faster, until they get to a point where each oscillation is happening in an interval that is shorter than the resolution of the graphing device. Go ahead and graph $f(x) = \sin(1/x)$ on Desmos, and see what happens (in particular as you zoom in to $(0, 0)$). This should convince you that a graph is not always a useful tool to understand a function.

Another function that illustrates the limitations of graphing is *Dirichlet's function*:

$$f(x) = \begin{cases} 1 & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational.} \end{cases}$$

Because the rationals are "dense" in the reals — there are rationals arbitrarily close to any real — and the irrationals are also dense, any attempt at a graph of f is going to end up looking like two parallel straight lines, one along the x-axis (corresponding to the irrational inputs) and the other one unit higher (corresponding to the rational inputs), and this is certainly a picture that fails the vertical line test.

Going back to $f(x) = \sin(1/x)$, let's consider a related function, $g(x) = x \sin(1/x)$ (again on domain $\mathbb{R} = \{0\}$). Again this has oscillations that get arbitrarily close together as x gets close to 0, but now these oscillations are "pinched" by the lines $y = x$ and $y = -x$, so as we get closer to zero, the amplitudes of the oscillations (difference between highest and lowest point reached) get smaller and smaller. We will soon discuss the significant difference between f and g in their behavior close to 0. For now, let's ask the question

> how do f and g behave for very large positive inputs?

It's not hard to see that f should be getting closer to 0 as the input x gets larger — for large x, $1/x$ is close to 0 and $\sin 0 = 0$. It's less clear what happens to g. The $\sin(1/x)$ part is going to 0, while the x part is going to infinity. What happens when these two parts are multiplied together?

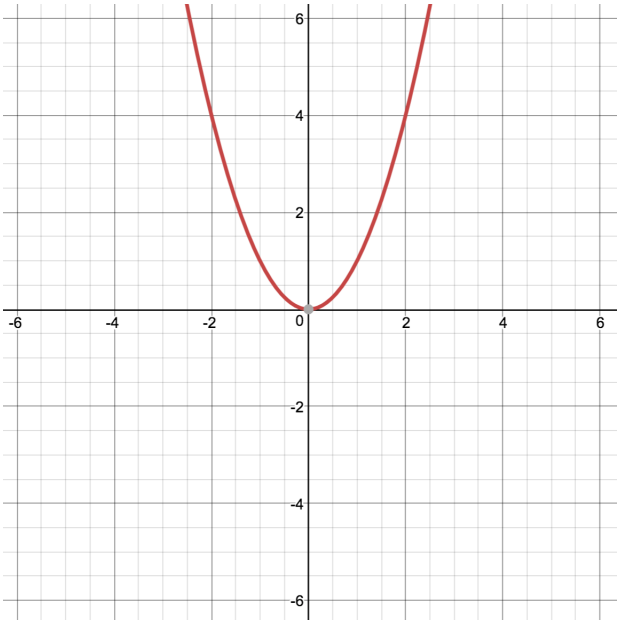
- Is the x part going to infinity faster than the $\sin(1/x)$ part is going to 0, leading to the product g going to infinity?
- Or is the x part going to infinity slower than the $\sin(1/x)$ part is going to 0, leading to the product g going to zero?
- Or are they both going to their respective limits at roughly the same rate, so that in the product they balance each other out, and g gets closer to some fixed number?
- Or is g oscillating as x grows, not moving towards some limit?

A look at the graph of g on a graphing calculator suggests the answer. To mathematically pin down the answer, we need to introduce a concept that is central to calculus, and has been central to a large portion of mathematics for the last 200 years, namely the concept of a limit.

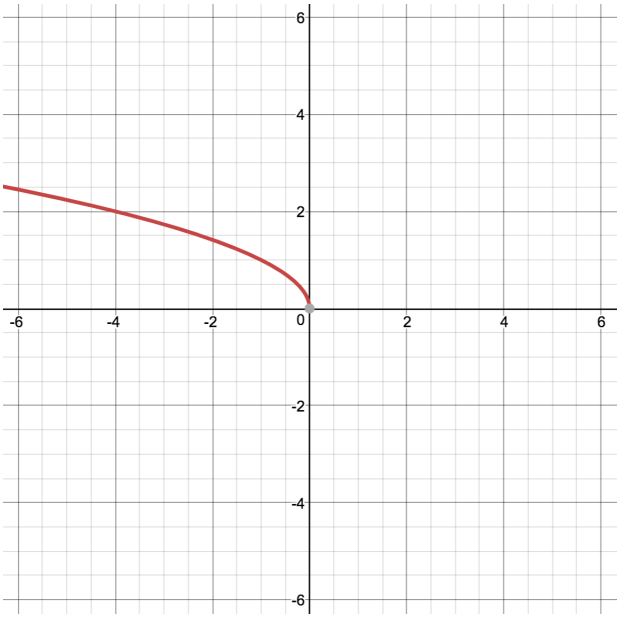
6 Limits

What does a function "look like" as the inputs "approach" a particular input? We'll formalize this vague question, already brought up at the end of the last section, using the notion of a limit. To begin, let us note that there are many possible behaviors a function might exhibit as the inputs approach a particular value a. We illustrate ten possible such behaviors here.

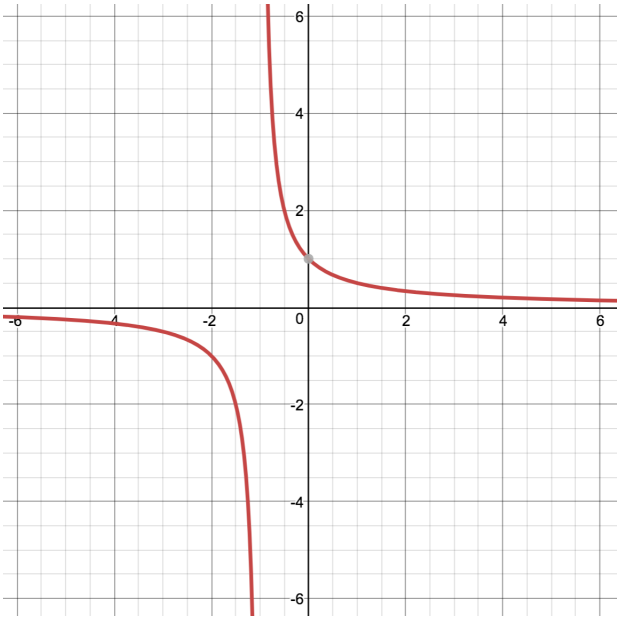
1. Function exhibits no problems at a $f_1(x) = x^2$ at $a = 1$.



2. Function defined nowhere near a $f_2(x) = \sqrt{-x}$ at $a = 1$.

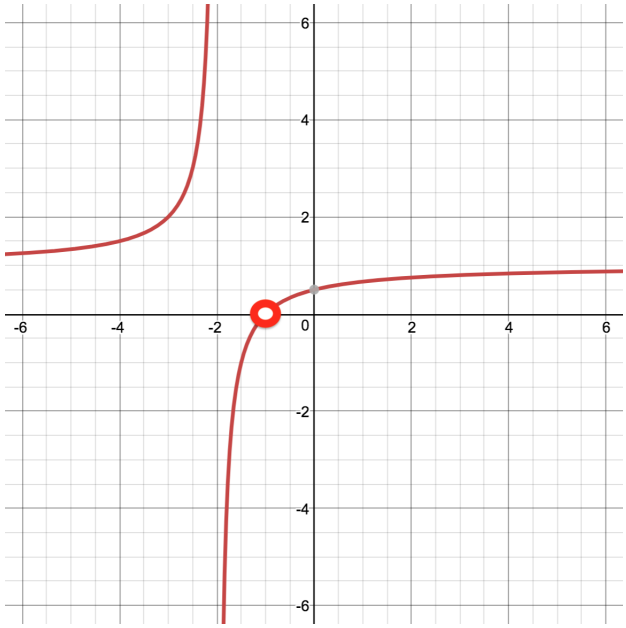


3. **Function blows up to infinity approaching a** $f_3(x) = 1/(1+x)$ at $a = -1$.

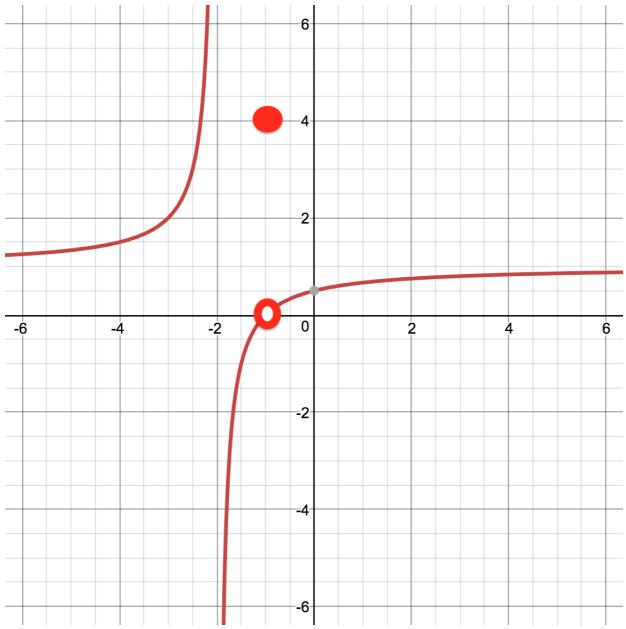


4. **Function not defined at a, but otherwise unremarkable** $f_4(x) = 1/(1 + (1/(1 + x)))$, $a = -1$. This situation, a function with a "hole", might seem odd, but it can arise naturally. Notice here that $f_4 = f_3 \circ f_3$, and that the expression on the right-hand side of the definition of f_4 can be re-written as $(1+x)/(2+x)$, which *does* make sense at $x = -1$. So the function $f_4 = f_3 \circ f_3$, (with natural domain $\mathbb{R} - \{-1, -2\}$), is identical to the function that sends x to $(1+x)/(2+x)$ (which has natural domain $\mathbb{R} - \{-2\}$), *except* at -1, where f_4 has a "naturally occurring" hole.

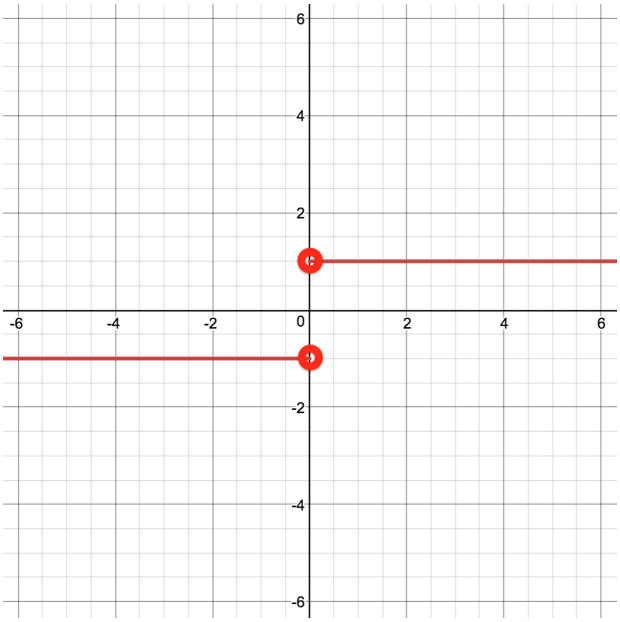
   Notice also the graphical notation that we use to indicate the "hole" at -1: literally, a hole.



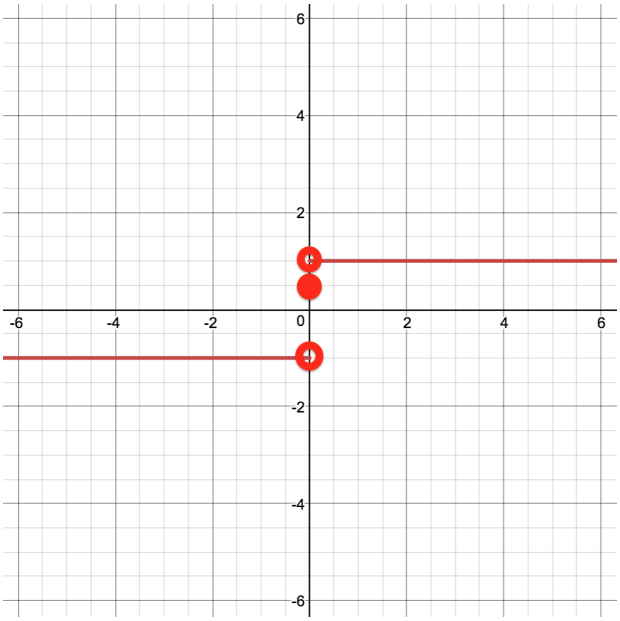
5. Function with "wrong value" at a $f_5(x) = \begin{cases} f_4(x) & \text{if } x \neq -1 \\ 4 & \text{if } x = -1 \end{cases}$ at $a = -1$.



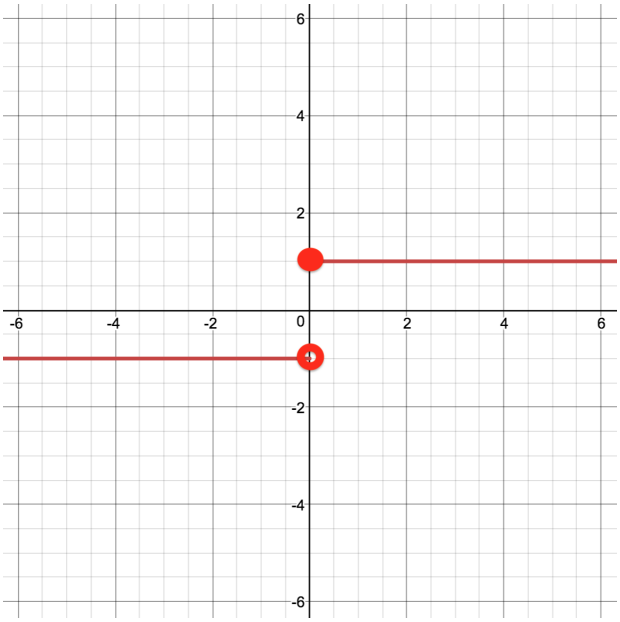
6. Function with a "jump" at a (1) $f_6(x) = \frac{x}{|x|}$ at $a = 0$. The natural domain here is $\mathbb{R} - \{0\}$, and for positive x, $x/|x| = 1$ while for negative x, $x/|x| = -1$. Notice that we graphically indicate the failure of the function to be defined at 0 by *two* holes, one at the end of each of the intervals of the graph that end at 0.



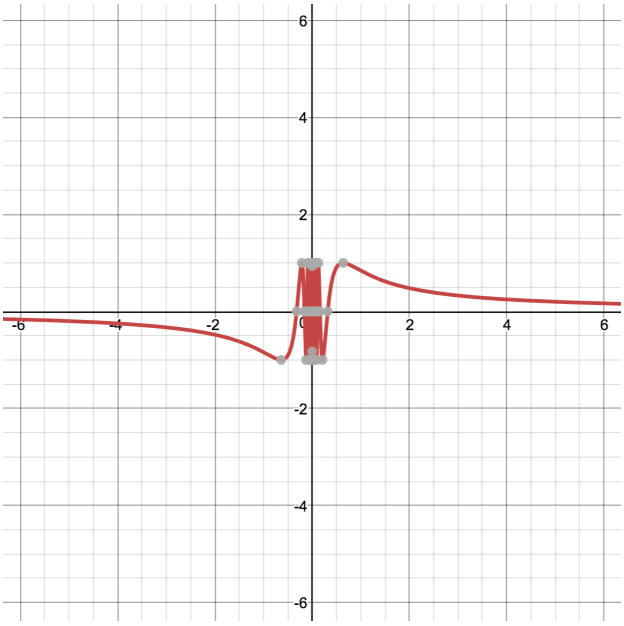
7. Function with a "jump" at a (2) $f_7(x) = \begin{cases} f_6(x) & \text{if } x \neq 0 \\ 1/2 & \text{if } x = 0 \end{cases}$ at $a = 0$. Notice that we graphically indicate the value of the function at 0 with a *solid* holes at the appropriate height.



8. Function with a "jump" at a (3) $f_8(x) = \begin{cases} f_6(x) & \text{if } x \neq 0 \\ 1 & \text{if } x = 0 \end{cases}$ at $a = 0$. Notice that here we graphically indicate the behavior of the function around its jump with an appropriate combination of holes and solid holes.



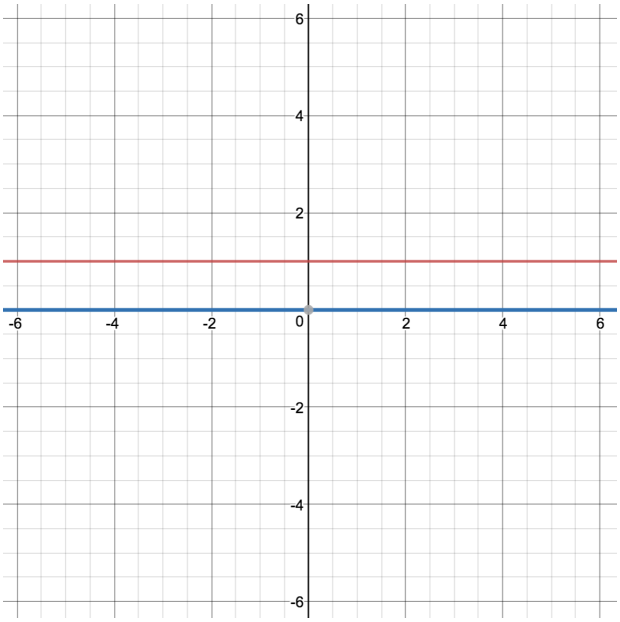
9. Oscillatory function near a $f_9(x) = \sin(1/x)$ at $a = 0$. The natural domain here is $\mathbb{R} - \{0\}$. Notice the complete failure of the graph to convey the behavior of the function!



10. Chaotic function near a $f_{10}(x) = \begin{cases} 1 & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational} \end{cases}$ at $a = 0$. This function is often called the *Dirichlet* function[73]. Because the rationals are dense in the reals, and so are the irrationals (given any real, there are rationals arbitrarily close to it, and

[73] After the German mathematician Peter Dirichlet, https://en.wikipedia.org/wiki/Peter_Gustav_Lejeune_Dirichlet.

irrationals arbitrarily close to it), the graph of f_{10} just looks like two horizontal lines, and appears to completely fail the vertical line test!



Other behaviors are possible, too, and of course, we could have one kind of behavior on one side of a, and another on the other side.

6.1 Definition of a limit

We would like to develop a definition of the notion "f approaches a limit near a", or "the outputs of f approach a limit, as the inputs approach a", that accounts for our intuitive understanding of the behavior of each of f_1 through f_{10}. Here is an intuitive sense of what is going on in each of the examples:

- f_1 approaches 1 near 1 (as input values get closer to 1, outputs values seem to get closer to 1).
- f_2 doesn't approach a limit near 1 (it isn't even defined near 1).
- f_3 doesn't approach a limit near -1 (or, it approaches some infinite limit — as input values get closer to -1, output values either get bigger and bigger positively, or bigger and bigger negatively).
- Even though f_4 is not defined at -1, it appears that f_4 approaches a limit of 0 near -1 (as input values get closer to -1, outputs values seem to get closer to 0).
- Even though $f_5(-1)$ is not 0, it seems reasonable still to say that f_5 approaches a limit of 0 near -1 (as input values get closer to -1, outputs values seem to get closer to 0).

- f_6 doesn't approach a limit near 0 (as input values get closer to 0 from the right, the outputs values seem to get closer to 1, but as input values get closer to 0 from the left, the outputs values seem to get closer to -1; this ambiguity suggests that we should not declare there to be a limit).
- f_7 doesn't approach a limit near 0 (exactly as f_6: specifying a value for the function at 0 doesn't change the behavior of the function as we approach 0).
- f_8 doesn't approach a limit near 0 (exactly as f_7).
- f_9 doesn't approach a limit near 0 (the outputs oscillate infinitely in the interval $[-1, 1]$ as the inputs approach 0, leading to an even worse ambiguity than that of f_6).
- f_{10} doesn't approach a limit near 0 (the outputs oscillate infinitely between -1 and 1 as the inputs approach 0, again leading to an worse ambiguity than that of f_6).

What sort of definition will capture these intuitive ideas of the behavior of a function, near a potential input value? As a provisional definition, we might take what is often considered the "definition" of a limit:

> **Provisional definition of function tending to a limit**: A function f tends to a limit near a, if there is some number L such that f can be made arbitrarily close to L by taking input values sufficiently close to a.

This definition seems to work fine for f_1 through f_4. For f_4, for example, it seems very clear that we can get the function to take values arbitrarily close to 0, by only considering input values that are pinned to be sufficiently close to -1 (on either side); and for f_3, no candidate L that we might propose for the limit will work — as soon as we start considering inputs that are too close to -1, the values of the outputs will start to be very far from L (they will either have the wrong sign, or have far greater magnitude than L).

It breaks down a little for f_5: we can't make output values of f_5 be arbitrary close to 0 by choosing input values sufficiently close to -1, because -1 surely fits the "sufficiently close to -1" bill (nothing could be closer!), and $f_5(-1) = 4$, far from 0. The issue here is that we want to capture the sense of how the function is behaving *as inputs get close to* a, and so we really should *ignore* what happens *exactly at* a. There's an easy fix for this: add "(not including a itself)" at the end of the provisional definition.

f_6 presents a more serious problem. We can certainly make the outputs of f_6 be arbitrarily close to 1, by taking inputs values sufficiently close to 0 — indeed, *any* positive input value has output *exactly* 1. But by the same token, we can make the outputs of f_6 be arbitrarily close to -1, by taking inputs values sufficiently close to 0 — *any* negative input value has output *exactly* -1.

The issue is that we are "cherry picking" the inputs that are sufficiently close to 0 — positive inputs to get the limit to be 1, negative inputs to get the limit to be -1. In f_9 the situation is even more dramatic. If we pick any L between -1 and 1, we can find a sequence

of numbers (one in each oscillation of the function) that get arbitrarily close to 0, such that when f_9 is evaluated at each of these numbers, the values are always *exactly* L (not just getting closer to L) — just look at the infinitely many places where that line $y = L$ cuts across the graph of f_9. So we can make, with our provisional definition, a case for *any* number between -1 and 1 being a limit of the function near 0! This runs at odds to intuition.

We need to remove the possibility of "cherry-picking" values of the input close to a to artificially concoct a limit that shouldn't really be a limit. The way we will do that is best described in terms of a game, played by Alice and Bob.

Suppose Alice and Bob are looking at the function $f : \mathbb{R} \to \mathbb{R}$ given by $x \mapsto 3x$. Alice believes that as x approaches 1, f approaches the limit 3. Bob is skeptical, and needs convincing. So:

- Bob says "1", and challenges Alice to show that for *all* values of the input sufficiently close to 3, f is within 1 of 9 (asking for *all* values is what eliminates the possibility of cherry-picking values). Think of "1" as a "window of tolerance".

- Alice notices that as x goes between 22/3 and 31/3, $f(x)$ goes between 8 and 10; that is, as long as x is within 1/3 of 3, $f(x)$ is within 1 of 9. So she convinces Bob that output values can be made to be within 1 of 9 by telling him to examine values of x within 1/3 of 1.

- Bob is ok with this, but now wants to see that f can be forced to be even closer to 1. He says "1/10", a smaller window of tolerance, and challenges Alice to show that for all values of the input sufficiently close to 3, f is within 1/10 of 9. Alice repeats her previous calculations with the new challenge number, and responds by saying "1/30": all values of x within 1/30 of 3 give values of $f(x)$ within 1/10 of 9.

- Bob ups the ante, and says "1/1000". Alice responds by saying "1/3000": all values of x within 1/3000 of 3 give values of $f(x)$ within 1/1000 of 9.

- Bob keeps throwing values at Alice, and Alice keeps responding. But Bob won't be fully convinced, until he knows that Alice can make a valid response for *every* possible window of tolerance. So, Bob says "ε: an arbitrary number greater than 0". Now Alice's response must be one that depends on ε, and is such that for each particular choice of $\varepsilon > 0$, evaluates to a valid response. She notices that as x goes between $3 - \varepsilon/3$ and $3 + \varepsilon/3$, $f(x)$ goes between $9 - \varepsilon$ and $9 + \varepsilon$; that is, as long as x is within $\varepsilon/3$ of 3, $f(x)$ is within ε of 9. She tells this to Bob, who is now convinced that as x approaches 1, f approaches the limit 3.

This leads to the definition of a limit.

Definition of function tending to a limit: A function f tends to a limit near a, if

- f is defined near a, meaning that for some small enough number b, the set $(a-b, a+b) \setminus \{a\}$ is in domain of f,

and

- there is some number L such that
- for all positive numbers ε
- there is a positive number δ such that
- whenever x is within a distance δ of a (but is not equal to a)
- f is within ε of L.

More succinctly, f tends to a limit near a, if f is defined near a and there is some number L such that for all $\varepsilon > 0$ there is $\delta > 0$ such that for all x, $0 < |x-a| < \delta$ implies $|f(x) - L| < \varepsilon$.

We write $f(x) \to L$ as $x \to a$ or $\lim_{x\to a} f(x) = L$.

6.2 Examples of calculating limits from the definition

Here's a simple example. Consider the constant function $f(x) = c$ for some real c. It seems clear that for any real a, $\lim_{x\to a} f(x) = c$. To formally verify this, let $\varepsilon > 0$ be given. We need to find a $\delta > 0$ such that if $0 < |x-a| < \delta$, then $|f(x) - c| < \varepsilon$. But $|f(x) - c| = 0 < \varepsilon$ for *every* x; so we can choose *any* $\delta > 0$ and the implication will be true. In particular, it will be true when we take, for example, $\delta = 1$.

Here's another simple example. Consider the linear function $f(x) = x$. It seems clear that for any real a, $\lim_{x\to a} f(x) = a$. To formally verify this, let $\varepsilon > 0$ be given. We need to find a $\delta > 0$ such that if $0 < |x-a| < \delta$, then $|f(x) - a| < \varepsilon$. But $|f(x) - a| = |x-a|$; so we are looking for a $\delta > 0$ such that if $0 < |x-a| < \delta$, then $|x-a| < \varepsilon$. It is clear that we will succeed in this endeavor by taking $\delta = \varepsilon$; note that since $\varepsilon > 0$, this choice of δ is positive.

The next simplest example is the function $f(x) = x^2$. It seems clear that for any real a, $\lim_{x\to a} f(x) = a^2$. The verification of this from the definition will be considerably more involved than the first two examples.

Let $\varepsilon > 0$ be given. We need to find a $\delta > 0$ such that if $0 < |x-a| < \delta$, then $|x^2 - a^2| < \varepsilon$. Since the only leverage we have is the choice of δ, and δ is related to $|x-a|$, it seems like it will be very helpful to somehow rewrite $|x^2 - a^2| < \varepsilon$ in a way that brings the expression $|x-a|$ into play. We have such a way, since

$$|x^2 - a^2| = |(x-a)(x+a)| = |x-a||x+a|.$$

We want to make the product of these two things small (less than ε). We can easily make $|x-a|$ small — in fact, we get a completely free hand in choosing how small this term is. We don't get to make $|x+a|$ small, however, and in fact we shouldn't expect to be able to make it small: near a, $|x+a|$ is near $|2a|$, which isn't going to be arbitrarily small.

This is an easily resolved problem. We only need to make $|x+a|$ *slightly* small. We can then use the freedom we have to make $|x-a|$ as small as we want, to make it so small that, even when multiplied by $|x+a|$, the product is still smaller than ε.

Here's a first attempt: as we've said, near a, $|x+a|$ is near $|2a|$, so $|x-a||x+a|$ is near $|2a||x-a|$. So we should make $|x-a|$ be smaller than $\varepsilon/|2a|$, to get $|x-a||x+a|$ smaller than ε.

One problem here is that $|a|$ might be 0, and so we are doing an illegal arithmetic operation. Another problem is that we are vaguely saying that "near a", $|x+a|$ is "close to" $|2a|$, which is not really an acceptable level of precision.

Here's a more rigorous approach: let's start by promising that whatever δ we choose, it won't be bigger than 1 (this is a completely arbitrary choice). With this promise, we know that when $0<|x-a|<\delta$ we definitely have $|x-a|<1$, so x is in the interval $(a-1,a+1)$. That means that $x+a$ is in the interval $(2a-1,2a+1)$. At most how big can $|x+a|$ be in this case? At most the maximum of $|2a-1|$ and $|2a+1|$. By the triangle inequality, $|2a-1|\le|2a|+1$ and $|2a+1|\le|2a|+1$, and so, as long as we stick to our promise that $\delta\le 1$, we have $|x+a|<|2a|+1$. This makes $|x^2-a^2|<(2|a|+1)|x-a|$. We'd like this to be at most ε, so we would like to choose δ to be no bigger than $\varepsilon/(2|a|+1)$ (thus forcing $|x-a|<\varepsilon/(2|a|+1)$ and $|x^2-a^2|<\varepsilon$ whenever $0<|x-a|<\delta$).

We don't want to simply say "ok, take δ to be any positive number $\le\varepsilon/(2|a|+1)$" (note that $\varepsilon/(2|a|+1)>0$, so there *is* such a positive δ). Our choice here was predicated on our promise that $\delta\le 1$. So what we really want to do, is choose δ to be any positive number no bigger than *both* $\varepsilon/(2|a|+1)$ *and* 1. We can do this, for example, by taking δ to be the minimum of $\varepsilon/(2|a|+1)$ and 1, or, symbolically,

$$\delta=\min\left\{\frac{\varepsilon}{2|a|+1},1\right\}.$$

Going back through the argument with this choice of δ, we see that all the boxes are checked: suppose $0<|x-a|<\delta$. Then in particular we have $|x-a|<1$, and we also have $|x-a|<\varepsilon/(2|a|+1)$. From $|x-a|<1$ we deduce $a-1<x<a+1$, so $2a-1<x+a<2a+1$, so $|x+a|<\max\{|2a-1|,|2a+1|\}\le 2|a|+1$. From this and $|x-a|<\varepsilon/(2|a|+1)$ we deduce

$$|x^2-a^2|=|x+a||x-a|<\frac{\varepsilon}{2|a|+1}(2|a|+1)=\varepsilon,$$

and so, since ε was arbitrarily, we deduce that indeed $\lim_{x\to a}f(x)=a^2$.

We do one more example: $\lim_{x\to 2}\frac{3}{x}$. It seems clear that this limit should be $3/2$. Given $\varepsilon>0$, we need $\delta>0$ such that $0<|x-2|<\delta$ implies $|(3/x)-(3/2)|<\varepsilon$. We have

$$\left|\frac{3}{x}-\frac{3}{2}\right|=\left|\frac{6-3x}{2x}\right|=\frac{3}{2}\frac{|x-2|}{|x|}.$$

We want to make this small, which requires making $|x|$ *large*. If $\delta\le 1$ then $0<|x-2|<\delta$ implies $x\in(1,3)$, so $|x|>1$ and $3/(2|x|)<3/2$. So if both $\delta\le 1$ *and* $\delta\le 2\varepsilon/3$, we have

$$\left|\frac{3}{x}-\frac{3}{2}\right|=\frac{3}{2}\frac{|x-2|}{|x|}<\frac{3}{2}\cdot\frac{2\varepsilon}{3}=\varepsilon$$

as long as $0 < |x-2| < \delta$. Taking δ to be $\min\{1, 2\varepsilon/3\}$ verifies $\lim_{x\to 2} 3/x = 3/2$.

Notice that we initially choose $\delta \leq 1$ to get a lower bound on $|x|$. Any δ would have worked, as long as we avoided have 0 in the possible range of values for x (if we allowed 0 to be in the possible range of values for x we would have *no* upper bound on $1/|x|$).

Essentially all examples of proving claimed values of limits directly from the definition follow the path of these last two examples:

- do some algebraic manipulation on the expression $|f(x)-L|$ to isolate $|x-a|$ (a quantity we have complete control over);
- by putting a preliminary bound on δ, put some bound $B > 0$ on the part of $|f(x) - L|$ that does not involve $|x - a|$;
- choose δ to be the smaller of ε/B and the preliminary bound on δ.

6.3 Limit theorems

To streamline the process of computing limits, we prove a few general results. The first is a result that says that the limits of sums, products and ratios of functions, are the sums, products and ratios of the corresponding limits.

Theorem 6.1. *(Sum/product/reciprocal theorem) Let f, g be functions both defined near some a. Suppose that $\lim_{x\to a} f(x) = L$ and $\lim_{x\to a} g(x) = M$ (that is, both limits exist, and they take the claimed values). Then*

- $\lim_{x\to a}(f+g)(x)$ *exists and equals $L + M$;*
- $\lim_{x\to a}(fg)(x)$ *exists and equals LM; and,*
- *if $M \neq 0$ then $\lim_{x\to a}(1/g)(x)$ exists and equals $1/M$.*

Proof: We begin with the sum statement. Since f, g are defined near a, so is $f + g$. Let $\varepsilon > 0$ be given. Because $\lim_{x\to a} f(x) = L$, there is $\delta_1 > 0$ such that $0 < |x - a| < \delta_1$ implies $|f(x) - L| < \varepsilon/2$, and because $\lim_{x\to a} g(x) = M$, there is $\delta_2 > 0$ such that $0 < |x - a| < \delta_2$ implies $|g(x) - M| < \varepsilon/2$. Now if $\delta = \min\{\delta_1, \delta_2\}$, we have that if $0 < |x - a| < \delta$ then

$$\begin{aligned} |(f+g)(x) - (L+M)| &= |(f(x)+g(x)) - (L+M)| \\ &= |(f(x)-L) + (g(x)-M)| \\ &\leq |f(x)-L| + |g(x)-M| \quad \text{by triangle inequality} \\ &< \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned}$$

This shows that $\lim_{x\to a}(f+g)(x) = L + M$.

We now move on to the product statement, which is a little more involved. Again, since f, g are defined near a, so is fg. Let $\varepsilon > 0$ be given. We have[74]

$$\begin{aligned} |(fg)(x) - LM| &= |f(x)g(x) - LM)| \\ &= |f(x)g(x) - Lg(x) + Lg(x) - LM)| \\ &= |g(x)(f(x) - L) + L(g(x) - M)| \\ &\leq |g(x)||f(x) - L| + |L||g(x) - M| \quad \text{by triangle inequality.} \end{aligned}$$

We can make $|f(x) - L|$ and $|g(x) - M|$ as small as we like; we would like to make them small enough that $|g(x)||f(x) - L| < \varepsilon/2$ and $|L||g(x) - M| < \varepsilon/2$. The second of those is easy to achieve. There's $\delta_1 > 0$ such that $0 < |x - a| < \delta_1$ implies $|g(x) - M| < \varepsilon/(2(|L| + 1))$, so $|L||g(x) - M| < |L|(\varepsilon/(2(|L| + 1))) < \varepsilon/2$.[75]

The first is less easy. We need an upper bound on $|g(x)|$. We know that there is a $\delta_2 > 0$ such that $0 < |x - a| < \delta_2$ implies $|g(x) - M| < 1$ so $|g(x)| < |M| + 1$. There's also a $\delta_3 > 0$ such that $0 < |x - a| < \delta_3$ implies $|f(x) - L| < \varepsilon/(2(|M| + 1))$.

As long as δ is at most the minimum of δ_1, δ_2 and δ_3, we have that $0 < |x - a| < \delta$ implies all of

- $|L||g(x) - M| < \varepsilon/2$
- $|g(x)| < |M| + 1$, so $|g(x)||f(x) - L| < (|M| + 1)||f(x) - L||$
- $|f(x) - L| < \varepsilon/(2(|M| + 1))$, so $|g(x)||f(x) - L| < \varepsilon/2$,
- so, combining first and fourth points, $|g(x)||f(x) - L| + |L||g(x) - M| < \varepsilon$.

It follows from the chain of inequalities presented at the start of the proof that $0 < |x - a| < \delta$ implies

$$|(fg)(x) - LM| < \varepsilon,$$

and so $\lim_{x \to a}(fg)(x) = LM$.

We now move on to the reciprocal statement. Here we have to do some initial work, simply to show that $(1/g)$ is defined near a. To show this, we need to establish that near a, g is not 0. The fact that g approaches M near a, and $M \neq 0$, strongly suggests that this is the case. To verify it formally, we make (and prove) the following general claim, that will be of some use to us in the future.

[74] We use a trick here — adding and subtracting the same quantity. The motivation is that we want to introduce $|f(x) - L|$ into the picture, so we subtract $Lg(x)$ from $f(x)g(x)$. But to maintain equality, we then need to add $Lg(x)$; this conveniently allows us to bring $|g(x) - M|$ into the picture, also. We'll see this kind of trick many times.

[75] Why did we want $2(|L| + 1)$ in the denominator, rather than $2|L|$? This was an overkill designed to avoid the possibility of dividing by 0.

Claim 6.2. *Let g be defined near a, and suppose $\lim_{x\to a} g(x)$ exists and equals M. If $M > 0$, then there is some δ such that $0 < |x-a| < \delta$ implies $g(x) \geq M/2$. If $M < 0$, then there is some δ such that $0 < |x-a| < \delta$ implies $g(x) \leq M/2$. In particular, if $M \neq 0$ then there is some δ such that $0 < |x-a| < \delta$ implies $|g(x)| \geq |M|/2$ and $g(x) \neq 0$.*

Proof of claim: Suppose $M > 0$. Applying the definition of $\lim_{x\to a} g(x) = M$ with $\varepsilon = M/2$ we find that there is some δ such that $0 < |x-a| < \delta$ implies $|g(x) - M| < M/2$, which in turn implies $g(x) \geq M/2$. On the other hand, if $M < 0$, then applying the definition of $\lim_{x\to a} g(x) = M$ with $\varepsilon = -M/2$ we find that there is some δ such that $0 < |x-a| < \delta$ implies $|g(x) - M| < -M/2$, which in turn implies $g(x) \leq -M/2$. □

We have established that $1/g$ is defined near a, and in fact that if $M > 0$ then g is positive near a, while if $M < 0$ then g is negative near a. We next argue that $\lim_{x\to a}(1/g)(x) = 1/M$. Given $\varepsilon > 0$, choose $\delta_1 > 0$ such that $0 < |x-a| < \delta_1$ implies $|g(x)| \geq |M|/2$ (which we can do by the claim). We have

$$\begin{aligned}
\left|\left(\frac{1}{g}\right)(x) - \frac{1}{M}\right| &= \left|\frac{1}{g(x)} - \frac{1}{M}\right| \\
&= \left|\frac{M - g(x)}{Mg(x)}\right| \\
&= \frac{|g(x) - M|}{|M||g(x)|} \\
&\leq \frac{2}{|M|^2}|g(x) - M|.
\end{aligned}$$

We would *like* to make $|(1/g)(x) - (1/M)| < \varepsilon$. One way to do this is to force $(2/|M|)^2|g(x) - M|$ to be smaller than ε, that is, to force $|g(x) - M|$ to be smaller than $(|M|^2\varepsilon)/2$.

Since $g \to M$ as $x \to a$, and since $(|M|^2\varepsilon)/2 > 0$, there is a $\delta_2 > 0$ such that $0 < |x-a| < \delta_2$ indeed implies $|g(x) - M| < (|M|^2\varepsilon)/2$.

So, if we let δ be the smaller of δ_1 and δ_2 then $0 < |x-a| < \delta$ implies $|(1/g)(x) - 1/M| < \varepsilon$, so that indeed $\lim_{x\to a}(1/g)(x) = 1/M$. □

An obvious corollary of the above is the following, which we give a proof of as a prototype of proofs of this kind.

Corollary 6.3. *For each $n \geq 1$, let $f_1, \ldots, f_n$ be functions all defined near some a. Suppose that $\lim_{x\to a} f_i(x) = L_i$ for each $i \in \{1, \ldots, n\}$. Then*

- *$\lim_{x\to a}(f_1 + \cdots + f_n)(x)$ exists and equals $L_1 + \cdots + L_n$.*

Proof: We proceed by induction on n, with the base case $n = 1$ trivial (it asserts that if $\lim_{x\to a} f_1(x) = L_1$ then $\lim_{x\to a} f_1(x) = L_1$).

For the induction step, suppose the result is true for some $n \geq 1$, and that we are given $n+1$ functions $f_1, \ldots, f_{n+1}$, all defined near a, with $f_i \to L_i$ near a for each i. We have

$$\lim_{x\to a}(f_1 + \cdots + f_n)(x) = L_1 + \cdots + L_n$$

by the induction hypothesis, and $\lim_{x\to a} f_{n+1}(x) = L_{n+1}$ by hypothesis of the corollary. By the sum/product/reciprocal theorem, we have that $\lim_{x\to a}((f_1 + \cdots + f_n) + f_{n+1})(x)$ exists and equals $(L_1 + \cdots + L_n) + L_{n+1}$; but since $((f_1 + \cdots + f_n) + f_{n+1})(x) = (f_1 + \cdots + f_{n+1})(x)$ and $(L_1 + \cdots + L_n) + L_{n+1} = L_1 + \cdots + L_n + L_{n+1}$, this immediately says that

$$\lim_{x\to a}(f_1 + \cdots + f_{n+1})(x) = L_1 + \cdots + L_{n+1}.$$

The corollary is proven, by induction.[76] □

We may similarly prove that for each $n \geq 1$, if $f_1, \ldots, f_n$ are functions all defined near some a, and if $\lim_{x\to a} f_i(x) = L_i$ for each $i \in \{1, \ldots, n\}$, then

- $\lim_{x\to a}(f_1 \cdot \cdots \cdot f_n)(x)$ exists and equals $L_1 \cdot \cdots \cdot L_n$.

This has an important consequence. Starting from the basic results that for any a, c, $\lim_{x\to a} c = c$ and $\lim_{x\to a} x = a$, by repeated applications of the sum/product/reciprocal theorem, together with its corollaries, we obtain the following important labor-saving results:

- Suppose that P is a polynomial. Then for any a, $\lim_{x\to a} P(x)$ exists and equals $P(a)$.
- Suppose that R is a rational function, say $R = P/Q$ where P, Q are polynomials. If a is in the domain of R, that is, if $Q(a) \neq 0$, then $\lim_{x\to a} R(x)$ exists and equals $R(a)$, that is,
$$\lim_{x\to a} \frac{P(x)}{Q(x)} = \frac{P(a)}{Q(a)}.$$

For example, we can immediately say

$$\lim_{x\to 1} \frac{2x^2 - 4x}{x^3 - 8} = \frac{2(1)^2 - 4(1)}{(1)^3 - 8} = -\frac{2}{7},$$

piggy-backing off our general theorems, and avoiding a nasty derivation from first principles.

What about $\lim_{x\to 2}(2x^2 - 4x)/(x^3 - 8)$? Here a direct evaluation is not possible, because 2 is not in the domain of $(2x^2 - 4x)/(x^3 - 8)$. But because 2 is not in the domain, we can algebraic manipulate $(2x^2 - 4x)/(x^3 - 8)$ by dividing above and below the line by $x - 2$ — this operation is valid exactly when $x \neq 2$! Formally we can say

$$\frac{2x^2 - 4x}{x^3 - 8} = \frac{2x(x-2)}{(x-2)(x^2 + 2x + 4)} = \frac{2x}{x^2 + 2x + 4},$$

valid on the entire domain of $(2x^2 - 4x)/(x^3 - 8)$. So

$$\lim_{x\to 2} \frac{2x^2 - 4x}{x^3 - 8} = \lim_{x\to 2} \frac{2x}{x^2 + 2x + 4} = \frac{4}{14} = \frac{2}{7}.$$

[76] Notice that in the induction step, dealing with deducing $p(n+1)$ from $p(n)$, we needed to invoke the $n = 2$ case. This occurs frequently when extending a result concern two objects to the obvious analog result concerning many objects. Examples include the general distributive law, and the general triangle inequality.

One last note on the limit. We have been implicitly assuming throughout all of this section that if f approaches a limit L near a, then L is the *only* limit that it approaches. We can easily prove this.

Claim 6.4. *Suppose f is defined near a and that $\lim_{x\to a} f(x) = L$, and also $\lim_{x\to a} f(x) = M$. Then $L = M$.*

Proof: Suppose for a contradiction that $L \neq M$. Assume, without any loss of generality[77], that $L > M$. Set $\varepsilon = (L - M)/4$. There is a $\delta > 0$ such that $0 < |x - a| < \delta$ implies $|f(x) - L| < \varepsilon$ and $|f(x) - M| < \varepsilon$. The first of these inequalities says that $f(x) > L - \varepsilon$, and the second says $f(x) < M + \varepsilon$, so together they imply that $L - \varepsilon < M + \varepsilon$, or $L - M < 2\varepsilon$, or $(L - M)/4 < \varepsilon/2$, or $\varepsilon < \varepsilon/2$, a contradiction. We conclude that $L = M$. □

6.4 Non-existence of limits

What does it mean for a function f *not* to tend to a limit L near a? For a function f to tend to a limit L near a, two things must happen:

1. f must be defined near a, and
2. for all $\varepsilon > 0$ there is $\delta > 0$ such that for all x, if $0 < |x - a| < \delta$ then $|f(x) - L| < \varepsilon$.

So for f not to tend to L, *either* the first clause above fails, so f is not defined near a, or the second clause fails. To understand what it means for the second clause to fail, it's helpful to write it symbolically, and then use the methods we have discussed earlier to negate it. The clause is

$$(\forall\varepsilon)(\exists\delta)(\forall x)((0 < |x - a| < \delta) \Rightarrow (|f(x) - L| < \varepsilon))$$[78]

and its negation is

$$(\exists\varepsilon)(\forall\delta)(\exists x)((0 < |x - a| < \delta) \wedge (|f(x) - L| \geq \varepsilon)).$$

So, unpacking all this, we get:

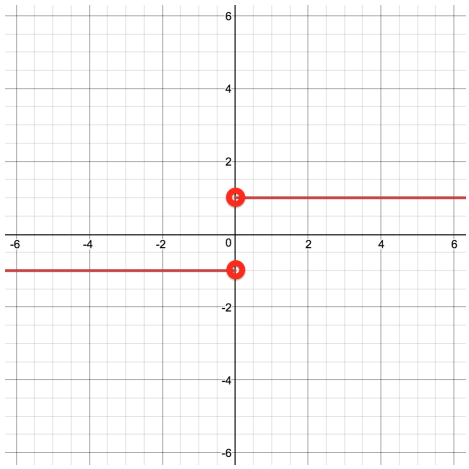
> **Definition of a function not tending to a limit L near a**: f does not approach the limit L near a if either
>
> - f is not defined near a (meaning, in any open interval that includes a, there are points that are not in the domain of f)
>
> or

[77] A handy phrase, but one to be used only when you are really saying that no generality is lost.

[78] Notice that we have included the quantification $\forall x$. Without this, the clause would be a predicate (depending on the variable x), rather than a statement.

- there's an $\varepsilon > 0$ (a window of tolerance around L presented by Bob) such that
- for all $\delta > 0$ (no matter what window of tolerance around a that Alice responds with)
- there is an x with $0 < |x - a| < \delta$ (an $x \neq a$ that is within δ of a)
- but with $|f(x) - L| \geq \varepsilon$ ($f(x)$ is at least ε away from L).

As an example, consider the function f_6 defined previously, that is given by $f_6(x) = x/|x|$.



It's seems quite clear that f_6 does not approach a limit near 0; the function gets close to both 1 and -1 in the vicinity of 0, so there isn't a single number that the function gets close to (and we know that if the limit exists, it is unique).

We use the definition just given of a function *not* tending to a limit, to verify that $\lim_{x\to 0} f_6(x) \neq 3/4$. Take $\varepsilon = 1/10$ (this is fairly arbitrary). Now consider any $\delta > 0$. We need to show that there is an $x \neq 0$, in the interval $(-\delta, \delta)$, with $|f_6(x) - 3/4| \geq 1/10$. There are many such x's that work. For example, consider $x = \delta/2$; for this choice of x, $|f_6(x) - 3/4| = |(\delta/2)/(|\delta/2|) - 3/4| = |1 - 3/4| = |1/4| = 1/4 \geq 1/10$[79]

Why did we choose $\varepsilon = 1/10$? We intuited that output values of f_6 could be made arbitrarily close to 1 by cherry-picking values of x close to 0. So to show that values of the output can't be made *always* arbitrarily close to 3/4 by choosing values of the input close enough to 0, we choose an ε so that the interval $(3/4 - \varepsilon, 3/4 + \varepsilon)$ did not get too close to 1 — that allowed us to choose an x close to 0 for which $f_6(x)$ was not close to 3/4. Any ε less that 1/4 would have worked.[80]

More generally, what does it mean for f not to tend to *any* limit near a? It means that for every L, f does not tend to limit L near a.

[79] We could have equally well picked $x = -\delta/2$; then $|f_6(x) - 3/4| = 7/4 \geq 1/10$.

[80] In fact, any ε less than 11/4 would have worked — we could have noticed that output values of f_6 could be made arbitrarily close to -1 by cherry-picking values of x close to 0.

Definition of a function not tending to a limit near a: f does not approach a limit near a if for *every* L it is the case that f does not approach the limit L near a.

Going back to our previous example: we claim that $\lim_{x\to 0} f_6(x)$ does not exist. Indeed, suppose that L is given, and proposed as a (the) limit. We want to find an $\varepsilon > 0$ such that for any $\delta > 0$, we can find at least one value of $x \neq 0$ that is within δ of 0, but that $f_6(x)$ is not within ε of L. We notice that by cherry-picking values of x arbitrarily close to 0, we can get $f_6(x)$ arbitrarily close to *both* -1 *and* to 1. This suggests the following strategy:

- If $L \geq 0$: take $\varepsilon = 1/2$. Given $\delta > 0$, consider $x = -\delta/2$. That's certainly within δ of 0 (and is certainly not equal to 0). But $f_6(x) = -1$, so $f_6(x)$ is distance at least 1 from L, and so not distance less than $1/2$.

- If $L < 0$: again take $\varepsilon = 1/2$. Given $\delta > 0$, consider $x = \delta/2$. It's non-zero and within δ of 0, but $f_6(x) = 1$, so $f_6(x)$ is distance more than 1 from L, and so not distance less than $1/2$.

One more example: we claim that $\lim_{x\to 0} |\sin(1/x)|$ does not exist. The intuition behind this is the same as for the previous example: by cherry picking values of x, we can get $\sin(1/x)$ to take the value 1, arbitrarily close to 0, and we can get it to take the value 0. Specifically, $|\sin(1/x)|$ takes the value 1 at $1/x = \pm\pi/2, \pm 3\pi/2, \pm 5\pi/2, \ldots$, so at $x = \pm 2/\pi, \pm 2/3\pi, \pm 2/5\pi, \ldots$, or more succinctly at $x = \pm 2/((2n+1)\pi)$, $n = 0, 1, 2, 3, \ldots$; and $|\sin(1/x)|$ takes the value 0 at $1/x = \pm\pi, \pm 2\pi, \pm 3\pi, \ldots$, so at $x = \pm 1/(n\pi)$, $n = 0, 1, 2, 3, \ldots$. So, given L (a proposed limit for $|\sin(1/x)|$ near 0), we can again treat two cases, depending on whether L is far from 0 or far from 1.

- If $L \geq 1/2$: take $\varepsilon = 1/4$. Given $\delta > 0$, there is some n large enough that $x := 1/(n\pi)$ is in the interval $(-\delta, \delta)$[81] (and is non-zero). For this x, $|\sin(1/x)| = 0$, which is *not* in the interval $(L - 1/4, L + 1/4)$.

- If $L < 1/2$: again take $\varepsilon = 1/4$. Given $\delta > 0$, there is some n large enough that $x := 2/((2n+1)\pi)$ is in the interval $(-\delta, \delta)$ (and is non-zero). For this x, $|\sin(1/x)| = 1$, which is *not* in the interval $(L - 1/4, L + 1/4)$.

We conclude that $\lim_{x\to 0} |\sin(1/x)|$ does not exist.

In the homework, you'll deal with another situation where a limit doesn't exist: where the output values don't approach a specific value, because they get arbitrarily large in magnitude near the input. We'll return to these "infinite limits" later.

One last comment for the moment about limits not existing: while $\lim_{x\to 0} |\sin(1/x)|$ does not exist, the superficially similar $\lim_{x\to 0} x|\sin(1/x)|$ does, and it's easy to prove that it takes the value 0. Indeed, given $\varepsilon > 0$, take $\delta = \varepsilon$. If $0 < |x| < \delta$ then $|x|\sin(1/x)|| \leq |x| < \delta = \varepsilon$,

[81] Is there???

so the limit is 0. This illustrates that while oftentimes computing limits directly from the definition is a slog, it can sometimes be surprisingly easy.

There's a general phenomenon that this last example — $f(x) = x|\sin(1/x)|$ near 0 — is a special case of. The function $f(x) = x|\sin(1/x)|$ is "squeezed" between two other functions that are quite easy to understand. If g_ℓ, g_u are defined by

$$g_\ell(x) = \begin{cases} x & \text{if } x < 0 \\ 0 & \text{if } x \geq 0 \end{cases}$$

and

$$g_u(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$$

then we easily have that

$$g_\ell(x) \leq f(x) \leq g_u(x)$$

for all real x. Indeed, for $x \geq 0$ we have, using $0 \leq |\sin(1/x)| \geq 1$, that $0 \leq x|\sin(1/x)| \leq x$, while if $x < 0$ then $0 \leq |\sin(1/x)| \geq 1$ implies $0 \geq x|\sin(1/x)| \geq x$ or $x \leq x|\sin(1/x)| \leq 0$; and these two inequalities together say that $g_\ell(x) \leq f(x) \leq g_u(x)$.

We also have that $g_\ell \to 0$ near 0, and that $g_u \to 0$ near 0. We verify the first of these now (the second is left as an exercise). Given $\varepsilon > 0$ we seek $\delta > 0$ so that $x \in (-\delta, \delta)$ (and $x \neq 0$) implies $g_\ell(x) \in (-\varepsilon, \varepsilon)$. Consider $\delta = \varepsilon$. If non-zero x is in $(-\delta, \delta)$ and is negative, then $g_\ell(x) = x \in (-\delta, \delta) = (-\varepsilon, \varepsilon)$, while if it is positive then $g_\ell(x) = 0 \in (-\delta, \delta) = (-\varepsilon, \varepsilon)$. This shows that $g_\ell \to 0$ near 0.

If both g_ℓ and g_u are approaching 0 near 0, and f is sandwiched between g_ell and g_u, then it should come as no surprise that f is *forced* to approach 0 (the common limit of its upper and lower bounds) near 0. The general phenomenon that this example illustrates is referred to as a *squeeze theorem.*

Theorem 6.5. *(Squeeze theorem) Let f, g, h be three functions, and let a be some real number. Suppose that $f, g,$ are all defined near a, that is, that there is some number $\Delta > 0$ such that on the interval $(a - \Delta, a + \Delta)$ it holds that $f(x) \leq g(x) \leq h(x)$ (except possibly at a, which might or might not be in the domains of any of the three functions). Suppose further that $\lim_{x\to a} f(x)$ and $\lim_{x\to a} h(x)$ both exist and both equal L. Then $\lim_{x\to a} g(x)$ exists and equals L.*

You will be asked for a proof of this in the homework.

6.5 One-sided limits

When discussing the squeeze theorem we saw the function

$$g_u(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0, \end{cases}$$

defined by cases, with different behavior to the right and left of 0 on the number line. When establishing $\lim_{x\to 0} g_u(x)$ we need to consider separately what happens for positive x and negative x. This strongly suggests that there could be some value in a refinement of the definition of limit, that considers separately what happens for x values that are larger a, and smaller than a. The natural refinement is referred to as a *one-sided limit*.

Definition of f approaching L near a *from the right* or *from above*: A function f approaches a limit L from the right near a from the right (or from above)[82] if

- f is defined near a, to the right, meaning that there is some $\delta > 0$ such that all of $(a, a + \Delta)$ is in the domain of f,

and

- for all $\varepsilon > 0$ there is $\delta > 0$ such that $0 < x - a < \delta$ implies $|f(x) - L| < \varepsilon$; that is, whenever x is within δ of a, and x is greater than a ("above" a in magnitude, "to the right of" a on the number line), then $f(x)$ is within ε of L.

We write

- $\lim_{x\to a^+} f(x) = L$, or $\lim_{x\searrow a} f(x) = L$
- $f \to L$ (or $f(x) \to L$) as $x \to a^+$ (or as $x \searrow a$).

Definition of f approaching L near a *from the left* or *from below*: A function f approaches a limit L near a from the left (or from below) if

- f is defined near a, to the left, meaning there is $\delta > 0$ with $(a - \Delta, a)$ in the domain of f,

and

- for all $\varepsilon > 0$ there is $\delta > 0$ such that $-\delta < x - a < 0$ implies $|f(x) - L| < \varepsilon$; that is, whenever x is within δ of a, and x is less than a ("below" a in magnitude, "to the left of" a on the number line), then $f(x)$ is within ε of L.

We write

- $\lim_{x\to a^-} f(x) = L$, or $\lim_{x\nearrow a} f(x) = L$
- $f \to L$ (or $f(x) \to L$) as $x \to a^-$ (or as $x \nearrow a$).

As an example consider the familiar old function $f_6(x) = x/|x|$. We know that $\lim_{x\to 0} f_6(x)$ does not exist. But this coarse statement seems to miss something about f_6 — that the function seems to approach limit 1 near 0, if we are only looking at positive inputs, and seems to approach limit -1 near 0, if we are only looking at negative inputs.

[82] Note well: as you'll see from the definition, it is a that is being approached from above, not L

The notion of one-sided limits just introduced captures this. We claim that $\lim_{x\to 0^+} f_6(x)$ exists, and equals 1. Indeed, given $\varepsilon > 0$, take $\delta = 1$. if $0 < x - 0 < \delta$ then $x > 0$ so $f_6(x) = 1$, and so in particular $|f_6(x) - 1| = 0 < \varepsilon$. Similarly, it's easy to show $\lim_{x\to 0^-} f_6(x) = -1$.

This example shows that both the one-sided limits can exist, while the limit may not exist. It's also possible for one one-sided limit to exist, but not the other (consider the function which takes value $\sin(1/x)$ for positive x, and 0 for negative x, near 0), or for both not to exist (consider $\sin(1/x)$ near 0). So, in summary, if the limit doesn't exist, then at least three things can happen with the one-sided limits:

- both exist, but take different values,
- one exists, the other doesn't, or
- neither exists.

There's a fourth possibility, that both one-sided limits exist and take the same value. But that *can't* happen when the limit does not exist, as we are about to see; and as we are also about to see, if the limit exists then there is one one possibility for the two one-sided limits, namely that they both exist and are equal.

Theorem 6.6. *For a f be a function defined near a, $\lim_{x\to a} f(x)$ exists and equals L if and only if both of $\lim_{x\to a+} f(x)$, $\lim_{x\to a-} f(x)$ exist and equal L.*

Proof: Suppose $\lim_{x\to a} f(x)$ exists and equals L. Let $\varepsilon > 0$ be given. There is $\delta > 0$ such that $0 < |x - a| < \delta$ implies $|f(x) - L| < \varepsilon$. In particular that means that $0 < x - a < \delta$ implies $|f(x) - L| < \varepsilon$, so that $\lim_{x\to a+} f(x)$ exists and equal L, and $-\delta < x - a < 0$ implies $|f(x) - L| < \varepsilon$, so that $\lim_{x\to a-} f(x)$ exists and equal L.

Conversely, both of $\lim_{x\to a+} f(x)$, $\lim_{x\to a-} f(x)$ exist and equal L. Given $\varepsilon > 0$ there is $\delta_1 > 0$ such that $0 < x - a < \delta_1$ implies $|f(x) - L| < \varepsilon$, and there is $\delta_2 > 0$ such that $-\delta_2 < x - a < 0$ implies $|f(x) - L| < \varepsilon$. If $\delta = \min\{\delta_1, \delta_2\}$ then $0 < |x - a| < \delta$ implies that either $0 < x - a < \delta \leq \delta_1$, or $-\delta_2 \leq -\delta < x - a < 0$. In either case $|f(x) - L| < \varepsilon$, so $\lim_{x\to a} f(x)$ exists and equal L. □

6.6 Infinite limits, and limits at infinity

A minor deficiency of the real numbers, is the lack of an "infinite" number. The need for such a number can be seen from a very simple example. We have that

$$\lim_{x\to 0} \frac{1}{x^2} \text{ does not exist,}$$

but not because the expression $1/x^2$ behaves wildly near 0. On the contrary, it behaves very predictably: the closer x gets to zero, from either the positive or the negative side, the larger (more positive) $1/x^2$ gets, without bound. It would be helpful to have an "infinite" number,

one that is larger than all positive numbers; such a number would be an ideal candidate for the limit of $1/x^2$ near 0.

There is no such real number. But it useful to introduce a symbol that can be used to encode the behavior of expressions like $\lim_{x\to 0} 1/x^2$.

Definition of an infinite limit Say that f *approaches the limit infinity*, or *plus infinity*, near a, denoted

$$\lim_{x\to a} f(x) = \infty$$[83]

(or sometimes $\lim_{x\to a} f(x) = +\infty$) if f is defined near a, and if

- for all real numbers M
- there is $\delta > 0$
- such that for all real x,

$$0 < |x - a| < \delta \quad \text{implies} \quad f(x) > M.$$[84]

Similarly, say that f *approaches the limit minus infinity* near a, denoted

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

if f is defined near a, and if for all real numbers M there is $\delta > 0$ such that for all real x,

$$0 < |x - a| < \delta \quad \text{implies} \quad f(x) < M.$$

Before doing an example, we make the following labor-saving observation. Suppose that we are trying to show $\lim_{x\to a} f(x) = \infty$, and that, for some M_0, we have found $\delta_0 > 0$ such that $0 < |x-a| < \delta_0$ implies $f(x) > M_0$. Then for *any* $M \leq M_0$ we have that $0 < |x-a| < \delta_0$ implies $f(x) > M$. The consequence of this is that in attempting to prove $\lim_{x\to a} f(x) = \infty$, we can start by picking an arbitrary real M_0, and then only attempt to verify the condition in the definition for $M \geq M_0$; this is enough to establish the limit statement. Often in practice, this observation is employed by assuming that $M > 0$, which assumption allows us to divide or multiply an inequality by M without either flipping the direction of the inequality, or having to worry about dividing by 0.

A similar observation can be made about showing $\lim_{x\to a} f(x) = -\infty$ (we need only verify the condition for all $M \leq M_0$; in practice this is often $M < 0$), and analogous observations can be made for establishing one-sided infinite limits (see below).

[83] The symbol "∞" here is just that — a *symbol*. It is not, **not**, a *number*. It has *no* place in any arithmetic calculation involving real numbers!

[84] Note that this is saying that $f(x)$ can be forced to be arbitrarily large and positive, by taking values of x sufficiently close to a.

Now we move on to an example, $\lim_{x\to 0} 1/x^2$. We claim that this limit is plus infinity. Indeed, let $M > 0$ be given. We would like to exhibit a $\delta > 0$ such that $0 < |x| < \delta$ implies $1/x^2 > M$. Now because x and M are both positive, we have that

$$1/x^2 > M \text{ is equivalent to } x^2 < 1/M, \text{ which is equivalent to } |x| < 1/\sqrt{M}.$$

So we may simply take $\delta = 1/\sqrt{M}$ (which is positive).

As with the ordinary limit definition, it is sometimes very helpful to be able to consider separately what happens as we approach a from each of the two possible sides.

Definitions of one-sided infinite limits Say that f approaches the limit (plus) infinity near a *from above*, or *from the right*, denoted

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

if f is defined near a from above (in some interval $(a, a+\delta)$, $\delta > 0$), and if

- for all real numbers $M > M_0$[85]
- there is $\delta > 0$
- such that for all real x,

$$0 < x - a < \delta \quad \text{implies} \quad f(x) > M.$$

To get the definition of f approaching the limit minus infinity near a from above ($\lim_{x\to a^+} f(x) = -\infty$), change "$M > M_0$" and "$f(x) > M$" above to "$M < M_0$" and "$f(x) < M$".

To get the definition of f approaching the limit plus infinity near a *from below*, or *from the left* ($\lim_{x\to a^-} f(x) = (+)\infty$), change "$0 < x - a < \delta$" above to "$-\delta < x - a < 0$".

To get the definition of f approaching the limit minus infinity near a from below ($\lim_{x\to a^-} f(x) = -\infty$), change "$M > M_0$", "$f(x) > M$" and "$0 < x - a < \delta$" above to "$M < M_0$", "$f(x) < M$" and "$-\delta < x - a < 0$".

As an example, we verify formally the intuitively clear result that

$$\lim_{x\to 1^-} \frac{1}{x-1} = -\infty.$$

Given $M < 0$, we seek $\delta > 0$ such that $x \in (1-\delta, 1)$ implies $1/(x-1) < M$. Now for $x < 1$ we have $x - 1 < 0$, so in this range $1/(x-1) < M$ is equivalent to $1 > M(x-1)$, and for $M < 0$ this is in turn equivalent to $1/M < x - 1$, or $x > 1 + 1/M$. From this it is clear that if we take $\delta = -1/M$ (note that this is positive, since we are assuming $M < 0$[86]), then $x \in (1-\delta, 1)$ indeed implies $1/(x-1) < M$.

[85] As observed after the definition of an infinite limit, this M_0 can be completely arbitrary.

[86] Without this (valid) assumption, the limit calculation would be rather more awkward.

As well as infinite limits, a very natural notion that slightly generalizes our concept of a limit is that of a "limit at infinity", capturing the behavior of a function as the input grows unboundedly large in magnitude, either positively or negatively.

Definition of a function approaching a limit at infinity Suppose that f is *defined near infinity* (or *near plus infinity*), meaning that there is some real number M such that f is defined at every point in the interval (M, ∞). Say that f *approaches the limit* L *near infinity* (or *near plus infinity*), denoted

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

if

- for all $\varepsilon > 0$
- there is a real number M
- such that for all x,

$$x > M \text{ implies } |f(x) - L| < \varepsilon.$$

Formulating precise definitions of

- $\lim_{x \to -\infty} = L$
- $\lim_{x \to \infty} = \infty$
- $\lim_{x \to \infty} = -\infty$
- $\lim_{x \to -\infty} = \infty$ and
- $\lim_{x \to -\infty} = -\infty$

are left as an exercise.

Here's an example. We claim that $\lim_{x \to \infty} \frac{x}{x+1} = 1$. To prove this entails showing that for all $\varepsilon > 0$ there is an M such that $x > M$ implies $x/(x+1) \in (1-\varepsilon, 1+\varepsilon)$. Let us initially commit to choosing that $M \geq -1$, so that for $x > M$ we have $x + 1 > 0$, and we do not run into any issues with attempting to divide by 0.

Now for all x we have $x < x + 1$, and so for those x satisfying $x + 1 > 0$ we have $x/(x+1) < 1$; so our goal is to ensure $x/(x+1) > 1 - \varepsilon$. But (again remembering that $1 + x > 0$, and also using $\varepsilon > 0$) we have that $x/(x+1) > 1 - \varepsilon$ is equivalent to $x > (1/\varepsilon) - 1$. So if we take M to be anything that is at least as large as both -1 and $(1/\varepsilon) - 1$, for example, $M = \max\{-1, (1/\varepsilon) - 1\}$, then $x > M$ implies $x/(x+1) \in (1-\varepsilon, 1+\varepsilon)$, as required[87].

Here are some general facts about limits at infinity, all of which you should be able to prove, as the proofs are very similar to related statements about ordinary limits (limits near a finite number).

[87] In fact, for $\varepsilon > 0$ it holds that $(1/\varepsilon) - 1 > -1$, so we could have simple said "take $M = (1/\varepsilon) - 1$".

Theorem 6.7.

- *If* $\lim_{x\to\infty} f(x) = L$ *and* $\lim_{x\to\infty} g(x) = M$*, then*
  - $\lim_{x\to\infty}(f+g)(x) = L + M$*;*
  - $\lim_{x\to\infty}(fg)(x) = LM$*;*
  - $\lim_{x\to\infty} cf(x) = cL$*; and*
  - $\lim_{x\to\infty}(f/g)(x) = L/M$*, provided* $M \neq 0$*.*
- *For* $n \in \mathbb{N} \cup \{0\}$*,*
  - $\lim_{x\to\infty} x^n = \begin{cases} 1 & \text{if } n = 0 \\ \infty & \text{if } n > 0 \end{cases}$ *and*
  - $\lim_{x\to\infty} \frac{1}{x^n} = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{if } n > 0. \end{cases}$
- *Suppose* $p(x)$ *is the polynomial* $p(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_1x + a_0$*, and* $q(x)$ *is the polynomial* $q(x) = x^m + b_{m-1}x^{m-1} + \ldots + b_1x + b_0$[88] *(*$n, m \geq 0$*). Then*

$$\lim_{x\to\infty} \frac{p(x)}{q(x)} = \begin{cases} 1 & \text{if } n = m \\ \infty & \text{if } n > m \\ 0 & \text{if } n < m. \end{cases}$$

Proof: We'll just prove two of the statements above, leaving the rest as exercises. First, suppose $\lim_{x\to\infty} f(x) = L$ and $\lim_{x\to\infty} g(x) = M$. We will consider $\lim_{x\to\infty}(fg)(x)$, and show that it equals LM. We have

$$\begin{aligned} |(fg)(x) - LM| &= |f(x)g(x) - Lg(x) + Lg(x) - LM| \\ &\leq |f(x) - L||g(x)| + |L||g(x) - M|. \end{aligned}$$

Since $\lim_{x\to\infty} g(x) = M$ we know that there is X_1[89] such that $x > X_1$ implies $g(x) \in (M-1, M+1)$, so $|g(x)| \leq |M|+1$. Now let $\varepsilon > 0$ be given. Since $\lim_{x\to\infty} f(x) = L$ we know that there is X_2 such that $x > X_2$ implies $|f(x) - L| < \varepsilon/(|M|+1)$. Since $\lim_{x\to\infty} g(x) = M$ we know that there is X_3 such that $x > X_3$ implies $|g(x) - M| < \varepsilon/(|L|+1)$[90]. It follows that if $x > \max\{X_1, X_2, X_3\}$ then

$$\begin{aligned} |(fg)(x) - LM| &\leq |f(x) - L||g(x)| + |L||g(x) - M| \\ &\leq |f(x) - L|(|M|+1) + (|L|+1)|g(x) - M| \\ &< \varepsilon/2 + \varepsilon/2 \\ &= \varepsilon, \end{aligned}$$

[88] This corollary of the previous parts could have been formulated for more general polynomials, with arbitrary (positive of negative) leading coefficients; but the statement would be messy, and in any case by pulling out an appropriate constant, the ratio of two arbitrary polynomials can always be reduced to the form presented above.

[89] We have to change notation slightly from the definition, since M is now being used for something else.

[90] We bound by $\varepsilon/(|L|+1)$ here, rather than $\varepsilon/|L|$, to avoid the possibility of dividing by 0

so $\lim_{x\to\infty}(fg)(x) = LM$, as claimed.

Let's also prove $\lim_{x\to\infty} x^n = \infty$ if $n > 0$. We haven't formulated the relevant definition, but of course what this must mean is that for all M (and, if we wish, we can take this M to be positive, or bigger than any fixed constant M_0) there is an N such that $x > N$ implies $x^n > M$.

Let's commit to only considering $M \geq 1$. If we take $N = M$, then $x > N$ implies $x > M$, which in turn implies (because $M \geq 1$) that $x^n > M$, and we have the required limit. □

Returning to the previous example, $\lim_{x\to\infty} \frac{x}{x+1}$: that the limit exists and is 1 follows easily, from the above theorem. Formulating an analogous result for limits near minus infinity is left as an exercise.[91]

[91] For plenty of exercises on the kinds of limits introduced in this section, see Spivak, Chapter 5, questions 32-41.

7 Continuity

Looking back at the ten functions that we used at the beginning of Section 6 to motivate the definition of the limit, we see that

- some of them — f_2, f_3, f_6, f_7, f_8, f_9 and f_{10} — did not approach a limit near the particular a's under consideration,
- while the rest of them — f_1, f_4 and f_5 — did.

These last three are definitely "nicer" near the particular a's under consideration than the first seven. But even among these last three, there is a further split:

- two of them — f_4 and f_5 — either have the property that the function is not defined *at* a, or that the function is defined, but the function value at a is different from the limit that the function is approaching near a,
- while the third — f_1 — has the function defined at a, *and* the function value equally the limit that the function is approaching near a.

This last is definitely "very nice" behavior near a; we capture precisely what's going on with the central definition of this section, that of continuity of a function at a point.

Definition of f being continuous at a A function f is *continuous* at a if

- f is defined at and near a (meaning there is $\Delta > 0$ such that all of $(a - \Delta, a + \Delta)$ is in Domain(f)), and
- $\lim_{x \to a} f(x) = f(a)$.

The sense of the definition is that near a, small changes in the input to f lead to only small changes in the output, or (quite informally), "near a, the graph of f can be drawn with taking pen off paper".

Unpacking the ε-δ definition of the limit, the continuity of a function f (that is defined at and near a) at a can be expressed as follows:

- for all $\varepsilon > 0$
- there is $\delta > 0$
- such that $|x - a| < \delta$
- implies $|f(x) - f(a)| < \varepsilon$.

Note that this is just the definition of $\lim_{x \to a} f(x) = f(a)$ with $0 < |x - a| < \delta$ changed to just $|x - a| < \delta$, or $x \in (a - \delta, a + \delta)$; we can make this change because at the one new value of x that is introduced into consideration, namely $x = a$, we certainly have $|f(x) - f(a)| < \varepsilon$ for all $\varepsilon > 0$, since in fact we have $|f(x) - f(a)| = 0$ at $x = a$. This ε-δ statement is often taken as the definition of continuity of f at a.

7.1 A collection of continuous functions

Here we build up a a large collection of functions that are continuous at all points of their domains. We have done most of the work for this already, when we discussed limits.

Constant function Let $f : \mathbb{R} \to \mathbb{R}$ be the constant function $f(x) = c$ (where $c \in \mathbb{R}$ is some constant). Since we have already established that $\lim_{x\to a} f(x) = c = f(a)$ for all a, we immediately get that f is continuous at all points in its domain.

Linear function Let $g : \mathbb{R} \to \mathbb{R}$ be the linear function $g(x) = x$. Since we have already established that $\lim_{x\to a} g(x) = a = g(a)$ for all a, we immediately get that g is continuous at all points in its domain.

Sums, products and quotients of continuous functions Suppose that f and g are both continuous at a. Then

- $f + g$ is continuous at a (proof: $f + g$ is certainly defined at and near a, if both f and g are, and by the sum/product/reciprocal theorem for limits,

$$\lim_{x\to a}(f + g)(x) = \lim_{x\to a} f(x) + \lim_{x\to a} g(x) = f(a) + g(a) = (f + g)(a));$$

- fg is continuous at a (proof: fg is certainly defined at and near a, if both f and g are, and by the sum/product/reciprocal theorem for limits,

$$\lim_{x\to a}(fg)(x) = \lim_{x\to a} f(x) \lim_{x\to a} g(x) = f(a)g(a) = (fg)(a));$$

- as long as $g(a) \neq 0$, $1/g$ is continuous at a (proof: that $1/g$ is defined at and near a follows from Claim 6.2, and for the limit part of the continuity definition, we have from the reciprocal part of sum/product/reciprocal theorem for limits that

$$\lim_{x\to a}(1/g)(x) = 1/\lim_{x\to a} g(x) = 1/g(a) = (1/g)(a));$$

- as long as $g(a) \neq 0$, f/g is continuous at a (proof: combine the last two parts).

Polynomials If P is a polynomial function, then P is continuous at all reals. For any *particular* polynomial, this follows by lots of applications of the the observations above about sums and products of continuous functions, together with the continuity of the constant and linear functions (to get things started); for polynomials *in general* this follows from the same ingredients as for the particular case, together with lots of applications of prove by induction.

Rational functions If R is a rational function, then R is continuous at all points in its domain; so in particular, if $R = P/Q$ where P, Q are polynomials and Q is not the constantly zero polynomial, then R is continuous at all reals x for which $Q(x)$ is not 0. This is an application of the continuity of polynomials, as well as the reciprocal part of the sum/product/reciprocal observation.

This gives us already a large collection of continuous functions. The list becomes even larger when we include the trigonometric functions:

Working assumption The functions sin and cos are continuous at all reals.[92]

This is a reasonable assumption; if we move only slightly along the unit circle from a point $(x, y) = (\cos\theta, \sin\theta)$, the coordinates of our position only move slightly, strongly suggesting that sin and cos are both continuous.

Armed with this working assumption, we can for example immediately say (appealing to our previous observations) that

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

is continuous, as long as $x \neq -1$ or $x \neq n\pi$ for $n \in \mathbb{Z}$ (i.e., it's continuous as long as it's defined); indeed, f is nothing more than a combination of known continuous functions, with the means of combination being addition, subtraction, multiplication and division, all of which we have discussed vis a vis continuity.

What about a superficially similar looking function like $f(x) = \sin(1/x)$? This is clearly not continuous at $x = 0$ (it is not even defined there), but it seems quite clear that is it continuous at all other x. None of the situations we have discussed so far apply to this particular function, though, because it is constructed from simpler functions not by addition, subtraction, multiplication and division, but rather by composition.

We could try to compute $\lim_{x\to a}\sin(1/x)$ and see if it is equal to $\sin(1/a)$, but that would almost certainly be quite messy. Instread, we appeal to one more general result about continuity:

Theorem 7.1. *If f, g are functions, and if g is continuous at a and f is continuous at $g(a)$ (so in particular, g is defined at and near a, and f is defined at and near $g(a)$, then $(f \circ g)$ is continuous at a.*

Proof: Unlike previous proofs involving continuity, this one will be quite subtle. Already we have to work a little to verify that $(f \circ g)$ is defined at and near a. That it is defined *at* a is obvious. To see that it is defined near a, note that f is continuous at $g(a)$, so there is some $\Delta' > 0$ such that f is defined at all points in the interval $(g(a) - \Delta', g(a) + \Delta')$. We want to show that there is a $\Delta > 0$ such that for all $x \in (a - \Delta, a + \Delta)$, we have $g(x) \in (g(a) - \Delta', g(a) + \Delta')$ (so that then for all $x \in (a - \Delta, a + \Delta)$, we have that $(f \circ g)(x)$ is defined). But this follows from the continuity of g at a: apply the ε-δ definition of continuity, with Δ' as the input tolerance ε, and take the output δ to be Δ.

[92] This is a "working assumption" rather than a theorem; we haven't yet formally defined the trigonometric functions, and without a precise and formal definition of the functions, there is no point in even attempting a proof of continuity.

Next we move on to showing that $(f \circ g)(x) \to f(g(a))$ as $x \to a$. Given $\varepsilon > 0$, we want to say that if x is sufficiently close to a then $|f(g(x)) - f(g(a))| < \varepsilon$.

Here's the informal idea: by choosing x close enough to a, we can make $g(x)$ close to $g(a)$ (since g is continuous at a). But then, since $g(x)$ is close to $g(a)$, we must have $f(g(x))$ close to $f(g(a))$ (since f is continuous at $g(a)$).

Formally: given $\varepsilon > 0$ there is $\delta' > 0$ such that $|X - g(a)| < \delta'$ implies $|f(X) - f(g(a))| < \varepsilon$ (this is applying the definition of the continuity of f at $g(a)$, with input ε).

Now use that δ' as the input for the definition of g being continuous at a, i.e., for $g(x) \to g(a)$ as $x \to a$: we get that there is some $\delta > 0$ such that $|x - a| < \delta$ implies $|g(x) - g(a)| < \delta'$, which, by definition of δ', implies $|f(g(x)) - f(g(a))| < \varepsilon$.[93] □

From this theorem, we can conclude that any function that is built from known continuous functions (such as polynomial and rational functions, or sin and cos) using addition, subtraction, multiplication, division and composition, is continuous at every point in its domain. So, for example, all of

- $\sin(1/x)$
- $x \sin(1/x)$
- $\sin^3(2x^2 + \cos x) - \frac{3x}{\cos^2 x - \sin(\sin x)}$

are all continuous wherever they are defined.

What about *discontinuous* functions? It's easy to come up with examples of functions that are discontinuous at sporadic points:

- $f(x) = x/|x|$ is discontinuous at $x = 0$ (it's not defined at 0, but even if we augment the definition of f to give it a value, it will still be discontinuous at 0, since $\lim_{x \to 0} f(x)$ does not exist);
- $f(x) = [x]$[94] is defined for all reals, but is discontinuous at infinitely many places, specifically at the infinitely many integers. Indeed, for any integer t there are values of x arbitrarily close to t for which $f(x) = t$ (any x slightly larger than t), and values of x arbitrarily close to t for which $f(x) = t - 1$ (any x slightly smaller than t), so it's an easy exercise that $\lim_{x \to t} f(x)$ doesn't exist;
- $f(x) = [1/x]$ is defined for all reals other than 0. Arbitrarily close to 0, it is discontinuous at infinitely many points (so there is a "clustering" of discontinuities close to 0). Indeed, f is easily seen to be discontinuous at 1 (across which it jumps from 2 to 1), at 1/2 (across which it jumps from 3 to 2), and more generally at $\pm 1/k$ for *every* integer k.

[93] There were only two things we could have used in this proof: the continuity of f at $g(a)$ and the continuity of g at a. The only question was, which one to use *first*? Using the continuity of g at a first would have lead us nowhere.

[94] "$[x]$" is the *floor*, or *integer part*, of x — the largest integer that is less than or equal to x. So for example $[2.1] = [2.9] = [2] = 2$ and $[-0.5] = [-.001] = [-1] = -1$.

There are even easy examples of functions that has $\mathbb{R}$ as its domain, and is discontinuous *everywhere*. One such is the *Dirichlet function* f_{10} defined earlier:

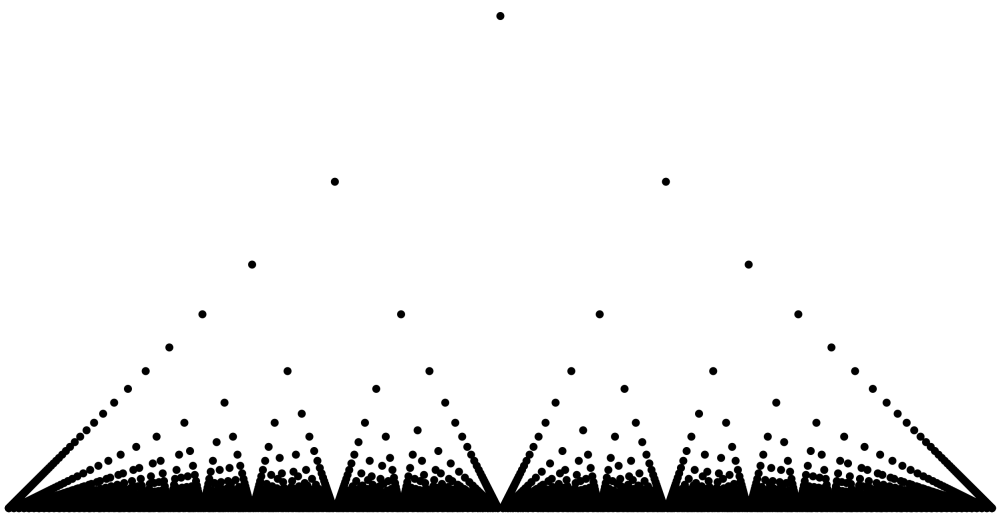
$$f_{10}(x) = \begin{cases} 1 & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational.} \end{cases}$$

Indeed, fix $a \in \mathbb{R}$. We claim that $\lim_{x\to a} f_{10}(x)$ does not exist. Let L be given. It must be the case that at least one of $|0-L|, |1-L|$ is greater than, say, $1/10$. Suppose $|1-L| > 1/10$. Take $\varepsilon = 1/10$. Given any $\delta > 0$, in the interval $(a-\delta, a+\delta)$ there must be[95] some irrational x (other than a, which may or may not be irrational; but we don't consider a when checking for a limit existing or not). We have $f_{10}(x) = 1$, so $|f_{10}(x) - L| > 1/10 = \varepsilon$. If on the other hand $|0-L| > 1/10$, again take $\varepsilon = 1/10$. Given any $\delta > 0$, in the interval $(a-\delta, a+\delta)$ there must be[96] some rational x (other than a, which may or may not be rational). We have $f_{10}(x) = 0$, so $|f_{10}(x) - L| > 1/10 = \varepsilon$. In either case we have the necessary witness to $\lim_{x\to a} f_{10}(x) \neq L$, and since L was arbitrary, the limit does not exist.

A rather more interesting example is the *Stars over Babylon* function.[97] We define it here just on the open interval $(0, 1)$:

$$f(x) = \begin{cases} 1/q & \text{if } x \text{ is rational, } x = p/q,\ p, q \in \mathbb{N},\ p, q \text{ have no common factors} \\ 0 & \text{if } x \text{ is irrational.} \end{cases}$$

Here's the graph of the Stars over Babylon function:



It takes the value $1/2$ at $1/2$; at $1/3$ and $2/3$ it takes the value $1/3$; at $1/4$ and $3/4$ it takes the value $1/4$ (but not at $2/4$; that was already covered by $1/2$); at $1/5$, $2/5$, $3/5$ and $4/3$ it

[95] Musn't there be?

[96] Again, musn't there be?

[97] So named by John Conway, for it's unusual graph; it is also called *Thomae's* function, or the *popcorn* function.

takes the value 1/5; at 1/6 and 5/6 it takes the value 1/6 (but not at 2/6, 3/6 or 4/6; these were already covered by 1/3, 1/2 and 2/3); et cetera.

We claim that for all $a \in (0,1)$, f approaches a limit near a, and specifically f approaches the limit 0. Indeed, given $a \in (0,1)$, and given $\varepsilon > 0$, we want to find a $\delta > 0$ such that $0 < |x - a| < \delta$ implies $|f(x)| < \varepsilon$.

Now there are only *finitely many* $x \in (0,1)$ with $f(x) \geq \varepsilon$, namely

$$1/2, 1/3, 2/3, 1/4, 3/4, \ldots, 1/n, \ldots, (n-1)/n$$

where $1/n$ is the largest natural number with $1/n \geq \varepsilon$. There are certainly no more than n^2 of these numbers; call them $x_1, x_2, \ldots, x_m$, written in increasing order. As long as none of these numbers satisfy $0 < |x - a| < \delta$, then for x satisfying this bound we have $|f(x)| < \varepsilon$.

So, let δ be any positive number that is smaller than

- the distance from a to 0
- the distance from a to 1 and
- the distance from a to the closest of the x_i to a (other than a itself, which may or may not be one of the x_i; but we don't care, because we don't consider a when checking for a limit existing or not).

If $0 < |x - a| < \delta$, then, because of the first two clauses above, we have that $x \in (0,1)$, so in the domain of f; and, because of the third clause, the only number in $(a - \delta, a + \delta)$ that could be among the x_i's is a itself; so, combining, if $0 < |x - a| < \delta$ then x is *not* among the x_i's, so $|f(x)| < \varepsilon$.

This completes the proof that $\lim_{x \to a} f(x) = 0$. An interesting consequence brings us back to the topic at hand, continuity: since $f(x) = 0$ exactly when x is irrational,

> Stars over Babylon is continuous at all irrationals, discontinuous at all rationals.

7.2 Continuity on an interval

Continuity at a *point* can say something about a function on an *interval*. Indeed, we have the following extremely useful fact about functions:

Claim 7.2. *Suppose f is continuous at a, and that $f(a) \neq 0$. Then there is some interval around a on which f is non-zero. Specifically, there is a $\delta > 0$ such that*

- *if $f(a) > 0$, then for all $x \in (a - \delta, a + \delta)$, $f(x) > f(a)/2$, and*
- *if $f(a) < 0$, then for all $x \in (a - \delta, a + \delta)$, $f(x) < f(a)/2$.*

We won't give a proof of this, as it is an immediate corollary of Claim 6.2, taking $M = f(a)$.

This moves us nicely along to our next main point, which is thinking about what can be said about a function that is known to be continuous not just at a point, but on an entire interval. We start with open intervals.

- Say that $f : (a, b) \to \mathbb{R}$ is *continuous on* (a, b) if it is continuous at all $c \in (a, b)$;
- say that $f : (-\infty, b) \to \mathbb{R}$ is *continuous on* $(-\infty, b)$ if it is continuous at all $c \in (-\infty, b)$;
- say that $f : (a, \infty) \to \mathbb{R}$ is *continuous on* (a, ∞) if it is continuous at all $c \in (a, \infty)$.

So, for example, the function $f(x) = 1/(x-1)(x-2)$ is continuous on the intervals $(-\infty, 1)$, $(1, 2)$ and $(2, \infty)$.

For functions defined on closed intervals, we have to be more careful, because we cannot talk about continuity at the end-points of the interval. Instead we introduce notions of one-sided continuity, using our previous notions of one-sided limits:

Definition of f being *right continuous* or *continuous from above* at a: A function f is right continuous or continuous from above at a if $\lim_{x \to a^+} f(x) = f(a)$.

Definition of f being *left continuous* or *continuous from below* at b: A function f is left continuous or continuous from below at b if $\lim_{x \to b^-} f(x) = f(b)$.

- Say that $f : [a, b] \to \mathbb{R}$ is *continuous on* $[a, b]$ if it is continuous at all $c \in (a, b)$, is right continuous at a and is left continuous at b;
- say that $f : (-\infty, b] \to \mathbb{R}$ is *continuous on* $(-\infty, b]$ if it is continuous at all $c \in (-\infty, b)$ and is left continuous at b;
- say that $f : [a, \infty) \to \mathbb{R}$ is *continuous on* $[a, \infty)$ if it is continuous at all $c \in (a, \infty)$ and is right continuous at a;
- say that $f : [a, b) \to \mathbb{R}$ is *continuous on* $[a, b)$ if it is continuous at all $c \in (a, b)$ and is right continuous at a;
- say that $f : (a, b] \to \mathbb{R}$ is *continuous on* $(a, b]$ if it is continuous at all $c \in (a, b)$ and is left continuous at b.

So, for example (easy exercises),

- the function that is defined by $f(x) = x/|x|$ away from 0 and is defined to be 1 at 0 is continuous on the intervals $(-\infty, 0)$ and $[0, \infty)$; the function, and
- $f(x) = [x]$ is continuous on all intervals of the form $[k, k+1)$, $k \in \mathbb{Z}$.

An important fact about right and left continuity is that the process of checking continuity at a is equivalent to the process of checking right and left continuity; this is an immediate corollary of Theorem 6.6:

Claim 7.3. *f is continuous at a if and only if it is both right continuous and left continuous at a.*

A quick corollary of this gives us another way to form new continuous functions from old: splicing. Suppose that f and g are defined on (a, b), and $c \in (a, b)$ has $f(c) = g(c)$. Define a new function $h : (a, b) \to \mathbb{R}$ by[98]

$$h(x) = \begin{cases} f(x) & \text{if } x \leq c \\ g(x) & \text{if } x \geq c \end{cases}$$

Corollary 7.4. *(of Claim 7.3) If f and g are both continuous at c, then h is continuous at c (and so if f, g are both continuous on (a, b), so is h).*

Proof: On $(a, c]$ h agrees with f. f is continuous at c, so is left continuous at c, and so h is left continuous at c. On $[c, b)$ h agrees with g, so right continuity of h at c follows similarly from continuity of g at c. Since h is both right and left continuous at c, it is continuous at c. □

As an example, consider the function $h(x) = |x|$. This is a splice of $f(x) = -x$ and $g(x) = x$, the splicing done at 0 (where f and g agree). Both f and g are continuous on $\mathbb{R}$, so h is continuous on $\mathbb{R}$.

7.3 The Intermediate Value Theorem

An "obvious" fact about continuous functions is that if f is continuous on $[a, b]$, with $f(a) < 0$ and $f(b) > 0$, then there must be some $c \in (a, b)$ such that $f(c) = 0$; a continuous function cannot "jump" over the x-axis.

But is this really obvious? We think of continuity at a point as meaning that the graph of the function near that point can be drawn without taking pen off paper, but the Stars over Babylon function, which is continuous at each irrational, but whose's graph near any irrational certainly *can't* be drawn without taking pen off paper, show us that we have to be careful with that intuition. The issue here, of course, is that when we say that a continuous function cannot jump over the x-axis, we are thinking about functions which are continuous at *all* points in an interval.

Here is a stronger argument for the "obvious" fact not necessarily being so obvious. Suppose that when specifying the number system we work with, we had stopped with axiom P12. Just using axioms P1-P12, we have a very nice set of numbers that we can work with — the rational numbers $\mathbb{Q}$ — inside which all usual arithmetic operations can be performed.

[98] Notice that there's no problem with the overlap of clauses here, since $f(c) = g(c)$.

If we agreed to just do our mathematics in $\mathbb{Q}$, we could still define functions, and still define the notion of a function approaching a limit, and still define the notion of a function being continuous — all of those definitions relied only on arithmetic operations (addition, subtraction, multiplication, division, comparing magnitudes) that make perfect sense in $\mathbb{Q}$. All the theorems we have proven about functions, limits and continuity would still be true.

Unfortunately, the "obvious" fact would *not* be true! The function $f : \mathbb{Q} \to \mathbb{Q}$ given by $f(x) = x^2 - 2$ is a continuous function, in the $\mathbb{Q}$-world, has $f(0) = -2 < 0$ and $f(2) = 2 > 0$, but in the $\mathbb{Q}$-world there is *no* $x \in (0, 2)$ with $x^2 = 2$ (as we have proven earlier), and so there is *no* $x \in (0, 2)$ with $f(x) = 0$: f goes from negative to positive without ever equalling 0.

So, if our "obvious" fact is true, it is as much a fact about real numbers as it is a fact about continuity, and it's proof will necessarily involve an appeal to the one axiom we introduced after P1-P12, namely the completeness axiom.

The "obvious" fact is indeed true in the $\mathbb{R}$-world, and goes under a special name:

Theorem 7.5. *(Intermediate Value Theorem, or IVT) Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function defined on a closed interval. If $f(a) < 0$ and $f(b) > 0$ then there is some $c \in (a, b)$ (so $a < c < b$) with $f(c) = 0$.*

We'll defer the proof for a while, and first make some remarks. The first remark to make is on the necessity of the hypothesis.[99]

- *Is IVT still true if f is not continuous on all of $[a, b]$?* No. Consider

$$f(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0. \end{cases}$$

  Viewed as, for example, a function on the closed interval $[-2, 2]$, f is continuous at all points on the interval $[-2, 2]$ *except* at 0. Also, $f(-2) < 0$ while $f(2) > 0$. But there is no $x \in (-2, 2)$ with $f(x) = 0$.

- *What if f is continuous on all of (a, b), just not at a and/or b?* Still No. Consider

$$f(x) = \begin{cases} -1 & \text{if } x = 0 \\ 1/x & \text{if } x > 0. \end{cases}$$

  Viewed as, for example, a function on the closed interval $[0, 1]$, f is continuous at all points on the interval $(0, 1)$. It's also left continuous at 1. The only place where (right) continuity fails is at 0. Also, $f(0) < 0$ while $f(1) > 0$. But there is no $x \in (0, 1)$ with $f(x) = 0$.

[99] Most important theorem come with hypotheses — conditions that must be satisfied in order for the theorem to be valid (for the IVT, the hypothesis is that f is *continuous on the whole closed interval* $[a, b]$). Most of the theorems we will see have been refined over time to the point where the hypotheses being assumed are the bare minimum necessary to make the theorem true. As such, it should be possible to come up with counterexamples to the conclusions of these theorems, whenever the hypothesis are even slightly weakened. You should get into the habit of questioning the hypotheses of every big theorem we see, specifically asking yourself "is this still true if I weaken any of the hypotheses?". Usually, it will not be true anymore.

A second remark is that the IVT quickly gives us the existence of a unique square root of any positive number:

Claim 7.6. *For each $a \geq 0$ there is a unique number $a' \geq 0$ such that $(a')^2 = a$. We refer to this number as the* square root *of a, and write it either as $\sqrt{a}$ or as $a^{1/2}$.*

Proof: If $a = 0$ then we take $a' = 0$. This is the unique possibility, since as we have earlier proven, if $a' \neq 0$ then $(a')^2 > 0$, so $(a')^2 \neq 0$.

Suppose $a > 0$. Consider the function $f_a : [0, a+1] \to \mathbb{R}$ given by $f_a(x) = x^2 - a$. This is a continuous function at all points on the interval, as we have previously proven. Also $f_a(0) = -a < 0$ and $f_a(a+1) = (a+1)^2 - a = a^2 + a + 1 > 0$. So by IVT, there is $a' \in (0, a+1)$ with $f_a(a') = 0$, that is, with $(a')^2 = a$.

To prove that this a' is the *unique* possibility for the positive square root of a, note that if $0 \leq a'' < a'$ then $0 \leq (a'')^2 < (a')^2$ (this was something we proved earlier), so $(a'')^2 \neq a$, while if $0 \leq a' < a''$ then $0 \leq (a')^2 < (a'')^2$, so again $(a'')^2 \neq a$. Hence a' is indeed unique. □

We can go further, with essentially no extra difficulty:

Claim 7.7. *Fix $n \geq 2$ a natural number. For each $a \geq 0$ there is a unique number $a' \geq 0$ such that $(a')^n = a$. We refer to this number as the* nth root *of a, and write it either as $\sqrt[n]{a}$ or as $a^{1/n}$.*

Proof: If $a = 0$ then we take $a' = 0$. This is the unique possibility, since if $a' \neq 0$ then $(a')^n \neq 0$.

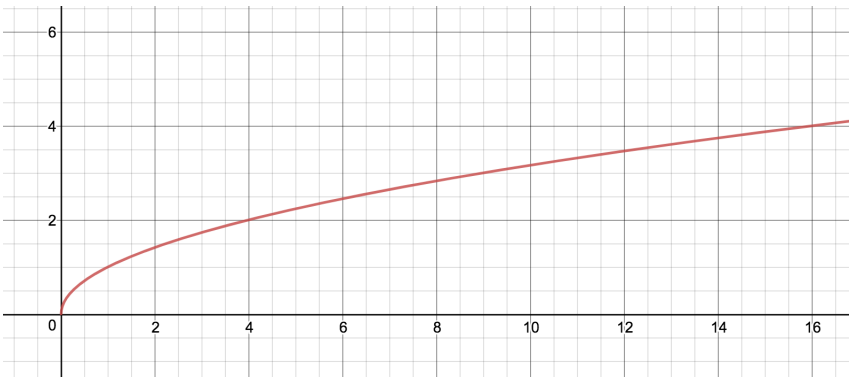
Suppose $a > 0$. Consider the function $f_a : [0, a+1] \to \mathbb{R}$ given by $f_a(x) = x^n - a$. This is a continuous function. Also $f_a(0) = -a < 0$ and (using the binomial theorem)

$$f_a(a+1) = (a+1)^n - a = a^n + \binom{n}{n-1} a^{n-1} + \cdots + \binom{n}{n-k} a^{n-k} + \cdots + \left(\binom{n}{1} - 1 \right) a + 1 > 0.$$

So by IVT, there is $a' \in (0, a+1)$ with $f_a(a') = 0$, that is, with $(a')^n = a$.

To prove that this a' is the *unique* possibility for the positive nth root of a, note that if $0 \leq a'' < a'$ then $0 \leq (a'')^n < (a')^n$ while if $0 \leq a' < a''$ then $0 \leq (a')^n < (a'')^n$. □

Define, for natural numbers $n \geq 2$, a function $f_n : [0, \infty) \to [0, \infty)$ by $x \mapsto x^{1/n}$. The graph of the function f_2 is shown below; it looks like it is continuous on its whole domain, and we would strongly expect f_n to be continuous on all of $[0, \infty)$, too.



Claim 7.8. *For all $n \geq 2$, $n \in \mathbb{N}$, the function f_n is continuous on $[0, \infty)$.*

Proof: As a warm=up, we deal with $n = 2$. Fix $a > 0$. Given $\varepsilon > 0$ we want to find $\delta > 0$ such that $|x - a| < \delta$ implies $|x^{1/2} - a^{1/2}| < \varepsilon$.

As usual, we try to manipulate $|x^{1/2} - a^{1/2}|$ to make an $|x - a|$ pop out. The good manipulation here is to multiply above and below by $|x^{1/2} + a^{1/2}|$, and use the difference-of-two-squares factorization, $X^2 - Y^2 = (X - Y)(X + Y)$, to get

$$\begin{aligned} |x^{1/2} - a^{1/2}| &= |x^{1/2} - a^{1/2}| \frac{|x^{1/2} + a^{1/2}|}{|x^{1/2} + a^{1/2}|} \\ &= \frac{|x^{1/2} - a^{1/2}||x^{1/2} + a^{1/2}|}{|x^{1/2} + a^{1/2}|} \\ &= \frac{|(x^{1/2} - a^{1/2})(x^{1/2} + a^{1/2}|}{|x^{1/2} + a^{1/2}|} \\ &= \frac{|x - a|}{|x^{1/2} + a^{1/2}|} \\ &= \frac{|x - a|}{x^{1/2} + a^{1/2}}, \end{aligned}$$

the last equality valid since $x^{1/2} \geq 0, a^{1/2} > 0$.

Now $x^{1/2} \geq 0$ so $x^{1/2} + a^{1/2} \geq a^{1/2}$ and sp

$$|x^{1/2} - a^{1/2}| \leq \frac{|x - a|}{a^{1/2}}.$$

Choose any δ at least as small as the minimum of a (to make sure that $|x - a| < \delta$ implies $x > 0$, so x is in the domain of f_2) and $a^{1/2}\varepsilon$. Then $|x - a| < \delta$ implies

$$|x^{1/2} - a^{1/2}| \leq \frac{|x - a|}{a^{1/2}} < \varepsilon.$$

That proves continuity of f_2 at all $a > 0$; right continuity at 0 (i.e., $\lim_{x \to a^+} x^{1/2} = 0$) is left as an exercise.

For the case of general n, we replace $X^2 - Y^2 = (X - Y)(X + Y)$ with

$$X^n - Y^n = (X - Y)(X^{n-1} + X^{n-2}Y + \cdots + XY^{n-2} + Y^{n-1}).$$

In the case $a > 0$, repeating the same argument as in the case $n = 2$ leads to

$$|x^{1/n} - a^{1/n}| = \frac{|x - a|}{(x^{1/n})^{n-1} + (x^{1/n})^{n-2}(a^{1/n}) + \cdots + (x^{1/n})(a^{1/n})^{n-2} + (a^{1/n})^{n-1}} \leq \frac{|x - a|}{(a^{1/n})^{n-1}},$$

and so continuity of f_n at $a > 0$ follows as before, this time taking any $\delta > 0$ at least as small as the minimum of a and $(a^{1/n})^{n-1}\varepsilon$. Again, right continuity at 0 is left as an exercise. □

We know that $a^{1/2}$ cannot make sense (i.e., cannot be defined) for $a < 0$: if there was a real number $a^{1/2}$ for negative a, we would have $(a^{1/2})^2 \geq 0$ (since squares of reals are always

positive), but also $(a^{1/2})^2 = a < 0$, a contradiction. By the same argument, we don't expect $a^{1/n}$ to make sense for negative a for any even natural number n.

But for odd n, we do expect that $a^{1/n}$ should make sense for negative a, and that is indeed the case.

Claim 7.9. *Fix $n \geq 3$ an odd natural number. For each $a \in \mathbb{R}$ there is a unique number $a' \in \mathbb{R}$ such that $(a')^n = a$. We refer to this number as the* nth root *of a, and write it either as $\sqrt[n]{a}$ or as $a^{1/n}$.*

Extending the function f_n defined above to all real numbers, we have that $f_n : \mathbb{R} \to \mathbb{R}$ given by $x \mapsto x^{1/n}$ is continuous for all reals.

We will not prove this, but rather leave it as an exercise. The main point is that if we define, for odd integer n and for *any* real a, the (continuous) function $f_a : \mathbb{R} \to \mathbb{R}$ via $f_a(x) = x^n - a$, then we can find $a' < a''$ for which $f_a(a') < 0 < f_a(a'')$. Once we have found a', a'' (which is a little tricky), the proof is very similar to the proofs we've already seen.

But in fact we will prove something more general than the existence of a', a''. From the section on graphing function, we have a sense that if $P(x)$ is an odd-degree polynomial of degree n, for which the coefficient of x^n is positive, then for all sufficiently negative numbers x we have $P(x) < 0$, while for all sufficiently positive x we have $P(x) > 0$. Since P is continuous, that would say (applying the IVT on any interval $[a', a'']$ where a' is negative and satisfies $P(a') < 0$, and a'' is positive and satisfies $P(a'') > 0$) that there is some $a \in \mathbb{R}$ with $P(a) = 0$ (and in particular applying this to $P(x) = x^n - a$ yields an nth root of a for every real a).

Claim 7.10. *Let $P(x) = x^n + a_1 x^{n-1} + \cdots + a_{n-1}x + a_n$ be a polynomial, with n* odd. *There are numbers x_1 and x_2 such that $P(x) < 0$ for all $x \leq x_1$, and $P(x_2) > 0$ for all $x \geq x_2$. As a consequence (via IVT) there is a real number c such that $P(c) = 0$.*

Proof: The idea is that for large x the term x^n "dominates" the rest of the polynomial — if x is sufficiently negative, then x^n is very negative, so much so that it remains negative after $a_1 x^{n-1} + \cdots + a_{n-1}x + a_n$ is added to it; while if x is sufficiently positive, then x^n is very positive, so much so that it remains positive after $a_1 x^{n-1} + \cdots + a_{n-1}x + a_n$ (which may itself be negative) is added to it.

To formalize this, we use the triangle inequality to bound $|a_1 x^{n-1} + \cdots + a_{n-1}x + a_n|$. Setting $M = |a_1| + |a_2| + \cdots + |a_n| + 1$ (the $+1$ at the end to make sure that $M > 1$), and considering only those x for which $|x| > 1$ (so that $1 < |x| < |x|^2 < |x|^3 < \cdots$), we have

$$\begin{aligned} |a_1 x^{n-1} + \cdots + a_{n-1}x + a_n| &\leq |a_1 x^{n-1}| + \cdots + |a_{n-1}x| + |a_n| \\ &= |a_1||x|^{n-1} + \cdots + |a_{n-1}||x| + |a_n| \\ &\leq |a_1||x^{n-1}| + \cdots + |a_{n-1}||x|^{n-1} + |a_n||x|^{n-1} \\ &< M|x|^{n-1}. \end{aligned}$$

It follows that for any x satisfying $|x| > 1$,

$$x^n - M|x|^{n-1} < P(x) < x^n + M|x|^{n-1}$$

Now take $x_2 = 2M$ (note $|x_2| > 1$). For $x \geq x_2$ (so in particular $x > 0$) we have

$$P(x) > x^n - M|x|^{n-1} = x^n - Mx^{n-1} = x^{n-1}(x - M) \geq 2^{n-1}M^n > 0$$

(using $x - M \geq M$ in the next-to-last inequality).

On the other hand, taking $x_1 = -2M$ (note $|x_1| > 1$) we have that for $x \leq x_1$,

$$P(x) < x^n + M|x|^{n-1} = x^n + Mx^{n-1} = x^{n-1}(x + M) \leq -2^{n-1}M^n < 0$$

(note that in the first equality above, we use $|x|^{n-1} = x^{n-1}$, valid since $n - 1$ is even). □

If the coefficient of x^n in P is not 1, but some positive real $a_0 > 0$, then an almost identically proof works to demonstrate the same conclusion ($P(x)$ is negative for all sufficiently negative x, and positive for all sufficiently positive x, and so $P(c) = 0$ for some c); and if the coefficient of x^n in P is instead some negative real $a_0 < 0$ then, applying the theorem just proven to the polynomial $-P$, we find that $P(x)$ is positive for all sufficiently negative x, and negative for all sufficiently positive x, and so again by the IVT $P(c) = 0$ for some c. In other words:

every odd degree polynomial has a real root.

Note that no such claim can be proved for *even* n; for example, the polynomial $P(x) = x^2 + 1$ never takes the value 0. We will return to even degree polynomials when we discuss the Extreme Value Theorem.

We now turn to the proof of IVT. As we have already observed, necessarily the proof will involve the completeness axiom. The informal idea of the proof is: "the first point along the interval $[a, b]$ where f stops being negative, must be a point at which f is zero". We will formalize this by considering the set of numbers x such that f is negative on the entire closed interval from a to x. This set is non-empty (a is in it), and is bounded above (b is an upper bound), so by completeness (P13), the set has a least upper bound. We'll argue that that least upper bound is strictly between a and b, and that that function evaluates to 0 at that point.

Proof (of Intermediate Value Theorem): Let $A \subseteq [a, b]$ be

$$\{x \in [a, b] : f \text{ is negative on } [a, x]\}.$$

We have $a \in A$ (since $f(a) < 0$), so A is not empty. We have that b is an upper bound for a (since $f(b) > 0$), so by the completeness axiom (P13), A has a least upper bound, call it c. Recall that this means that

- c is an upper bound for A ($x \leq c$ for all $x \in A$), and that

- c is the least such number (if c' is any other upper bound then $c' \geq c$).

We will argue that $a < c < b$, and that $f(c) = 0$. That $c > a$ follows from left continuity of f at a, and $f(a) < 0$ (the proof that if f is *continuous* and negative at a, then there's some $\delta > 0$ such that f is negative on all of $(a - \delta, a + \delta)$, can easily be modified to show that if f is *right continuous* and negative at a, then there's some $\delta > 0$ such that f is negative on all of $[a, a + \delta)$, so certainly $a + \delta/2 \in A$). Similarly, that $c < b$ follows from right continuity of f at b, and $f(b) > 0$ (there's $\delta > 0$ such that f is positive on all of $(b - \delta, b]$, so certainly $b - \delta/2$ is an upper bound for A).

Next we argue that $f(c) = 0$, by showing that assuming $f(c) > 0$ leads to a contradiction, and similarly assuming $f(c) < 0$ leads to a contradiction.

Suppose $f(c) > 0$. There's $\delta > 0$ such that f is positive on $(c - \delta, c + \delta)$, so $c - \delta/2$ is an upper bound for A — no number in $[c - \delta/2, c]$ can be in A, because f is positive at all these numbers — contradicting that c is the *least* upper bound for A.

Suppose $f(c) < 0$. There's $\delta > 0$ such that f is negative on $(c - \delta, c + \delta)$. In fact, f is negative on all of $[a, c + \delta)$ — if f was positive at any $c' < c$, c' would be an upper bound on A, contradicting that c is the *least* upper bound for A — and so $c + \delta/2 \in A$, contradicting that c is even an *upper bound* for A. □

There are a few obvious variants of the Intermediate Value Theorem that are worth bearing in mind, any require virtually no work to prove once we have the version we have already proven.

- If f is continuous on $[a, b]$, and if $f(a) > 0$, $f(b) < 0$, then there is some $c \in (a, b)$ with $f(c) = 0$. (To prove this, apply the IVT as we have proven it to the function $-f$; the c thus produced has $(-f)(c) = 0$ so $f(c) = 0$.)

- If f is continuous on $[a, b]$, with $f(a) \neq f(b)$, and if t is any number that lies between $f(a)$ and $f(b)$, then there is $c \in (a, b)$ with $f(c) = t$. (To prove this in the case where $f(a) < f(b)$, apply the IVT as we have proven it to the function $x \mapsto f(x) - t$, and to prove it in the case where $f(a) > f(b)$, apply the IVT as we have proven it to the function $x \mapsto t - f(x)$.)

- If f is a continuous function on an interval, and f takes on two different values, then it takes on all values between those two values.[100] (To prove this, let a and b be the two inputs on which f is seen to take on different values, where, without loss of generality, $a < b$, and apply the version of the IVT in the second bullet point above to f on the interval $[a, b]$.)

7.4 The Extreme Value Theorem

We begin this section with some definitions. In each of these definitions, we want to think about a function not necessarily on its whole natural domain, but rather on some specific

[100] This is often taken as the statement of the Intermediate Value Theorem.

subset of the domain. For example, we may wish to consider the function $x \mapsto 1/x$ not at being defined on all reals except 0, but rather being defined on all positive reals, or on the open interval $(0, 1)$. One way to do that is to artificially define the function only on the particular set of reals that we are interested in; but this is a little restrictive, as we may want to think about the same function defined on many different subsets of its natural domain. The approach taken in this definitions, while it may seem a little wordy at first, allows us this flexibility, and will be very useful in other situations too.

Definition of a function being bounded from above f is *bounded from above* on a subset S of Domain(f) if there is some number M such that $f(x) \leq M$ for all $x \in S$; M is an *upper bound* for the function *on* S.

Definition of a function being bounded from below f is *bounded from below* on S if there is some number m such that $m \leq f(x)$ for all $x \in S$; m is a *lower bound* for the function on S.

Definition of a function being bounded f is *bounded* on S if it is bounded from above **and** bounded from below on S.

Definition of a function achieving its maximum f *achieves* its maximum on S if there is a number $x_0 \in S$ such that $f(x) \leq f(x_0)$ for all $x \in S$. (Notice that this automatically implies that f is bounded from above on S: $f(x_0)$ is an upper bound.)

Definition of a function achieving its mimimum f achieves its minimum on S if there is a number $x_0 \in S$ such that $f(x_0) \leq f(x)$ for all $x \in S$. (Notice that this automatically implies that f is bounded from below on S: $f(x_0)$ is a lower bound.)

It's an easy exercise that f is bounded on S if and only if there is a *single* number M such that $|f(x)| < M$ for all $x \in S$.

Basically anything can happen vis a vis upper and lower bounds, depending on the specific choice of f and S. For example:

- $f(x) = 1/x$ is bounded on $[1, 2]$, and achieves both maximum and minimum;
- $f(x) = 1/x$ is bounded on $(1, 2)$, but achieves *neither* maximum *nor* minimum;
- $f(x) = 1/x$ is bounded on $[1, 2)$, does not achieve its maximum, but does achieve its minimum;
- $f(x) = 1/x$ is not bounded from above on $(0, 2)$, is bounded from below, and does not achieve its minimum;
- $f(x) = 1/x$ is not bounded from above or from below on its natural domain.

The second important theorem of continuity (IVT was the first) says that a continuous function *on a closed interval* is certain to be as well-behaved as possible with regards bounding.

Theorem 7.11. *(Extreme Value Theorem, or EVT for short) Suppose $f : [a, b] \to \mathbb{R}$ is continuous. Then*

- *f is bounded on $[a, b]$*[101]*, and*
- *f achieves both its maximum and minimum on $[a, b]$.*

We will see many applications of the EVT throughout this semester and next, but for the moment we just give one example. Recall that earlier we used the IVT to prove that if P is an *odd degree* polynomial then there must be c with $P(c) = 0$, and we observed that no such general statement could be made about *even degree* polynomials. Using the EVT, we can say something about the behavior of even degree polynomials.

Claim 7.12. *Let $P(x) = x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n$ be a polynomial, with n* even*. There is a number $x^\star$ such that $P(x^\star)$ is the minimum of P on $(-\infty, \infty)$, that is, such that $p(x^\star) \leq p(x)$ for all real x.*

Proof: Here's the idea: we have $p(0) = a_n$. We'll try to find numbers $x_1 < 0 < x_2$ such that

$$P(x) > a_n \text{ for all } x \leq x_1 \text{ and for all } x \geq x_2. \ (\star)$$

We then apply the EVT on the interval $[x_1, x_2]$ to conclude that there is a number $x^\star \in [x_1, x_2]$ such that $P(x^\star) \leq P(x)$ for all $x \in [x_1, x_2]$. Now since $0 \in [x_1, x_2]$, we have $P(x^\star) \leq P(0) = a_n$, and so also $P(x^\star) \leq P(x)$ for all $x \in (-\infty, x_1]$ and $[x_2, \infty)$ (using $(\star)$). So, $P(x^\star) \leq P(x)$ for all $x \in (-\infty, \infty)$.

To find x_1, x_2, we use a very similar strategy to the one used in the proof of Claim 7.10, to show that if $M = |a_1| + \cdots + |a_n| + 1$ then there are numbers x_1, x_2 with $x_1 < 0 < x_2$ such that $P(x) \geq 2^{n-1}M^n$ for all $x \leq x_1$ and for all $x \geq x_2$ (the details of this step are left as an exercise).

Because M is positive and at least 1, and because n is at least 2, we have

$$2^{n-1}M^n \geq M \geq |a_n| + 1 \geq a_n + 1 > a_n,$$

and so we are done. □

We now turn to the proof of the Extreme Value Theorem. We begin with a preliminary observation, that if f is continuous at a point c then it is "locally bounded": there is a $\delta > 0$ such that f is bounded above, and below, on $(a - \delta, a + \delta)$. Indeed, apply the definition of continuity at c with $\varepsilon = 1$ in order to get such a δ, with specifically $f(c) - 1 < f(x) < f(c) + 1$ for all $x \in (a - \delta, a + \delta)$.

The intuition of the proof we give is that we can stringing together local boundedness at each point in the interval $[a, b]$ to get that f is bounded on $[a, b]$. We have to do it

[101] In other words, a continuous function on a closed interval cannot "blow up" to infinity (or negative infinity).

carefully, though, to avoid the upper bounds growing unbounded larger, and the lower bounds unbounded smaller. The approach will be similar to our approach to the IVT: this time, we find the longest closed interval starting at a on which f is bounded, and try to show that the interval goes all the way to b, by arguing that it it falls short of b, getting only as far as some $c < b$, *one* application of "local boundedness" allows us to stretch the interval a little further, contradicting that the interval stopped at c.

Proof (of Extreme Value Theorem): We start with the first statement, that a function f that is continuous on $[a, b]$ is bounded on $[a, b]$. We begin by showing that f is bounded from above. Let

$$A = \{x : a \leq x \leq b \text{ and } f \text{ is bounded above on } [a, x]\}.$$

We have that $a \in A$ and that b is an upper bound for A, so $\sup A := \alpha$ exists.

We cannot have $\alpha < b$. For suppose this was the case. Since f is continuous at α, it is bounded on $(\alpha - \delta, \alpha + \delta)$ for some $\delta > 0$. Now we consider two cases.

Case 1, $\alpha \in A$ Here f is bounded on $[a, \alpha]$ (by M_1, say) and also on $[\alpha - \delta/2, \alpha + \delta/2]$ (by M_2, say), so it is bounded on $[a, \alpha + \delta/2]$ (by $\max\{M_1, M_2\}$), so $\alpha + \delta/2 \in A$, contradicting that α is the least upper bound of A.

Case 2, $\alpha \notin A$ Here it must be that some $c \in (\alpha - \delta, \alpha)$ is in A; if not, $\alpha - \delta$ would be an upper bound for A, contradicting that α is the least upper bound of A. As in Case 1, f is bounded on $[a, c]$ and also on $[c, \alpha + \delta/2]$, so it is bounded on $[a, \alpha + \delta/2]$, again a contradiction.

We conclude that $\alpha = b$, so it seems like we are done; but, we wanted f bounded on $[a, b]$, and $\sup A = b$ doesn't instantly say this, because the supremum of a set doesn't have to be in the set.[102] So we have to work a little more.

Since f is right continuous at b, f is bounded on $(b - \delta, b]$ for some $\delta > 0$. If $b \notin A$, then, since $b = \sup A$, we must have $x_0 \in A$ for some $x_0 \in (b - \delta, b)$ (otherwise $b - \delta$ would work as an upper bound for A). So f is bounded on $[a, x_0]$ and also on $[x_0, b]$, so it is bounded on $[a, b]$, so $b \in A$, a contradiction. So in fact $b \in A$, and f is bounded from above on $[a, b]$.

and since f bounded on $[a, x_0]$ for some $x_0 \in (b - \delta, b)$ (our fact again — $b \notin A$), have f bdd on $[a, b]$.

A similar proof, using the equivalent form of the Completeness axiom introduced earlier (a non-empty set with a lower bound has a greatest lower bound) can be used to show that f is also bounded from below on $[a, b]$; or, we can just apply what we have just proven about upper bounds to the (continuous) function $-f$ defined on $[a, b]$ — $-f$ has some upper bound M on $[a, b]$, so $-M$ is a lower bound for f on $[a, b]$.

[102] A easy example: $\sup(0, 1) = 1$ which is not in $(0, 1)$. An example more relevant to this proof: consider $g(x) = 1/(1 - x)$ on $[0, 1)$, and $g(1) = 0$ at 1. If $A = \{x : g \text{ bounded on } [0, x]\}$, then $\sup A = 1$ but $1 \not\in A$. The problem here of course is that g is not *continuous* at 1

We now move on to the second part of the EVT: if $f : [a, b] \to \mathbb{R}$ is continuous, it achieves both its maximum and its minimum; there are $y, z \in [a, b]$ such that $f(z) \leq f(x) \leq f(y)$ for all $x \in [a, b]$. We just show that f achieves its maximum; the trick of applying this result to $-f$ will again work to show that f also achieves its minimum.

Consider $A = \{f(x) : x \in [a, b]\}$ (notice that now we are looking at a "vertical" set; a set of points along the y-axis of the graph of f). A is non-empty ($f(a) \in A$), and has an upper bound (by previous part of the EVT, that we have already proven). So $\sup A = \alpha$ exists. We have $f(x) \leq \alpha$ for all $x \in [a, b]$, so to complete the proof we just need to find a y such that $f(y) = \alpha$.

Suppose there is no such y. Then the function $g : [a, b] \to \mathbb{R}$ given by

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

is continuous function (the denominator is never 0). So, again by the previous part of the EVT, g is bounded above on $[a, b]$, say by some $M > 0$. So on $[a, b]$ we have $1/(\alpha - f(x)) \leq M$, or $\alpha - f(x) \geq 1/M$, or $f(x) \leq \alpha - 1/M$. But this contradicts that $\alpha = \sup A$.

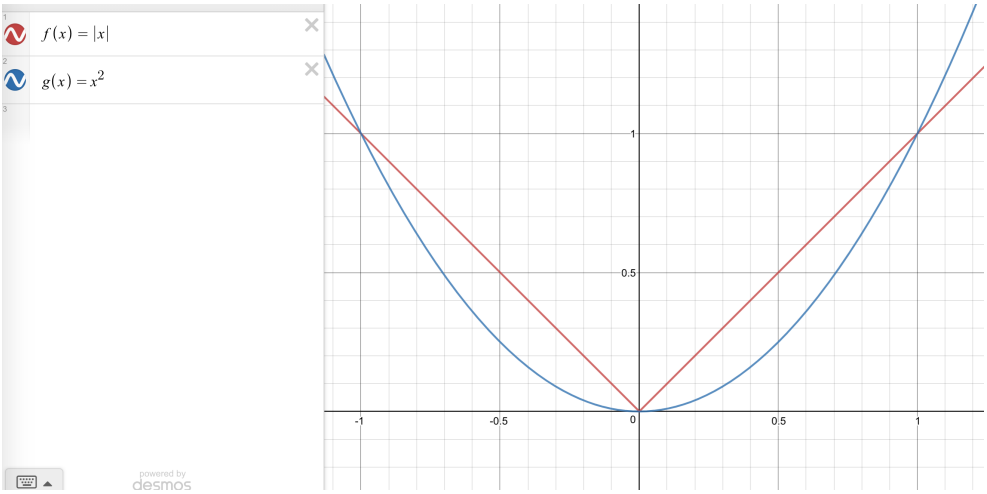
We conclude[103] that there must be a y with $f(y) = \alpha$, completing the proof of the theorem. □

[103] Somewhat miraculously — the function g was quite a rabbit-out-of-a-hat in this proof.

8 The derivative

We think, informally, as continuity as being a measure of "smoothness": if a function f is continuous at a, then small changes in the input to f near a lead only to small changes in the output.

But there are definitely "degrees of smoothness". The functions $f(x) = |x|$ and $g(x) = x^2$ (see figure) are both continuous at 0, and both achieve their minima at 0, but their graphs behave very differently near 0 — g curves gently, while f has a sharp point.



The tool we introduce now, that (among many many other things) distinguishes these two behaviors, is the familiar tool of the *derivative*.

8.1 Two motivating examples

Instantaneous velocity Suppose that a particle is moving along a line, and that its distance from the origin at time t is given by the function $s(t)$.

It's easy to calculate the *average velocity* of the particle over a time interval for, say, time $t = a$ to time $t = b$: it's the total displacement of the particle, $s(b) - s(a)$, divided by the total time, $b - a$.[104]

But what is the *instantaneous velocity* of the particle at a certain time t? To make sense of this, we might do the following: over a small time interval $[t, t + \Delta t]$ (starting

[104]Remember that velocity is a *signed* quantity: if a particle starts 10 units to the right of the origin, and two seconds later is 14 units to the right of the origin, then its average velocity over those two seconds is $(14 - 10)/2 = 2$ units per second, positive because the particle has progressed further from the origin. If, on the other hand, it starts 14 units to the right of the origin, and two seconds later is 10 units to the right of the origin, then its average velocity over those two seconds is $(10 - 14)/2 = -2$ units per second, negative because the particle has progressed *closer* to the origin. In both cases the average *speed* is the same — 2 units per second — speed being the absolute value of velocity.

at time t, ending at time $t + \Delta t$), with $\Delta t > 0$, the average velocity is

$$\frac{\text{displacement}}{\text{time}} = \frac{s(t + \Delta t) - s(t)}{\Delta t}.$$

Similarly over a small time interval $[t + \Delta t, t]$, with $\Delta t < 0$, the average velocity

$$\frac{s(t) - s(t + \Delta t)}{-\Delta t} = \frac{s(t + \Delta t) - s(t)}{\Delta t}.$$

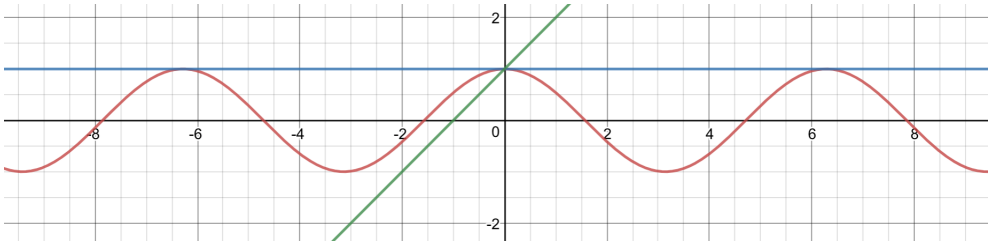
If this common quantity, $(s(t + \Delta t) - s(t))/\Delta t$, is approaching a limit as Δt approaches 0, then it makes sense to *define* instantaneous velocity at time t to be that limit, that is, to be

$$\lim_{\Delta t \to 0} \frac{s(t + \Delta t) - s(t)}{\Delta t}.$$

Tangent line : What is the equation of the tangent line to the graph of function f at some point $(a, f(a))$ on the graph? To answer that, we must answer the more fundamental question, "what do we mean by 'tangent line'?". A preliminary definition might be that

> a tangent line to a graph at a point on the graph is a straight line that touches the graph only at that point.

This is a fairly crude definition, and fairly clearly doesn't work: the line $y = 1$ touches the graph of $y = \cos x$ infinitely many times, at $x = 0, \pm\pi, \pm 2\pi, \ldots$, but clearly should be declared to be a tangent line to $y = \cos x$ at $(0, 1)$; on the other hand, the line $y = 10x$ touches the graph of $y = \cos x$ only once, at $(0, 1)$, but clearly should *not* be declared to be a tangent line to $y = \cos x$ at $(0, 1)$.



What we really want to say, is that a tangent line to a graph at a point on the graph is a straight line that passes through the point, and that just "glances off" the graph at that point, or is "going in the same direction as the graph" at that point, or "has the same slope as the graph does" at that point.

Clearly these phrases in quotes need to be made more precise. What do we mean by "the slope of a graph, at a point"? We can make this precise, in a similar way to the way we made precise the notion of instantaneous velocity.

A *secant line* of a graph is a straight line that connects two points $(x_1, f(x_1))$, $(x_2, f(x_2))$ on the graph. It makes perfect sense to talk of the "slope of a secant line": it is

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

The define the slope at a point $(a, f(a))$, we can consider the slope of the secant line between $(a, f(a))$ and $(a + h, f(a + h))$ for small $h > 0$, or between $(a + h, f(a + h))$ and $(a, f(a))$ for small $h < 0$. In both cases, this is

$$\frac{f(a + h) - f(a)}{h}.$$

This secant slope seems like it should be a reasonable approximation of the slope of the graph *at* the point $(a, f(a))$; and in particular, if the slopes of the secant lines approach a limit as h approaches 0, then it makes a lot of sense to *define* the slope at $(a, f(a))$ to be that limit:

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

Going back to the original question, once we have found the slope, call it m_a, we can easily figure out the equation of the tangent line to the graph at $(a, f(a))$, since it is the unique straight line that passes through $(a, f(a))$ and has slope m_a:

$$y - f(a) = m_a(x - a) \quad \text{or} \quad y = m_a(x - a) + f(a) \quad \text{or} \quad y = m_a x - m_a a + f(a).$$

The two expressions we have obtained from these two examples — $\lim_{\Delta t\to 0}(s(t + \Delta t) - s(t))/\Delta t$ and $\lim_{h\to 0}(f(a + h) - f(a))/h$ — are of exactly the same form. Since the same expression has cropped up in two rather different-looking applications, it makes sense to look at the expression as an interesting object in its own right, and study its properties. That is exactly what we will do in this section.

8.2 The definition of the derivative

Let f be a function, and let f be defined at and near some number a (i.e., suppose there is some $\Delta > 0$ such that all of the interval $(a - \Delta, a + \Delta)$ is in the domain of f.

Definition of derivative Say that f is *differentiable* at a if

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

exists. If f is differentiable at a, then write $f'(a)$ for the value of this limit; $f'(a)$ is referred to as the *derivative* of f at a[105].

[105]There are many alternate notations for the derivative:

From the previous section, we obtain immediately two interpretations of the quantity $f'(a)$:

Velocity if $s(t)$ measures the position at time t of a particle that is moving along the number line, then $s'(a)$ measures the velocity of the particle at time a.

Slope $f'(a)$ is the slope of the tangent line of the graph of function f at the point $(a, f(a)$. Consequently the equation of the tangent line is

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

Once we have the notion of the derivate of a function at a point, it's a very short leap to considering the derivative as a *function*.

Definition of the derivative function If $f : D \to \mathbb{R}$ is a function defined on some domain D, then the *derivative* of f is a function, denoted f'[106], whose domain is $\{a \in D : f \text{ differentiable at } a\}$[107], and whose value at a is the derivative of f at a.

As we will see in a series of examples, the domain of f' may the same as the domain of f, or slightly smaller, or *much* smaller.

Before going on to the examples, we mention an alternate definition for of the definition of derivative:

$$f'(a) = \lim_{b \to a} \frac{f(b) - f(a)}{b - a}.$$

Indeed, suppose $\lim_{b\to a}(f(b)-f(a))/(b-a)$ exists and equal L. Then for all $\varepsilon > 0$ there is $\delta > 0$ such that whenever b is within δ of a (but not equal to a), we have that $(f(b) - f(a))/(b - a)$ is within ε of L. Rewriting b as $a + h$ (so $b - a = h$), this says that whenever $a + h$ is within δ of a (but not equal to a), that is, whenever h is within δ of 0 (but not equal to 0), we have that $(f(a + h) - f(a))/h$ is within ε of L. This says $\lim_{h\to 0}(f(a + h) - f(a))/h$ exists and equal L. The converse direction goes along the same lines.

- $f'(a)$
- $\dot{f}(a)$
- $\frac{d}{dx} f(x) \,|_{x=a}$
- $\frac{df(x)}{dx} \,|_{x=a}$
- $\frac{df}{dx} \,|_{x=a}$
- $\frac{dy}{dx} \,|_{x=a}$ (if y is understood to be another name for f)
- $\dot{y}(a)$ (again, if y is another name for f).

We will almost exclusively use the first of these.

[106] or $\frac{df}{dx}$, or $\dot{f}$.

[107] We will shortly modify this definition slightly, to deal with functions which are defined on closed intervals such as $[0, 1]$; we will introduce a notion of "differentiable from the right" and "differentiable from the left" so as to be able to talk about what happens at the end points of the interval.

8.3 Some examples of derivatives

Given the work we have done on limits and continuity, calculating the derivatives of many simple function, even directly from the definition, is fairly straightforward. We give a bunch of examples here.

Constant function $f(x) = c$, where c is some fixed real number. Presumably, the derivative of this function is 0 at any real a, that is,

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

Notice that we can't verify this instantly by appealing to continuity of the expression $(f(a+h) - f(a))/h$, viewed as a function of h, at $h = 0$, and then just evaluating the expression at $h = 0$; the expression is not only not continuous at $h = 0$, it is not even defined at $h = 0$! This will be a common theme in computing derivatives: the expression $(f(a+h) - f(a))/h$ (viewed as a function of h), regardless of the f under consideration, will *always* not be defined at $h = 0$, since the numerator and the denominator both evaluate to 0 at $h = 0$. So here, and in all other examples that we do, we will have to engage in some algebraic manipulation of the expression $(f(a+h) - f(a))/h$. The goal of the manipulation is to try and find an alternate expression, that is equal to $(f(a+h) - f(a))/h$ for all h except (possibly) $h = 0$ (the one value of h we do not really care about); and then see if we can use some of our previous developed techniques to evaluate the limit as h goes to 0 of the new expression.

For any real a we have, for $h \neq 0$,

$$\frac{f(a+h) - f(a)}{h} = \frac{c - c}{h} = \frac{0}{h} = 0,$$

and so

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

from which we conclude that f is differentiable at all a, with derivative 0. (Of course: the line $y = c$ is clearly the tangent line to f at *any* point, and this line has slope 0; or, if a particle is located at the same position, c, on the line at all times, its velocity at all times is 0.)

In this example, f' is the constant 0 function, on the same domain ($\mathbb{R}$) as f.

This example is really simple, but it is worth doing in detail for two reasons. First, a philosophical reason: to act as a reality check for the definition, and our understanding of the definition. Second, a practical reason: to illustrate a subtlety of writing up proofs from first principles of derivatives of functions. It's very tempting to argue that $f'(a) = 0$ by writing

$$\text{“}f'(a) = \lim_{h\to 0} \frac{f(a+h) - f(a)}{h} = \lim_{h\to 0} \frac{c - c}{h} = \lim_{h\to 0} \frac{0}{h} = \lim_{h\to 0} 0 = 0.\text{”}$$

But this presentation, starting with the expression $f'(a)$, presupposes that the limit that defines the derivative actually exists. We'll come across *plenty* of examples where the limit doesn't exists. The more correct mathematical approach is to do the algebraic manipulation to $(f(a+h)-f(a))/h$ *first*, and then, when a nicer expression has been arrived at, whose limit near 0 can easily be computed, introduce the limit symbol. That was how we approached the write-up above, although frequently in the future we will be sloppy and write "lim $---=$" before we've formally decided that the limit exists.[108]

Linear function Let $f(x) = mx+b$ for real constants m, b. Since the graph of f is a straight line with slope m, it should be its own tangent at all points, and so the derivative at all points should be m. We verify this. As discussed in the last example, we will do this slightly sloppily, beginning by assuming that the limit exists.

For each real a we have

$$\begin{aligned} f'(a) &= \lim_{h\to 0} \frac{f(a+h)-f(a)}{h} \\ &= \lim_{h\to 0} \frac{m(a+h)+b-(ma+b))}{h} \\ &= \lim_{h\to 0} \frac{mh}{h} \\ &= \lim_{h\to 0} m \\ &= m, \end{aligned}$$

as we suspected. The key line was the second from last — dividing above and below by h was valid, because we never consider $h = 0$ when calculating the limit near 0. In the last line, we use that the constant function is continuous everywhere, so the limit can be evaluated by direct evaluation.

In this example, f' is the constant m function, on the same domain ($\mathbb{R}$) as f.

Quadratic function Let $f(x) = x^2$. There is every reason to expect that this function is differentiable everywhere — its graph, on any graphing utility, appears smooth. There is little reason to expect a particular value for the derivative, as we did in the last two examples[109]. We just go through the calculation, and see what comes out. This time, we'll do it in what might be called the "proper" way, not initially assuming the existence of the derivative.

[108] We actually already did this above, when we wrote "and so

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

[109] Not entirely true — when we motivate the product rule for differentiation, we see a good reason to expect that the derivative of $f(x) = x^2$ at a is $2a$.

For each real a, and for $h \neq 0$, we have

$$\begin{aligned}\frac{(a+h)^2 - a^2}{h} &= \frac{a^2 + 2ah + h^2 - a^2}{h} \\ &= \frac{2ah + h^2}{h} \\ &= 2a + h.\end{aligned}$$

Since $\lim_{h\to 0}(2a+h)$ evidently exists and equals $2a$, we conclude that $\lim_{h\to 0}((a+h)^2 - a^2)/h$ exists and equals $2a$, and so for all real a,

$$f'(a) = 2a.$$

In this example, f' is the linear function $x \mapsto 2x$, on the same domain ($\mathbb{R}$) as f.

Power function In general, calculating the derivative of $f(x) = x^n$ for $n \in \mathbb{N}$ at arbitrary real a is not much harder than in the special case of $n = 2$, just as long as we bring the right tool to the algebraic manipulation. Since we'll be faced with the expression $(a+h)^n - a^n$, it seems that the Binomial Theorem is probably the[110] right tool.

For each real a, and for $h \neq 0$, we have

$$\begin{aligned}\frac{(a+h)^n - a^n}{h} &= \frac{a^n + \binom{n}{1}a^{n-1}h + \binom{n}{2}a^{n-2}h^2 + \cdots + \binom{n}{n-1}ah^{n-1} + h^n - a^n}{h} \\ &= \frac{\binom{n}{1}a^{n-1}h + \binom{n}{2}a^{n-2}h^2 + \cdots + \binom{n}{n-1}ah^{n-1} + h^n}{h} \\ &= \binom{n}{1}a^{n-1} + \binom{n}{2}a^{n-2}h + \cdots + \binom{n}{n-1}ah^{n-2} + h^{n-1}.\end{aligned}$$

Now

$$\lim_{h\to 0}\binom{n}{1}a^{n-1} = \binom{n}{1}a^{n-1} = na^{n-1},$$

while

$$\lim_{h\to 0}\binom{n}{2}a^{n-2}h = \lim_{h\to 0}\binom{n}{3}a^{n-3}h^2 = \cdots = \lim_{h\to 0}\binom{n}{n-1}ah^{n-2} = \lim_{h\to 0}h^{n-1} = 0,$$

all these facts following from our previous work on continuity. So by the sum part of the sum/product/reciprocal theorem for limits, we conclude that

$$\lim_{h\to 0}\binom{n}{1}a^{n-1} + \binom{n}{2}a^{n-2}h + \cdots + \binom{n}{n-1}ah^{n-2} + h^{n-1} = na^{n-1}.$$

But then it follows that

$$\lim_{h\to 0}\frac{(a+h)^n - a^n}{h} = na^{n-1};$$

[110] or at least a

in other words, f is differentiable for all real a, with

$$f'(a) = na^{n-1}.$$

In this example, f' is the power function $x \mapsto nx^{n-1}$, on the same domain ($\mathbb{R}$) as f.

Quadratic reciprocal One final example in the vein of the previous ones: $f(x) = 1/x^2$. As long as $a \neq 0$, we have

$$\begin{aligned} f'(a) &= \lim_{h\to 0} \frac{f(a+h) - f(a)}{h} \\ &= \lim_{h\to 0} \frac{\frac{1}{(a+h)^2} - \frac{1}{a^2}}{h} \\ &= \lim_{h\to 0} \frac{a^2 - (a+h)^2}{(a+h)^2 a^2 h} \\ &= \lim_{h\to 0} \frac{-2ah - h^2}{(a+h)^2 a^2 h} \\ &= \lim_{h\to 0} \frac{-2a - h}{(a+h)^2 a^2} \\ &= \frac{-2a}{a^2 a^2} \\ &= \frac{-2}{a^3}. \end{aligned}$$

In this example, f' is the function $x \mapsto -2x^3$, on the same domain ($\mathbb{R} \setminus \{0\}$) as f.

Absolute value function Here we consider $f(x) = |x|$. We would strongly expect that for $a > 0$, we have f differentiable at a, with derivative 1, because a little neighborhood around such a, we have that $f(x) = x$; indeed, for $a > 0$ we have that for all sufficiently small h (say, for all $h < a/2$)

$$\frac{|a+h| - |a|}{h} = \frac{a+h-a}{h} = \frac{h}{h} = 1,$$

and so $\lim_{h\to 0}(|a+h| - |a|)/h = \lim_{h\to 0} 1 = 1$. We can similarly verify that for all $a < 0$, $f'(a) = -1$. But at $a = 0$, something different happens:

$$\frac{|0+h| - |0|}{h} = \frac{|h|}{h},$$

and we know that $\lim_{h\to 0} |h|/h$ *does not exist*. So, this is our first example of a function that is *not* always differentiable; the domain of f' here is $\mathbb{R} \setminus \{0\}$ while the domain of f is $\mathbb{R}$.

We should not have expected $f(x) = |x|$ to be differentiable at 0, as there is no coherent "direction" that the graph of the function is going near 0 — if we look to the right of

zero, it is increasing consistently at rate 1, while if we look to the left of zero, it is *decreasing* consistently at rate 1. Nor is there obviously an unambiguous tangent line.

The comments in the previous paragraph suggest that it might be useful to define notions of right and left derivatives, as we did with continuity. Say that f is *right differentiable* at a, or *differentiable from the right*, or *differentiable from above*, if

$$\lim_{h\to 0^+} \frac{f(a+h)-f(a)}{h}$$

exists, and if it does, denote by $f'_+(a)$ the value of the limit. Say that f is *left differentiable* at a, or *differentiable from the left*, or *differentiable from below*, if

$$\lim_{h\to 0^-} \frac{f(a+h)-f(a)}{h}$$

exists, and if it does, denote by $f'_-(a)$ the value of the limit. It's a (hopefully routine, by now) exercise to check that

> f is differentiable at a if and only if f is both left and right differentiable at a, and the two one-sided derivatives have the same value; in this case that common value is the value of the derivative at a.

So, for example, with $f(x)=|x|$ we have

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

while

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

so that f is not differentiable at 0.

Piecewise defined functions Consider

$$f(x) = \begin{cases} x^2 & \text{if } x<1 \\ ax+b & \text{if } x\geq 1 \end{cases}$$

where a, b are some constants. What choices of a, b make f both continuous and differentiable on all reals?

Well, clearly f is both continuous and differentiable on all of $(-\infty, 1)$ and on all of $(1,\infty)$. What about at 1? We have

$$\lim_{x\to 1^+} f(x) = \lim_{x\to 1^+} ax+b = a+b$$

and

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

so in order for f to be continuous at 1, we require $a+b=1$. For differentiability, at 1, we have

$$\lim_{h\to 0^+}\frac{f(1+h)-f(1)}{h}=\lim_{h\to 0^+}\frac{a+ah+b-(a+b)}{h}=\lim_{h\to 0^+}\frac{ah}{h}=\lim_{h\to 0^+}a=a,$$

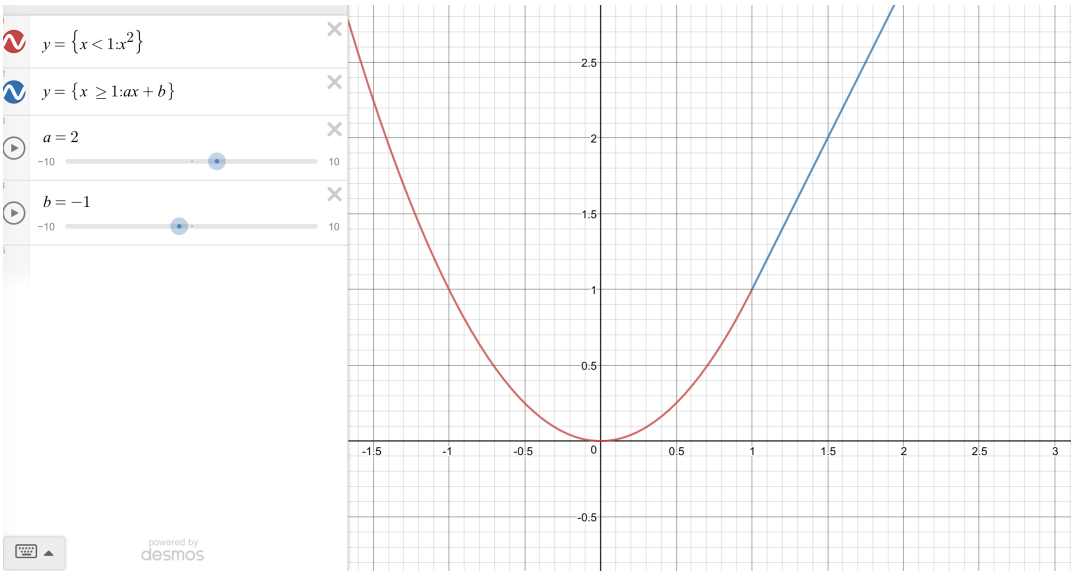
and (recalling that $a+b=1$, since we require f to be continuous at 1)

$$\lim_{h\to 0^-}\frac{f(1+h)-f(1)}{h}=\lim_{h\to 0^-}\frac{(1+h)^2-(a+b)}{h}=\lim_{h\to 0^-}\frac{2h+h^2}{h}=\lim_{h\to 0^-}(2+h)=2.$$

So, for f to be differentiable at 1 we require $a=2$; and since $a+b=1$ this says $b=-1$. The function we are considering is thus

$$f(x)=\begin{cases} x^2 & \text{if } x<1\\ 2x-1 & \text{if } x\geq 1.\end{cases}$$

Here is the graph. It shows the two pieces not justing fitting together at 1, but fitting together *smoothly*.



The square root function Consider $f(x)=\sqrt{x}$, defined on $[0,\infty)$. Tho compute its derivative at any $a\in(0,\infty)$ we proceed in the usual way:

$$\begin{aligned}\lim_{h\to 0}\frac{\sqrt{a+h}-\sqrt{a}}{h} &= \lim_{h\to 0}\left(\frac{\sqrt{a+h}-\sqrt{a}}{h}\right)\left(\frac{\sqrt{a+h}+\sqrt{a}}{\sqrt{a+h}+\sqrt{a}}\right)\\ &= \lim_{h\to 0}\frac{(a+h)-a}{h(\sqrt{a+h}+\sqrt{a})}\\ &= \lim_{h\to 0}\frac{1}{(\sqrt{a+h}+\sqrt{a})}\\ &= \frac{1}{2\sqrt{a}}.\end{aligned}$$

So f is differentiable on $(0,\infty)$, with derivative $f'(a) = 1/2\sqrt{a}$.

What about at 0? Because f is not defined for negative inputs, we must consider a one sided derivative, in particular the right derivative, and we have

$$\lim_{h\to 0^+} \frac{\sqrt{0+h}-\sqrt{0}}{h} = \lim_{h\to 0^+} \frac{1}{\sqrt{h}}.$$

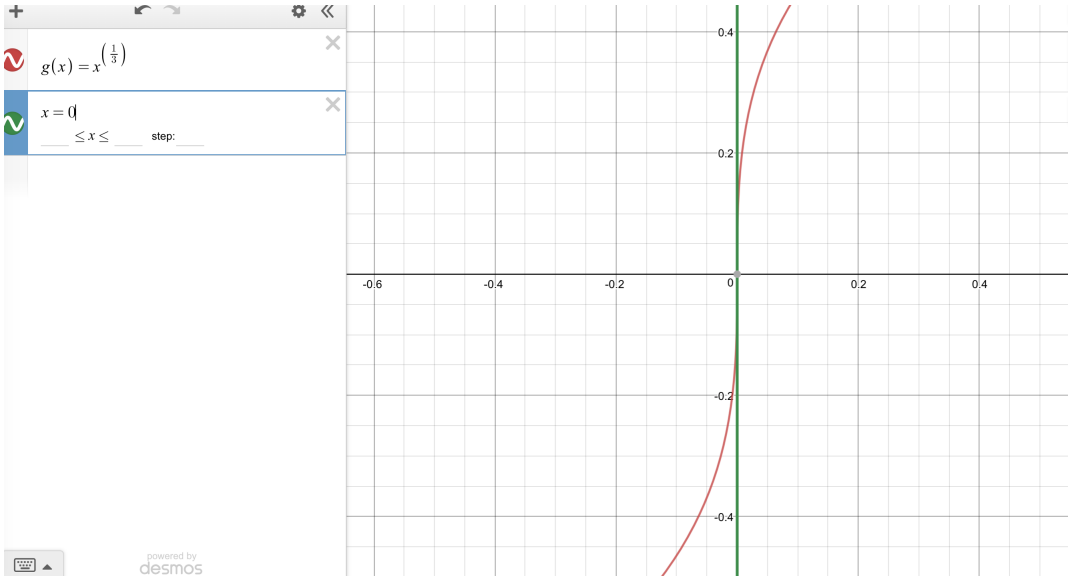
This limit does not exist, so f is not left differentiable at 0.

A more dramatic example in a similar vein comes from considering $g(x) = x^{1/3}$, which has all of $\mathbb{R}$ as its domain. By a similar calculation to above, we get that f is differentiable at all $a \neq 0$, with derivative $f'(a) = 1/(3a^{2/3})$. At $a = 0$ we have

$$\lim_{h\to 0} \frac{(0+h)^{1/3}-0^{1/3}}{h} = \lim_{h\to 0} \frac{1}{h^{2/3}},$$

which again does not exist, so g is not differentiable at 0.

What's odd about this is that from a drawing of the graph of g, it seems that g has an unambiguous slope/tangent line at the point $(0,0)$:



It is the *vertical* line, $x = 0$. We are failing to see this in the math, because the vertical line has infinite slope, and we have no real number that captures that.[111]

$\sin(1/x)$ and variants Consider the three functions

$$f_1(x) = \sin(1/x),\ x \neq 0$$
$$f_2(x) = x\sin(1/x),\ x \neq 0$$
$$f_3(x) = x^2\sin(1/x),\ x \neq 0,$$

[111] Shortly we will talk about "infinite limits" and rectify this deficiency.

with $f_1(0) = f_2(0) = f_3(0) = 0$.

All three of these functions have domain $\mathbb{R}$. What about the domains of their derivatives? Presumably they are all differentiable at all non-zero points.[112]

What about at 0? For f_1, using $f_1(0) = 0$ we have (if the limit exists)

$$f'(0) = \lim_{h \to 0} \frac{\sin(1/h)}{h}.$$

It's a slightly tedious, but fairly instructive, exercise to verify that this limit does not exist; so f_1 is not differentiable at 0 (and maybe we shouldn't have expected it to be: it's not even continuous at 0).

For f_2, which is continuous at 0, we have a better chance. But

$$\lim_{h \to 0} \frac{f_2(0+h) - f_2(0)}{h} = \lim_{h \to 0} \frac{h \sin(1/h)}{h} = \lim_{h \to 0} \sin(1/h),$$

which does not exist; so f_2 is not differentiable at 0, either.

For f_3, however, we have

$$f_3'(0) = \lim_{h \to 0} \frac{h^2 \sin(1/h)}{h} = \lim_{h \to 0} h \sin(1/h) = 0,$$

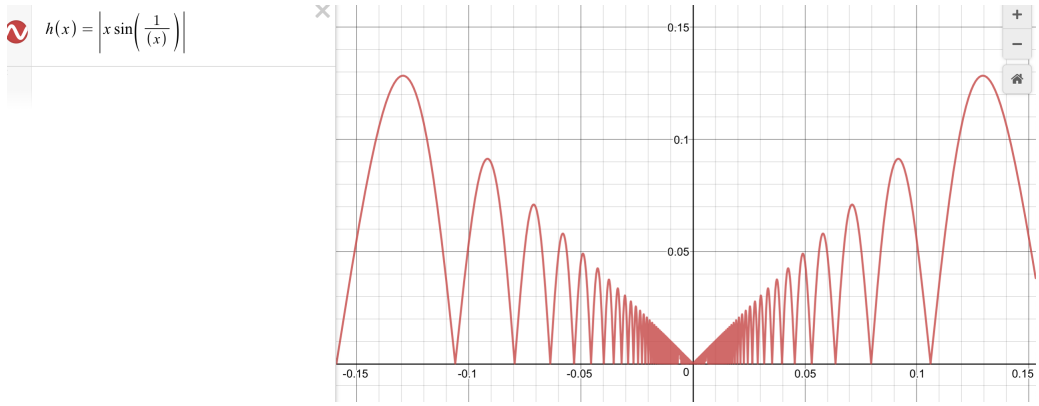
so f_3 *is* not differentiable at 0, with derivative 0.

We will return to this example when considering a function which is k times differentiable, but not $(k+1)$ times differentiable.

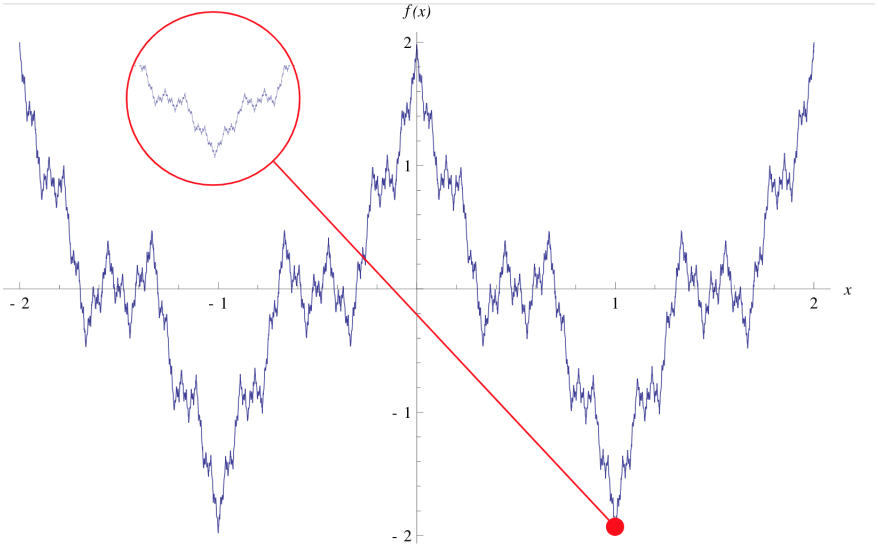
Weierstrass function It's easy to come up with an example of a function that is defined *everywhere* differentiable *nowhere* — the Dirichlet function works well. But this seems a cheat. The Dirichlet function is nowhere continuous, and if we imagine differentiability to be a more stringent notion of smoothness than continuity, then we might expect that non-continuous functions are for some fairly trivial reason non-differentiable.[113]

So what about a *continuous* function that is nowhere differentiable? It's fairly easy to produce an example of a function that is continuous everywhere, but that has infinitely many points of non-differentiability; even an infinite collection of points that get arbitrarily close to each other. For example, the function $f(x) = x \sin(1/x)$ is continuous everywhere on its domain $\mathbb{R} \setminus \{0\}$, and presumably differentiable everywhere on its domain; but if we take its absolute value we get a function that is still continuous, but has a sharp point (so a point of non-differentiability) at each place where it touches the axis (see figure below). In other words, if $h(x) = |x \sin(1/x)|$ then while $\text{Domain}(h) = \mathbb{R} \setminus \{0\}$ we have $\text{Domain}(h') = \mathbb{R} \setminus \{0, \pm 1/\pi, \pm 2/\pi, \pm 3/\pi, \ldots\}$.

[112] We will verify this informally soon, using or informal/geometric definition of the trigonometric functions.
[113] This is true, as we'll see in a moment.



It is far less easy to come with an example of a function which is continuous *everywhere*, but differentiable *nowhere*; nor is it easy to imagine what such a function could look like. There *are* examples[114], but they are not as easy to explain as the Dirichlet function (our example of a function that is defined everywhere but continuous nowhere). The first such example was found by Karl Weiestrass in 1872, and so is traditionally called the *Weierstrass function*. It is infinitely jagged, and displays a self-similarity or fractal behavior: zoom in on any portion of the graph, and you see something very similar to the global picture (see figure below).



Higher derivatives Let f be a function on some domain D. As we have been discussing in these examples, there may be some points in the domain of f at which f is differentiable, leading to a function f', the derivative function, which might have a smaller domain than D. But the function f' may itself be differentiable at some points, leading to a function $(f')'$ (which might have a smaller domain than that of f'). Rather than

[114] In fact, in a quite precise sense *most* continuous function are nowhere differentiable.

working with this ungainly notation, we denote the second derivative by f''. Formally, the second derivative of a function f at a point a is defined to be

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

assuming that limit exists — which presupposes that f is both defined at and near a, and is differentiable at and near a.

We may further define the third derivative function, denoted f''', as the derivative of the second derivative function f''. And we can go on; but even without the parentheses, this "prime" notation gets a little ungainly, quickly. We use the notation $f^{(k)}$ to denote the kth derivative of f, for and natural number k (so $f^{(3)} = f'''$ and $f^{(1)} = f'$). By convention, $f^{(0)} = f$.

Physically, if $f(t)$ is the position of a particle at time t, then

- $f'(t)$ is velocity at time t (rate of change of position with respect to time);
- $f''(t)$ is the acceleration at time t (rate of change of velocity with respect to time);
- $f'''(t)$ is the *jerk* at time t (rate of change of acceleration with respect to time), and so on.

Consider, for example, $f(x) = 1/x$, with domain all reals except 0. We have

- $f'(x) = -1/x^2$, domain $\mathbb{R} \setminus \{0\}$;
- $f''(x) = 2/x^3$, domain $\mathbb{R} \setminus \{0\}$;
- $f'''(x) = -6/x^3$, domain $\mathbb{R} \setminus \{0\}$, and so on.

As another example, consider the function that is obtained by splicing the cube function and the square function, i.e.

$$f(x) = \begin{cases} x^3 & \text{if } x \leq 0 \\ x^2 & \text{if } x \geq 0. \end{cases}$$

By looking at one sides limits, it is easy to check that f is continuous at 0, differentiable at 0, and even twice differentiable at 0, but *not* thrice differentiable. A homework problem asks for an example of a function that, at least at some points, is differentiable k times, but not $k+1$ times.

Before moving on to some more theoretical properties of the derivative, we mention one more motivation. The tangent line to a curve at a point, as we have defined it, seems to represent a good approximation to the curve, at least close to the point of tangency. Now the tangent line is a *straight* line, and it is relatively easy to calculate exact values of points along a straight line, while the graph of a typical function near a point may well be *curved*, and the

graph may be that of a function whose values are hard to compute (we may be dealing with $f(x) = x \sin x / \cos^2(\pi(x + 1/2))$, for example).

This suggests that it might be fruitful to use the point (x, y) on the tangent line at $(a, f(a))$ to the graph of function f, as an approximation for the point $(x, f(x))$ (that's actually on the graph); and to use y as an approximation for $f(x)$. It seems like thais might be particularly useful if x is relatively close to a.

Recalling that the equation of the tangent line at $(a, f(a))$ to the graph of function f is $y = f(a) + f'(a)(x - a)$, we make the following definition:

Linearization of a function at a point The *linearization* $L_{f,a}$ of a function f at a at which the function is differentiable is the function $L_{f,a} : \mathbb{R} \to \mathbb{R}$ given by

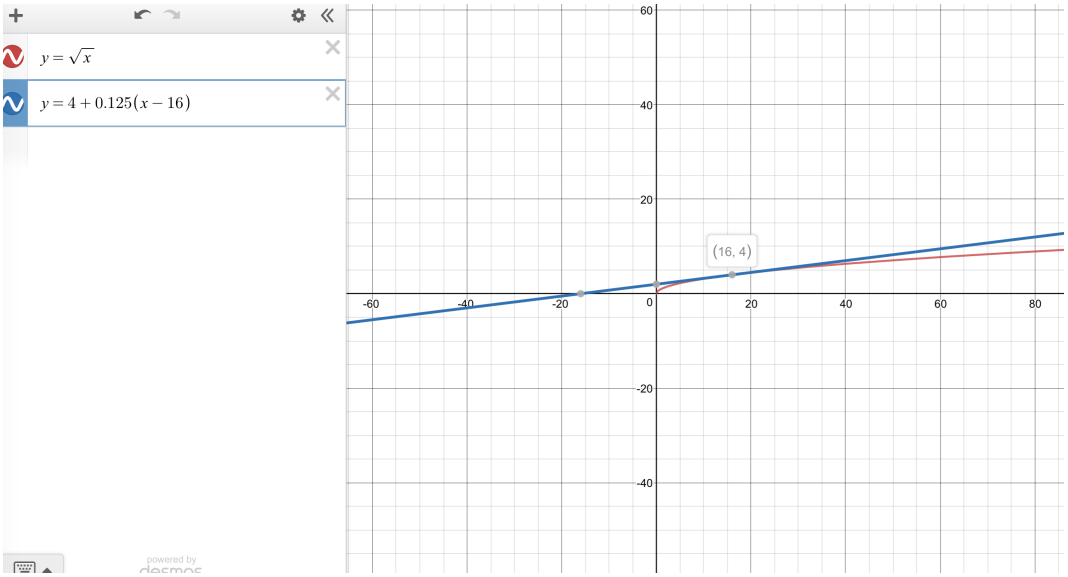
$$L_{f,a}(x) = f(a) + f'(a)(x - a).$$

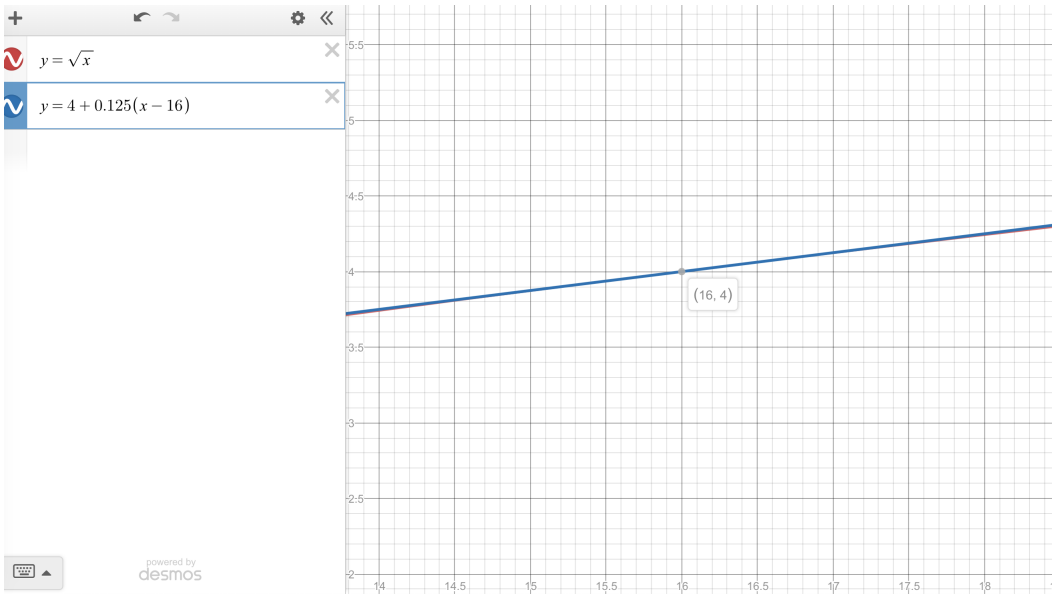
Notice that the linearization of f at a agrees with f at a: $L_{f,a}(x) = f(a)$. The intuition of the definition is that *near* a, $L_{f,a}(x)$ is a good approximation for $f(x)$, and is (often) much easier to calculate than $f(x)$.

The linearization is particularly useful if the point a is such that both $f(a)$ and $f'(a)$ are particularly "nice". Here's an example: consider $f(x) = \sqrt{x}$. It's not in general easy to calculate values of f, but there are some places where it *is* easy, namely at those x which are perfect squares of integers $(1, 4, 9, \ldots)$. So, take $a = 4$. We have $f(a) = f(16) = 4$, and, since $f'(x) = 1/(2\sqrt{x})$, we have $f'(a) = f'(16) = 1/8 = 0.125$. That means that the linearization of f at 16 is the function

$$L_{f,16}(x) = 4 + 0.125(x - 16).$$

Here are two pictures showing the graphs of both f and $L_{f,16}$, one from a wide view, and the other focussing in on what happens near the point $(16, 4)$. Notice that near to 16 on the x-axis, the two graphs are very close to each other; this is especially apparent from the second, close-up, picture, where is is almost impossible to tell the two graphs apart.





If we use $L_{f,16}$ to approximate $\sqrt{14}$, we get

$$\sqrt{14} = f(14) \approx L_{f,16}(14) = 4 + 0.125(14 - 16) = 3.75.$$

This is not too bad! A calculator suggests that $\sqrt{14} = 3.7416\cdots$, so the linearization gives an answer with an absolute error of around 0.0083, and a relative error of around 2.2%.

Of course, the situation won't always be so good: if we use $L_{f,16}$ to approximate $\sqrt{100}$, we get an estimate of $4 + 0.125(100 - 16) = 14.5$, which differs from the true value (10) by a large amount[115]; and if we use it to estimate $\sqrt{-8}$ we get an estimate of $4 + 0.125(-8 - 16) = 1$ for a quantity that doesn't exist!

This leads to the first of two natural questions to ask about the linearization (the second, you are probably already asking yourself):

- *How good is the linearization as an approximation tool, precisely?*: It's easy to approximate *any* function, at *any* possible input: just say "7". An approximation is only useful if it comes with some guarantee of its accuracy, such as "$\sqrt{14}$ is approximately 3.75; and this estimate is accurate to error ± 0.2", meaning that "$\sqrt{14}$ is certain to lie in the interval $(3.55, 3.95)$". The linearization does come with a guarantee of accuracy, but we will not explore it until next semester, when we consider the much more general (and powerful) Taylor polynomial.

- *Why use a scheme like this, to estimate the values of complicated functions, when we could just use a calculator?*: To answer this, ask another question: how does a *calculator* figure out the values of complicated functions?!?

Here's a theoretical justification for the linearization as a tool for approximating the values of a function, near the point around which we are linearizing: it's certainly the case

[115] Not too surprising, since by most measures 100 is *not* close to 16.

that

$$\lim_{x\to a}(f(x)-L_{f,a}(x))=\lim_{x\to a}f(x)-\lim_{x\to a}L_{f,a}(x)=f(a)-f(a)=0,$$

which says that as x approaches a, the linearization gets closer and closer to f (makes smaller and smaller error). But this is true of lots and lots of candidates for a simple approximating function; in particular it's true about the constant function $f(a)$, but something as naive as that can hardly be considered as a good tool for approximating the function f away from a (it takes into account nothing except the value of the function at a). The linearization takes a little more into account about the function; it consider the direction in which the graph of the function is moving, at the point $(a, f(a))$. As a consequence of this extra data being built into the linearization, we have the following fact:

$$\begin{aligned}\lim_{x\to a}\frac{f(x)-L_{f,a}(x)}{x-a} &= \lim_{x\to a}\frac{f(x)-f(a)-f'(a)(x-a)}{x-a}\\ &= \lim_{x\to a}\left(\frac{f(x)-f(a)}{x-a}-f'(a)\right)\\ &= \lim_{x\to a}\frac{f(x)-f(a)}{x-a}-\lim_{x\to a}f'(a)\\ &= f'(a)-f'(a)\\ &= 0.\end{aligned}$$

In other words, not only does the value of the linearization get closer and closer to the value of f as x approaches a, but also

> the linearization get closer and closer to f as x approaches a, even when the error is measured relative to $x-a$

(a stronger statement, since $x-a$ is getting smaller as x approaches a).[116]

8.4 The derivative of sin

Here we go through an informal calculation of the derivative of the sin function. It is informal, because we have only informally defined sin. Next semester, we will give a proper definition of sin (via an integral), from which all of its basic properties will emerge quite easily.

Along the way, we will derive the important and non-obvious trigonometric limit

$$\lim_{h\to 0}\frac{\sin h}{h}=?.$$

Because we haven't yet rigorously defined sin, the treatment here will be quite casual and intuitive. But at least it will give a sense of the behavior of the trigonometric functions vis a vis the derivative, and allow us to add sin and cos to the army of functions that we can differentiate.

Recall how we (informally, geometrically) defined the trigonometric functions sin and cos:

[116] The linearization is actually the *unique* linear function with this property. We'll have much more to say about this next semester, when we look at Taylor series.

If P is a point on the unit circle $x^2 + y^2 = 1$, that is a distance θ from $(1, 0)$, measured along the circle in a counterclockwise direction (starting from P), then the x-coordinate of P is $\cos\theta$, and the y-coordinate is $\sin\theta$.

It's typical to refer to the angle made at $(0, 0)$, in going from P to $(0, 0)$ to $(1, 0)$, as θ; see the picture below.

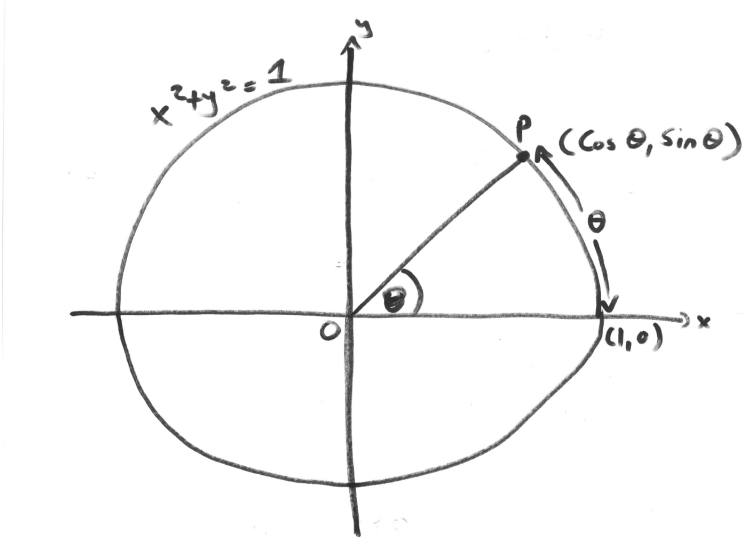


FIGURE 1 → DEF. OF SIN, COS.

And once we have said what "angle" means, it is easy to see (by looking at ratios of side-lengths of similar triangles) that this definition of sin and cos coincides with the other geometric definition you've seen: if triangle ABC has a right angle at B, and an angle θ at A, then

$$\sin\theta = \frac{\text{BC}}{\text{AC}} = \frac{\text{opposite}}{\text{hypothenuse}}, \qquad \cos\theta = \frac{\text{BA}}{\text{AC}} = \frac{\text{adjacent}}{\text{hypothenuse}}.$$

What is the derivative of sin? By definition, $\sin'\theta$ is

$$\lim_{h\to 0} \frac{\sin(\theta + h) - \sin\theta}{h}.$$

There's no obvious algebraic manipulation we can do here to make this limit easy to calculate. We need the *sine sum formula*:

for any angles α, β, $\sin(\alpha + \beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta$.

Here is a picture, that leads to a proof of this formula (square brackets indicate right angles):

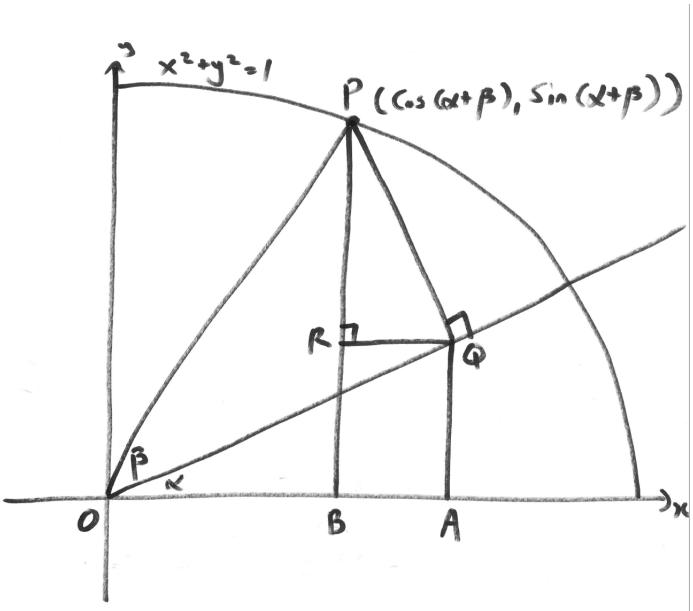


FIGURE 2 → SIN SUM FORMULA

Question 1: Why does this prove

$$\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta?$$

Answer:

- First argue that angle RPQ is α (look first at OQA, then RQO, then RQP, then RPQ).
- Argue (from the definition of sin, and similar triangles) that PQ is $\sin\beta$, and so PR is $\cos\alpha\sin\beta$.
- Argue similarly that OQ is $\cos\beta$, and so AQ is $\sin\alpha\cos\beta$.
- Since AQ is the same as RB, and since PB is known, conclude that

$$\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta.$$

Of course, this geometric proof only works for $\alpha, \beta \geq 0$, $\alpha+\beta \leq \pi/2$; but similar pictures can be drawn for all other cases.[117]

Now using the sin sum formula, we have (throughout assuming that all the various limits

[117]Here is another picture that justifies the trigonometric sum formulae, due to Tamás Görbe:

in fact exist):

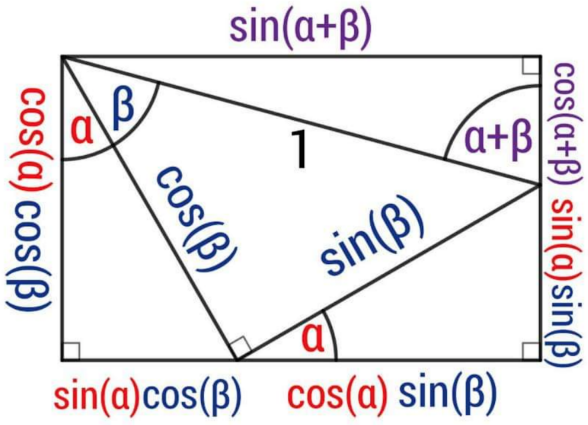
$$\begin{aligned}
\sin'\theta &= \lim_{h\to 0}\frac{\sin(\theta+h)-\sin\theta}{h}\\
&= \lim_{h\to 0}\frac{\sin\theta\cos h+\cos\theta\sin h-\sin\theta}{h}\\
&= \lim_{h\to 0}\frac{\sin\theta(\cos h-1)+\cos\theta\sin h}{h}\\
&= \lim_{h\to 0}\frac{\sin\theta(\cos h-1)}{h}+\lim_{h\to 0}\frac{\cos\theta\sin h}{h}\\
&= \sin\theta\lim_{h\to 0}\frac{\cos h-1}{h}+\cos\theta\lim_{h\to 0}\frac{\sin h}{h}.
\end{aligned}$$

We have reduced to two limits, neither of which look any easier than the one we started with! But, it turns out they they are essentially the same limit:

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

In the second from last line we used continuity of sin and cos, and in the last line, we used the (as yet unjustified) fact that $(\sin h)/h$ actually tends to a limit, as h nears 0. On this assumption, we get

$$\sin'\theta = \cos\theta\lim_{h\to 0}\frac{\sin h}{h}.$$



So now we have one limit left to consider, and it is a *little* bit simpler than the limit we started with.

We now claim that

$$\lim_{h\to 0} \frac{\sin h}{h} = 1.$$

Here is a picture, that leads to a proof of this claim:

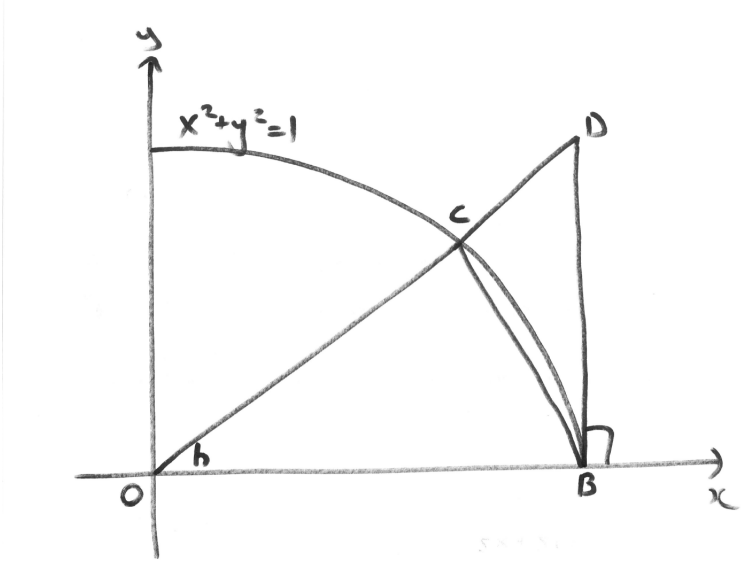


FIGURE 3 → LIMIT OF $\frac{\sin h}{h}$

Question 2: Why does this prove

$$\lim_{h\to 0} \frac{\sin h}{h} = 1?$$

Answer:

- What is the area of the triangle OBC?
- What is the area of the wedge of the circle between OC and OB?
- What is the area of the triangle OBD?
- What is the inequality relation between these three areas?
- Conclude that
$$\frac{\sin h}{2} \leq \frac{h}{2} \leq \frac{\sin h}{2\cos h}.$$
- Conclude that
$$\cos h \leq \frac{\sin h}{h} \leq 1.$$

- Use continuity of cos, and the squeeze theorem, to get the result.

Of course, the picture only shows that $\lim_{h\to 0^+}(\sin h)/h = 1$; but a similar picture gives the other one-sided limit.

We now get to conclude that

$$\sin'\theta = \cos\theta.$$

What about the derivative of cos? We could play the same geometric game to derive

$$\cos'\theta = -\sin\theta;$$

after we've seen the chain rule, we'll give an alternate derivation.

8.5 Some more theoretical properties of the derivative

In this section, rather than looking at specific examples to bring out properties of the derivative, we derive some more general properties (that, incidentally, will allow us to discover many more specific examples of derivatives).

We have observed intuitively that differentiability at a point is a more stringent "smoothness" condition that simple continuity; in other words, it's possible for a function to be continuous at a point but not differentiable there ($f(x) = |x|$ at $x = 0$ is an example), but it should not be possible for a function to be differentiable at a point without first being continuous. We'll now turn this intuition into a proven fact.

The hard way to do this is to start with a function which is defined at some point a, but not continuous there, and then argue that it cannot be differentiable at that point (the non-existence of the continuity limit somehow implying the non-existence of the differentiability limit). This is the hard way, because there are many different ways that a function can fail to be continuous, and we would have to deal with all of them in this approach.

The soft way is to go via the contrapositive, and prove that if f is differentiable at a point, then it most also be continuous there (the existence of the differentiability limit somehow implying the existence of the continuity limit). This is easier, because there's only one way for a limit to exist; and it immediately implies that failure of continuity implies failure of differentiability.

Claim 8.1. *Suppose that f is defined at and near a, and is differentiable at a. Then f is continuous at a.*

Proof: Since f is differentiable at a, we have that for some real number a,

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

But also, $\lim_{h\to 0} h = 0$. By the product part of the sum/product/reciprocal theorem for limits, we can conclude that

$$\begin{aligned}\lim_{h\to 0}(f(a+h)-f(a)) &= \lim_{h\to 0}\left(\frac{f(a+h)-f(a)}{h}\right)h \\ &= \left(\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}\right)\left(\lim_{h\to 0} h\right) \\ &= f'(a)\cdot 0 \\ &= 0,\end{aligned}$$

so (by the sum part of the sum/product/reciprocal theorem for limits)

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

which says that f is continuous at a.[118] □

The take-away from this is:

continuity is necessarily for differentiability, but not sufficient.

We now derive some identities that will allow us to easily compute some new derivatives from old ones.

Claim 8.2. *Suppose f and g are functions that are both differentiable at some number a, and that c is some real constant. Then both $f+g$ and cf are differentiable at a, with*

$$(f+g)'(a) = f'(a) + g'(a)$$

and

$$(cf)'(a) = cf'(a).$$

Proof: Both statements follow quickly from previously established facts about limits. We have that

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

and

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

and so

$$\begin{aligned}f'(a)+g'(a) &= \lim_{h\to 0}\frac{f(a+h)-f(a)}{h} + \lim_{h\to 0}\frac{g(a+h)-g(a)}{h} \\ &= \lim_{h\to 0}\left(\frac{f(a+h)-f(a)}{h} + \frac{g(a+h)-g(a)}{h}\right) \\ &= \lim_{h\to 0}\frac{f(a+h)+g(a+h)-f(a)-g(a)}{h} \\ &= \lim_{h\to 0}\frac{(f+g)(a+h)-(f+g)(a)}{h},\end{aligned}$$

[118] This is not exactly the definition of continuity at a; but you can prove that this is an equivalent definition, just as we proved earlier that $\lim_{b\to a}(f(b)-f(a))/(b-a)$ is the same as $\lim_{h\to 0}(f(a+h)-f(a))/h$.

which exactly says that $f+g$ is differentiable at a, with derivative $f'(a)+g'(a)$.

Similarly,

$$\begin{aligned} cf'(a) &= c\lim_{h\to 0}\frac{f(a+h)-f(a)}{h} \\ &= \lim_{h\to 0}\frac{c(f(a+h)-f(a))}{h} \\ &= \lim_{h\to 0}\frac{(cf)(a+h)-(cf)(a)}{h} \end{aligned}$$

which exactly says that cf is differentiable at a, with derivative $cf'(a)$. □

Note that the second part of the above claim should alert us to an important insight: we should *not* expect (as we might have, by analogy with similar properties for limits) that the derivative of the product of a pair of functions is the product of the derivatives. If that *were* the case, then, since the derivate of the constant function c is 0, we would have that the derivative of cf at any point is also 0.

Another reason we should not expect that the derivative of the product of a pair of functions is the product of the derivatives is through a dimension analysis. Suppose that $f(x)$ is measuring the height (measured in meters) of a square at time x (measured in seconds), and that $g(x)$ is measuring the width of the square. Then $(fg)(a)$ is measuring the area of the square (measured in meter squared) at time a. Now, the derivative of a function at a can be thought of as measuring the *instantaneous* rate at which the function is changing at a, as the input variable changes — indeed, for any $h\neq 0$ the quantity $(f(a+h)-f(a))/h$ is measuring the average change of f over the time interval $[a,a+h]$ (or $[a+h,a]$, if $h<0$), so if the limit of this ratio exists at h approaches 0, it makes sense to declare that limit to be the instantaneous rate of change at a.

So $(fg)'(a)$ is measuring the rate at which the area of the square is changing, at time a. This is measured in meters squared per second . But $f'(a)g'(a)$, being the product of two rates of changes of lengths, is measured in meters squared per second *squared*. The conclusion is that $(fg)'(a)$ and $f'(x)g'(a)$ have different dimensions, so we should not expect them to be equal in general.

What *should* we expect $(fg)'(a)$ to be? The linearizations of f and g provide a hint. For x near a, we have

$$f(x)\approx L_{f,a}(x)=f(a)+f'(a)(x-a)$$

and

$$g(x)\approx L_{f,a}(x)=g(a)+g'(a)(x-a).$$

Using the linearizations to approximate f and g at $a+h$ we get

$$\begin{aligned} (fg)(a+h) &= f(a+h)g(a+h) \\ &\approx (f(a)+f'(a)h)(g(a)+g'(a)h) \\ &= f(a)g(a)+f'(a)g(a)h+f(a)g'(a)h+f'(a)g'(a)h^2 \end{aligned}$$

and so

$$\frac{(fg)(a+h)-(fg)(a)}{h} \approx f'(a)g(a)+f(a)g'(a)+f'(a)g'(a)h.$$

Considering what happens as $h \to 0$, this strongly suggests that $(fg)'(a) = f'(a)g(a) + f(a)g'(a)$, and this is indeed the case.

Claim 8.3. *(Product rule for differentiation) Suppose f and g are functions that are both differentiable at some number a. Then both fg is differentiable at a, with*

$$(fg)'(a) = f'(a)g(a) + f(a)g'(a).$$

Proof: The proof formalizes the intuition presented above. We begin by assuming that fg is indeed differentiable at a, and try to calculate its derivative.

$$\begin{aligned}
(fg)'(a) &= \lim_{h\to 0} \frac{f(a+h)g(a+h)-f(a)g(a)}{h} \\
&= \lim_{h\to 0} \frac{f(a+h)g(a+h)-f(a+h)g(a)+f(a+h)g(a)-f(a)g(a)}{h} \\
&= \lim_{h\to 0} \frac{f(a+h)\left(g(a+h)-g(a)\right)+\left(f(a+h)-f(a)\right)g(a)}{h} \\
&= \lim_{h\to 0} \frac{f(a+h)\left(g(a+h)-g(a)\right)}{h} + \lim_{h\to 0} \frac{\left(f(a+h)-f(a)\right)g(a)}{h} \\
&= \lim_{h\to 0} f(a+h) \lim_{h\to 0} \frac{g(a+h)-g(a)}{h} + \left(\lim_{h\to 0} \frac{f(a+h)-f(a)}{h}\right) g(a) \\
&= f(a)g'(a)+f'(a)g(a).
\end{aligned}$$

Going backwards through the chain of equalities we see that all manipulations with limits are justified, by

- repeated applications of the sum/product/reciprocal theorem for limits,
- the differentiability of f and g at a, and
- the continuity of f at a (needed for $\lim_{h\to 0} f(a+h) = f(a)$), which itself follows from the differentiability of f at a.

□

As a first application of the product rule, we re-derive the fact that if $f_n : \mathbb{R} \to \mathbb{R}$ ($n \in \mathbb{N}$) is given by $f_n(x) = x^n$, then f_n' (the derivative, viewed as a function) has domain $\mathbb{R}$ and is given by $f_n'(x) = nx^{n-1}$. We prove this by induction on n; we've already done the base case $n = 1$. For $n \geq 2$ we have that $f_n = f_1 f_{n-1}$, so, for each real a, by the product rule

(applicable by the induction hypothesis: f_{n-1} and f_1 are both differentiable at a) we have

$$\begin{aligned} f_n'(a) &= (f_1 f_{n-1})'(a) \\ &= f_1'(a) f_{n-1}(a) + f_1(a) f_{n-1}'(a) \quad \text{(product rule)} \\ &= 1 \cdot a^{n-1} + a \cdot (n-1)a^{n-2} \quad \text{(induction)} \\ &= a^{n-1} + (n-1)a^{n-1} \\ &= na^{n-1}, \end{aligned}$$

which completes the induction step.

We extend this now to negative exponents, giving the presentation in a slightly more streamlined way. For $n \in \mathbb{N}$, define g_n by $g_n(x) = 1/x^n$ (so the domain of g_n is $\mathbb{R} \setminus \{0\}$). We claim that

$$g_n'(x) = \frac{-n}{x^{n+1}}$$

(again with domain $\mathbb{R} \setminus \{0\}$). We prove this by induction on n. The base case $n = 1$ claims that if g_1 is defined by $g_1(x) = 1/x$ then for all non-zero x g_1 is differentiable with derivative $-1/x^2$. This is left as an exercise — it's very similar to a previous example.

For the induction step, assume that $g_{n-1}' = -(n-1)/x^n$. We have

$$\begin{aligned} g_n'(x) &= (g_1 g_{n-1})'(x) \\ &= g_1'(x) g_{n-1}(x) + g_1(x) g_{n-1}'(x) \quad \text{(product rule)} \\ &= \frac{(-1)(1)}{(x^2)(x^{n-1})} + \frac{(1)(-(n-1))}{(x)(x^n)} \quad \text{(induction)} \\ &= \frac{-n}{x^{n+1}}, \end{aligned}$$

which completes the induction step.

Both the sum rule for derivatives and the product rule extend to sums and products of multiple functions. For sums, the conclusion is obvious:

if $f_1, f_2, \ldots, f_n$ are all differentiable at a, then so is $\sum_{k=1}^n f_k$, and

$$\left(\sum_{k=1}^n f_k \right)'(a) = \sum_{k=1}^n f_k'(a).$$

The proof is an easy induction on n, as it is left as an exercise.

For products, the conclusion is less obvious. But once we apply the product rule multiple times to compute the derivative of the product of *three* functions, a fairly clear candidate conclusion emerges. We have (dropping reference to the particular point a, to keep the notation readable)

$$(fgh)' = ((fg)h)' = (fg)'h + (fg)h' = ((f'g) + (fg'))h + (fg)h' = f'gh + fg'h + fgh'.$$

This suggests:

Claim 8.4. *(Product rule for product of n functions)*[119] *If $f_1, f_2, \ldots, f_n$ are all differentiable at a, then so is $\prod_{k=1}^n f_k$, and*

$$\left(\prod_{k=1}^{n} f_k\right)'(a) = \sum_{k=1}^{n} f_1(a)f_2(a)\cdots f_{k-1}(a)f_k'(a)f_{k+1}(a)\cdots f_{n-1}(a)f_n(a).$$

Proof: Strong induction on n, with $n = 1$ trivial and $n = 2$ being the product rule. For $n > 2$, write $f_1 f_2 \cdots f_n$ as $(f_1 \cdots f_{n-1})(f_n)$, and apply the product rule to get

$$(f_1 f_2 \cdots f_n)'(a) = (f_1 \cdots f_{n-1})'(a)f_n(a) + (f_1 \cdots f_{n-1})(a)f_n'(a).$$

The second term here, $(f_1 \cdots f_{n-1})(a)f_n'(a)$, gives the term corresponding to $k = n$ in $\sum_{k=1}^n f_1(a) \cdots f_k'(a) \cdots f_n(a)$. By induction the first term is

$$\begin{aligned}(f_1 \cdots f_{n-1})'(a)f_n(a) &= \left(\sum_{k=1}^{n-1} f_1(a) \cdots f_k'(a) \cdots f_{n-1}(a)\right) f_n(a) \\ &= \sum_{k=1}^{n-1} f_1(a) \cdots f_k'(a) \cdots f_{n-1}(a)f_n(a),\end{aligned}$$

and this gives the remaining terms ($k = 1, \ldots, n-1$) in $\sum_{k=1}^n f_1(a) \cdots f_k'(a) \cdots f_n(a)$. This completes the induction. □

As application of this generalized product rule, we give a cute derivation of

$$h_n'(x) = \frac{1}{n}x^{\frac{1}{n}-1}$$

where $h_n(x) = x^{1/n}$, $n \in \mathbb{N}$. Recalling the previously used noatation $f_1(x) = x$, we have

$$h_n(x)h_n(x)\cdots h_n(x) = f_1(x),$$

where there are n terms in the product. Differentiating both sides, using the product rule for the left-hand side — and noting that, since all the terms in the product are the same, it follows that all n terms in the sum that determines the derivative of the product are the same — we get

$$nh_n'(x)h_n(x)^{n-1} = 1,$$

which, after a little algebra, translates to

$$h_n'(x) = \frac{1}{n}x^{\frac{1}{n}-1}.$$

[119] In words: the derivative of a product of n functions is the derivative of the first times the product of the rest, plus the derivative of the second times the product of the rest, and so on, up to plus derivative of the last times the product of the rest.

Although this is cute, it's a little flawed — we assumed that $h_n'(x)$ exists. So essentially what we have done here is argued that *if* the nth root function is differentiable *then* its derivative must be what we expect it to be. To actually verify that the root function is differentiable, we need to go back to the definition, as we did with the square root function.

Another way to generalize the product rule is to consider higher derivatives of the product of two functions. We have

$$(fg)^{(0)} = fg = f^{(0)}g^{(0)},$$
$$(fg)^{(1)} = (fg)' = fg' + f'g = f^{(0)}g^{(1)} + f^{(1)}g^{(0)},$$

and

$$(fg)^{(2)} = (fg)'' = (fg' + f'g)' = fg'' + 2f'g' + f''g'' = f^{(0)}g^{(2)} + 2f^{(1)}g^{(1)} + f^{(2)}g^{(0)}.$$

There seems to be a pattern here:

$$(fg)^{(n)} = \sum_{k=0}^{n} (\text{SOME COEFFICIENT DEPENDING ON } n \text{ and } k)\ f^{(k)}g^{(n-k)}.$$

A homework problem asks you to find the specific pattern, and prove that is correct for all $n \geq 0$.

After the product rule, comes the quotient rule. We work up to that by doing the reciprocal rule first.

Claim 8.5. *(Reciprocal rule for differentiation) Suppose g is differentiable at some number a. If $g(a) \neq 0$ Then $(1/g)$ is differentiable at a, with*

$$\left(\frac{1}{g}\right)'(a) = -\frac{g'(a)}{g^2(a)}.$$

Proof: Since g is differentiable at a, it is continuous at a, and since $g(a) \neq 0$, $g(x) \neq 0$ for all x in some interval around a. So $1/g$ is defined in an interval around a.

We have

$$\begin{aligned}
\left(\frac{1}{g}\right)'(a) &= \lim_{h\to 0} \frac{\frac{1}{g(a+h)} - \frac{1}{g(a)}}{h} \\
&= \lim_{h\to 0} \frac{g(a) - g(a+h)}{hg(a+h)g(a)} \\
&= -\left(\lim_{h\to 0} \frac{g(a+h) - g(a)}{h}\right)\left(\lim_{h\to 0} \frac{1}{g(a+h)g(a)}\right) \\
&= -\frac{g'(a)}{g^2(a)},
\end{aligned}$$

where, as usual, going backwards through the chain of equalities we see that all manipulations with limits are justified (and all claimed limits exist), by

- repeated applications of the sum/product/reciprocal theorem for limits,
- the differentiability of g at a, and
- the continuity of g at a.

□

The reciprocal rule allows an alternate derivation of the derivative of $g_n(x) = x^{-n}$ ($n \in \mathbb{N}$). Since $g_n(x) = 1/f_n(x)$ (where $f_n(x) = x^n$) and $f_n'(x) = nx^{n-1}$, we have

$$g_n'(x) = -\frac{nx^{n-1}}{x^{2n}} = -\frac{n}{x^{n+1}}.$$

The rule for differentiating the quotient of functions follows quickly from the reciprocal, by combining it with the product rule:

Claim 8.6. *(Quotient rule for differentiation) Suppose f and g are differentiable at some number a. If $g(a) \neq 0$ then (f/g) is differentiable at a, with*

$$\left(\frac{f}{g}\right)'(a) = \frac{f'(a)g(a) - f(a)g'(a)}{g^2(a)}.$$

Proof: We view f/g as $(f)(1/g)$, and apply product and quotient rules to get

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

□

With all of these rules, we can easily differentiate any rational function, and all root functions, but we cannot yet differentiate a function like $f(x) = \sqrt{x^2+1}$, or like $f(x) = \sin^2 x$ (unless we go back to the definition, which is nasty). What we need next is rule saying what happens when we try to differentiate the *composition* of two known functions. This rule is probably the most important one in differential calculus, so we give it its own section.

8.6 The chain rule

Suppose that g is differentiable at a, and that f is differentiable at $g(a)$. We would expect that $f \circ g$, the composition of f and g, should be differentiable at a, but what should the derivative be? To get an intuition, we do what we did before deriving the product rule, and

consider the linearizations of g and f near a and $g(a)$, respectively. We have, for any number a at which g is differentiable,

$$g(a+h) - g(a) \approx g'(a)h, \ (\star)$$

and for any number A at which f is differentiable,

$$f(A+k) - f(A) \approx f'(A)k, \ (\star\star)$$

both approximations presumably reasonable when h and k are small (and in particular, getting better and better as h and k get smaller). So

$$\begin{aligned}
\frac{(f\circ g)(a+h) - (f\circ g)(a)}{h} &= \frac{f(g(a+h)) - f(g(a))}{h} \\
&\approx \frac{f(g(a)+g'(a)h) - f(g(a))}{h} \text{ (applying } (\star)) \\
&\approx \frac{f'(g(a))g'(a)h}{h} \text{ (applying } (\star\star) \text{ with } A = g(a) \text{ and } k = g'(a)h) \\
&= f'(g(a))g'(a).
\end{aligned}$$

This suggest an answer to the question "what is the derivative of a composition?", and it turns out to be the correct answer.

Claim 8.7. *(Chain rule for differentiation) Suppose that g is differentiable at a, and that f is differentiable at $g(a)$. Then $f\circ g$ is differentiable at a, and*

$$(f\circ g)'(a) = f'(g(a))g'(a).$$

A word of warning: $f'(a)$ means the derivative of f at input a; so $f'(g(a))$ is the derivative of f evaluated at $g(a)$, **NOT** the derivative of (f composed with g) evaluated at a (that's $(f\circ g)'(a)$). These two things — $f'(g(a))$ and $(f\circ g)'(a)$ — are usually different. Indeed,

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

while

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

Usually $g(a)+h \neq g(a+h)$ (consider, for example, $g(x) = x^3$: we have

$$g(a)+h = a^3 + h \neq (a+h)^3 = g(a+h)).$$

There is one exception: when $g(x) = x + c$, for some constant c, we have

$$g(a)+h = (a+c)+h = (a+h)+c = g(a+h).$$

One upshot of this is that if $h(x) = f(x+c)$ then $h'(x) = f'(x+c)$. But in general if a function h is defined as the composition $f\circ g$ (here g was $g(x) = x+c$), you need to use the chain rule to evaluate the derivative of h.

Before giving the proof of the chain rule, we present some examples.

- $h_1(x) = \sin^2 x$. This is the composition $h = \text{square} \circ \sin$, where $\text{square}(x) = x^2$. By the chain rule
$$h_1'(x) = \text{square}'(\sin(x)) \sin'(x) = 2 \sin x \cos x.$$

- $h_2(x) = \sin(x^2)$. This is the composition $h = \sin \circ \text{square}$. By the chain rule
$$h_2'(x) = \sin'(\text{square}(x))\square'(x) = (\cos(x^2))2x = 2x \cos x^2.$$
Notice that $h_1'(x) \neq h_2'(x)$ (in general); but since composition is not commutative, there is no particular reason to expect that these two functions h_1, h_2 would end up having the same derivative.

- $f(x) = 1/x^n$. We can view this as the composition "reciprocal after nth power, and find, via chain rule (and the fact that we have already computed the derivatives of both the reciprocal function and the nth power function), that
$$f'(x) = \frac{-1}{(x^n)^2} n x^{n-1} = \frac{-n}{x^{n+1}}.$$
Or, we can view f as the composition "nth power after reciprocal" to get
$$f'(x) = n(1/x)^{n-1} \cdot \frac{-1}{x^2} = \frac{-n}{x^{n+1}}.$$
Either way we get the same answer.

- The derivative of $\cos x$. We have observed that the derivative of cos can be obtained by geometric arguments in a manner similar to the way we derived the derivative of sin. Another approach is to consider the equation $\sin^2 x + \cos^2 x = 1$. The right- and left-hand sides here are both functions of f, so both can be differentiated as functions of x. Using the chain rule for the right-hand side, we get
$$2 \sin x \cos x + 2 \cos x \cos' x = 0$$
or, dividing across by $\cos x$[120],
$$\cos' x = -\sin x$$
(as expected).

- Composition of three functions. Consider $f(x) = (\sin x^3)^2$. This is the composition of squaring (on the outside), sin (in the middle), cubing (inside), so $f_1 \circ f_2 \circ f_3$, where f_1 is the square function, f_2 the sin function, and f_3 the cube function. The chain rule says
$$\begin{aligned}(f_1 \circ f_2 \circ f_3)'(a) &= (f_1 \circ (f_2 \circ f_3))'(a) \\ &= f_1'((f_2 \circ f_3)(a))((f_2 \circ f_3)'(a)) \\ &= f_1'((f_2 \circ f_3)(a)) f_2'(f_3(a)) f_3'(a) \\ &= f_1'(f_2(f_3(a))) f_2'(f_3(a)) f_3'(a).\end{aligned}$$

[120] This is a little informal; it really only works in some interval around 0 where $\cos x \neq 0$. But that's ok; we only have informal definitions of sin and cos to start with. We will shore all this up next semester.

So we get

$$f'(x) = 2(\sin x^3)(\cos x^3)3x^2 = 6x^6(\sin x^3)(\cos x^3).$$

The chain rule pattern (what happens for the derivative of a composition of four, or five, or six, or more, functions, should be fairly clear from this example. In applying the chain rule on a complex composition, you should get used to "working from the outside in".

We now present a formalization of the heuristic argument given above for the chain rule. We will give *two* approaches; the second is perhaps the easier. These are somewhat different justification to the one presented by Spivak.

First proof of chain rule: Note that given our definition of differentiation, saying that g is differentiable at a automatically says that it is defined in some interval (b, c) with $b < a < c$, and saying that f is differentiable at $g(a)$ automatically says that it is defined in some interval (b', c') with $b' < g(a) < c'$.

We start the proof by observing that since g is differentiable at a, we have (for some number $g'(a)$)

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

which says that as h approaches 0, the expression

$$\frac{g(a+h) - g(a)}{h} - g'(a)$$

approaches 0. Denoting this expression by $\varepsilon(h)$ (a function of h, named ε), we get that

$$g(a+h) - g(a) = g'(a)h + \varepsilon(h)h \tag{3}$$

where $\varepsilon(h) \to 0$ as $h \to 0$. The function $\varepsilon(h)$ is defined *near* 0, but not *at* 0; however, the fact that $\varepsilon(h) \to 0$ as $h \to 0$ means that if we extend ε by declaring $\varepsilon(0) = 0$ then not only is ε defined at 0, but it is continuous at 0.

Similarly

$$f(g(a) + k) - f(g(a)) = f'(g(a))k + \eta(k)k \tag{4}$$

where $\eta(k) \to 0$ as $k \to 0$; as before, we extend η to a function that is continuous at 0 by declaring $\eta(0) = 0$. Notice that (4) remains true at $k = 0$.

We now study the expression $f(g(a+h)) - f(g(a))$. Applying (3) we get

$$f(g(a+h)) - f(g(a)) = f(g(a) + g'(a)h + \varepsilon(h)h) - f(g(a)). \tag{5}$$

Now, for notational convenience, set $k = g'(a)h + \varepsilon(h)h$ (notice that this depends on h, so we really should think of k as $k(h)$; but to keep the notation manageable, we will mostly just write k). Applying (4) to (5) we get

$$f(g(a+h)) - f(g(a)) = f(g(a)) + f'(g(a))k + \eta(k)k - f(g(a)) = (f'(g(a)) + \eta(k))k.$$

Now notice that k consists of two terms both of which are multiples of h, so we may divide through by h to obtain

$$\begin{aligned}\frac{f(g(a+h))-f(g(a))}{h} &= (f'(g(a))+\eta(k))(g'(a)+\varepsilon(h)) \\ &= f'(g(a))g'(a)+f'(g(a))\varepsilon(h)+\eta(k)(g'(a)+\varepsilon(h)).\end{aligned}$$

As h approaches 0, it is certainly that case that $f'(g(a))\varepsilon(h)$ approaches 0, since $\varepsilon(h)$ does. If we could show that $\eta(k) \to 0$ as $h \to 0$, then we would also have $\eta(k)(g'(a)+\varepsilon(h)) \to 0$ as $h \to 0$, and so we could conclude that

$$\frac{f(g(a+h))-f(g(a))}{h} \to f'(g(a))g'(a) \quad \text{as} \quad h \to 0,$$

which is exactly what the chain rule asserts.

It seems clear that $\eta(k) \to 0$ as $h \to 0$, since we know that $\eta(k)$ approaches 0 as the argument k approaches 0, and we can see from the equation $k = g'(a)h + \varepsilon(h)h$ that k approaches 0 as h approaches 0. Making this precise requires an argument very similar to the one that we used to show that the composition of continuous functions is continuous.

Let $\varepsilon > 0$ be given (this has nothing to do with the function $\varepsilon(h)$ introduced earlier). Since $\eta(x) \to 0$ as $x \to 0$, there is a $\delta > 0$ such that $0 < |x| < \delta$ implies $|\eta(x)| < \varepsilon$. But in fact, since η is continuous at 0 (and takes the value 0) we can say more: we can say that $|x| < \delta$ implies $|\eta(x)| < \varepsilon$. (In a moment we'll see why this minor detail is important).

Now consider $k = k(h) = g'(a)h + \varepsilon(h)h$. As observed earlier, $k(h)$ approaches 0 as h approaches 0, so, using the definition of limits, there is a $\delta' > 0$ such that $0 < |h| < \delta'$ implies that $|k(h)| < \delta$ (the same δ from the last paragraph). From the last paragraph we conclude that $0 < |h| < \delta'$ in turn implies $|\eta(k(h))| < \varepsilon$, and since $\varepsilon > 0$ was arbitrary, this shows that $\eta(k) \to 0$ as $h \to 0$, finishing the proof of the chain rule.

Notice that if we only knew that $0 < |x| < \delta$ implies $|\eta(x)| < \varepsilon$ (i.e., if we didn't have continuity of η at 0), then knowing that $0 < |h| < \delta'$ implies that $|k(h)| < \delta$ would allow us to conclude *nothing* — for those h for which $k(h) = 0$ (for which $g'(a) = -\varepsilon(h)$), we would be unable to run this argument.

Second proof of chain rule: We begin with a preliminary observation about the linearization of a function. Suppose that a function f is differentiable at a. Then we can write, for all h,

$$f(a+h) = L_{f,a}(a+h) + \mathrm{err}_{f,a}(h),$$

where $\mathrm{err}_{f,a}(h)$, the error in using $L_{f,a}(a+h)$ to estimate $f(a+h)$, is defined by

$$\mathrm{err}_{f,a}(h) := f(a+h) - f(a) - f'(a)h.$$

Notice that (by continuity of f at a) we have that $\mathrm{err}_{f,a}$ approaches limit 0 near 0, i.e.,

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

This is no surprise — it's just saying that the linearization of f agrees with f at a.

But more is true. We have that

$$\frac{\text{err}_{f,a}(h)}{h} = \frac{f(a+h)-f(a)}{h} - f'(a) \to f'(a) - f'(a) = 0 \text{ as } h \to 0 \quad (\star)$$

(the last step valid because f is not just continuous but *differentiable* at a). In other words,

> the error we make in using $L_{f,a}(a+h)$ to approximate $f(a+h)$ goes to zero as h approaches zero, even when scaled relative to h.

We will return to this idea next semester, when discussing Taylor series.

Armed with $(\star)$, we can formally prove the chain rule, without using a ε-δ proof. Recall that we are assuming that g is differentiable at a and that f is differentiable at $g(a)$. Using first the linearization of g at a (with error $\text{err}_{g,a}$) and then the linearization of f at $g(a)$ (with error $\text{err}_{f,g(a)}$), we have that $\frac{f(g(a+h))-f(g(a))}{h}$

$$\begin{aligned}
&= \frac{f(g(a)+hg'(a)+\text{err}_{g,a}(h)) - f(g(a))}{h} \\
&= \frac{f(g(a)) + (hg'(a)+\text{err}_{g,a}(h))f'(g(a)) + \text{err}_{f,g(a)}(hg'(a)+\text{err}_{g,a}(h)) - f(g(a))}{h} \\
&= f'(g(a))g'(a) + \frac{\text{err}_{g,a}(h)}{h}f'(g(a)) + \frac{\text{err}_{f,g(a)}(hg'(a)+\text{err}_{g,a}(h))}{h}. \quad (\star\star)
\end{aligned}$$

We claim that $(\star\star) \to f'(g(a))g'(a)$ as $h \to 0$; if we can show this then we have proved the chain rule. Indeed, by $(\star)$ we have that

$$\frac{\text{err}_{g,a}(h)}{h} \to 0 \quad \text{as } h \to 0.$$

To deal with the last term in $(\star\star)$ (which we also want to tend to 0) we have to work a little harder. We have

$$\begin{aligned}
\frac{\text{err}_{f,g(a)}(hg'(a)+\text{err}_{g,a}(h))}{h} &= \left(\frac{\text{err}_{f,g(a)}(hg'(a)+\text{err}_{g,a}(h))}{h\left(g'(a)+\frac{\text{err}_{g,a}(h)}{h}\right)}\right)\left(g'(a)+\frac{\text{err}_{g,a}(h)}{h}\right) \\
&= \left(\frac{\text{err}_{f,g(a)}(hg'(a)+\text{err}_{g,a}(h))}{hg'(a)+\text{err}_{g,a}(h)}\right)\left(g'(a)+\frac{\text{err}_{g,a}(h)}{h}\right)
\end{aligned}$$

The second term on the right-hand side goes to $g'(a)$ as $h \to 0$, by $(\star)$. The first term can be viewed as a composition $(a \circ b)(h)$ where

- $b(h) = hg'(a) + \text{err}_{g,a}(h)$ — a function which is defined and continuous everywhere
- $a(x) = \frac{\text{err}_{f,g(a)}(x)}{x}$ — a function which is defined everywhere *except* $x = 0$, and is continuous everywhere it is defined.

However, from $(\star)$ we know that $\lim_{x\to 0} a(x) = 0$, so we can extend a to a function that is defined *and continuous* everywhere, by setting $a(0) = 0$.

We have previously proven that the composition of continuous functions is continuous, so $(a \circ b)$ is a function that is defined and continuous everywhere. In particular that means that

$$\lim_{h\to 0} \frac{\mathrm{err}_{f,g(a)}(hg'(a) + \mathrm{err}_{g,a}(h))}{hg'(a) + \mathrm{err}_{g,a}(h)} = \lim_{h\to 0} (a \circ b)(h) = (a \circ b)(0) = a(b(0)) = a(0) = 0,$$

and the proof of the chain rule is complete.

9 Applications of the derivative

In this section, we discuss some applications of the derivative. All of these related, loosely, to getting information about the "shape" of a function (or more correctly about the shape of the graph of a function) from information about the derivative; but as we will see in examples, the applications go well beyond this limited scope.

9.1 Maximum and minimum points

At a very high level, this is what our intuition suggests: if a function f is differentiable at a, then, since we interpret the derivative of f at a as being the slope of the tangent line to the graph of f at the point $(a, f(a))$, we should have:

- if $f'(a) = 0$, the tangent line is horizontal, and at a f should have either a "local maximum" or a "local minimum";
- if $f'(a) > 0$, the tangent line has positive slope, and f should be "locally increasing" near a; and
- if $f'(a) < 0$, the tangent line has negative slope, and f should be "locally decreasing" near a.

This intuition is, unfortunately, *wrong*: for example, the function $f(x) = x^3$ has $f'(0) = 0$, but f does not have a local maximum at a; in fact, it is increasing as we pass across $a = 0$.

More correctly, the intuition is only partly correct. What we do now is formalize some of the vague terms presented in quotes in the intuition above, and salvage it somewhat by presenting Fermat's principle.

Let f be a function, and let A be a subset of the domain of f.

Definition of maximum point Say that x is a *maximum point* for f on A if

- $x \in A$ and
- $f(x) \geq f(y)$ for all $y \in A$.

In this case, say that $f(x)$ is *the*[121] *maximum value* of f on A.

Definition of minimum point Say that x is a *minimum point* for f on A if

- $x \in A$ and
- $f(x) \leq f(y)$ for all $y \in A$.

In this case, say that $f(x)$ is *the minimum value* of f on A.

[121] "the": the maximum value is easily checked to be unique.

Maximum/minimum points are not certain to exist: consider $f(x) = x$, with $A = (0, 1)$; f has neither a maximum point nor a minimum point on A. And if they exist, they are not certain to be unique: consider $f(x) = \sin x$ on $[0, 2\pi]$, which has maximum value 1 achieved at two maximum points, namely $\pi/2$ and $3\pi/2$, and minimum value 1 achieved at three minimum points, namely 0, π and 2π.

While having derivative equal to 0 doesn't ensure being at a max point, something is true in the converse direction: under certain conditions, being a maximum or minimum point, and being differentiable at the point, ensures that the derivative is 0. The specific conditions are that the function is defined on an *open interval*.

Claim 9.1. *(Fermat principle, part 1) Let $f : (a, b) \to \mathbb{R}$. If*

- *x is a maximum point for f on (a, b), or a minimum point, and*
- *f differentiable at x*

then $f'(x) = 0$.

Before giving the proof, some remarks are in order:

- As observed earlier via the example $f(x) = x^3$ at 0, the converse to Fermat principle is not valid: a function f may be differentiable at a point, with zero derivative, but not have a maximum or minimum at that point.
- The claim becomes false if the function f is considered on a *closed* interval $[a, b]$. For example, the function $f(x) = x$ on $[0, 1]$ has a maximum at 1 and a minimum at 0, is differentiable at both points[122], but at neither point in the derivative zero.[123]
- Fermat principle makes no assumptions about the function f — it's not assumed to be differentiable everywhere, or even continuous. It's just a function.

Proof: Suppose x is a maximum point, and that f is differentiable at x. Consider the derivative of f from below at x. We have, for $h < 0$,

$$\frac{f(x+h) - f(x)}{h} \geq 0$$

since $f(x + h) \leq f(x)$ (x is a maximum point), so the ratio has non-positive numerator and negative denominator, so is positive. It follows that

$$f'_-(x) = \lim_{h \to 0^-} \frac{f(x+h) - f(x)}{h} \geq \lim_{h \to 0^-} 0 = 0.$$

[122] As differentiable as it can be ... differentiable from above at 0 and from below at 1.

[123] The state of Connecticut provides a real-world example: the highest point in the state is on a slope up to the summit of Mt. Frissell, whose peak is in Massachusetts.

Now consider the derivative of f from above at x. We have, for $h > 0$,

$$\frac{f(x+h) - f(x)}{h} \leq 0$$

since $f(x+h) \leq f(x)$ still, and so the ratio has non-positive numerator and positive denominator, so is negative. It follows that

$$f'_+(x) = \lim_{h \to 0^+} \frac{f(x+h) - f(x)}{h} \leq \lim_{h \to 0^+} 0 = 0.$$

Since f is differentiable at x (by hypothesis), we have $f'(x) = f'_+(x) = f'_-(x)$, so $f'(x) \leq 0 \leq f'(x)$, making $f'(x) = 0$.

An almost identical argument works if x is a minimum point. □

Fermat principle extends to "local" maxima and minima — points where a function has a maximum point or a minimum point, if the domain on which the function is viewed is made sufficient small around the point. Again let f be a function, and let A be a subset of the domain of f.

Definition of local maximum point Say that x is a *local maximum point* for f on A if

- $x \in A$ and
- there is a $\delta > 0$ such that $f(x) \geq f(y)$ for all $y \in (x - \delta, x + \delta) \cap A$.

In this case, say that $f(x)$ is *a*[124] *local maximum value* of f on A.

Definition of local minimum point Say that x is a *local minimum point* for f on A if

- $x \in A$ and
- there is a $\delta > 0$ such that $f(x) \leq f(y)$ for all $y \in (x - \delta, x + \delta) \cap A$.

In this case, say that $f(x)$ is *a local minimum value* of f on A.

Just like maximum/minimum points, local maximum/minimum points are not certain to exist: consider $f(x) = x$, with $A = (0, 1)$; f has neither a local maximum point nor a local minimum point on A. And if they exist, they are not certain to be unique: consider $f(x) = 2x^2 - x^4$ defined on $[-2, 3]$. A look at the graph of this function shows that it has local maxima at both -1 and 1 (both taking value 1, although of course it isn't necessarily the case that multiple local maxima have to share the same value, in general[125]). It also has a local minima at -2, 0 and 3, with values -8, 0 and -63. Notice that there are local minima at the endpoints of the interval, even though if the interval was extended slightly neither would be a local minimum. This is because the definition of x being a local minimum

[124] "a": a local maximum value is clearly not necessarily unique; see examples below.

[125] Physically, a local maximum is the summit of a mountain, and of course different mountains in general have different heights.

of a set A specifies that we should compare the function at x to the function at all y nearby to x *that are also in* A.

There is an analog of the Fermat principle for local maxima and minima.

Claim 9.2. *(Fermat principle, part 2) Let $f : (a, b) \to \mathbb{R}$. If*

- *x is a local maximum point for f on (a, b), or a local minimum point, and*
- *f differentiable at x*

then $f'(x) = 0$.

We do not present the proof here; it is in fact just a corollary of Claim 9.1. Indeed, if x is a local maximum point for f on (a, b), then from the definition of local maximum and from the fact that the interval (a, b) is open at both ends, it follows that there is some $\delta > 0$ small enough that $(x - \delta, x + \delta)$ is completely contained in (a, b), *and* that x is a *maximum point* (as opposed to local maximum point) for f on $(x - \delta, x + \delta)$; then if f is differentiable at x with derivative zero, Claim 9.1 shows that $f'(x) = 0$.[126]

As with Claim 9.1, Claim 9.2 fails if f is defined on a *closed* interval $[a, b]$, as the example $f(x) = 2x^2 - x^4$ on $[-2, 3]$ discussed above shows.[127]

Fermat principle leads to an important definition.

Definition of a critical point x is a *critical point* for a function f if f is differentiable at x, and if $f'(x) = 0$.[128] The value $f(x)$ is then said to be a *critical value* of f.

Here's the point of critical points. Consider $f : [a, b] \to \mathbb{R}$. Where could a maximum point or a minimum point of f be? Well, maybe at a or b. If not at a or b, then somewhere in (a, b). And by Fermat principle, the only possibilities for a maximum point or a minimum point in (a, b) are those points where f is not differentiable or (and this is where Fermat principle comes in) where f is differential and has derivative equal to 0; i.e., the critical point of f.

The last paragraph gives a proof of the following.

Theorem 9.3. *Suppose $f : [a, b] \to \mathbb{R}$. If a maximum point, or a minimum point, of f exists (on $[a, b]$), then x must be one of*

- *a or b*
- *a critical point in (a, b) or*
- *a point of non-differentiability in (a, b).*

[126] I wrote "We do not present the proof here"; but then it seems I went and gave the proof.

[127] And as does Connecticut: the south slope of Mt. Frissell, crossing into Massachusetts, is a local maximum high point of Connecticut, but not a point with derivative zero. The highest *peak* in Connecticut, the highest point with derivative zero, is the summit of Mt. Bear, a little south of Mt. Frissell.

[128] Many authors also say that x is a critical point for f if f is not differentiable at x.

In particular, if f is continuous on $[a, b]$ (and so, a maximum point and a minimum point exists, by the Extreme Value Theorem), then to locate a maximum point and/or a minimum point of f it suffices to consider the values of f at a, b, the critical points of f in (a, b) and the points of non-differentiability of f in (a, b).

Often this theorem reduces the task of finding the maximum or minimum value of a function on an interval (*a priori* a task that involves checking infinitely many values) to that of find the maximum or minimum of a finite set of values. We give three examples:

- $f : [-1, 1] \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 1/3 & \text{if } x = -1 \text{ or } x = 1 \\ 1/2 & \text{if } x = 0 \\ |x| & \text{otherwise.} \end{cases}$$

  Here the endpoints of the closed interval on which the function is defined are -1 and 1. We have $f(-1) = f(1) = 1/3$. There are no critical points in $(-1, 1)$, because where the function is differentiable (on $(-1, 0)$ and $(0, 1)$) the derivative is never 0. There is one point of non-differentiability in $(-1, 1)$, namely the point 0, and $f(0) = 1/2$. It might seem that the theorem tells us that the maximum value of f on $[-1, 1]$ is $1/2$ and the minimum value is $1/3$. But this is clearly wrong, on both sides: $f(3/4) = 3/4 > 1/2$, for example, and $f(-1/4) = 1/4 < 1/3$. The issue is that the function f *has no maximum* on $[-1, 1]$ (it's not hard to check that $\sup\{f(x) : x \in [-1, 1]\} = 1$ and $\inf\{f(x) : x \in [-1, 1]\} = 0$, but that there are no x's in $[-1, 1]$ with $f(x) = 1$ or with $f(x) = -1$), and so the hypotheses of the theorem are not satisfied.

- $f : [0, 4] \to \mathbb{R}$ defined by

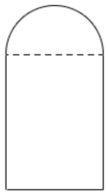
$$f(x) = \frac{1}{x^2 - 4x + 3}.$$

  Here the endpoints of the closed interval on which the function is defined are 0 and 4. We have $f(0) = f(4) = 1/3$. To find the critical points in $(0, 4)$, we differentiate f and set the derivative equal to 0:

$$f'(x) = \frac{-(2x - 4)}{(x^2 - 4x + 3)^2} = 0 \text{ when } 2x - 4 = 0, \text{ or } x = 2.$$

  So there is one critical point (at 2), and $f(2) = -1$. It might seem that the theorem tells us that the maximum value of f on $[0, 4]$ is $1/3$ and the minimum value is -1. But a quick look at the graph of the functions shows that this is quite wrong; the function takes arbitrarily large and arbitrarily small values on $[1, 4]$, in particular near to 1 and near to 2. What went wrong was that, as with the last example, we did not verify the hypotheses of the theorem. The function f may be written as $f(x) = 1/((x-1)((x-3)))$, and so is not defined at either 1 or 3, rendering the starting statement "$f : [0, 4] \to \mathbb{R}$" meaningless. not satisfied.

- (A genuine example, a canonical example of a "calculus optimization problem") A piece of wire of length L is to be bent into the shape of a Roman window — a rectangle below with a semicircle on top (see the figure below).



What is the maximum achievable area that can be enclosed by the wire with this shape?

We start by introducing names for the various variables of the problem. There are two reasonable variables: x, the base of the rectangle, and y, the height (these two values determine the entire shape). There is a relationship between these two numbers, namely $x + 2y + \pi(x/2) = L$ (the window has a base that is a straight line of length x, two vertical straight line sides of length y each, and a semicircular cap of radius $x/2$, so length $\pi(x/2)$). The total area enclosed may be expressed as $A = xy + \pi x^2/8$ (the area of the rectangular base, plus the area of the semicircular cap). We use $x(1 + \pi/2) + 2y = L$ to express y in terms of x: $y = (L - (1 + \pi/2)x)/2$, so that the area A becomes a function $A(x)$ of x, namely

$$A(x) = (x/2)(L - (1 + \pi/2)x) + \pi x^2/8.$$

Clearly the smallest value of x that we need consider is 0. The largest value is the one corresponding to $y = 0$, so $x = L/(1 + \pi/2)$. Therefore we are considering the problem of finding the maximum value of a continuous function A on the closed interval $[0, L/(1 + \pi/2)]$. Because A is continuous, we know that the maximum value exists. Because A is everywhere differentiable, the theorem tells us that we need only consider A at 0, $L/(1 + \pi/2)$, and any point between the two where $A'(x) = 0$. There is one such point, at $L/(2 + \pi/2)$.

We have

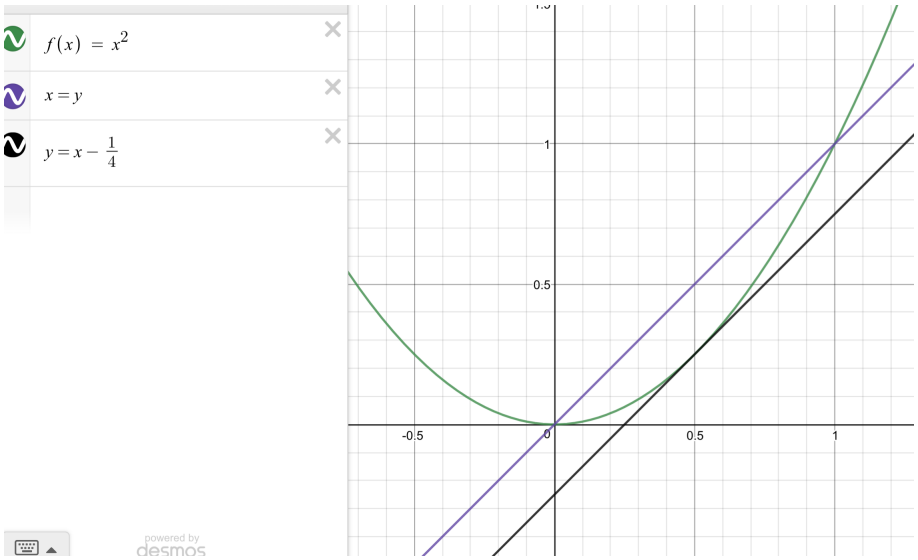
- $A(0) = 0$,
- $A(L/(2 + \pi/2)) = \frac{L^2}{2(2+\pi/2)^2} + \frac{\pi L^2}{8(2+\pi/2)^2}$ and
- $A(L/(1 + \pi/2)) = \frac{\pi L^2}{8(1+\pi/2)^2}$.

The second of these is the largest, so is the largest achievable area.

9.2 The mean value theorem

To go any further with the study of the derivative, we need a tool that is to differentiation as the Intermediate and Extreme Value Theorems are to continuity. That tool is called the Mean Value Theorem (MVT). The MVT says, mathematically, that if a function is differentiable on an interval, then at some point between the start and end point of the interval, the slope of the tangent line to the function should equal the average slope over the whole interval, that is, the slope of the secant line joining the initial point of the interval to the terminal point. Informally, it says that if you travel from South Bend to Chicago averaging 60 miles per hour, then at some point on the journey you must have been traveling at exactly 60 miles per hour.

By drawing a graph of a generic differentiable function, it is fairly evident that the MVT *must* be true. The picture below shows the graph of $f(x) = x^2$. Between 0 and 1, the secant line is $y = x$, with slope 1, and indeed there is a number between 0 and 1 at which the slope of the tangent line to f is 1, i.e., at which the derivative is 1, namely at 1/2.



However, we need to be careful. If we choose to do our mathematics only in the world of rational numbers, then the notions of limits, continuity and differentiability make perfect sense; and just as it was possible to come up with examples of continuous functions in this "$\mathbb{Q}$-world" that satisfy the hypotheses of IVT and EVT, but do not satisfy their conclusions, it is also possible to come up with an example of a function on a closed interval that is differentiable in the $\mathbb{Q}$-world, but for which there is no point in the interval where the derivative is equal the slope of the secant line connecting the endpoints of the interval.[129] This says that to prove the MVT, the completeness axiom will be needed. But in fact we'll bypass completeness, and prove MVT using EVT (which itself required completeness).

Theorem 9.4. *(Mean value theorem) Suppose $f : [a, b] \to \mathbb{R}$ is*

- *continuous on $[a, b]$, and*

[129] Find one! (It will be on the homework $\cdots$).

- *differentiable on* (a, b).

Then there is $c \in (a, b)$ *with*

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

Proof: We begin with the special case $f(a) = f(b)$. In this case, we require $c \in (a, b)$ with $f'(c) = 0$.[130]

By the Extreme Value Theorem, f has a maximum point and a minimum point in $[a, b]$. If there is a maximum point $c \in (a, b)$, then by Fermat principle, $f'(c) = 0$. If there is a minimum point $c \in (a, b)$, then by Fermat principle, $f'(c) = 0$. If neither of these things happen, then the maximum point and the minimum point must both occur at one (or both) of a and b. In this case, both the maximum and the minimum of f on $[a, b]$ are 0, so f is constant on $[a, b]$, and so $f'(c) = 0$ for *all* $c \in (a, b)$.

We now reduce the general remaining case, $f(a) \neq f(b)$, to the case just considered. Set

$$L(x) = f(a) + (x - a)\frac{f(b) - f(a)}{b - a};$$

notice that the graph of this function is the line that passes through $(a, f(a))$ and $(b, f(b))$. Now let $h(x)$ be the (vertical) distance from the point $(x, f(x))$ to the point $(x, L(x))$, so

$$h(x) = f(x) - f(a) - \left(\frac{f(b) - f(a)}{b - a}\right)(x - a).$$

We have $h(a) = h(b) = 0$, and h is continuous on $[a, b]$, and differentiable on (a, b). So by the previous case, there is $c \in (a, b)$ with $h'(c) = 0$. But

$$h'(x) = f'(x) - \left(\frac{f(b) - f(a)}{b - a}\right),$$

so $f'(c) = \frac{f(b)-f(a)}{b-a}$. □

Note that both Rolle's theorem and the MVT fail if f is not assumed to be differentiable on the whole of the interval (a, b): consider the function $f(x) = |x|$ on $[-1, 1]$.

In the proof of the MVT, we used the fact that if $f : (a, b) \to \mathbb{R}$ is constant, then it is differentiable at all points, with derivative 0. What about the converse of this? If $f : (a, b) \to \mathbb{R}$ is differentiable at all points, with derivative 0, can we conclude that f is constant? This seems a "fact" so obvious that it barely requires a proof: physically, it is asserting that if a particle has 0 velocity at all times, then it must always be located in the same position.

[130] This special case is often referred to as *Rolle's theorem*. It is traditional to make fun of Rolle's theorem; see e.g. this XKCD cartoon: https://xkcd.com/2042/. Before dismissing Rolle's theorem as a triviality, though, remember this: in $\mathbb{Q}$-world, it is false, and so its proof requires the high-level machinery of the completeness axiom.

But of course, it is not[131] be obvious. Indeed, if true, it must be a corollary of the completeness axiom, because in $\mathbb{Q}$-world, the function $f : [0,2] \to \mathbb{Q}$ given by

$$f(x) = \begin{cases} 0 & \text{if } x^2 < 2 \\ 1 & \text{if } x^2 > 2 \end{cases} \quad cc$$

is continuous on $[0,2]$, differentiable on $(0,2)$, has derivative 0 everywhere, but certainly is not constant.

We will establish this converse, not directly from completeness, but from MVT.

Claim 9.5. *If $f : (a,b) \to \mathbb{R}$ is differentiable at all points, with derivative 0, then f is constant.*

Proof: Suppose that f satisfies the hypotheses of the claim, but is not constant. Then there are $a < x_0 < x_1 < b$ with $f(x_0) \neq f(x_1)$. But then, applying MVT on the interval $[x_0, x_1]$, we find $c \in (x_0, x_1) \subseteq (a,b)$ with

$$f'(c) = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \neq 0,$$

a contradiction. We conclude that f is constant on (a,b). □

Corollary 9.6. *If $f, g : (a,b) \to \mathbb{R}$ are both differentiable at all points, with $f' = g'$ on all of (a,b), then there is a constant such that f and g differ by that (same) constant at every point in (a,b) (i.e., there's c with $f(x) = g(x) + c$ for all $x \in (a,b)$).*

Proof: Apply Claim 9.5 on the function $f - g$. □

Our next application of MVT concerns the notions of a function increasing/decreasing on an interval. Throughout this definition, I is some interval (maybe an open interval, like (a,b) or (a,∞), or $(-\infty,b)$ or $(-\infty,\infty)$, or maybe a closed interval, like $[a,b]$, or maybe a mixture, like $(a,b]$ or $[a,b)$ or $(-\infty,b]$ or $[a,\infty)$).

Definition of a function increasing Say that f is *increasing* on I, or *strictly increasing*[132], if whenever $a < b$ in I, $f(a) < f(b)$. Say that f is *weakly increasing* on I if whenever $a < b$ in I, $f(a) \leq f(b)$.

[131] at least, should not

[132] There is a truly annoying notational issue here. To some people, "increasing" means just what it has been defined to mean here, namely that as the input to the function increases, the output of the function genuinely increases, too. In this interpretation, the constant function is *not* increasing (it's *weakly* increasing). To other people, "increasing" means that as the input to the function increases, the output of the function either increases or stays the same. In this interpretation, the constant function *is* increasing. There is no resolution to this ambiguity, as both usages are firmly established in mathematics. So you have to be *very* careful, when someone talks about increasing/decreasing, that you know which interpretation they mean.

Definition of a function decreasing Say that f is *decreasing* on I, or *strictly decreasing*, if whenever $a < b$ in I, $f(a) > f(b)$. Say that f is *weakly decreasing* on I if whenever $a < b$ in I, $f(a) \geq f(b)$.

Definition of a function being monotone Say that f is *monotone* on I, or *strictly monotone*, if it is either increasing on I or decreasing on I. Say that f is *weakly monotone* on I if it is either weakly increasing on I or weakly decreasing on I.

Claim 9.7. *If $f'(x) > 0$ for all x in some interval I, then f is strictly increasing on I. If $f'(x) < 0$ for all $x \in I$, then f is strictly decreasing on I.*

If $f'(x) \geq 0$ for all x in some interval I, then f is weakly increasing on I. If $f'(x) \leq 0$ for all $x \in I$, then f is weakly decreasing on I.

Proof: Suppose $f'(x) > 0$ for all $x \in I$. Fix $a < b$ in I. By the MVT, there's $c \in (a, b)$ with

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

By hypothesis, $f'(c) > 0$, so $f(b) > f(a)$, proving that f is strictly increasing on I.

All the other parts of the claim are proved similarly. □

The converse of this claim is not (entirely) true: if f is strictly increasing on an interval, and differentiable on the whole interval, then it is not necessarily the case that $f'(x) > 0$ on the interval. The standard example here is $f(x) = x^3$, defined on the whole of the real line; it's strictly increasing, differentiable everywhere, but $f'(0) = 0$. On the other hand, we do have the following converse, which doesn't require MVT; it just comes from the definition of the derivative (similar to the proof of the Fermat principle). The proof is left as an exercise.

Claim 9.8. *If f is weakly increasing on an interval, and differentiable on the whole interval, then $f'(x) \geq 0$ on the interval.*[133]

Now that we have established a way of identifying intervals on which a function is increasing and/or decreasing, we can develop some effective tools for identifying where functions have local minima/local maxima. The first of these gives a partial converse to the Fermat principle. Recall that Fermat principle says that if f is defined on (a, b), with f differentiable at some $x \in (a, b)$, then if $f'(x) = 0$ x *might* be a local minimum or local maximum; while if $f'(x) \neq 0$, f cannot possibly be a local minimum or local maximum. This next claim gives some conditions under which we can say that x *is* a local minimum or local maximum, when its derivative is 0.

Claim 9.9. *(First derivative test) Suppose f is defined on (a, b), and that f is differentiable at $x \in (a, b)$, with $f'(x) = 0$. Suppose further that f is differentiable near x*[134]*. If $f'(y) < 0$*

[133] And, since strictly increasing implies weakly increasing, it follows that if f is strictly increasing on an interval, and differentiable on the whole interval, then $f'(x) \geq 0$ on the interval.

[134] Recall that "near x" means: in some interval $(x - \delta, x + \delta)$, $\delta > 0$.

for all y in some small interval to the left of x, and $f'(y) > 0$ in some small interval to right of x, then x is a local minimum for f on (a, b); in fact, x is a strict *local minimum, meaning $f(x) < f(y)$ for all y close to x. If, on the other hand, $f'(y) > 0$ for all y in some small interval to the left of x, and $f'(y) < 0$ in some small interval to right of x, then x is a strict local maximum for f on (a, b).*

Proof: We consider only the case where x is claimed to be a strict local minimum (the other is very similar). We have that on some small interval $(x - \delta, x]$, f has non-positive derivative (positive on $(x - \delta, x)$ and 0 at x), so, by Claim 9.7, f is weakly decreasing on this interval. By the same token, f is weakly increasing on $[x, x + \delta)$. This immediately says that x is a local minimum point for f on (a, b).

To get the strictness: f is strictly decreasing on $(x - \delta, x))$. For any y in this interval, pick any y' with $y < y' < x$. We have $f(y) > f(y')$ (because f is strictly decreasing between y and y'), and $f(y') \geq f(x)$ (because f is weakly decreasing between y' and x), so $f(y) > f(x)$; and by the same token $f(x) < f(y)$ for all y in a small interval to the right of x. □

Claim 9.10. *(Second derivative test) Suppose f is defined on (a, b), and that f is* twice *differentiable at $x \in (a, b)$, with $f'(x) = 0$.*

- *If $f''(x) > 0$, then a is a (strict) local minimum for f on (a, b).*
- *If $f''(x) < 0$, then a is a (strict) local maximum for f on (a, b).*
- *If $f''(x) = 0$ then anything can happen.*

Proof: We first consider the case where $f''(x) > 0$. We have

$$0 < f''(x) = f''_-(x) = \lim_{h \to 0^-} \frac{f'(a + h) - f'(a)}{h} = \lim_{h \to 0^-} \frac{f'(a + h)}{h}.$$

The denominator in the fraction at the end is negative. For the limit to be positive, the numerator must be negative for all sufficiently small (close to 0 and negative) h; in other words, $f'(y)$ must be negative on some small interval to the left of x. By a similar argument, $f'(y)$ must be positive on some small interval to the right of x. By Claim 9.9, f has a strict local minimum on (a, b) at x.

The case $f''(x) < 0$ is similar. To show that no conclusion can be reached when $f''(x) = 0$, consider the functions $f(x) = x^3$, $g(x) = x^4$ and $h(x) = -x^4$ at $x = 0$. In all three cases the functions have derivative 0 at 0, and second derivative 0 at 0. For f, 0 is neither a local maximum nor a local minimum point. For g, 0 is a local minimum. For h, 0 is a local maximum. □

9.3 Curve sketching

How do you get a good idea of the general appearance of the graph of a "reasonable" function (one which is continuous and differentiable at "most" points)? An obvious strategy is to use a graphing tool (such as Desmos.com or WolframAlpha.com). Here we'll describe a "by-hand" approach, that mostly utilizes information gleaned from the derivative. With powerful graphing tools available, this might seem pointless; but it's not. Here are two reasons why we might want to study curve sketching from first principles.

- It's a good exercise in reviewing the properties of the derivative, before applying them in situations where graphing tools may not be as helpful, and
- sometimes, graphing tools get things *very* wrong[135], and it's helpful to be able to do things by hand yourself, so that you can trouble-shoot when this happens.

The basic strategy that is often employed to sketch graphs of "reasonable" functions is as follows.

Step 1 Identify the domain of the function. Express it as a union of intervals.

Step 2 Identify the limiting behavior of the function at any *open* endpoints of intervals in the domain; this will usually involve one sided limits and/or limits at infinity, as well as possible infinite limits).

Step 3 Find the derivative of the function, and identify critical points (where the derivative is 0), intervals where the derivative is positive (and so the function is increasing), and intervals where the derivative is negative (and so the function is decreasing).

Step 4 Use the first derivative test to identify local maxima and minima.

Step 5 Plot some obvious points (such as intercepts of axes, local minima and maxima, and points where the derivative does not exist).

Step 6 Interpolate the graph between all these plotted points, in a manner consistent with the information obtained from the first four points.

There is also a zeroth step: check if the function is even, or is odd. This typically halves the work involved in curve sketching: if the function is even, then the graph is symmetric around the y-axis, and if it is odd, then the portion of the graph corresponding to negative x is obtained from the portion corresponding to positive x by reflection through the origin.

Our first example is $f(x) = x^3 + 3x^2 - 9x + 12$, which is neither even nor odd.

Step 1 The domain of f is all reals, or $(-\infty, \infty)$.

Step 2 $\lim_{x\to\infty} x^3 + 3x^2 - 9x + 12 = \infty$ and $\lim_{x\to-\infty} x^3 + 3x^2 - 9x + 12 = -\infty$.

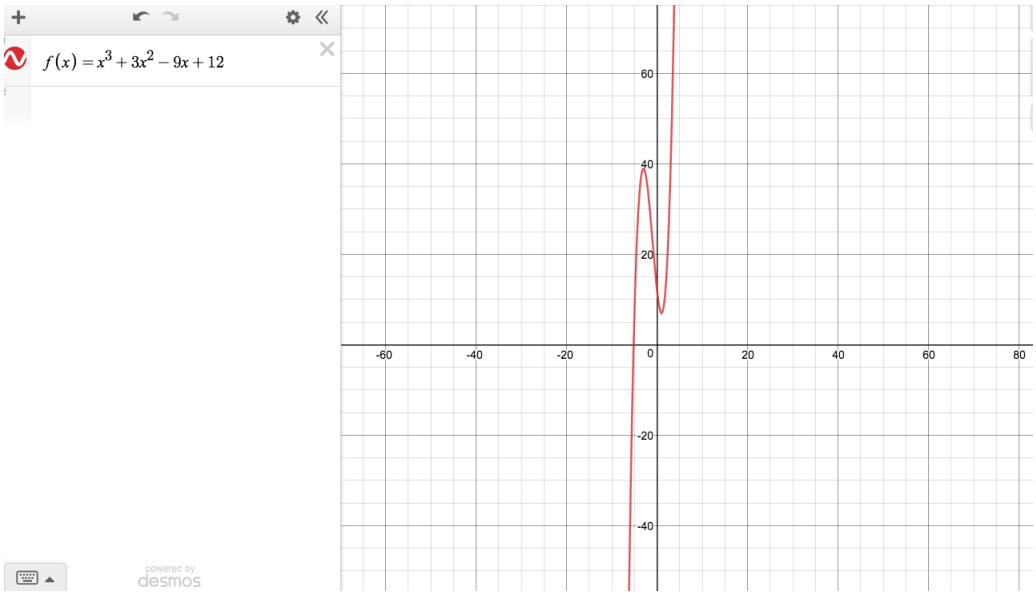
[135] Ask Desmos to graph the function $f(x) = [x \cdot (1/x)]$.

Step 3 $f'(x) = 3x^2 + 6x - 9$. This is defined, and continuous, on all of $\mathbb{R}$, so to find intervals where it is positive or negative, it is enough to find where it is 0 — $3x^2 + 6x - 9 = 0$ is the same as $x^2 + 2x - 3 = 0$ or $x = (-2 \pm \sqrt{4 + 12})/2 = 1$ or -3. Removing these two numbers from $\mathbb{R}$ leaves intervals $(-\infty, -3)$, $(-3, 1)$ and $(1, \infty)$. By the IVT, on each of these intervals f' must be either always positive or always negative (if f' is both positive and negative on any of the intervals then by continuity of f', f' must be 0 somewhere on that interval, but it can't be since we have removed to points where f' is 0). So we need to just test *one* point in each of $(-\infty, -3)$, $(-3, 1)$ and $(1, \infty)$, to determine the sign of f' on the entire interval. Since $f'(-100) > 0$, $f'(0) < 0$ and $f'(100) > 0$, we find that f is increasing on $(-\infty, -3)$, decreasing on $(-3, 1)$, and increasing on $(1, \infty)$.

Step 4 By the first derivative test, there is a local maximum at $x = -3$ (to the left of -3 the derivative is positive, to the right it is negative, at -3 it is 0), a local minimum at $x = 1$, and no other local extrema.

Step 5 At $x = 0$, $f(x) = 12$, so $(0, 12)$ is on the graph. The local maximum at $x = -3$ is the point $(-3, 39)$, and the local minimum at $x = 1$ is the point $(1, 7)$. The equation $f(x) = 0$ isn't obviously easy to solve, so we don't try to calculate any point at which the graph crosses the x-axis.

Step 6 We are required to plot a curve that's defined on all reals. As we move from $-\infty$ in the positive direction, the curve increases from $-\infty$ until it reaches a local maximum at $(-3, 39)$. Then it drops to a local minimum at $(1, 7)$, passing through $(0, 12)$ along the way. From the local minimum at $(1, 7)$ it increases to $+\infty$ at $+\infty$. This is a verbal description of the graph; here's what is looks like visually, according to Desmos:



With what we know so far, we couldn't have sketched such an accurate graph; we know, for example, that f decreases from -3 to 1, but how do we know that it decreases in the manner

that it does (notice how it "bulges": between -3 and 1, for a while the graph is lying to the right of the straight line joining $(-3, 39)$ to $(1, 7)$, and then it moves to being on the left)? To get this kind of fine detail, we need to study the *second* derivative, and specifically the topic of *convexity*; that will come in a later section.

As a second example, consider $f(x) = x^2/(1 - x^2)$. This is an even function — $f(-x) = f(x)$ for all x — so we only consider it on the interval $[0, \infty)$.

Step 1 The domain of the function (with our attention restricted to $[0, \infty)$) is all non-negative numbers except $x = 1$, that is, $[0, 1) \cup (1, \infty)$.

Step 2 We have

$$\lim_{x \to 0^-} \frac{x^2}{1 - x^2} = +\infty,$$

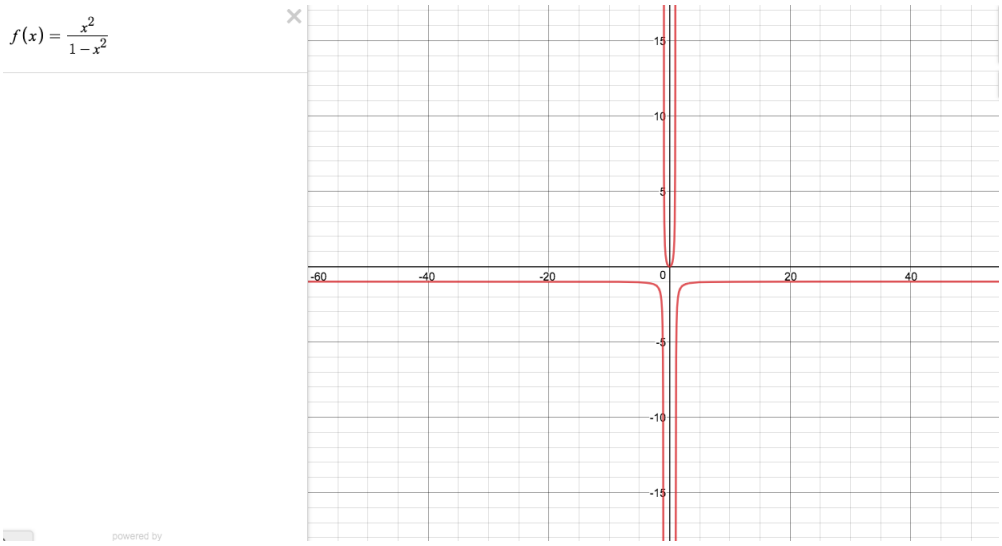
$$\lim_{x \to 0^+} \frac{x^2}{1 - x^2} = -\infty$$

and

$$\lim_{x \to \infty} \frac{x^2}{1 - x^2} = -1.$$

Steps 3, 4, 5 We have $f'(x) = 2x/(1 - x^2)^2$, and the domain of f' is the same at that of f: $[0, 1) \cup (1, \infty)$. The derivative is only equal to 0 at 0; at all other points it is positive. We conclude that f is strictly increasing on $(0, 1)$ and on $(1, \infty)$, and it is weakly increasing on $[0, 1)$. The graph passes through the point $(0, 0)$, and it does not seem like there are any other obviously easy-to-identify points.

Step 6 Moving from 0 to infinity: the graph starts at $(0, 0)$, and increases to infinity as x approaches 1 (the line $x = 1$ is referred to as a *vertical asymptote* of the graph). To the right of 1, it (strictly) increases from $-\infty$ to -1 as x moves from (just to the right of) 1 to ("just to the left of") ∞. (The line $y = -1$, that the graph approaches near infinity but doesn't reach, is referred to as a *horizontal asymptote* of the graph). To the left of the origin, the graph is the mirror image (the mirror being the y-axis) of what we have just described. Here is Desmos' rendering (for clarity, the aspect ratio has been changed from $1 : 1$):



9.4 L'Hôpital's rule

What is $\lim_{x\to 1} \frac{x^2-1}{x^3-1}$? The function $f(x) = (x^2-1)/(x^3-1)$ is not continuous at 1 (it is not even defined at 1) so we cannot asses the limit by a direct evaluation. We can figure out the limit, via a little bit of algebraic manipulation, however: away from 1

$$\frac{x^2-1}{x^3-1} = \frac{(x-1)(x+1)}{(x-1)(x^2+x+1)} = \frac{x+1}{x^2+x+1}.$$

Using our usual theorems about limits, we easily have $\lim_{x\to 1} \frac{x+1}{x^2+x+1} = 2/3$ (the function $g(x) = (x+1)/(x^2+x+1)$ is continuous at 1, with $g(1) = 2/3$, and g agrees with f at all reals other than 1).

We have calculated many such awkward limits using this kind of algebraic trickery. A common feature to many of these limits, is that the expression we are working with is a ratio, where both the numerator and denominator approach 0 near the input being approached in the limit calculation; this leads to the meaningless expression "0/0" when we attempt a "direct evaluation" of the limit as $0/0$[136]. Using the derivative, there is a systematic way of approaching all limits of this kind, called *L'Hôpital's rule*.

Suppose that we want to calculate $\lim_{x\to a} f(x)/g(x)$, but a direct evaluation is impossible because $f(a) = g(a) = 0$. We can approximate both the numerator and the denominator of the expression, using the linearization. The linearization of f near a is $L_f(x) = f(a) +$

[136] A meaningless expression, that can take on any possible value, or no value. Consider the following examples:

- $\lim_{x\to 0} \frac{cx}{x} = c$, c any real number;
- $\lim_{x\to 0} \frac{\pm x^2}{x} = \pm\infty$; and
- $\lim_{x\to 0} \frac{x\sin(1/x)}{x}$, which does not exist.

$f'(a)(x-a) = f'(a)(x-a)$, and the linearization of g near a is $L_g(x) = g(a) + g'(a)(x-a) = g'(a)(x-a)$.[137] Assuming that the linearization is a good approximation to the function it's linearizing, especially near the point of interest a, we get that near (but not at) a,

$$\frac{f(x)}{g(x)} \approx \frac{L_f(a)}{L_g(a)} = \frac{f'(a)(x-a)}{g'(a)(x-a)} = \frac{f'(a)}{g'(a)}$$[138]

This strongly suggests that if f, g are both differentiable at a, with $g'(a) \neq 0$ (and with $f(a) = g(a) = 0$), then

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

For example, with $f(x) = x^2 - 1$, $g(x) = x^3 - 1$, $a = 1$, so $f(a) = g(a) = 0$, $f'(x) = 2x$, $g'(x) = 3x^2$, so $f'(a) = 2$ and $g'(a) = 3$, we have

$$\lim_{x\to 1} \frac{x^2-1}{x^3-1} = \frac{2}{3}.$$

Before doing some examples, we try to formalize the linearization proof described above; along the way we keep track of all the various hypotheses we need to make on f and g.

So, suppose $f(a) = g(a) = 0$. We have, if all the various limits exist,

$$\begin{aligned}
\lim_{x\to a} \frac{f(x)}{g(x)} &= \lim_{x\to a} \frac{f(x)-f(a)}{g(x)-g(a)} \quad (f(a) = g(a) = 0) \\
&= \lim_{x\to a} \frac{\frac{f(x)-f(a)}{x-a}}{\frac{g(x)-g(a)}{x-a}} \\
&= \frac{\lim_{x\to a} \frac{f(x)-f(a)}{x-a}}{\lim_{x\to a} \frac{g(x)-g(a)}{x-a}} \quad \text{(adding assumption here: bottom limit is non-zero)} \\
&= \frac{f'(a)}{g'(a)}.
\end{aligned}$$

Going backwards through this chain of equalities yields a proof of the following result, what turns out to be a fairly weak form of what we will ultimately call L'Hôpital's rule.

Claim 9.11. *Suppose that f and g are both differentiable at a (so, in particular, defined in some small neighborhood around a, and also continuous at a), and that $g'(a) \neq 0$. If $f(a) = g(a) = 0$, then*

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

Here are a few examples.

[137] We're making the assumption here that f, g are both differentiable at a.

[138] We're making another assumption here — that $g'(a) \neq 0$.

$\lim_{x\to 0} \frac{\sin x}{x}$ Here $f(x) = \sin x$, $g(x) = x$, all hypotheses of the claim are clearly satisfied, and

$$\lim_{x\to 0} \frac{\sin x}{x} = \frac{\cos 0}{1} = 1,$$

as we already knew.[139]

$\lim_{x\to 0} \frac{x}{\tan x}$ Recall(?) that $\tan x = \frac{\sin x}{\cos x}$, so by the quotient rule,

$$\tan' x = \frac{(\sin' x)(\cos x) - (\sin x)(\cos' x)}{(\cos x)^2} = \frac{(\cos x)^2 + (\sin x)^2}{(\cos x)^2} = \frac{1}{(\cos x)^2}.$$

It follows that all hypotheses of the claim are satisfied, and so

$$\lim_{x\to 0} \frac{x}{\tan x} = \frac{1}{1/(\cos 0)^2} = 1.$$

Alternately we could write $x/\tan x = (x\cos x)/(\sin x)$, and, since the derivative of $x\cos x$ is $-x\sin x + \cos x$, obtain

$$\lim_{x\to 0} \frac{x}{\tan x} = \lim_{x\to 0} \frac{x\cos x}{\sin x} = \frac{-0\sin 0 + \cos 0}{\cos 0} = 1.$$

What we have so far is a very weak form of L'Hôpital's rule. It is not capable, for example, of dealing with

$$\lim_{x\to 1} \frac{x^3 - x^2 - x + 1}{x^3 - 3x + 2},$$

because although f and g are both 0 at 1, and both differentiable at 1, the derivative of g at 1 is 0. We can, however, deal with this kind of expression using simple algebraic manipulation: away from 1

$$\frac{x^3 - x^2 - x + 1}{x^3 - 3x + 2} = \frac{(x+1)(x-1)^2}{(x+2)(x-1)^2} = \frac{x+1}{x+2}$$

so

$$\lim_{x\to 1} \frac{x^3 - x^2 - x + 1}{x^3 - 3x + 2} = \lim_{x\to 1} \frac{x+1}{x+2} = \frac{2}{3}.$$

The issue L'Hôpital's rule is running into here is that what's causing g to be zero at 1 is somehow "order 2"; one pass of differentiating only half deals with the problem.

There is a much more powerful version of L'Hôpital's rule that gets around this issue by making *far* fewer assumptions on f and g: differentiability of f and g at a is dropped (and so, continuity, and even existence), and replaced with the hypothesis that near a, the limit of $f'(x)/g'(x)$ exists (and so, at least, we are demanding that f and g be differentiable and continuous *near* a). Here is the strongest statement of L'Hôpital's rule.[140]

[139] But note that this is more of a reality check than an example. We used this particular limit to discover that the derivative of sin is cos, so using L'Hôpital (which requires knowing the derivative of sin) to calculate the limit, is somewhat circular!

[140] The proof is quite messy, and will only appear in these notes, not in class.

Theorem 9.12. *(L'Hôpital's rule) Let f and g be functions defined and differentiable near a*[141]*. Suppose that*

- $\lim_{x\to a} f(x) = 0$,
- $\lim_{x\to a} g(x) = 0$, *and*
- $\lim_{x\to a} \frac{f'(x)}{g'(x)}$ *exists.*[142]

Then $\lim_{x\to a} \frac{f(x)}{g(x)}$ *exists, equals* $\lim_{x\to a} \frac{f'(x)}{g'(x)}$.

This version of L'Hôpital's rule is ideal for iterated applications. Consider, for example,

$$\lim_{x\to 1} \frac{x^3 - x^2 - x + 1}{x^3 - 3x + 2}.$$

Does this exist? By L'Hôpital's rule, it does if

$$\lim_{x\to 1} \frac{3x^2 - 2x - 1}{3x^2 - 3}$$

exists (and if so, the two limits have the same value). Does this second limit exist? Again by L'Hôpital's rule, it does if

$$\lim_{x\to 1} \frac{6x - 2}{6x}$$

exists (and if so, all three limits have the same value). But this last limit clearly exists and equals 2/3, so we conclude

$$\lim_{x\to 1} \frac{x^3 - x^2 - x + 1}{x^3 - 3x + 2} = \frac{2}{3}.$$

In practice, we would be more likely to present the argument much more compactly as follows:

$$\begin{aligned} \text{“}\lim_{x\to 1} \frac{x^3 - x^2 - x + 1}{x^3 - 3x + 2} &= \lim_{x\to 1} \frac{3x^2 - 2x - 1}{3x^2 - 3} \quad \text{(by L'Hôpital's rule)} \\ &= \lim_{x\to 1} \frac{6x - 2}{6x} \quad \text{(by L'Hôpital's rule)} \\ &= \frac{2}{3}, \end{aligned}$$

where all limits are seen to exist, and all applications of L'Hôpital's rule are seen to be valid, by considering the chain of equalities from bottom to top.

The proof of L'Hôpital's rule relies on a generalization of the Mean Value Theorem, known as the *Cauchy Mean Value Theorem*, that considers slopes of parameterized curve.

Definition of a parameterized curve A *parameterized curve* is a set of points of the form $(f(t), g(t))$, where f and g are functions; specifically it is $\{(f(t), g(t)) : t \in [a, b]\}$ where $[a, b]$ is (some subset of) the domain(s) of f and of g.

[141] But not necessarily even defined *at* a.

[142] Note that we don't require $g'(a) \neq 0$: $g'(a)$ might not even exist!

Think of a particle moving in 2-dimensional space, with $f(t)$ denoting the x-coordinate of the point at time t, and $g(t)$ denoting the y-coordinate. Then the parameterized curve traces out the location of the particle as time goes from a to b.

The graph of the function $f : [a, b] \to \mathbb{R}$ can be viewed as a parameterized curve — for example, it is $\{(t, f(t)) : t \in [a, b]\}$[143] On the other hand, not every parameterized curve is the graph of a function. For example, the curve $\{(\cos t, \sin t) : t \in [0, 2\pi]\}$ is a circle (the unit radius circle centered at $(0, 0)$), but is not the graph of a function.

We can talk about the *slope* of a parameterized curve at time t: using the same argument we made to motivate the derivative being the slope of the graph of a function, it makes sense to say that the slope of the curve $\{(f(t), g(t)) : t \in [a, b]\}$ at some time $t \in (a, b)$ is

$$\lim_{h\to 0} \frac{f(t+h) - f(t)}{g(t+h) - g(t)} = \lim_{h\to 0} \frac{(f(t+h) - f(t))/h}{(g(t+h) - g(t))/h} = \frac{\lim_{h\to 0}(f(t+h) - f(t))/h}{\lim_{h\to 0}(g(t+h) - g(t))/h} = \frac{f'(t)}{g'(t)},$$

assuming $f'(t)$, $g'(t)$ exist and $g'(t) \neq 0$.

We can also talk about the *average* slope of the curve, across the time interval $[a, b]$; it's

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

assuming $g(a) \neq g(b)$. The Cauchy Mean Value Theorem says that if the parameterized curve is suitably smooth, there is some point along the curve where the slope is equal to the average slope.

Theorem 9.13. *(Cauchy Mean Value Theorem) Suppose that $f, g : [a, b] \to \mathbb{R}$ are both continuous on $[a, b]$ and differentiable on (a, b). There is $t \in (a, b)$ with*

$$(f(b) - f(a))g'(t) = (g(b) - g(a))f'(t).$$

Before turning to the (short) proof, some remarks are in order.

- If $g(b) \neq g(a)$ then the theorem says that

  $$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(t)}{g'(t)}$$

  for some $t \in (a, b)$; that is, there is a point of the parameterized curve $\{(f(t), g(t)) : t \in [a, b]\}$ where the slope equal the average slope (as promised).

- If g is the identity ($g(x) = x$) then the Cauchy Mean Value theorem says that if $f : [a, b] \to \mathbb{R}$ is continuous on $[a, b]$ and differentiable on (a, b), then there is $t \in (a, b)$ with

  $$f(b) - f(a) = (b - a)f'(t);$$

  this is *exactly* the Mean Value Theorem.

[143] But this representation is not unique. For example, $\{(t, f(t)\} : t \in [0, 1]$ and $\{(t^2, f(t^2)) : t \in [0, 1]\}$ both trace out the same graph, that of the squaring function on domain $[0, 1]$; but they are different parameterized curves, since the particles are moving at different speeds in each case.

- Using the Mean Value Theorem, we can quickly get something that looks a little like the Cauchy MVT: there's $t_1 \in (a,b)$ with

$$\frac{f(b)-f(a)}{b-a} = f'(t_1)$$

and $t_2 \in (a,b)$ with

$$\frac{g(b)-g(a)}{b-a} = g'(t_2),$$

from which it follows that

$$(f(b)-f(a))g'(t_2) = (g(b)-g(a))f'(t_1).$$

The power of the Cauchy MVT is that it is possible to take $t_1 = t_2$, and this can't be obviously deduced from the Mean Value Theorem.

Proof (of Cauchy Mean Value Theorem): Define

$$h(x) = (g(b)-g(a))f(x) - (f(b)-f(a))g(x).$$

This is continuous on $[a,b]$ and differentiable on (a,b). Also,

$$\begin{aligned} h(a) &= (g(b)-g(a))f(a) - (f(b)-f(a))g(a) \\ &= g(b)f(a) - f(b)g(a) \\ &= (g(b)-g(a))f(b) - (f(b)-f(a))g(b) \\ &= h(b). \end{aligned}$$

By Rolle's theorem (or MVT) there is $t \in (a,b)$ with $h'(t) = 0$. But

$$h'(t) = (g(b)-g(a))f'(t) - (f(b)-f(a))g'(t)$$

so $h'(t) = 0$ says $(f(b)-f(a))g'(t) = (g(b)-g(a))f'(t)$, as required. □

Proof (of L'Hôpital's rule)[144]: To begin the proof of L'Hôpital's rule, note that a number of facts about f and g are implicit from the facts that $\lim_{x\to a} f(x) = 0$, $\lim_{x\to a} g(x) = 0$, and $\lim_{x\to a} f'(x)/g'(x)$ exists:

[144]Here's a sketch of the argument. f and g are both continuous on some interval $(a, a+\Delta)$ (because they are differentiable near a). Since $f, g \to 0$ near a, we can declare $g(a) = f(a) = 0$ to make the functions both continuous on $[a, \Delta$ (this may change the value of f, g at a, but won't change any of the limits involved in L'Hôpital's rule). Now for each $b < \Delta$ we have (since f, g are continuous on $[a,b]$ and differentiable on (a,b)) that $f(b)/g(b) = (f(b)-f(a))/(g(b)-g(a)) = f'(c)/g'(c)$ for some $c \in (a,b)$; this is Cauchy MVT. As b approaches a from above, the c that comes out of CMVT approaches a, so near a (from above) $f(b)/g(b)$ approaches $\lim_{c\to a^+} f'(c)/g'(c)$. A very similar argument gives the limit from below. Because f, g are not known to be differentiable at a, CMVT can't be applied in any interval that has a in its interior, which is why the argument gets split up into a "from above" and "from below" part.

- both f and g are differentiable and hence continuous in some open interval around a, except possibly at a itself (neither f nor g are necessarily even defined at a) and
- there is some open interval around a on which the derivative of g is never 0 (again, we rule out considering the derivative of g at a here, as this quantity may not exist).

Combining these observations, we see that there exists a number $\delta > 0$ such that on $(a-\delta, a+\delta) \setminus \{a\}$ both f and g are continuous and differentiable and g' is never 0.

Redefine f and g by declaring $f(a) = g(a) = 0$ (this may entailing increasing the domains of f and/or g, or changing values at one point). Notice that after f and g have been re-defined, the hypotheses of L'Hôpital's rule remain satisfied, and if we can show the conclusion for the re-defined functions, then we trivially have the conclusion for the original functions (all this because in considering limits approaching a, we never consider values at a). Notice also that f and g are now both continuous at a, so are in fact continuous on the whole interval $(a-\delta, a+\delta)$.

In particular, this means that we can apply both the Mean Value Theorem and Cauchy's Mean Value Theorem on any interval of the form $[a, b]$ for $b < a+\delta$ or $[b, a]$ for $b > a-\delta$ (we have to split the argument into a consideration of two intervals, one to the right of a and one to the left, because we do not know whether f and/or g are differentiable at a).

Given any b, $a < b < a+\delta$, we claim that $g(b) \neq 0$. Indeed, if $g(b) = 0$ then applying the Mean Value Theorem to g on the interval $[a, b]$ we find that there is c, $a < c < b$, with $g'(c) = (g(b) - g(a))/(b-a) = 0$, but we know that g' is never 0 on $(a, a+\delta)$. Similarly we find that $g(b) \neq 0$ for any b, $a-\delta < b < a$.

Fix an x, $a < x < a+\delta$. Applying Cauchy's Mean Value Theorem on the interval $[a, x]$ we find that there is an α_x, $a < \alpha_x < x$, such that

$$\frac{f(x)}{g(x)} = \frac{f(x) - f(a)}{g(x) - g(a)} = \frac{f'(\alpha_x)}{g'(\alpha_x)}.$$

(Here we use $g(a) = f(a) = 0$ and the fact that $g(x) \neq 0$).

Since $\alpha_x \to a^+$ as $x \to a^+$, and since $\lim_{x\to a^+} f'(x)/g'(x)$ exists, it seems clear that

$$\lim_{x\to a^+} \frac{f(x)}{g(x)} = \lim_{x\to a^+} \frac{f'(x)}{g'(x)}, \tag{6}$$

and by similar reasoning on the interval $(a-\delta, a)$ we should have

$$\lim_{x\to a^-} \frac{f(x)}{g(x)} = \lim_{x\to a^-} \frac{f'(x)}{g'(x)}. \tag{7}$$

L'Hôpital's rule follows from a combination of (6) and (7).

Thus to complete the proof of L'Hôpital's rule we need to verify (6). Fix $\varepsilon > 0$. There is a $\delta' > 0$ such that $a < x < a+\delta'$ implies $|f'(x)/g'(x) - L| < \varepsilon$, where $L = \lim_{x\to a^+} f'(x)/g'(x)$. We may certainly assume that $\delta' < \delta$. But then $a < x < a+\delta$, and so we have that

$f(x)/g(x) = f'(\alpha_x)/g'(\alpha_x)$ where $a < \alpha_x < x < a + \delta'$. Since α_x is close enough to a we have $|f'(\alpha_x)/g'(\alpha_x) - L| < \varepsilon$ and so $|f(x)/g(x) - L| < \varepsilon$. We have shown that $a < x < a + \delta'$ implies $|f(x)/g(x) - L| < \varepsilon$, which is the statement that $L = \lim_{x\to a^+} f(x)/g(x)$. This completes the verification of (6). □

The expressions that L'Hôpital's rule helps calculate the limits of, are often referred to as "indeterminates of the form 0/0" (for an obvious reason). There is a more general form of L'Hôpital's rule, that can deal with more "indeterminate" forms. In what follows, we use "lim" to stand for any of the limits

- $\lim_{x\to a}$,
- $\lim_{x\to a^-}$,
- $\lim_{x\to a^+}$,
- $\lim_{x\to\infty}$, or
- $\lim_{x\to-\infty}$,

and in interpreting the following claim, we understand that whichever version of "lim" we are thinking of for the first limit ($\lim f$), we are thinking of the *same* version for all the others ($\lim g$, $\lim f'/g'$ and $\lim f/g$).

Claim 9.14. *(General form of L'Hôpital's rule)*[145] *Suppose that* $\lim f(x)$ *and* $\lim g(x)$ *are either both* 0 *or are both* $\pm\infty$*. If*

$$\lim \frac{f'}{g'}$$

has a finite value, or if the limit is $\pm\infty$ *then*

$$\lim \frac{f}{g} = \frac{f'}{g'}.$$

We won't give a prove of this version of L'Hôpital's rule, but here's a sketch of how one of the variants goes. Suppose $\lim_{x\to\infty} f(x) = \infty$, $\lim_{x\to\infty} g(x) = \infty$, and $\lim_{x\to\infty} f'(x)/g'(x) = L$. Then we claim that $\lim_{x\to\infty} f(x)/g(x) = L$.

To show this we first have to argue a number of properties of f and g, most of which are implicit in, or can be read out of, the statement that $\lim_{x\to\infty} f'(x)/g'(x)$ exists and is finite; verifying them all formally may be considered a good exercise in working with the definitions.

[145] As with the earlier version of L'Hôpital's rule, indeterminates of the form ∞/∞ can have any limit, finite or infinite, or no limit. Consider, for example,

- $\lim_{x\to\infty} \frac{cx}{x} = c$, where c can be any real;
- $\lim_{x\to\infty} \frac{\pm x^2}{x} = \pm\infty$; and
- $\lim_{x\to\infty} \frac{x(2+\sin x)}{x}$, which does not exist.

- At all sufficiently large number, f is continuous;
- the same for g;
- for all sufficiently large x, $g'(x) \neq 0$; and
- if $x > N$ and N is sufficiently large, then $g(x) - g(N) \neq 0$ (this follows from Rolle's theorem: if N is large enough that $g'(c) \neq 0$ for all $c > N$, then if $g(x) = g(N)$ Rolle's theorem would imply that $g'(y) = 0$ for some $c \in (N, x)$, a contradiction).

Now write

$$\frac{f(x)}{g(x)} = \frac{f(x) - f(N)}{g(x) - g(N)} \cdot \frac{f(x)}{f(x) - f(N)} \cdot \frac{g(x) - g(N)}{g(x)}. \quad (\star)$$

For each fixed N, the fact that $\lim_{x\to\infty} f(x) = \infty$ says that eventually (for all sufficiently large x) $f(x) - f(N) \neq 0$, so it makes sense to talk about $\lim_{x\to\infty} f(x)/(f(x) - f(N))$; and (again since $\lim_{x\to\infty} f(x) = \infty$) we have $\lim_{x\to\infty} f(x)/(f(x) - f(N)) = 1$. Similarly (since $\lim_{x\to\infty} g(x) = \infty$) we have $\lim_{x\to\infty}(g(x) - g(N))/g(x) = 1$. In both limits calculated here, we are using that N is *fixed*, so that $f(N), g(N)$ are just fixed numbers.

Now for any N that is large enough that f and g are both continuous on $[N, \infty)$ and differentiable on (N, ∞), with $g'(x) \neq 0$ for any $x > N$ and $g(x) - g(N) \neq 0$ for any $x > N$ (such an N exists, by our previous observations), the Cauchy Mean Value Theorem tells us that there is $c \in (N, x)$ with

$$\frac{f(x) - f(N)}{g(x) - g(N)} = \frac{f'(c)}{g'(c)}.$$

Because $\lim_{x\to\infty} f'(x)/g'(x) = L$, we can make the first term in $(\star)$ be as close as we want to L; and by then choosing x sufficiently large, we can make the second and third terms in $(\star)$ be arbitrarily close to 1. In this way, the product of the three terms can be made arbitrarily close to L.

Good examples of the use of this more general form of L'Hôpital's rule are not so easy to come by at the moment; the rule really shows its strength when we deal with the exponential, logarithm and power functions, which we won't see until later. If you know about these functions, then the following example will make sense; if not, just ignore it.

Consider $f(x) = (\log x)/x$[146] What does this look like for large x? It's an indeterminate of the form ∞/∞, so by L'Hôpital's rule the limit $\lim_{x\to\infty}(\log x)/x$ equals $\lim_{x\to\infty}(\log' x)/x' = \lim_{x\to\infty}(1/x)/1 = \lim_{x\to\infty} 1/x$, as long as this limit exists. Since this limit exists and equals

[146]Here $\log : (0, \infty) \to \mathbb{R}$ is the *natural logarithm* function, which has the property that if $\log x = y$ then $x = e^y$, where e is a particular real number, approximately 2.71828, called the *base of the natural logarithm*. We'll see why such an odd looking function is "natural" next semester. The properties of log that we'll use in this example are that $\lim_{x\to\infty} \log x = \infty$, and that $\log'(x) = 1/x$.

0, it follows that

$$\lim_{x\to\infty} \frac{\log x}{x} = 0.$$[147]

Going to Desmos and looking at the graphs of $f(x) = \log x$ (entered as `ln x`) and $g(x) = x$, it seems pretty clear that for even quite small x, $\log x$ is dwarfed by x, so it is not surprising that the limit is 0. On the other hand, looking at the graphs of $f(x) = (\log x)^2$ and $g(x) = \sqrt{x}$, it's less clear what the limit

$$\lim_{x\to\infty} \frac{(\log x)^2}{\sqrt{x}}$$

might be. Looking at x up to about, say, 180, it seems that $f(x)$ is growing faster than $\sqrt{x}$, but for larger values of x the trend reverses, and at about $x = 5,500$, $g(x)$ has caught up with $f(x)$, and from there on seems to outpace it. This suggests that the limit might be 0. We can verify this using L'Hôpital's rule. With the usual caveat that the equalities are valid as long as the limits actually exist (which they will all be seen to do, by applications of L'Hôpital's rule, working from the back) we have

$$\begin{aligned}
\lim_{x\to\infty} \frac{(\log x)^2}{\sqrt{x}} &= \lim_{x\to\infty} \frac{2(\log x)(1/x)}{1/(2\sqrt{x})} \\
&= \lim_{x\to\infty} \frac{4\log x}{\sqrt{x}} \\
&= \lim_{x\to\infty} \frac{4/x}{1/(2\sqrt{x})} \\
&= \lim_{x\to\infty} \frac{8}{\sqrt{x}} \\
&= 0.
\end{aligned}$$

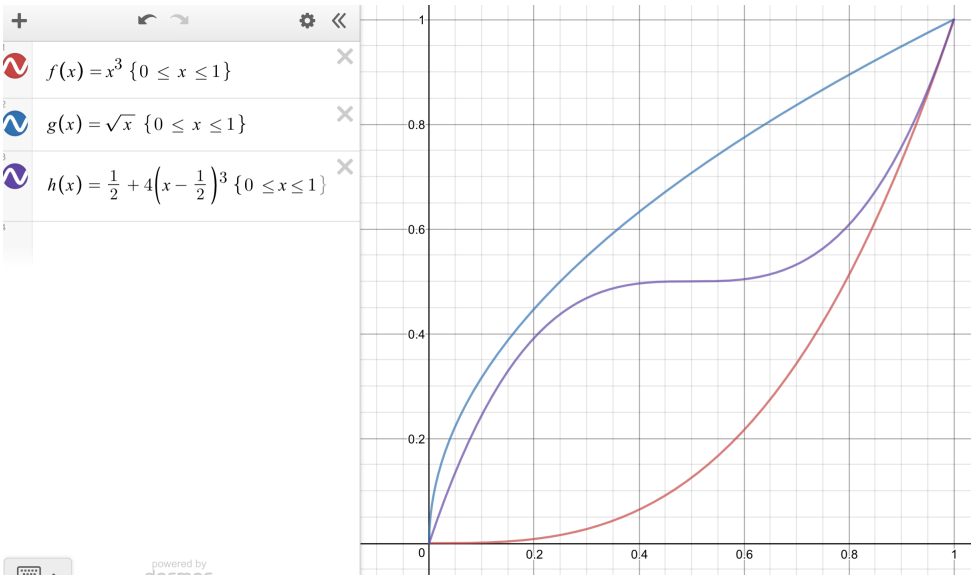
9.5 Convexity and concavity

Knowing that $f'(x) \geq 0$ for all $x \in [0,1]$ tells us that f is (weakly) increasing on $[0,1]$, but that doesn't tell the whole story. Below there is illustrated the graphs of three functions,

- $f(x) = x^3$
- $g(x) = \sqrt{x}$ and
- $h(x) = \frac{1}{2} + 4\left(x - \frac{1}{2}\right)^3$,

all of which are increasing on $[0,1]$, but that otherwise look very different from each other.

[147] What about $\lim_{x\to\infty} x^{1/x}$? Write $x^{1/x} = e^{\log(x^{1/x})} = e^{(\log x)/x}$. Since $(\log x)/x$ approaches 0 as x gets larger, it seems reasonable that $e^{(\log x)/x}$ approaches $e^0 = 1$; so $\lim_{x\to\infty} x^{1/x} = 1$. This is a very typical application of L'Hôpital's rule: we have two parts of a function that are competing with each other (in this case the x in the base, causing $x^{1/x}$ to grow larger as x grows, and the $1/x$ in the exponent, causing $x^{1/x}$ to grow smaller as x grows), and L'Hôpital's rule (often) allows for a determination of which of the two "wins" in the limit.



The fine-tuning of the graph of a function "bulges"' is captured by the second derivative. Before delving into that, we formalize what we mean by the graph "bulging".

Let f be a function whose domain includes the interval I.

Definition of a function being convex Say that f is *strictly convex*, or just *convex*[148], on I if for all $a, b \in I$, $a < b$, and for all t, $0 < t < 1$,

$$f((1-t)a + tb) < (1-t)f(a) + tf(b).$$

If instead $f((1-t)a + tb) \le (1-t)f(a) + tf(b)$ for all a, b and t, say that f is *weakly convex* on the interval.

Definition of a function being concave Say that f is *strictly concave*, or just *concave*[149], on I if for all $a, b \in I$, $a < b$, and for all t, $0 < t < 1$,

$$f((1-t)a + tb) > (1-t)f(a) + tf(b).$$

If instead $f((1-t)a + tb) \ge (1-t)f(a) + tf(b)$ for all a, b and t, we say that f is *weakly concave* on the interval.

Notice that as t varies from 0 to 1, the value of $(1-t)a + tb$ varies from a to b. The point

$$((1-t)a + tb, f((1-t)a + tb))$$

[148] Just as with "increasing", there is no universal convention about the meaning of the word "convex", without a qualifying adjective. By the word "convex", some people (including Spivak and me) mean what in the definition above is called "strictly convex", and others mean what above is called "weakly convex". It's a slight ambiguity that you have to learn to live with.

[149] Some authors, especially of 1000-page books called "Calculus and early transcendental functions, 45th edition", use "concave up" for what we are calling "convex", and "concave down" for what we are calling "concave". These phrases (the "up"-"down" ones) are almost never used in discussions among contemporary mathematicians.

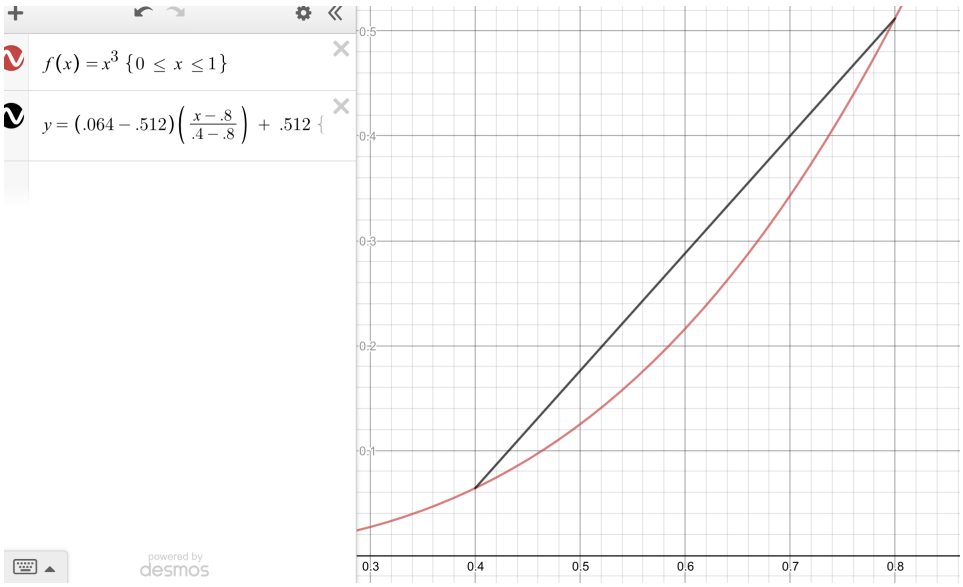
is a point on the graph of the function f, while the point

$$((1-t)a+tb, (1-t)f(a)+tf(b))$$

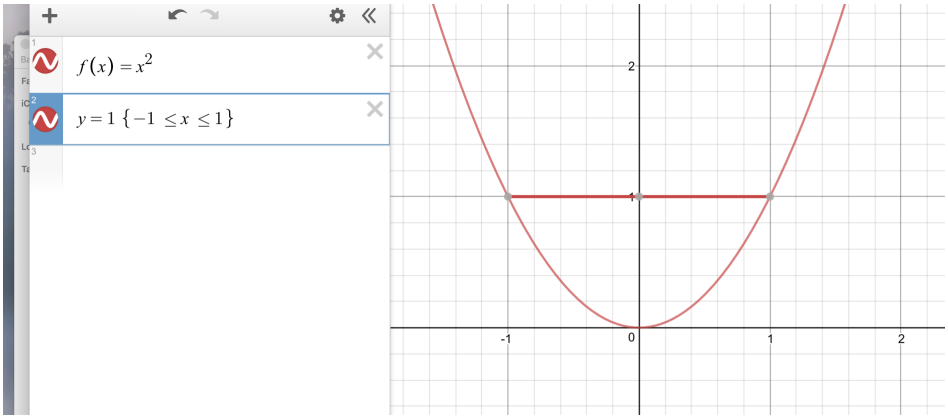
is a point on the secant line to the graph of the function f between the points $(a, f(a))$ and $(b, f(b))$. So the graphical sense of convexity is that

f is convex on I if the graph of f lies below the graphs of all its secant lines on I.

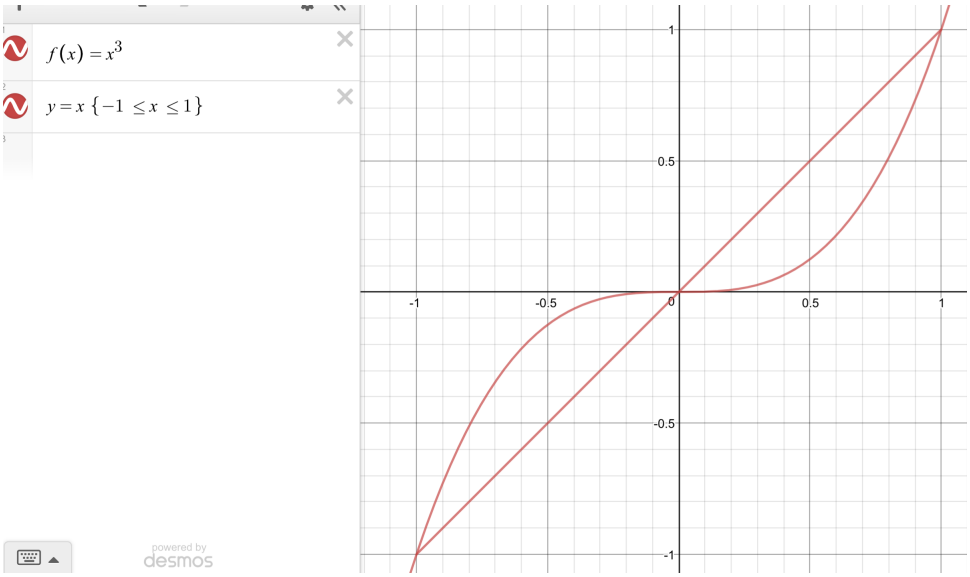
Illustrated below is the graph of $f(x) = x^3$, which lies below all its secant lines between 0 and 1, and so fairly clearly is convex on that interval. The picture below shows one secant line, from $(0.4, 0.64)$ to $(0.8, 0.512)$.



It is worth noting that convexity/concavity has nothing to do with f increasing or decreasing. It should be fairly clear from the graph of $s(x) = x^2$ that this function is convex on the entire interval $(-\infty, \infty)$, even though it is sometimes decreasing and sometimes increasing. (The picture below shows a secant line lying above the graph of s, that straddles the local minimum).



On the other hand, it's clear that $f(x) = x^3$ is concave on $(-\infty, 0]$ and convex on $[0, \infty)$, though it is increasing throughout. (The picture below shows one secant line, on the negative side, lying *below* the graph of f, and another, on the positive side, lying above).



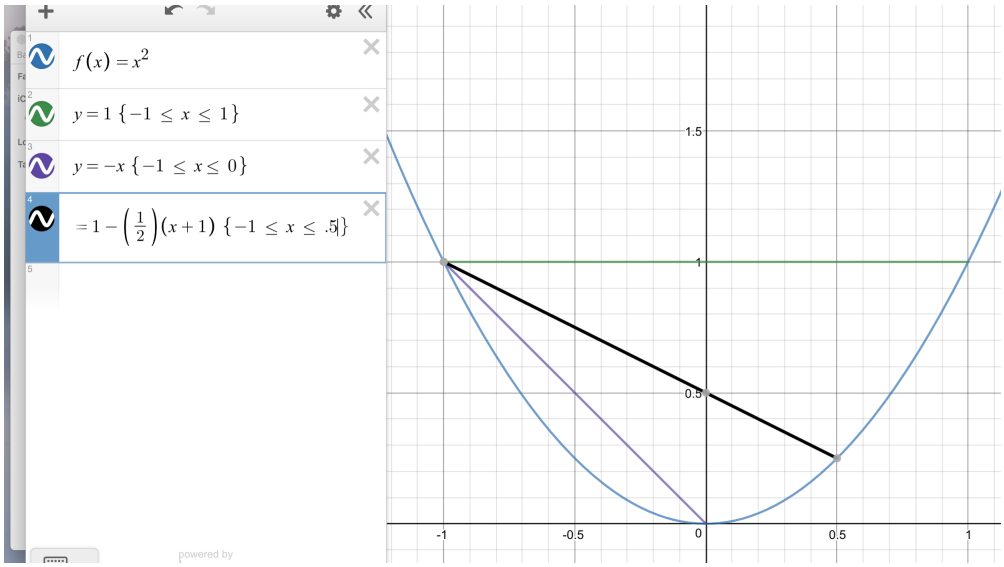
As was implicitly assumed in the last example discussed, just as convexity graphically means that secant lines lie above the graph, we have a graphical interpretation of concavity:

f is concave on I if the graph of f lies *above* the graphs of all its secant lines on I.

In terms of proving properties about convexity and concavity, there is an easier way to think about concavity. The proof of the following very easy claim is left to the reader; it is evident from thinking about graphs.

Claim 9.15. *f is concave of an interval I if and only if $-f$ is convex on I.*

There is an alternate algebraic characterization of convexity and concavity, that will be very useful when proving things. If f is concave on I, and $a, b \in I$ with $a < b$, then it seems clear from a graph that as x runs from a to b, the slope of the secant line from the point $(a, f(a))$ to the point $(x, f(x))$ is *increasing*. The picture below illustrates this with the square function, with $a = -1$ and $b = 1$. The slope of the secant line from $(-1, 1)$ to $(0, 0)$ is -1; from $(-1, 1)$ to $(1/2, 1/4)$ is $-1/2$; and from $(-1, 1)$ to $(1, 1)$ is 0.



We capture this observation in the following claim, which merely says that as x runs from a to b, the slopes of all the secant lines are *smaller* than the slope of the secant line from $(a, f(a))$ to $(b, f(b))$.

Claim 9.16.

- *f is convex on I if and only if for all $a, b \in I$ with $a < b$, and for all $x \in (a, b)$ we have*

$$\frac{f(x) - f(a)}{x - a} < \frac{f(b) - f(a)}{b - a}. \quad (\star)$$

Also[150]*, f is convex on I if and only if for all $a, b \in I$ with $a < b$, and for all $x \in (a, b)$ we have*

$$\frac{f(b) - f(a)}{b - a} < \frac{f(b) - f(x)}{b - x}.$$

- *f is concave on I if and only if for all $a, b \in I$ with $a < b$, and for all $x \in (a, b)$ we have*

$$\frac{f(x) - f(a)}{x - a} > \frac{f(b) - f(a)}{b - a} > \frac{f(b) - f(x)}{b - x}.$$

Proof: The key point is that any $x \in (a, b)$ has a unique representation as $x = (1 - t)a + tb$ with $0 < t < 1$, specifically with

$$t = \frac{x - a}{b - a}$$

(it is an easy check that this particular t works; that it is the unique t that works follows from the fact that for $t \neq t'$, $(1 - t)a + tb \neq (1 - t')a + t'b$). So, f being convex on I says *precisely* that for $a < x < b \in I$,

$$f(x) < \left(1 - \frac{x - a}{b - a}\right) f(a) + \left(\frac{x - a}{b - a}\right) f(b).$$

[150]This next clause of the claim says that convexity also means that as x runs from a to b, the slopes of the secant lines from $(x, f(x))$ to $(b, f(b))$ *increase*. This can also easily be motivated by a picture.

Subtracting $f(a)$ from both sides, and dividing across by $x-a$, this is seen to be equivalent to

$$\frac{f(x)-f(a)}{x-a} < \frac{f(b)-f(a)}{b-a},$$

as claimed.

But now also note that

$$\left(1-\frac{x-a}{b-a}\right)f(a)+\left(\frac{x-a}{b-a}\right)f(b) = \left(\frac{b-x}{b-a}\right)f(a)+\left(1-\frac{b-x}{b-a}\right)f(b),$$

so f being convex on I also says *precisely* that for $a<x<b\in I$,

$$f(x) < \left(\frac{b-x}{b-a}\right)f(a)+\left(1-\frac{b-x}{b-a}\right)f(b),$$

which after similar algebra to before is equivalent to

$$\frac{f(b)-f(a)}{b-a} < \frac{f(b)-f(x)}{b-x},$$

also as claimed.

We now move on to the concavity statements. f being concave means that $-f$ is convex, which (by what we have just proven) is equivalent to

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

for $a<x<b\in I$, and (multiplying both sides by -1) this is equivalent to

$$\frac{f(b)-f(a)}{b-a} > \frac{f(b)-f(x)}{b-x};$$

and the other claimed inequality for concavity is proved similarly. □

This alternate characterization of convexity and concavity allows us to understand the relationship between convexity and the derivative.

Theorem 9.17. *Suppose that f is convex on an interval. If f is differentiable at a and b in the interval, with $a<b$, then $f'(a)<f'(b)$ (and so, if f is differentiable everywhere on the interval, then f' is increasing on the interval).*

Proof: We will use our alternate characterization for convexity to show that

$$f'(a) < \frac{f(b)-f(a)}{b-a} < f'(b).$$

Pick any $b'\in(a,b)$. Applying our alternate characterization on concavity on the interval $[a,b']$, we have that for any $x\in(a,b')$,

$$\frac{f(x)-f(a)}{x-a} < \frac{f(b')-f(a)}{b'-a}.$$

Because f is differentiable at a, we have[151] that

$$f'(a) = f'_+(a) = \lim_{x \to a^+} \frac{f(x) - f(a)}{x - a} \leq \frac{f(b') - f(a)}{b' - a}.$$

But now, also applying our alternate characterization of convexity on the interval $[a, b]$, and noting that $a < b' < b$, we have

$$\frac{f(b') - f(a)}{b' - a} < \frac{f(b) - f(a)}{b - a}.$$

It follows that

$$f'(a) < \frac{f(b) - f(a)}{b - a}.$$

So far, we have only used the first part of the alternate characterization of convexity (the part marked ($\star$) above). Using the second part, an almost identical argument (which is left as an exercise) yields

$$\frac{f(b) - f(a)}{b - a} < f'(b),$$

and we are done. □

There is of course a similar theorem relating concavity and the derivative, which can be proven by using the fact that f is concave iff $-f$ is convex (it is left as an exercise).

Theorem 9.18. *Suppose that f is concave on an interval. If f is differentiable at a and b in the interval, with $a < b$, then $f'(a) > f'(b)$ (and so, if f is differentiable everywhere on the interval, then f' is decreasing on the interval).*

There is a converse to these last two theorems.

Theorem 9.19. *Suppose that f is differentiable on an interval. If f' is increasing on the interval, then f is convex, which if f' is decreasing, then f is concave.*

Before proving this, we make some comments.

- We are now in a position to use the first derivative to pin down intervals where a function is convex/concave — the intervals of convexity are precisely the intervals where f' is increasing, and the intervals of concavity are those where f' is decreasing. Of course, the easiest way to pin down intervals where f' is increasing/decreasing is to look at the derivative of f' (if it exists). That leads to the following corollary.

  Corollary 9.20. *If f is twice differentiable, then the intervals where f'' is positive (so f' is increasing) are the intervals of convexity, and the intervals where f'' is negative (so f' is decreasing) are the intervals of concavity.*

[151] In the next line, we use a fact that we may not have formally proved, but is easy to prove (and very useful): suppose that $f(x) < M$ (for some constant M) for all $x > a$, and that $\lim_{x \to a^+} f(x)$ exists. Then $\lim_{x \to a^+} f(x) \leq M$.

The places where f transitions from being concave to convex or vice-versa (usually, but not always, where f'' is zero), are called *points of inflection*.

- As an example, consider $f(x) = x/(1+x^2)$. Its domain is all reals. It goes to 0 as x goes to both $+\infty$ and to $-\infty$. We have

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

which is

  - negative for $x < -1$ (so x is decreasing on $(-\infty, -1)$),
  - positive for $-1 < x < 1$ (so x is increasing on $(-1, 1)$), and
  - negative for $x > 1$ (so x is decreasing on $(1, \infty)$).

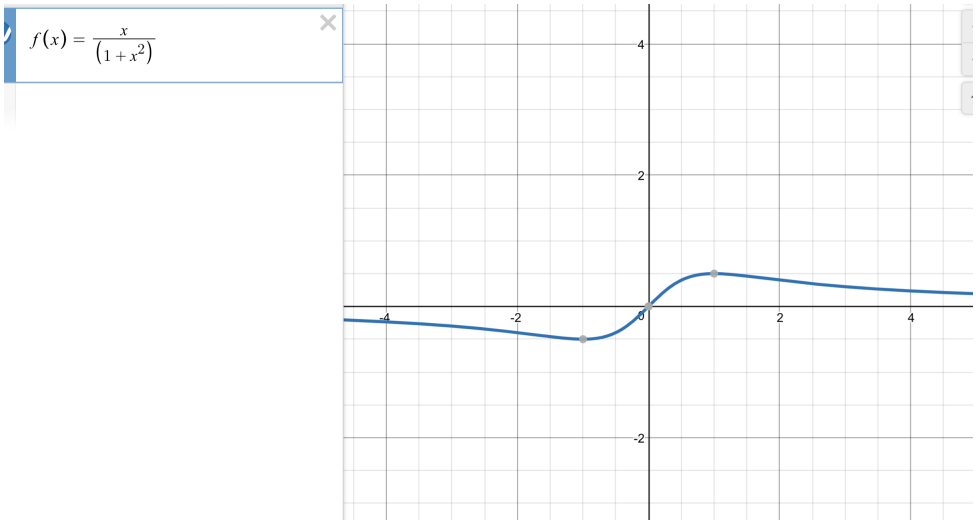
It follows that there is a local minimum at $(-1, -1/2)$ and a local maximum at $(1, 1/2)$. We also have

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

which is

  - negative for $x < -\sqrt{3}$ (so x is concave on $(-\infty, -\sqrt{3})$),
  - positive for $-\sqrt{3} < x < 0$ (so x is convex on $(-\sqrt{3}, 0)$),
  - negative for $0 < x < \sqrt{3}$ (so x is concave on $(0, \sqrt{3})$), and
  - positive for $\sqrt{3} < x < \infty$ (so x is convex on $(\sqrt{3}, \infty)$).

It follows that there is are points of inflection at $(-\sqrt{3}, -\sqrt{3}/4)$ and at $(\sqrt{3}, \sqrt{3}/4)$. Based on all of this information, it is not surprising to see that Desmos renders the graph of the function as follows.



Before proving Theorem 9.19, we need a preliminary lemma, the motivation for which which is that if f is convex between a and b, and $f(a) = f(b)$, then we expect the graph of f to always be below the line joining $(a, f(a))$ to $(b, f(b))$.

Lemma 9.21. *Suppose f is differentiable on an interval, with f' increasing on the interval. For $a < b$ in the interval, if $f(a) = f(b)$ then for all $x \in (a, b)$, $f(x) < f(a)$ and $f(x) < f(b)$.*

Proof: Suppose there is an $x \in (a, b)$ with $f(x) \geq f(a)$ (and so also $f(x) \geq f(b)$). Then there is a maximum point of f on $[a, b]$ at some particular $x \in (a, b)$. This is, of course, also a maximum point of f on (a, b). Since f is differentiable everywhere, by Fermat principle $f'(x) = 0$. By the Mean Value Theorem applied to $[a, x]$, there is $x' \in (a, x)$ with

$$f'(x') = \frac{f(x) - f(a)}{x - a}. \quad (\star)$$

Now f' is increasing on $[a, b]$ (by hypothesis), so $f'(x') < f'(x) = 0$. But $f(x) \geq f(a)$ (since x is a maximum point for f on $[a, b]$), so $(f(x) - f(a))/(x - a) \geq 0$. This contradicts the equality in $(\star)$ above.[152] □

Proof (of Theorem 9.19): Recall that we wish to show that if f is differentiable on an interval, and if f' is increasing on the interval, then f is convex (the concavity part is left as an exercise; it follows as usual from the observation that f is concave iff $-f$ is convex).

Given $a < b$ in the interval, set

$$g(x) = f(x) - \frac{f(b) - f(a)}{b - a}(x - a).$$

We have $g(a) = g(b)$ (both are equal to $f(a)$). We also have

$$g'(x) = f'(x) - \frac{f(b) - f(a)}{b - a},$$

which is increasing on the interval, since f' is. It follows from the preliminary lemma that for all $x \in (a, b)$, we have $g(x) < g(a)$ and $g(x) < g(b)$. The first of these says

$$f(x) - \frac{f(b) - f(a)}{b - a}(x - a) < f(a),$$

or

$$\frac{f(x) - f(a)}{x - a} < \frac{f(b) - f(a)}{b - a},$$

which is the characterization $(\star)$ of convexity from earlier; so f is convex on the interval.[153] □

[152] Note that we didn't actually use that $f(a) = f(b)$ in this proof, so what we actually showed was that if f is differentiable on an interval, with f' increasing on the interval, then for any $a < b$ in the interval, for all $x \in (a, b)$ we have $f(x) < \max\{f(a), f(b)\}$. The weaker result stated is, however, easier to comprehend visually, and it is the case that we will use in a moment.

[153] The second inequality, $g(x) < g(b)$, similarly reduces to the other characterization of convexity, but that isn't needed here.

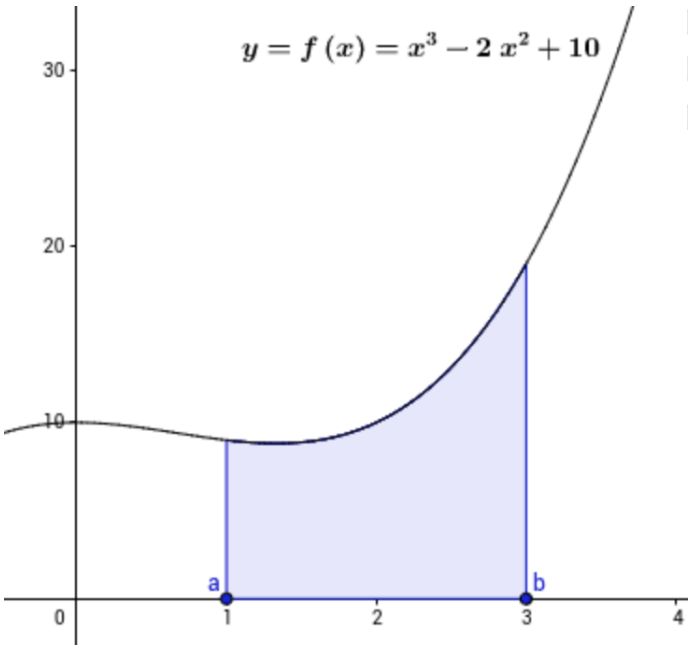
10 The Darboux integral

The main focus of the spring semester is on the integral. In contrast to the derivative, whose motivation was dynamic — measuring rate of change — the main motivation for the integral is static. Although there will be no real similarities between the definitions of the integral and the derivative, the two will quickly turn out to be intimately related.

Note: In this version of the notes, there are very few figures. It will be helpful as you read the notes, to supply graphs to illustrate the various examples, as we will do in class.

10.1 Motivation via area

Suppose $f : [a, b] \to \mathbb{R}$ is a non-negative function. What is the area of the region bounded by the curve $y = f(x)$, the x-axis, the lines $x = a$ and the line $x = b$? (See picture below.)



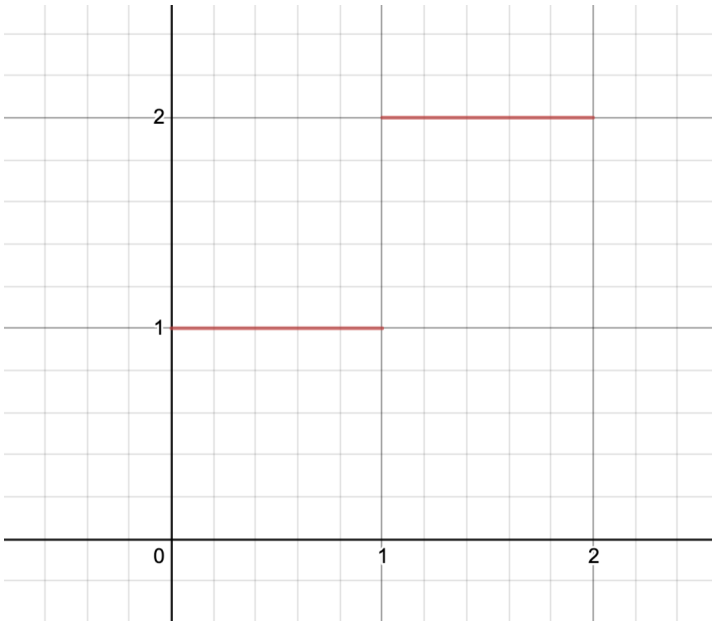
If f is the constant function, $f(x) = c$, the area is clearly $c(b-a)$ (the region is a rectangle with dimensions c and $b - a$). If f is the linear function $f(x) = mx + c$ then the area is equally easy to calculate, since the region is made up of a rectangle and a right triangle. More generally if f is such that the region we are trying to compute the area of is a polygon, then it is relatively straightforward (if somewhat tedious) to compute the area using standard geometric facts.

For general f, however, it is not clear how to calculate "area bounded by curve $y = f(x)$, x-axis, lines $x = a$, $x = b$", or even to interpret what we might mean by that phrase. If, for example, $f(x) = x^2 + 2$ with $a = 0$, $b = 1$, then one of the four lines bounding the relevant area is not a line as such, but a curve, and there are no obvious geometric rules to determine

areas bounded by curves. And if f is defined piecewise, say

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x < 1, \\ 2 & \text{if } 1 \leq x \leq 2, \end{cases}$$

(see picture below) then it is a little unclear what region is under discussion when, say, $a = 0$, $b = 2$. The obvious region (the polygon with vertices $(0,0)$, $(0,1)$, $(1,1)$, $(1,2)$, $(2,2)$ and $(2,0)$ (which has easily calculable area) isn't really the region "bounded" by $y = f(x)$, the x-axis, and the lines $x = a$, $x = b$, since the bounding line from $(1,1)$ to $(1,2)$ is missing.



And when f is something pathological like the stars over Babylon function, questions about area are more than just "a little" unclear.

We get over these issues by the following process, which doesn't calculate exact area, but rather approximates it. More precisely, the process we are about to describe approximates the region we are trying to study by a collection of disjoint rectangles, then calculates the *exact* area of this approximate region, and then attempts to understand what happens to this approximation in some suitable limit.

Mark a, b and a number of points, say $x_1 < \ldots < x_{n-1}$ that are between a and b. Draw vertical lines at $x = x_1$, $x = x_2$, et cetera. This divides the region under discussion into columns, with possibly curved caps. Replace each column by a rectangle, by replacing the curved cap by a straight line that is at approximately the same height as the cap. Then sum up the areas of the rectangles.

If n is small, and/or the gaps between the x_i's are large, then this collection of rectangular columns may not give a good approximation to the area being considered: the curved caps may have quite variable height, and so "replacing the curved cap by a straight line that is at approximately the same height as the cap" may lead to a substantial discrepancy between the area of the approximating rectangle, and the area of the column being approximated.

If n is large, however, and the gaps between the x_i's are small, then it is quite reasonable to expect that the curved caps are reasonably uniform, and that there is little error made in the approximation. In particular, it seems reasonable to imagine that in some kind of limit as n goes to infinity, and the gaps between the x_i's go to 0, the collection of approximate areas converge to something that can be declared to be the actual area.

The goal of this section is to formalize this idea, and to determine large classes of functions for which the idea can be successfully implemented.

Our formalization will involving something called the Darboux integral. It turns out that, unlike the notion of derivative, there are many competing definitions for the integral: there's the Darboux integral, the Riemann integral, the Lebesgue integral, the Riemann-Stieltjes integral, the Henstock integral, and many more. The definitions are all different, and each can be applied to a different class of functions[154]. If you have seen a definition of the integral before, it is almost certainly the Riemann integral. The Darboux integral is defined similarly, but there are significant differences. The advantage of the Darboux integral over the Riemann integral is that the notation is a lot simpler, and this leads to much simpler proofs of all the basic properties of the integral.[155]

10.2 Definition of the Darboux integral

Definition of a partition For real numbers $a < b$, a *partition* P of $[a, b]$ is a sequence of real numbers $(t_0, \ldots, t_n)$ with $a = t_0 < t_1 < \ldots < t_n = b$ (here $n \geq 1$). We refer to the numbers t_i as the *points* of the partition.

Definition of lower and upper Darboux sums Let $f : [a, b] \to \mathbb{R}$ (with $a < b$) be a **bounded** function. For any partition $P = ((t_0, \ldots, t_n))$ of $[a, b]$ the

- *lower Darboux sum* $L(f, P)$

is defined to be

$$L(f, P) = \sum_{i=1}^{n} m_i \Delta_i$$

where

$$m_i = \inf\{f(x) : x \in [t_{i-1}, t_i]\}$$

and

$$\Delta_i = t_i - t_{i-1},$$

and the

- *upper Darboux sum* $U(f, P)$

[154] Of course, whenever two of the definitions can be applied to the same function, they should give the same answer.

[155] It turns out the only difference between the Darboux and the Riemann integral is in the language of their definitions. The set of functions that the two definitions can be applied to end up being exactly the same.

is defined to be

$$U(f,P) = \sum_{i=1}^{n} M_i(t_i - t_{i-1})$$

where

$$M_i = \sup\{f(x) : x \in [t_{i-1}, t_i]\}.$$

Some remarks are in order.

- Recall that for a non-empty set A, $\inf A = \alpha$ means that α is a lower bound for A ($\alpha \leq a$ for all $a \in A$), and is the greatest such (if $\alpha' \leq a$ for all $a \in A$, then $\alpha' \leq \alpha$). The completeness axiom asserts that every non-empty set that has a lower bound, has a greatest lower bound. The set $A = \{f(x) : x \in [t_{i-1}, t_i]\}$ is non-empty, and by the assumption that f is a bounded function, A has a lower bound (any lower bound for f on $[a,b]$ is also a lower bound for A). So by the completeness axiom the number m_i exists, and the lower Darboux sum exists as a finite number. Similarly, the upper Darboux sum exists. If we did not assume that f is bounded, then one or other of $L(f,P)$, $U(f,P)$ might not exist.

- The extreme value theorem tells us that if $f : [a,b] \to \mathbb{R}$ is continuous then it is bounded. We don't initially want to make the restrictive assumption that f is continuous, though; the Darboux integral can handle many more than just continuous functions.

- Both the upper and lower Darboux sums depend on f and P, and this dependence is captured in the notation we use. They also depend on a and b; but since the partition P already determines what a and b are, we don't mention them in the notation.

- We've motivated the need for the integral by considering non-negative f, but in fact nothing we do in the lead-up to the definition of the Darboux integral requires non-negativity; so from here on you should think of f as an arbitrary (bounded) function.

- If you have seen the Riemann integral (Riemann sum) you may be used to the idea of a partition being an *equipartition*, in which all the Δ_i's (the lengths of the sub-intervals in the partition) are the same, namely $\Delta_i = (b-a)/n$. This is *not* (**not**, ***not***) a requirement of the Darboux approach to integration. While we will often consider equipartitions, it will be very helpful for developing general properties of the integral to allow more general partitions.

Convention going forward To avoid unnecessary extra text, from here until further notice we assume that whenever a function f is mentioned without further qualification, it is a bounded function defined on the closed interval $[a,b]$, where $a < b$ are real numbers, that whenever a partition P is mentioned without further qualification, it is the partition $(t_0, \ldots, t_n)$ of $[a,b]$, and that whenever numbers m_i, M_i, Δ_i are mentioned without further qualification, they are

$$m_i = \inf\{f(x) : x \in [t_{i-1}, t_i]\}, \quad M_i = \sup\{f(x) : x \in [t_{i-1}, t_i]\}, \quad \text{and} \quad \Delta_i = t_i - t_{i-1}.$$

A basic fact about Darboux sums is that for any f and any partition P,

$$L(f,P) \leq U(f,P).$$

This follows immediately from the fact the infimum of any set is no bigger than the supremum, so $m_i \leq M_i$ for all i.[156] A much less basic fact is that for any f and any two partitions P, Q (which might have nothing to do with each other),

$$L(f,P) \leq U(f,Q). \ (\star)$$

In other words,

every lower Darboux sum is no bigger than every upper Darboux sum.

This fact, which we will now prove in stages, drives the definition of the integral.

Lemma 10.1. *Suppose Q has one more point than P, say P is $(t_0,\ldots,t_k,t_{k+1},\ldots,t_n)$ and Q is $(t_0,\ldots,t_k,u,t_{k+1},\ldots,t_n)$. Then*

$$L(f,P) \leq L(f,Q) \leq U(f,Q) \leq U(f,P).$$

Proof: We'll show $L(f,P) \leq L(f,Q)$, and leave $U(f,Q) \leq U(f,P)$ as an exercise (the proof is *very* similar to that of $L(f,P) \leq L(f,Q)$). Since we've already observed that $L(f,Q) \leq U(f,Q)$, this gives all the claimed inequalities.

We need a basic fact about infima and suprema, that we will use numerous times as we go on, usually without much comment:

$$\text{If } \emptyset \neq B \subseteq A \text{ and } A \text{ is bounded then } \inf A \leq \inf B. \tag{8}$$

Indeed, suppose $\alpha = \inf A$ and $\beta = \inf B$ (α and β both exist; exercise!). We have $\alpha \leq a$ for all $a \in A$, so $\alpha \leq a$ for all $a \in B$, so α is a lower bound for B, so $\alpha \leq \beta$ (since β is the greatest lower bound for B). [157]

In $L(f,P)$ there is the summand

$$\inf\{f(x) : x \in [t_k, t_{k+1}]\}\Delta_{k+1}.$$

[156] Is it perfectly clear why the infimum of any set is no bigger than the supremum? If it is perfectly clear, that's great. If it's not, that's fine too. Treat verifying this fact (and other, similar facts that will get thrown around later) from the definitions as an exercise.

[157] Informally (8) is the (intuitively obvious) fact that "the infimum of a set can't get smaller if we make the set smaller". Similarly, the supremum can't get bigger if we make the set smaller, that is, if $\emptyset \neq B \subseteq A$ and A is bounded then $\sup A \geq \sup B$. This analogous statement is left as an exercise. In fact, as you'll see, almost every fact we encounter going forward will have an "infimum" part and a "supremum" part, and we will only proof the infima parts, leaving the suprema parts as exercises. Going through the entire derivation of the definition of the integral, only verifying the *suprema* parts (the parts that are not verified in these notes) is a great way of checking that you understand the derivation.

This equals

$$\inf\{f(x) : x \in [t_k, t_{k+1}]\}(t_{k+1} - u) + \inf\{f(x) : x \in [t_k, t_{k+1}]\}(u - t_k).$$

By two applications of (8) this is less than or equal to

$$\inf\{f(x) : x \in [u, t_{k+1}]\}(t_{k+1} - u) + \inf\{f(x) : x \in [t_k, u]\}(u - t_k).$$

This is the sum of two summands that appear in $L(f, Q)$. All other summands in $L(f, P)$ appear in $L(f, Q)$, unchanged, and $L(f, Q)$ has no other summands. It follows that $L(f, P) \leq L(f, Q)$. □

Lemma 10.1 says that if *one* point is added to a partition, it brings the upper and lower Darboux sums closer together. It easily follows that the same is true if *many* points are added.

Lemma 10.2. *If Q has all the points of P, and some more, then*

$$L(f, P) \leq L(f, Q) \leq U(f, Q) \leq U(f, P).$$

Proof: Apply Lemma 10.1 multiple times, adding one new point at a time. (The details are left as an exercise). □

We are now in a position to verify ($\star$).

Lemma 10.3. *If P and Q are* any *partitions of $[a, b]$, then $L(f, P) \leq U(f, Q)$.*

Proof: Let $P \cup Q$ be the partition of $[a, b]$ that includes every point that is either in P, or in Q (or in both). By Lemma 10.2 (applied twice)

$$L(f, P) \leq L(f, P \cup Q) \leq U(f, P \cup Q) \leq U(f, Q).$$

□

We now draw a key corollary.

Corollary 10.4. *Let $f : [a, b] \to \mathbb{R}$ be bounded, with $a < b$. The set of all possible lower Darboux sums for f (as P varies over all possible partitions of $[a, b]$) has a supremum,*

$$L(f) := \sup\{L(f, P) : P \text{ a partition of } [a, b]\},$$

the set of all possible upper Darboux sums for f has an infimum,

$$U(f) := \inf\{U(f, P) : P \text{ a partition of } [a, b]\},$$

and, for any partition P of $[a, b]$,

$$L(f, P) \leq L(f) \leq U(f) \leq U(f, P).$$

Proof: The set $\{L(f,P) : P \text{ a partition of } [a,b]\}$ of lower Darboux sums is non-empty (any partition gives an element), and by Corollary 10.4 it has an upper bound (any upper Darboux sum is an upper bound). So by the completeness axiom $L(f)$ exists. Similarly, $U(f)$ exists.

The inequality $L(f,P) \leq L(f)$ follows immediately from the definition of $L(f)$, and $U(f) \leq U(f,P)$ follows immediately from the definition of $U(f)$, so we just need to verify $L(f) \leq U(f)$. This follows from the basic fact that if non-empty sets A and B are such that everything in A is less than or equal to everything in B, then $\sup A \leq \inf B$ (that everything in $\{L(f,P) : P \text{ a partition of } [a,b]\}$ is less than or equal to everything in $\{U(f,P) : P \text{ a partition of } [a,b]\}$ follows from Corollary 10.4; the basic fact is left as an exercise[158]). □

We are now ready to define the (Darboux) integral.

Definition of (Darboux) integral Let $f : [a,b] \to \mathbb{R}$ be bounded, with $a < b$. Say that f is *(Darboux) integrable* on $[a,b]$ if

$$L(f) = U(f),$$

in which case the common value is the *integral* of f on $[a,b]$, denoted

$$\int_a^b f$$

or

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

If $L(f) < U(f)$ we say that f is *not integrable* on $[a,b]$.

We declare every function defined at a to be integral on the trivial interval $[a,a]$, and set

$$\int_a^a f = 0.$$

Note that if $f(x) \geq 0$ for all $x \in [a,b]$, then we can think of $\int_a^b f$ the *area* of the region bounded by under $y = f(x)$, the x-axis, $x = a$ and $x = b$. As we will shortly see, if f is not always non-negative then $\int_a^b f$ can still be thought of as area, except now those parts of the region that fall below the x-axis contribute negatively.

Note also that in the expression $\int_a^b f(x)\ dx$, x is a *dummy variable* — the numerical value of the expression is unchanged if all occurrences of x are replaced by r, or t, or ♕:

$$\int_a^b f(x)\ dx = \int_a^b f(r)\ dr = \int_a^b f(t)\ dt = \int_a^b f(♕)\ d♕.$$

The definition of the integral is somewhat involved, and not so easy to use for an arbitrary function (certainly, it is much less easy to work with than the definition of the derivative). We will develop some methods that allow for fairly easy calculations of integrals, but for now we just give two very simple examples direct from the definition.

[158] A familiar one: it was Question 4 in Homework 8 of the fall semester.

Example 1: constant function Let f be the constant function $f(x) = c$ for some constant c. For any $a < b$, and any partition $P = (t_0, \ldots, t_n)$ of $[a, b]$, we have (since f is constant on $[t_{i-1}, t_i]$)

$$m_i = M_i = c,$$

and so

$$L(f, P) = \sum_{i=1}^{n} c\Delta_i = c\sum_{i=1}^{n} \Delta_i = c(b - a)$$

and $U(f, P) = c(b-a)$ also[159]. So $L(f) = U(f) = c(b-a)$, f is integrable on $[a, b]$, and

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

(exactly as we would expect).

Example 2: the Dirichlet function Recall that the Dirichlet function $f : \mathbb{R} \to \mathbb{R}$ is defined by

$$f(x) = \begin{cases} 1 & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational.} \end{cases}$$

Given $a < b$, and any partition $P = (t_0, \ldots, t_n)$ of $[a, b]$, we have (since both the rationals and the irrationals are dense in $\mathbb{R}$) that in each interval $[t_{i-1}, t_i]$ there is both a rational and an irrational, and so

$$m_i = 0, \qquad M_i = 1.$$

It follows that $L(f, P) = 0$ and $U(f, P) = 1$, so that $L(f) = 0 < 1 = U(f)$. The conclusion is that the Dirichlet function is *not* integrable on any interval $[a, b]$ with $a < b$.[160]

---

[159] The equality $\sum_{i=1}^{n} \Delta_i = (b-a)$ can be interpreted geometrically: the partition P divides the interval $[a, b]$ into n non-overlapping subintervals, of lengths $\Delta_1, \Delta_2, \ldots, \Delta_n$, and the sum of the lengths of these subintervals is the length of the whole interval, of $b - a$. This is slightly non-rigorous, as we haven't really talked about "length" in any precise way. But this intuitively obvious result can be verified perfectly rigourously, via:

$$\begin{aligned}
\sum_{i=1}^{n} \Delta_i &= \Delta_1 + \Delta_2 + \cdots + \Delta_{n-1} + \Delta_n \\
&= (t_1 - t_0) + (t_2 - t_1) + \cdots + (t_{n-1} - t_{n-2}) + (t_n - t_{n-1}) \\
&= (-t_0 + t_1) + (-t_1 + t_2) + \cdots + (-t_{n-2} + t_{n-1}) + (-t_{n-1} + t_n) \\
&= -t_0 + (t_1 - t_1) + (t_2 - t_2) + \cdots + (t_{n-1} - t_{n-1}) + t_n \\
&= -t_0 + t_n \\
&= b - a.
\end{aligned}$$

This is an example of a *telescoping sum.*

[160] Is this as we might expect? Maybe, or maybe not. The question of the integrability of the Dirichlet function will return in Honors Analysis 1, as the Lebesgue integral is introduced.

To illustrate the issues surrounding using the definition of the Darboux integral to actually *calculate* an integral, consider an example only slightly more complicated that the two discussed above: $f(x) = x$ on $[0, 1]$. For a partition $P = (t_0, \ldots, t_n)$ of $[0, 1]$, we clearly have

$$m_i = \inf\{f(x) : x \in [t_{i-1}, t_i]\} = t_{i-1},$$ [161]

so

$$L(f, P) = \sum_{i=1}^{n} t_{i-1}(t_i - t_{i-1}) = \left(\sum_{i=1}^{n} t_{i-1}t_i\right) - \left(\sum_{i=1}^{n} t_{i-1}^2\right).$$

To compute $L(f)$, we have to maximize the above expression, over all choices of $(t_0, \ldots, t_n)$ satisfying $0 = t_0 < t_1 < \cdots < t_n = 1$; not an easy optimization problem!

On the other hand, the definition of the integral does give us an easy way of putting *bounds* on the value of an integral, once we know it exists: if m and M are lower and upper bounds, respectively, for f on $[a, b]$, then we certainly have $m \leq m_i \leq M_i \leq M$ for all i, and so if $\int_a^b f$ exists then

$$m(b - a) \leq \int_a^b f \leq M(b - a).$$

We now develop a useful criterion for integrability, that allows us to verify integrability not by considering all possible partitions, but instead by focusing on a few carefully chosen ones.

Lemma 10.5.

- *Bounded f is integrable on $[a, b]$ if and only if for every $\varepsilon > 0$, there is a partition P_ε with*
$$U(f, P_\varepsilon) - L(f, P_\varepsilon) < \varepsilon.$$
- *Let f be bounded on $[a, b]$. Suppose there is a number I such that for all $\varepsilon > 0$ there is a partition P_ε such that*
$$L(f, P_\varepsilon) \geq I - \varepsilon$$
and
$$U(f, P_\varepsilon) \leq I + \varepsilon$$
then f is integrable on $[a, b]$, and
$$\int_a^b f = I.$$

Proof: We start with item 1. Suppose that f is integrable, with integral I. There are partitions P_1 and P_2 with $I - \varepsilon/2 < L(f, P_1) \leq I$ and $I \leq U(f, P_2) < I + \varepsilon/2$. Let P be the partition $P_1 \cup P_2$. By Lemma 10.2 we have

$$I - \varepsilon/2 < L(f, P_1) \leq L(f, P) \leq I \leq U(f, P) \leq U(f, P_2) < I + \varepsilon/2$$

[161] There's a simple general fact here, that gets used again and again as we go on, mostly without comment: if f is an increasing function (weakly or strictly) on an interval $[a, b]$, and f is continuous on $[a, b]$, then $\inf\{f(x) : x \in [a, b]\} = f(a)$ and $\sup\{f(x) : x \in [a, b]\} = f(b)$. Similarly if f is decreasing and continuous then $\inf\{f(x) : x \in [a, b]\} = f(b)$ and $\sup\{f(x) : x \in [a, b]\} = f(a)$.

so $U(f,P) - L(f,P) < \varepsilon$.

Conversely, suppose f is *not* integrable on $[a,b]$. Then $U(f) - L(f) = \delta > 0$, and for any $\varepsilon < \delta$ and any P we have $U(f,P) - L(f,P) \geq \delta > \varepsilon$.

We now move on to item 2. From the hypotheses, for every $\varepsilon > 0$ there is a partition P_ε such that

$$L(f,P_\varepsilon) \geq I - \varepsilon/3 \quad \text{and} \quad U(f,P_\varepsilon) \leq I + \varepsilon/3.$$

Combining these inequalities yields

$$U(f,P_\varepsilon) - L(f,P_\varepsilon) \leq 2\varepsilon/3 < \varepsilon,$$

so by item 1 f is integrable.

Since for every $\varepsilon > 0$ there is a partition P_ε with $U(f,P_\varepsilon) \leq I + \varepsilon$, it follows that $U(f,P) \leq I + \varepsilon$ for every $\varepsilon > 0$, so[162] $U(f,P) \leq I$ and $\int_a^b f \leq I$; but by a similar argument $\int_a^b f \geq I$, so $\int_a^b f = I$.

□

We can do something a little bit better than item 1 of Lemma 10.5.

Corollary 10.6. *Bounded f is integrable on $[a,b]$ if and only if for every natural number n there is a partition P_n with*

$$U(f,P_n) - L(f,P_n) < 1/n.$$

Proof: To see that Corollary 10.6 follows from Lemma 10.5 (item 1), first notice that $1/n > 0$, so if f is integrable we can apply the forward direction of Lemma 10.5 (item 1) with $\varepsilon = 1/n$ to get the forward direction of Corollary 10.6.

To get the backward direction of Corollary 10.6, suppose that for every natural number n there is a partition P_n of $[a,b]$ with $U(f,P_n) - L(f,P_n) < 1/n$. Let $\varepsilon > 0$ be given. There's an n with $1/n \leq \varepsilon$; taking P_ε to be P_n we get $U(f,P_\varepsilon) - L(f,P_\varepsilon) < 1/n \leq \varepsilon$, so we can use the backward direction of Lemma 10.5 (item 1) to get the backward direction of Corollary 10.6. □

Corollary 10.6 is nice because it allows us to check integrability by exhibiting just one (carefully chosen) partition for each natural number — a much easier task, in general, than calculating suprema and infima over uncountably infinitely many partitions!

We illustrate this lemma with a few examples.

Example 1: linear function Consider $f(x) = x$ on $[0,1]$. For each n let P_n be the partition $(0, 1/n, 2/n, \ldots, 1)$. We have (using, half way down, an identity that is easy to prove by induction: for natural numbers m,

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

[162] easy exercise.

$$\begin{aligned} L(f,P_n) &= 0\cdot\left(\frac{1}{n}\right)+\frac{1}{n}\cdot\left(\frac{1}{n}\right)+\cdots+\frac{n-1}{n}\left(\frac{1}{n}\right) \\ &= \left(\frac{1}{n^2}\right)(0+1+2+\cdots+(n-1)) \\ &= \frac{n(n-1)}{2n^2} \\ &= \frac{n-1}{2n} \\ &= \frac{1}{2}-\frac{1}{2n}. \end{aligned}$$

Note that to get the second line, we factored $1/n^2$ out of each term. In the first term, we rewrote "0" as "$0/n^2$" to put it on the same footing as all the other terms. Similarly

$$U(f,P_n)=\frac{1}{n}\cdot\left(\frac{1}{n}\right)+\frac{2}{n}\cdot\left(\frac{1}{n}\right)+\cdots+\frac{n}{n}\left(\frac{1}{n}\right)=\frac{1}{2}+\frac{1}{2n}.$$

So $U(f,P_n)-L(f,P_n)=1/n$. Given any $\varepsilon>0$ there is n with $1/n<\varepsilon$, and so for every $\varepsilon>0$ there is a partition P with $U(f,P)-L(f,P)<\varepsilon$. From Lemma 10.5 it follows that $\int_0^1 x\,dx$ exists.[163]

What is the value of the integral? Since $L(f,P_n)=1/2-1/(2n)$, and this can be made arbitrarily close to $1/2$ by choosing n large enough, we get that $L(f)\geq 1/2$[164]. Similarly, $U(f)\leq 1/2$. Since $L(f)=U(f)$, the only possible common value is $1/2$, and so $\int_0^1 x\,dx=1/2$[165]. (We have essentially re-proved item 2 of Lemma 10.5 here.)

More generally, we can easily show that for any $b>0$,

$$\int_0^b x\,dx=\frac{b^2}{2}.$$

Example 2: quadratic function Consider $f(x)=x^2$ on $[0,b]$, $b>0$. Take P_n to be the partition that divides $[0,b]$ into n equal intervals: $P_n=(0,b/n,2b/n,\ldots,(n-1)b/n,b)$. We have (using another identity that is easily proved by induction: for any natural number m,

$$\sum_{k=1}^{m}k^2=\frac{m(m+1)(2m+1)}{6})$$

[163]Or, we could have directly applied Corollary 10.6 at this point, without introducing the ε — I chose to take the long-winded route to reinforce the proof of Corollary 10.6.

[164]This is another of those statement that might be completely clear to you, or might not. If it is completely clear, that is fine. If it is not, that is fine also. You have the tools to verify it formally — go ahead and do so as an exercise!

[165]As we expected: $\int_0^1 x\,dx$ is supposed to be the area of the triangle with vertices $(0,0)$, $(0,1)$ and $(1,1)$, and this is half the unit square.

$$\begin{aligned} L(f,P_n) &= 0^2 \cdot \left(\frac{b}{n}\right) + \frac{b^2}{n^2} \cdot \left(\frac{b}{n}\right) + \cdots + \frac{(n-1)^2 b^2}{n^2}\left(\frac{b}{n}\right) \\ &= \left(\frac{b^3}{n^3}\right)(0^2 + 1^2 + 2^2 + \cdots + (n-1)^2) \\ &= b^3 \frac{(n-1)n(2n-1)}{6n^3} \\ &= b^3 \frac{2n^3 - 3n^2 + n}{6n^3} \\ &= \frac{b^3}{3} - \frac{b^3}{2n} + \frac{b^3}{6n^2}. \end{aligned}$$

Similary

$$U(f,P_n) = \frac{b^2}{n^2} \cdot \left(\frac{b}{n}\right) + \frac{2b^2}{n^2} \cdot \left(\frac{b}{n}\right) + \cdots + \frac{n^2 b^2}{n^2}\left(\frac{b}{n}\right) = \frac{b^3}{3} + \frac{b^3}{2n} + \frac{b^3}{6n^2}.$$

So

$$U(f,P_n) - L(f,P_n) = \frac{b^3}{n}.$$

Exactly as in the case of the linear function, it now follows that $\int_0^b x^2\ dx$ exists and equals $b^3/3$.

Notice that this says something quite non-obvious about an area in the plane that is bounded by some non-linear curves.

More generally, we could push these ideas to show that for any natural number t, and any $b > 0$,

$$\int_0^b x^t\ dx = \frac{b^{t+1}}{t+1}.$$

This would involve a little more computation, and some reasonable expression for

$$\sum_{k=1}^{m} k^t$$

(the sum of the first m perfect tth powers). We'll derive this familiar integral in an easier way shortly, after developing some more theory.

Before doing that, here's one more example. So far all our examples of integrable functions have been continuous; and in fact we will see that *every* continuous function is integrable. But many more functions are integrable. Consider, for example, the function $f : [0,1] \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 1 & \text{if } x = 1, 1/2, 1/3, \ldots \text{ is rational} \\ 0 & \text{otherwise.} \end{cases}$$

This function is discontinuous at infinitely many points, including at a collection of points that are bunched arbitrarily closely together (around 0). Nevertheless, it is integrable on

$[0,1]$. For any partition P of $[0,1]$we have $L(f,P)=0$ (by density of $\mathbb{R}\setminus\{1,1/2,1/3,\ldots\}$), so $L(f)=0$. So to show that the function is integrable on $[0,1]$, we need to show $U(f)=0$ (and this will for free show that the value of the integral in 0). To show $U(f)=0$, it is enough to find, for every $\varepsilon>0$, a partition P_ε with $U(f,P_\varepsilon)<\varepsilon$) (note that we already know $U(f)\geq 0$, since $U(f)\geq L(f)$).

A partition of $[0,1]$ is determined by a finite collection of distinct points in $[0,1]$, that includes 0 and 1. Consider the following set of points, determining a partition P:

- 0 and 1,
- $1/N+1/2N^2$, for some N that will be determined later,
- $1-1/2N^2$,
- and, for each $k=2,3,\ldots,N-1$, the points $1/k-1/2N^2$ and $1/k+1/2N^2$.

This creates the following intervals in the partition:

- $[0,1/N+1/2N^2]$, of length $1/N+1/2N^2$, on which the supremum of f is 1,
- around 1/2, 1/3, ... $1/(N-1)$, intervals of length $1/N^2$, centered at 1/2, 1/3, et cetera, on each of which the supremum of f is 1. Notice that half of the gap from $1/N$ to $1/(N-1)$ is $1/(2N(N-1))$, which is bigger than $1(2N^2)$, so these intervals don't overlap.
- $[1-1/2N^2,1]$, of length $1/2N^2$, on which the supremum of f is 1,
- and many other intervals ($N-2$ many), on which the supremum of f is 0.

We have

$$U(f,P)=\left(\frac{1}{N}+\frac{1}{2N^2}\right)+\frac{N-2}{N^2}+\frac{1}{2N^2}.$$

This quantity can be made arbitrarily small by choosing N large enough; in particular there is an N such that $U(f,P)<\varepsilon$. This shows that

$$\int_0^1 f=0.$$

Notice that in this example we did not use a partition that divides $[a,b]$ into equal-length subintervals.

10.3 Some basic properties of the integral

In this section we gather together some basic facts about integrals, many of which will allow us to discovered the integrability of some new functions from the integrability of old functions.

Splitting an interval of integration

Lemma 10.7. *Fix $a < c < b$. Suppose that (bounded) f is integrable on $[a, b]$. Then it is integrable on both $[a, c]$ and $[c, b]$, and*

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

Proof: Fix $\varepsilon > 0$. By Lemma 10.5 there is a partition P of $[a, b]$ with $U(f, P) - L(f, P) < \varepsilon$. If c is not one of the partition points of P, then, letting P_c be the partition obtained from P by adding c, we have (by Lemma 10.1)

$$L(f, P) \leq L(f, P_c) \leq U(f, P_c) \leq U(f, P),$$

so $U(f, P_c) - L(f, P_c) < \varepsilon$. So we may in fact assume that c is one of the partition points.

We can break P into P', a partition of $[a, c]$ (by taking all the partition points of P between a and c, inclusive), and P'', a partition of $[c, b]$ (by taking all the partition points of P between c and b, inclusive). We have

- $L(f, P) = L(f, P') + L(f, P'')$ and
- $U(f, P) = U(f, P') + U(f, P'')$,

so $[U(f, P') - L(f, P')] + [U(f, P'') - L(f, P'')] < \varepsilon$. Each summand in square brackets is non-negative, so each individually is at most ε. The partitions P' and P'' witness, via Lemma 10.5, that both $\int_a^c f$ and $\int_c^b$ exist.

To get the summation identity ($\int_a^b f = \int_a^c f + \int_c^b f$), let P be any partition. Add c as a partition point (if necessary) to get partitions P_c (of $[a, b]$), P'_c (of $[a, c]$) and P''_c (of $[c, b]$). By the definition of the integral we have

$$L(f, P'_c) \leq \int_a^c f \leq U(f, P'_c) \quad \text{and} \quad L(f, P''_c) \leq \int_c^b f \leq U(f, P''_c)$$

and so, adding these two inequalities and applying Lemma 10.1 if necessary (i.e., if c was not a partition point of P) it follows that

$$L(f, P) \leq L(f, P_c) \leq \int_a^c f + \int_c^b f \leq U(f, P_c) \leq U(f, P)$$

These inequalities are true for any P, so we have

$$\int_a^b f = \sup L(f, P) \leq \int_a^c f + \int_c^b f \leq \inf U(f, P) = \int_a^b f,$$

and so indeed $\int_a^b f = \int_a^c f + \int_c^b f$. □

This result has a converse.

Lemma 10.8. *Fix $a < c < b$. If (bounded) f integrable on both $[a, c]$ and $[c, b]$ then it is integrable on $[a, b]$ (and so, by Lemma 10.7, again $\int_a^b f = \int_a^c f + \int_c^b f$).*

Proof: Fix $\varepsilon > 0$. There are partitions P' of $[a, c]$ and P'' of $[c, b]$ with

$$U(f, P') - L(f, P') < \varepsilon/2 \quad \text{and} \quad U(f, P'') - L(f, P'') < \varepsilon/2$$

Let $P = P' \cup P''$. This is a partition of $[a, b]$ that satisfies

$$U(f, P) - L(f, P) = [U(f, P') + U(f, P'')] - [(L(f, P') + L(f, P'')] < \varepsilon,$$

so by Lemma 10.5 f is is integrable on $[a, b]$. □

We have already encoded in our definition of the integral that for any function f and any real a, $\int_a^a f = 0$. This allows us to make sense of $\int_a^b$ whenever $a \leq b$. To deal with $a > b$, we make a further definition.

For $a > b$, say that $\int_a^b f$ exists if $\int_b^a f$ exists. If $\int_b^a f$ exists, then set $\int_a^b f := -\int_b^a f$.

A consequence of this definition (taken together with Lemmas 10.7 and 10.8) is the following:

Corollary 10.9. *If a, b, c are any three real numbers (distinct or not distinct, and in any relative order), and all the integrals involved exist, then*

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

"Proof": It is a tedious business to verify this in all possible cases[166]. The case $a < c < b$ is exactly Lemmas 10.7 and 10.8. We just verify one other case, and leave the rest as exercises. If $a < b < c$, then $\int_a^c f = \int_a^b f + \int_b^c f$, so $\int_a^b f = \int_a^c f - \int_b^c f$. Now by definition $-\int_b^c f = \int_c^b$, so indeed $\int_a^b f = \int_a^c f + \int_c^b f$. □

Closure under addition

We now move on to some basic closure lemmas for the integral. First, the set of integrable functions is closed under addition.

Lemma 10.10. *If f and g are both integrable on $[a, b]$, then so is $f + g$, and*

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

Proof: Let $P = (t_0, \ldots, t_n)$ be any partition of $[a, b]$. For each i let

- $m_i = \inf\{f(x) + g(x) : x \in [t_{i-1}, t_i]\}$

[166] Question for the curious: how many cases, exactly, have to be considered?

- $m_i^f = \inf\{f(x) : x \in [t_{i-1}, t_i]\}$
- $m_i^g = \inf\{g(x) : x \in [t_{i-1}, t_i]\}$

and define M_i, M_i^f, M_i^g similarly.

It is not unreasonable to expect that $m_i = m_i^f + m_i^g$ and $M_i = M_i^f + M_i^g$, but this is not actually true. What *is* true[167] is that

$$m_i \geq m_i^f + m_i^g \quad \text{and} \quad M_i \leq M_i^f + M_i^g.$$

It follows that

$$L(f, P) + L(g, P) \leq L(f + g, P) \leq U(f + g, P) \leq U(f, P) + U(g, P).$$

Now there is a partition partition P_1 with $U(f, P_1) - L(f, P_1) < \varepsilon/2$, and a partition P_2 with $U(g, P_2) - L(g, P_2) < \varepsilon/2$, so there is a partition P' with $U(f, P') - L(f, P') < \varepsilon/2$ and $U(g, P') - L(g, P') < \varepsilon/2$ (any partition that is a common refinement of P_1 and P_2, that is, that has all the points of P_1 and all the points of P_2, will work). For this partition P' we have

$$U(f + g, P') - L(f + g, P') < \varepsilon$$

so $f + g$ integrable.

What is the value of the integral? For the partition P' created above, we have

$$U(f, P') \leq \int_a^b f + \varepsilon/2$$

($U(f, P') \leq L(f, P') + \varepsilon/2$, and $L(f, P') \leq \int_a^b f$) and

$$U(g, P') \leq \int_a^b g + \varepsilon/2,$$

so

$$U(f, P') + U(g, P') \leq \int_a^b f + \int_a^b g + \varepsilon.$$

Since $U(f + g, P') \leq U(f, P') + U(g, P')$ it follows that

$$U(f + g, P') \leq \int_a^b f + \int_a^b g + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, it follows that the infimum of $U(f + g, P)$ over all partitions P is at most $\int_a^b f + \int_a^b g$, that is,

$$\int_a^b (f + g) \leq \int_a^b f + \int_a^b g.$$

[167] Verifying what follows, and showing by example that the inequality can occur strictly, is left as a homework exercise

But a similar argument using lower Darboux sums gives

$$\int_a^b (f+g) \geq \int_a^b f + \int_a^b g,$$

so that indeed

$$\int_a^b (f+g) = \int_a^b f + \int_a^b g$$

as claimed. □

By induction this result extends to the fact that if $f_1, \ldots, f_n$ are all integrable on $[a,b]$, then so is $\sum_{i=1}^n f_i$, and

$$\int_a^b \sum_{i=1}^n f_i = \sum_{i=1}^n \int_a^b f_i.$$

Closure under multiplication by a constant

Next, the set of integrable functions is closed under multiplication by a constant.

Lemma 10.11. *If f is integrable on $[a,b]$ and λ is any real number then λf is integrable on $[a,b]$, and*

$$\int_a^b \lambda f = \lambda \int_a^b f.$$

Proof: If $\lambda = 0$ then the result is easy. For $\lambda > 0$, fix $\varepsilon > 0$ and let P be a partition of $[a,b]$ with $U(f,P) - L(f,P) < \varepsilon/\lambda$. Using

$$\inf\{(\lambda f)(x) : x \in [t_{i-1}, t_i]\} = \lambda \inf\{f(x) : x \in [t_{i-1}, t_i]\} \quad (\star)$$

and

$$\sup\{(\lambda f)(x) : x \in [t_{i-1}, t_i]\} = \lambda \sup\{f(x) : x \in [t_{i-1}, t_i]\} \quad (\star\star)$$

we quickly get that $U(\lambda f, P) - L(\lambda f, P) < \varepsilon$, so λf is integrable on $[a,b]$. The value of the integral can easily be computed, using a trick similar to the one used in the proof of Lemma 10.10: for the partition P introduced above, $U(f,P) \leq \int_a^b f + \varepsilon/\lambda$, so $\lambda U(f,P) \leq \lambda \int_a^b f + \varepsilon$. But $\lambda U(f,P) = U(\lambda f, P)$, so

$$U(\lambda f, P) \leq \lambda \int_a^b f + \varepsilon$$

from which it follows that $\int_a^b \lambda f \leq \lambda \int_a^b f$. A similar argument gives $\int_a^b \lambda f \geq \lambda \int_a^b f$.

The case $\lambda < 0$ is left as an exercise.[168] □

[168] Here one has to be careful with the inequalities, since multiplying an inequality by a negative number flips the direction of the sign. The analogs of $(\star)$ and $(\star\star)$ that need to be used for $\lambda < 0$ are

$$\inf\{(\lambda f)(x) : x \in [t_{i-1}, t_i]\} = \lambda \sup\{f(x) : x \in [t_{i-1}, t_i]\}$$

and

$$\sup\{(\lambda f)(x) : x \in [t_{i-1}, t_i]\} = \lambda \inf\{f(x) : x \in [t_{i-1}, t_i]\}.$$

Closure under changing finitely many values

A nice corollary of Lemmas 10.10 and 10.11 is that if a function is integrable on an interval, and the values of the function are changed at finitely many places, then it remains integrable, and the value of the integral does not change. This is in sharp contrast to the derivative: changing the value of a differentiable function at a single point means that it is no longer differentiable at that point. The corollary presented below is particularly useful for calculating the integrals of piecewise defined functions, as it allows the values of the function at endpoints of the clauses of definition to be changed, arbitrarily. Some examples of this will appear on homework.

Corollary 10.12. *Suppose that f is integrable on $[a, b]$ and g, defined on $[a, b]$, differs from f at only finitely many values. Then g is integrable on $[a, b]$, and $\int_a^b g = \int_a^b f$.*

Proof: Consider the function[169] $\chi_c : \mathbb{R} \to \mathbb{R}$ defined by

$$\chi_c(x) = \begin{cases} 1 & \text{if } x = c \\ 0 & \text{if } x \neq c. \end{cases}$$

It is an easy check that χ_c is integrable on any interval, and that the integral is always 0.

Now suppose that g differs from f at the points $c_1, \ldots, c_n$. We have

$$g = f + \sum_{i=1}^{n} (g(c_i) - f(c_i))\chi_{c_i}.$$

Repeated applications of Lemmas 10.10 and 10.11 give that g is integrable on $[a, b]$, with

$$\int_a^b g = \int_a^b f + \sum_{i=1}^{n} (g(c_i) - f(c_i)) \int_a^b \chi_{c_i} = \int_a^b f.$$

□

Integral and area

If f is non-negative on $[a, b]$, then we can interpret $\int_a^b f$ as a measure of the area of the region bounded by the curve $y = f(x)$, the x-axis, and the vertical lines $x = a$ and $x = b$. Indeed, for a large class of curves $y = f(x)$ for which we do not have any known formulae for area, we tend to take $\int_a^b f$ as the *definition* of the area.

What if f is non-positive on $[a, b]$? Then $-f$ is non-negative, and so $\int_a^b (-f)$ is the area of the region bounded by the curve $y = -f(x)$, the x-axis, and the vertical lines $x = a$ and $x = b$. This region is the reflection across the x-axis of the region bounded by the curve $y = -f(x)$, the x-axis, and the vertical lines $x = a$ and $x = b$, and (by Lemma 10.11) the value of $\int_a^b (-f)$ is $-\int_a^b f$. So, in the case where f is non-positive on $[a, b]$,

[169] This is sometimes called an *indicator* function: it indicates whether or not the input is c.

$\int_a^b f$ is the negative area of the region bounded by the curve $y = f(x)$, the x-axis, and the vertical lines $x = a$ and $x = b$.

What if f crosses back and forth over the x-axis many times between a and b? Suppose there is a finite collection of numbers $a = c_0 < c_1 < c_2 < \cdots < c_{n-1} < c_n = b$ such that on each interval $[c_0, c_1], [c_1, c_2], \ldots, [c_{n-1}, c_n]$, f is either always non-negative or always non-positive. By Lemmas 10.7 and 10.8 (and an application of induction) we have

$$\int_a^b f = \sum_{i=1}^{n} \int_{c_{i-1}}^{c_i} f,$$

and by the discussion above this is the *signed* area bounded by $y = f(x)$, the x-axis, and the lines $x = a$ and $x = b$: area above the axis is counted positively, and area below the axis is counted negatively.

Although our definition of the integral is motivated by calculating area, it doesn't completely settle the issue. For example, in discussing non-positive functions, we implicitly assumed that if a region A below the x-axis is the reflection across the x-axis of a region B above the x-axis, then A and B have the same area. And in saying something quite reasonable like "the area between $y = x^2$ and $y = x^3$, between $x = 0$ and $x = 1$, is $\int_0^1 x^2\ dx - \int_0^1 x^3\ dx$", we're making another implicit assumption about area, namely that sum of areas of two regions that only overlap in a curve is equal to area of union of regions. There are lots of other subtle area/integral issues like this. In truth, the Darboux integral is not the best tool for dealing with areas of regions in the plane. A better tool is the *Lebesgue integral*, that's defined quite differently from the Darboux integral (but agrees with the Darboux integral on every Darboux integrable function) and can handle the integrals of many more functions, for example the Dirichlet function.

Nonetheless the Darboux integral is still incredibly useful, and there is no harm at all in thinking of it of (informally at least) as an area calculator.

Closure under absolute value and multiplication

Our final basic closure lemma for the integral is that the set of integrable functions is closed under multiplication. The proof involves many steps, which will mostly be left as exercises. Some of the steps are themselves useful lemmas that establish the integrability of certain functions that are obtained by modifying known integrable functions.

We first outline some preliminary steps. These will all be homework problems.

- Suppose that A is a bounded, non-empty set of real numbers. Let $|A| = \{|a| : a \in A\}$. Then
$$\sup |A| - \inf |A| \leq \sup A - \inf A. \quad (\star)$$
- From $(\star)$ it follows that
$$\text{if } f \text{ is integrable on } [a, b], \text{ then so is } |f|.$$

- Define $\max\{f,0\}$ to be the function which at input x takes the value $f(x)$ if $f(x) \geq 0$, and takes value 0 otherwise, and $\min\{f,0\}$ to be the function which at input x takes the value $f(x)$ if $f(x) \leq 0$, and takes value 0 otherwise. If f is integrable on $[a,b]$, then from the integrability of $|f|$ it follows that both $\max\{f,0\}$ and $\min\{f,0\}$ are integrable.

- The *positive part of* f is the function $f^+ = \max\{f,0\}$. Informally, think of the positive part of f as being obtained from f by pushing all parts of the graph of f that lie below the x-axis, up to the x-axis. The *negative part of* f is the function $f^- = -\min\{f,0\}$. Note that $f = f^- f^-$ is a representation of f as a linear combination of non-negative functions. It can be shown that f is integrable on $[a,b]$ if and only if f^+ and f^- are both integrable on $[a,b]$.

The preliminary steps lead to main point, the closure of integrable functions under multiplication.

Lemma 10.13. *Suppose that f and g are both integrable on $[a,b]$. Then so is fg.*

Proof (sketch): We begin by assuming that $f, g \geq 0$. This case is somewhat similar to the proof of closure under addition (Lemma 10.10). With the notation as in that proof, we begin by establishing

$$M_i \leq M_i^f M_i^g \quad \text{and} \quad m_i^f m_i^g \leq m_i$$

(these are left as exercises).

Now for any partition P we have

$$\begin{aligned} U(fg,P) - L(fg,P) &= \sum_{i=1}^{n} (M_i - m_i)(t_i - t_{i-1}) \\ &\leq \sum_{i=1}^{n} (M_i^f M_i^g - m_i^f m_i^g)(t_i - t_{i-1}) \\ &= \sum_{i=1}^{n} (M_i^f M_i^g - m_i^f M_i^g + m_i^f M_i^g - m_i^f m_i^g)(t_i - t_{i-1}) \\ &\leq M \left(\sum_{i=1}^{n} (M_i^f - m_i^f)(t_i - t_{i-1}) + \sum_{i=1}^{n} (M_i^g - m_i^g)(t_i - t_{i-1}) \right). \end{aligned}$$

There's a partition that makes both summands in the last line above at most $\varepsilon/2M$, so fg is integrable on $[a,b]$.

For general f and g, write $f = f^+ - f^-$ where $f^+ = \max\{f,0\}$, $f^- = -\min\{f,0\}$, and $g = g^+ - g^-$, so that

$$fg = f^+ g^+ - f^+ g^- - f^- g^+ + f^- g^-.$$

By the preliminary steps, all of f^+, g^+, f^-, g^- are integrable, and they are all non-negative, so by the early case of this proof, the various products of pairs of them are integrable, and so fg, being a linear combination of integrable functions, is integrable. □

Some integral inequalities

Notice that in the last section we asserted that if f is integrable, then so is $|f|$; and that if f and g are integrable, then so is fg. We did not, however, give a way of expressing $\int_a^b |f|$ in terms of $\int_a^b f$, or $\int_a^b fg$ in terms of $\int_a^b f$ and $\int_a^b g$ (as we did with, for example, closure under addition, where we showed $\int_a^b (f+g) = \int_a^b f + \int_a^b g$). This was not an oversight; there is no way of simply expressing $\int_a^b |f|$ in terms of $\int_a^b f$, and there is no equivalent of the product rule for differentiation, allowing us in general to express $\int_a^b fg$ in terms of $\int_a^b f$ and $\int_a^b g$.

There are, however, some *inequalities* that relate $\int_a^b |f|$ to $\int_a^b f$, and $\int_a^b fg$ to $\int_a^b f$ and $\int_a^b g$. We discuss two of them here. The proof of the first (the triangle inequality) is left as an exercise.

Lemma 10.14. *(Triangle inequality for integrals) If f is integrable on $[a,b]$ then*

$$\left| \int_a^b f \right| \leq \int_a^b |f| \,.$$

The second inequality, the *Cauchy-Schwarz-Bunyakovsky inequality*, is one of the most most important and frequently occurring in analysis.[170]

Lemma 10.15. *if f, and g are both integrable on $[a,b]$ then*

$$\left(\int_a^b fg \right)^2 \leq \left(\int_a^b f^2 \right) \left(\int_a^b g^2 \right).$$

Proof: Let $\alpha = \int_a^b g^2$, $\beta = \int_a^b fg$ and $\gamma = \int_a^b f^2$ (all of which we now know to exist, by Lemma 10.13). For any real number t

$$0 \leq \int_a^b (f - tg)^2 = \alpha t^2 - 2\beta t + \gamma.$$

Because $g^2 \geq 0$ we have $\alpha \geq 0$. If $\alpha > 0$, then the fact that the quadratic $\alpha t^2 - 2\beta t + \gamma$ is always non-negative means that it must have either repeated (real) roots or complex roots. The only way this can happen is if its discriminant $(-2\beta)^2 - 4\alpha\gamma$ is non-positive; but this says that $\beta^2 \leq \alpha\gamma$, which is exactly the claimed inequality.

The case that remains to be considered is $\alpha = 0$. In this case $0 \leq -2\beta t + \gamma$ for all t, which can only happen if also $\beta = 0$; but then the claimed inequality $\beta^2 \leq \alpha\gamma$ is trivial. □

Here's a special[171] (discrete) case of the Cauchy-Schwarz-Bunyakovsky inequality: if $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$ are real sequences then

$$\left(\sum_{i=1}^n a_i b_i \right)^2 \leq \left(\sum_{i=1}^n a_i^2 \right) \left(\sum_{i=1}^n b_i^2 \right).$$

[170] It has it own 300-page book, for example, the wonderful *The Cauchy-Schwarz Master Class (An Introduction to the Art of Mathematical Inequalities)*, by J. Michael Steele, Cambridge University Press, 2004.

[171] Exercise: figure out what choice of integrable functions f, g reduce the integral form to the inequality to the discrete form given here.

Conclusion

Together with the fact that the constant function and the linear function are both integrable, the various lemmas and corollaries presented in this section give a large class of integrable function: for example, all polynomial functions are integrable, as are all functions obtainable from polynomials by any combination of

- altering the function at finitely many values,
- taking absolute value
- taking positive or negative part
- multiplying.

One very large class of integrable functions remains to be explored — the set of continuous functions. That will require a slight digression, which we make in the next section.

10.4 Uniform continuity

Recall what it means for a function f to be *continuous* on an interval I (open, closed, or open at one end, closed at the other): it means that

- for all $c \in I$
- for all $\varepsilon > 0$
- there is $\delta > 0$ such that
- for all $x \in I$[172]

$$x \in (c - \delta, c + \delta) \quad \text{implies} \quad f(x) \in (f(c) - \varepsilon, f(c) + \varepsilon).$$

Notice that δ in this definition depends on *both* ε *and* c, and it may not be the case that the same δ works for a given ε for *all* c in the interval. Consider, for example, the function $f(x) = 1/x$ defined on $(0, 1)$. f is continuous on the interval, but for each ε it is not the case that there is a single δ that works for all $c \in (0, 1)$ in the definition of continuity. To see this, consider $\varepsilon = 1/10$. What does it take to force $f(x)$ to be in the interval

$$(f(c) - 1/10, f(c) + 1/10) = ((10 - c)/10c, (10 + c)/10c)?$$

[172]By saying "$c \in I$" here, we get around the problem of c being one of the endpoints of I. For example, suppose that $I = [c, c']$. In this case, saying $x \in I$ *and* $x \in (c - \delta, c + \delta)$ is the same as saying just $x \in [c, c + \delta)$, and we are where we want to be, asserting the continuity of f at c *from the right*.

If $x \in (c-\delta, c+\delta)$ then $f(x) \in (1/(c+\delta), 1/(c-\delta))$ (and as x ranges over $(c-\delta, c+\delta)$, $f(x)$ ranges over all of $(1/(c+\delta), 1/(c-\delta))$). For $f(x)$ also to always be in the interval $((10-c)/10c, (10+c)/10c)$, it is necessary that

$$\frac{1}{c+\delta} \geq \frac{10+c}{10c},$$

or $\delta < 10c/(10-c) - c$. As c approaches 0, the quantity $10c/(10-c) - c$ approaches 0 also, so as c gets smaller, the required δ to witness that the continuity statement is true (at $\varepsilon = 1/10$) needs to get smaller too, and approaches 0. So no single $\delta > 0$ will work for all c.

Sometimes, a function is continuous on an interval in a more "uniform" way: for every c and every ε there is a δ, but the δ does *not* depend on c, only on ε; it can be chosen after ε has been selected, *without* reference to c.

Consider, for example, $f(x) = x$ on $\mathbb{R}$. Given $\varepsilon > 0$, choose $\delta = \varepsilon$. Now, for each $c \in (0,1)$, suppose $x \in (c-\delta, c+\delta) = (c-\varepsilon, c+\varepsilon)$. Then $f(x) = x \in (c-\varepsilon, c+\varepsilon) = (f(c)-\varepsilon, f(c)+\varepsilon)$. This shows that f is continuous on $\mathbb{R}$, and notice that for each $\varepsilon > 0$ a *single* choice of $\delta > 0$ was enough for all $c \in \mathbb{R}$.

This leads to a new definition, that captures a slightly stronger notion of continuity.

Definition of uniform continuity on an interval Let f be a function defined on an interval I. Say that f is *uniformly continuous* on I if

- for all $\varepsilon > 0$
- there is $\delta > 0$ such that
- for all $c \in I$
- for all $x \in I$

$$x \in (c-\delta, c+\delta) \quad \text{implies} \quad f(x) \in (f(c)-\varepsilon, f(c)+\varepsilon).$$

Symbolically, f is uniformly continuous on I if

$$(\forall \varepsilon > 0)(\exists \delta > 0)(\forall c \in I)(\forall x \in I)([x \in (c-\delta, c+\delta)] \Rightarrow [f(x) \in (f(c)-\varepsilon, f(c)+\varepsilon)]).$$

Some remarks:

- Uniform continuity is a concept that only makes sense *on an interval*: it makes no sense to say that a function is uniformly continuous at a point.
- A function that is uniformly continuous on an interval, is necessarily continuous on that interval (easy exercise), but as we have seen from an example, the converse is not true.
- Compare the definition of "f is continuous on I:

$$(\forall c \in I)(\forall \varepsilon > 0)(\exists \delta > 0)(\forall x \in I)([x \in (c-\delta, c+\delta)] \Rightarrow [f(x) \in (f(c)-\varepsilon, f(c)+\varepsilon)]).$$

with the definition of "f is uniformly continuous on I:

$$(\forall \varepsilon > 0)(\exists \delta > 0)(\forall c \in I)(\forall x \in I)([x \in (c-\delta, c+\delta)] \Rightarrow [f(x) \in (f(c)-\varepsilon, f(c)+\varepsilon)]).$$

The difference is in the order of the quantifiers.

- There is an alternate form of the definition of uniform continuity, which will be useful when we prove the main theorem of this section (Theorem 10.17). The premise "$x \in (c-\delta, c+\delta)$" of the implication in the definition is exactly the same as the premise "$|x-c| < \delta$" ("x is within δ of c" is the same as "x and c are within δ of each other"). Similarly the conclusion "$f(x) \in (f(c)-\varepsilon, f(c)+\varepsilon$" is the same as "$|f(x)-f(c)| < \varepsilon$". The usefulness of the alternate premise and conclusion is that they are more symmetric than the premise and conclusion they replace. To further highlight this symmetry, we can replace "c" with "y", to get the following, equivalent, formulation of uniform continuity: f is *uniformly continuous* on I if

  - for all $\varepsilon > 0$
  - there is $\delta > 0$ such that
  - for all $x, y \in I$
$$|x-y| < \delta \quad \text{implies} \quad |f(x)-f(y)| < \varepsilon.$$

  Symbolically,

$$(\forall \varepsilon > 0)(\exists \delta > 0)(\forall x \in I)(\forall y \in I)([|x-y| < \delta] \Rightarrow [|f(x)-f(y)| < \varepsilon]).$$

Uniformly continuous functions are often easier to work with than continuous functions. Here's an illustration — the proof that if f is uniformly continuous on an interval $[a, b]$, then it is integrable on $[a, b]$.

Theorem 10.16. *If $f : [a, b] \to \mathbb{R}$ is uniformly continuous on $[a, b]$, then f is integrable on $[a, b]$.*

Proof: Let $\varepsilon > 0$ be given. We need to find a partition P of $[a, b]$ with $U(f, P) - L(f, P) < \varepsilon$.

Because f is uniformly continuous on $[a, b]$, there is a $\delta > 0$ such that for all $c \in [a, b]$, if $x \in [a, b]$ and $x \in (c-\delta, c+\delta)$ then $f(x) \in (f(c) - \varepsilon/(3(b-a)), f(c) + \varepsilon/(3(b-a)))$.

Let P be any partition of $[a, b]$ in which the length $t_i - t_{i-1}$ of every interval $[t_{i-1}, t_i]$ in the partition is less than 2δ. Let c_i be the midpoint of $[t_{i-1}, t_i]$. If $x \in [t_{i-1}, t_i]$, then $x \in (c-\delta, c+\delta)$, so $f(x) \in f(c_i) - \varepsilon/(3(b-a)), f(c_i) + \varepsilon/(3(b-a)))$. It follows that

- $m_i \geq f(c_i) - \varepsilon/(3(b-a))$,
- $M_i \leq f(c_i) + \varepsilon/(3(b-a))$, and
- $M_i - m_i \leq 2\varepsilon/(3(b-a))$,

so

$$U(f,P) - L(f,P) = \sum_{i=1}^{n} (M_i - m_i)(t_i - t_{i-1}) \leq (2\varepsilon/(3(b-a)))(b-a) < \varepsilon.$$

□

This raises a natural question: *which* continuous functions on $[a,b]$ are *uniformly* continuous? All of them, it turns out!

Theorem 10.17. *If $f : [a,b] \to \mathbb{R}$ is continuous, it is uniformly continuous.*

Corollary 10.18. *If $f : [a,b] \to \mathbb{R}$ is continuous, it is integrable.*

So, the set of functions on $[a,b]$ that are differentiable everywhere, sits strictly inside the set of functions on $[a,b]$ that are continuous everywhere, and this in turn sits strictly inside the set of functions on $[a,b]$ that are integrable.

Proof (of Theorem 10.17): Let $\varepsilon > 0$ be given. We would like to show that there is a $\delta > 0$ such that if $x, y \in [a,b]$ are within δ of each other, $f(x)$ and $f(y)$ are within ε of each other. Following Spivak, we say that f is *ε-good* on $[a,b]$ if this happens.

Let $A = \{x \in [a,b] : f \text{ is } \varepsilon\text{-good on } [a,x]\}$. We have $a \in A$ (and $\delta > 0$ will work to witness that f is ε-good on $[a,a]$), and b is an upper bound for A, so by the completeness axiom A has a least upper bound, $\alpha = \sup A$. We will argue first that $\alpha = b$, and then that $\alpha \in A$, to conclude that f is ε-good on $[a,b]$; since $\varepsilon > 0$ was arbitrary, this completes the proof.

Suppose $\alpha < b$. Since f is continuous at α, there is a $\delta > 0$ such that $|y - \alpha| < \delta$ implies $|f(y) - f(\alpha)| < \varepsilon/2$ (and we can choose δ small enough that $|y - \alpha| < \delta$ implies $y \in [a,b]$). So if y, z are both in $(\alpha - \delta, \alpha + \delta)$ then

$$|f(y) - f(z)| \leq |f(y) - f(\alpha)| + |f(z) - f(\alpha)| < \varepsilon.$$

From this we can conclude that f is ε-good on $[\alpha - \delta/2, \alpha + \delta/2]$. But also, because $\alpha = \sup A$, there is some $x \in (\alpha - \delta/3, \alpha]$ with f ε-good on $[a,x]$, and so also f is ε-good on $[a, \alpha - \delta/2]$.

We now need a lemma:

> Suppose $p < q < r$ and f is continuous on $[p,r]$. If f is ε-good on $[p,q]$, and on $[q,r]$, then it is ε-good on $[p,r]$.

With this lemma we conclude that f is ε-good on $[a, \alpha + \delta/2]$, contradicting that $\alpha = \sup A$; so we conclude that $\alpha = b$.

We finish the proof of the main theorem, by establishing that $b \in A$, before proving the lemma. Since f is left-continuous at b, there is a $\delta > 0$ with f is ε-good on $[b - \delta/2, b]$; and since $b = \sup A$, f is ε-good on $[a, b - \delta/2]$ (both of these steps are very similar to our arguments about α). From the lemma it follows that f is ε-good on $[a,b]$, as required.

We now turn to the proof of the lemma. There is $\delta_1 > 0$ such that

$$\text{for } x, y \in [p, q],\ |x - y| < \delta_1 \text{ implies } |f(x) - f(y)| < \varepsilon,\ (\star)$$

and there is $\delta_2 > 0$ such that

$$\text{for } x, y \in [q, r],\ |x - y| < \delta_2 \text{ implies } |f(x) - f(y)| < \varepsilon.\ (\star\star)$$

Since f is continuous at q there is also $\delta_3 > 0$ such that $|x-q| < \delta_3$ implies $|f(x)-f(q)| < \varepsilon/2$, so

$$\text{for } x, y \text{ with } |x - q| < \delta_3 \text{ and } |y - q| < \delta_3, \text{ we have } |f(x) - f(y)| < \varepsilon.\ (\star\star\star)$$

Let $\delta = \min\{\delta_1, \delta_2, \delta_3\}$. Suppose we have $x, y \in [p, r]$ with $|x - y| < \delta$.

- If $x, y \in [p, q]$ then $|f(x) - f(y)| < \varepsilon$, by $(\star)$.
- If $x, y \in [q, r]$ then $|f(x) - f(y)| < \varepsilon$, by $(\star\star)$.
- If $x < q < y$ or $y < q < x$ then both $|x-q| < \delta$ and $|y-q| < \delta$ hold, and $|f(x)-f(y)| < \varepsilon$ by $(\star\star\star)$.

□

10.5 The fundamental theorem of calculus, part 1

Suppose that a function f is integrable on some interval I. Fix $a \in I$. We can define a new function $F : I \to \mathbb{R}$ by

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

We can think of F as a function which "accumulates area": if f is non-negative, then (by definition) $F(a) = 0$, and for $x > a$ $F(x)$ is the area of the region bounded by the graph of f, the x-axis, and vertical lines at a and x. As x increases, so does F, as more and more area is accumulated. On the other hand if $x < a$ then $F(x)$ is a negative area, and as x gets further from a in the negative direction, F gets more negative. In what follows we will (somewhat informally) refer to F as the "integral" of f.

Recall that we saw that often the derivative of a function is less well behaved than the original function. For example, the function $f(x) = |x|$ is continuous everywhere, but not differentiable everywhere; and the function

$$f(x) = \begin{cases} x^2/2 & \text{if } x \geq 0 \\ -x^2/2 & \text{if } x \leq 0 \end{cases}$$

is differentiable everywhere, but it's derivate ($f'(x) = |x|$) is *not* differentiable everywhere.

The integral of a function, on the other hand, tends to be better behaved than the original function, in the sense that it is typically "smoother". Consider, for example,

$$f(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0. \end{cases}$$

This function is not continuous on any interval that includes 0. It is integrable on any interval, however, and so we can define the function $F(x) = \int_0^x f(t)\ dt$[173]. Without delving into the gritty details of the definition of the integral, we can calculate the values of $F(x)$ by remember the "accumulating area" interpretation. For $x \geq 0$, $F(x)$ is the area of a square whose dimensions are x by 1, so $F(x) = x$. For $x < 0$ we have

$$F(x) = \int_0^x f(t)\ dt = -\int_x^0 f(t)\ dt = \int_x^0 (-f)(t)\ dt.$$

Since $-f \geq 0$ on $[x, 0]$, this last integral is the area of a square whose dimensions are $-x$ by 1, so $F(x) = -x$. Combining we find that $F(x) = |x|$. So in this example

$$f \text{ is not continuous everywhere, but its integral } F \text{ is.}$$

As a second example, consider $f(x) = |x|$. Again thinking about the integral function as accumulating area, we find that

$$F(x) := \int_0^x f(t)\ dt = \begin{cases} -x^2/2 & \text{if } x < 0 \\ x^2/2 & \text{if } x \geq 0. \end{cases}$$

Notice that f is continuous at 0, but not differentiable, but F is "smoother", being not just continuous at 0 but also differentiable. So in this example

$$f \text{ is not differentiable everywhere, but its integral } F \text{ is.}$$

These are just two example, and are not much on which to base a general hypothesis. But we could work through many more examples, and always find the same thing happening — if f is integrable, then integration "smooths out" f in the sense that F, the integral of f,

[173]A note on notation: when we say "$f : [a, b] \to \mathbb{R}$ is function", we mean that symbol f is standing for set a of ordered pairs, whose set of first coordinates is exactly $[a, b]$, and with no element of that set appearing twice as a first coordinate. If f is integrable on $[a, b]$, we use $\int_a^b f$ to denote integral.

Sometimes we represent a function by the expression "$f(x)$", which is usually a rule explaining how to compute the second co-ordinate of the pair whose first co-ordinate is x; for example, $f(x) = x^2$. In this case we write $\int_a^b f(x)\ dx$. Here "dx" means *nothing* on its own.

Note that the "x" in here (in the integral "$\int_a^b f(x)\ dx$") is *not* a variable of a function — $\int_a^b f(x)\ dx$ is not a function, it is a *number*. "x" is a *dummy variable*, and can be given any name we like: $\int_a^b f(x)\ dx = \int_a^b f(y)\ dy = \int_a^b f(\text{banana})\ d\text{banana}$.

Sometimes we are *forced* to use a name other than x in the presentation of an integral: For example here we want to define a function F, whose value at input x is the integral of f on the interval $[a, x]$. We could denote that by $F(x) = \int_a^x f$, or, using the "$f(\cdot)d\cdot$" notation, we could denote it by $F(x) = \int_a^x f(t)\ dt$, or $F(x) = \int_a^x f(r)\ dr$, or .... What we *cannot* do is write $F(x) = \int_1^x \frac{dx}{x}$ — we're already using the symbol x for the variable inputed into the function F (and it *is* a variable in this equation), and so we need another, different (and quite arbitrary) name for the dummy variable.

Sometimes a variable in the limits of integration *can* appear sensibly inside integral. For examle, the expression $\int_1^x xt\ dt$ makes perfect sense for all x. At $x = 2$, it is the number $\int_1^2 2t\ dt$, at $x = -1$ it is the number $\int_1^{-1}(-t)\ dt$, and so on.

is continuous even where f is not, and moreover is differentiable whenever f is continuous, even when f is not differentiable.

We now prove that these phenomena occur in general. We start with continuity.

Proposition 10.19. *Suppose $f : I \to \mathbb{R}$ (defined on an interval I) is integrable on any closed interval contained in I. Fix $a \in I$ and define $F : I \to \mathbb{R}$ by*

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

Then F is continuous at all points in I.

Proof: Fix $a < c < b \in I$. We will argue that F is continuous from the right at c, and from the left. We start with right-continuity.

Let $h > 0$. Define

- $M_h = \sup\{f(x) : x \in [c, c+h]\}$
- $m_h = \inf\{f(x) : x \in [c, c+h]\}$
- M to be any be such that $|f(x)| < M$ for all $x \in [a, b]$.

Observe that $-M \leq m_h \leq M_h \leq M$ (as long as h is small enough that $c + h \leq b$), and $M > 0$.

We have

$$F(c+h) - F(c) = \int_c^{c+h} f$$

so

$$-Mh \leq m_h h \leq F(c+h) - F(c) \leq M_h h \leq Mh.$$

By the squeeze theorem, as $h \to 0^+$ get $F(c+h) \to F(c)$, so F is right continuous at c.

For left continuity, again let $h > 0$. Re-define

- $M_h = \sup\{f(x) : x \in [c-h, c]\}$
- $m_h = \inf\{f(x) : x \in [c-h, c]\}$

We still have $-M \leq m_h \leq M_h \leq M$ (as long as $c - h \geq a$), and $M > 0$. Now

$$F(c) - F(c-h) = \int_{c-h}^c f$$

so

$$-Mh \leq hm_h \leq F(c) - F(c-h) \leq hM_h \leq Mh.$$

Again by the squeeze theorem, as $h \to 0^+$ get $F(c-h) \to F(c)$, so F is left continuous at c.

Notice that there might not *be* a number b such that $a < c < b$; this happens if c is the right-endpoint of I. It is left as an exercise to consider what happens in this case (and so only continuity from the left need be considered). The case $c = a$ is also left as an exercise.

What happens if $c < a$? Now $F(c) = \int_a^c f = -\int_c^a f$, and we can run a very similar argument to the one presented above to obtain the desired result. □

Returning to our two examples, recall that if

$$f(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0. \end{cases}$$

then $F(x) := \int_0^x f = |x|$. The function f is continuous for all x other than 0, and the integral F is *differentiable* at all those points. Moreover, the derivative of f at every non-zero point x, is exactly the value of f at x.

The same holds for our other example, $f(x) = |x|$, for which

$$F(x) := \int_0^x f(t)\, dt = \begin{cases} -x^2/2 & \text{if } x < 0 \\ x^2/2 & \text{if } x \geq 0 : \end{cases}$$

f is continuous everywhere (but not differentiable everywhere), and the integral F is differentiable everywhere, with $F'(x) = f(x)$ for all x.

It is not hard to intuitively see that this is a general phenomenon. Suppose that f is a continuous function, and define $F(x) = \int_a^x f$ (for some fixed a). We have, for $h > 0$

$$\frac{F(x+h) - F(x)}{h} = \frac{1}{h}\int_x^{x+h} f \approx f(x)$$

using that, since f is continuous at x its values close to x are close to $f(x)$, so $\int_x^{x+h} f$ is approximately $f(x)$ times the length of the interval $[x, x+h]$, or $hf(x)$. This strongly suggests that F is differentiable[174] with derivative $f(x)$.

That this is in fact true in general is the content of one of two theorems that are significant enough to warrant the adjective "fundamental".

Theorem 10.20. *(fundamental theorem of calculus, Part 1) Suppose $f : I \to \mathbb{R}$ (defined on an interval I) is integrable on any closed interval contained in I. Fix $a \in I$ and define $F : I \to \mathbb{R}$ by*

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

If f is continuous at some $c \in I$, then F is differentiable at c, and $F'(c) = f(c)$.

Proof: As with the proof of Proposition 10.19, we begin by considering $a < c < b \in I$. What happens when c is an endpoint of I is left as an exercise, as are the cases $c = a$ and $c < a$.

We begin by showing that F is differentiable from the right at c. With the notation from the proof of Proposition 10.19, we have from that proof that

$$m_h \leq \frac{F(c+h) - F(c)}{h} \leq M_h.$$

[174] at least, from the right; but we make make a similar argument to justify differentiability from the left

Now since f is continuous at c, we have

$$\lim_{h\to 0^+} M_h = \lim_{h\to 0^+} m_h = f(c) \quad (\star)$$

(and so, by the squeeze theorem, F is right differentiable at c with derivative $f(c)$). To see $(\star)$ note that:

- Given $\varepsilon > 0$ there's $\delta > 0$ such that $x \in (c-\delta, c+\delta)$ implies $f(x) \in (f(c)-\varepsilon, f(c)+\varepsilon)$.
- Choose $h = \min\{\delta, b-c\}$; for this h, $f(c) \le M_h \le f(c)+\varepsilon$ and $f(c)-\varepsilon \le m_h \le f(c)$, and indeed for all $0 < h' \le h$ we have $f(c) \le M_{h'} \le f(c)+\varepsilon$ and $f(c)-\varepsilon \le m_{h'} \le f(c)$.
- Letting $\varepsilon \to 0$, the squeeze theorem gives $(\star)$.

For left differentiability we have from the proof of Proposition 10.19 that

$$m_h \le \frac{F(c)-F(c-h)}{-h} \le M_h,$$

and the proof proceeds as before. □

We conclude this section with two remarks.

- It is not uncommon to see the Fundamental Theorem of Calculus, part 1 (from here on abbreviated FTOC1) presented as:

  > if f is continuous on I, $a \in I$ and F is defined on I by $F(x) = \int_a^x f$, then F is differentiable on I with $F' = f$.

  This *follows* from FTOC1 as we have given it (just apply FTOC1 at each point in I), but it doesn't obviously *imply* our FTOC1, which allows for the possibility that f is *not* continuous at lots of points of I, but still says something about F at those points where f happens to be continuous.

- When dealing with the case $c < a$ in the proofs of Proposition 10.19 and Theorem 10.20, rather than repeating the arguments for continuity and differentiability of F given for $c > a$, it's possible to take the following approach: for input $x < a$, choose a number $b < x$. We have

  $$F(x) = \int_a^x f = -\int_x^a f = -\left(\int_b^a f - \int_b^x f\right) = \int_b^x f - \int_b^a f = G(x) - \int_b^a f$$

  where G is defined by $G(x) = \int_b^x f$. Because $x > b$ we know (from the arguments given in the proofs of Proposition 10.19 and Theorem 10.20) that G is continuous at x, so F is also, and that if f is continuous at x then G is differentiable at x with $G'(x) = f(x)$, so F is also differentiable at x with $F'(x) = f(x)$.

10.6 The fundamental theorem of calculus, part 2

FTOC1 says (loosely) that if a new function is defined from an old function by integration, then (under appropriate circumstances) the derivative of that new function is the old function. So "differentiation undoes integration". The connection between differentiation and integration goes the other way, too. We start with a statement that is a fairly immediate corollary of FTOC1, that we will think of as a "weak" version of what we will eventually call the fundamental theorem of calculus, part 2.

Corollary 10.21. *(Weak FTOC2; a corollary of FTOC1) Suppose that f is continuous on $[a,b]$. If there is a function $g:[a,b]\to\mathbb{R}$ such that $g'=f$ on $[a,b]$, then*

$$\int_a^b f = g(b)-g(a).$$

Proof: Since f is continuous, it is integrable, so the function $F(x)=\int_a^x f$ exists; and moreover, by FTOC1, F is differentiable on $[a,b]$, with $F'=f=g'$. Since $F'=g'$ it follows that there is a number c such that $F(x)=g(x)+c$. Evaluating at $x=a$ we get $c=-g(a)$, and then evaluating at b we get $F(b)=g(b)-g(a)$, as claimed. □

The power of this corollary is that if a function f is continuous, and if we can find an *antiderivative* or *primitive* of f — a function g such that $g'=f$, then we can *very easily* evaluate integrals involving f.

For example, what is the area under $y=x^n$, above x-axis, between $x=a$ and $x=b$? In other words, what is

$$\int_a^b x^n\,dx?$$

Here $f(x)=x^n$, and since, as long as n is an integer other than -1

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

has $g'(x)=f(x)$, we get

$$\text{Area}=\int_a^b x^n\,dx=g(b)-g(a)=\frac{b^{n+1}-a^{n+1}}{n+1}. \tag{9}$$

Calculating this directly, using the definition of the integral, would be quite unpleasant.

If $n=-2$, $a=1$ and $b=10$, we have

$$\text{Area}=\int_1^{10}\frac{dx}{x^2}=\left(-\frac{1}{10}\right)-\left(-\frac{1}{1}\right)=\frac{9}{10},$$

a reasonable answer. If $n=-2$ and we try to use (9) to calculate the area under $y=x^2$, above the x-axis, and between $x=-a$ and $x=a$ for some positive a, we get

$$\int_{-a}^{a}\frac{dx}{x^2}=\left(-\frac{1}{a}\right)-\left(-\frac{1}{-a}\right)=-\frac{2}{a},$$

which is clearly wrong, since the function we are integrating is always positive!

What went wrong here? The issue is that the expression $1/x^2$ is not defined on all of $[-a, a]$ (specifically it is not defined at 0), and so the integral expression makes no sense. So we need to insert a caveat into (9): for negative n we have

$$\int_a^b x^n \, dx = g(b) - g(a) = \frac{b^{n+1} - a^{n+1}}{n+1}$$

only if either $0 < a \leq b$ or $a \leq b < 0$.

We specified above that $n \neq -1$; this is because there is no obvious g with $g'(x) = 1/x$. We know, though, that there *must be* such a function g: since $1/x$ is a continuous function away from 0, by FTOC1 if we define (for positive x)

$$g(x) = \int_1^x \frac{dt}{t}$$

then $g'(x) = 1/x$. But, if we know of no better expression for $g(x)$, then Corollary 10.21 (weak FTOC2) tells us nothing about the value of integrals of $1/x$: it tells us that

$$\int_1^a \frac{dx}{x} = \int_1^a \frac{dt}{t} - \int_1^1 \frac{dt}{t} = \int_1^a \frac{dt}{t},$$

a trivial tautology.[175]

We refer to Corollary 10.21 as "weak" because it includes as a hypothesis that f is continuous. As we have seen, this makes it little more than a corollary of FTOC1. We can in fact drop this hypothesis, or rather, replace it with the weaker hypothesis that f is merely integrable. The proof of what we will call the *fundamental theorem of calculus, part 2* cannot appeal to FTOC1 (because no assumption about continuity of f is made), and so has to go back to the definition of the integral. Nonetheless, it is quite a short proof.

Theorem 10.22. *(fundamental theorem of calculus, part 2) Suppose $f : [a, b] \to \mathbb{R}$ is integrable, and that there is some function $g : [a, b] \to \mathbb{R}$ such that $g' = f$. Then*

$$\int_a^b f = g(b) - g(a)$$

.

Proof: g is differentiable on $[a, b]$, so it is continuous. Given a partition P of $[a, b]$, there is an x_i in each $[t_{i-1}, t_i]$ with

$$g(t_i) - g(t_{i-1}) = g'(x_i)(t_i - t_{i-1}) = f(x_i)(t_i - t_{i-1}).$$

(The first equality here is an application of the Mean Value Theorem.) Now $f(x_i) \in [m_i, M_i]$, so from the above we get

$$m_i(t_i - t_{i-1}) \leq g(t_i) - g(t_{i-1}) \leq M_i(t_i - t_{i-1})$$

[175] We will (of course) return to $\int_1^x dt/t$ shortly!

Summing these inequalities from $i = 1$ to n we get

$$L(f, P) \leq g(b) - g(a) \leq U(f, P).$$

This is true for *all* partitions P. Since f is integrable, it follows that $\int_a^b f = g(b) - g(a)$. $\square$

Note that FTOC2 *does not* assert that if f is integrable, then there is a function g with $g' = f$; on the homework, there is example of an $f : [a, b] \to \mathbb{R}$ that is integrable, for which there is no g satisfying $g' = f$. Rather, FTOC2 assert that *if* f is integrable *and also* there is a function g with $g' = f$, then $\int_a^b f$ can be computed using g.

The idea of taking a known function f and creating from it a new function F via $F(x) = \int_a^x f$ is a very valuable one; it will allow us to properly define many of the basic transcendental (beyond ration) functions of mathematics, such as

- the trigonometric functions
- the hyperbolic trigonometric functions (sinh, cosh, et cetera)
- the logarithmic and exponential functions.

It's important to be able to work with fluently with this construction, and to manipulate functions defined in this way just as we manipulated functions defined in a more standard way. We present some examples here.

- Suppose $f(x) = \int_1^{x^2-x} g(t)\ dt$. What is $f'(x)$?

  Define $F(x) = \int_1^x g(t)\ dt$. Then $f(x) = F(x^2 - x) = (F \circ h)(x)$ where $h(x) = x^2 - x$, so, by the chain rule, $f'(x) = F'(h(x))h'(x)$. Now $F'(x) = g(x)$, so $F'(h(x)) = g(h(x))$, so

  $$f'(x) = g(x^2 - x)(2x - 1).$$

- Suppose $q(x) = \cos\left(\int_{\sqrt{x}}^a g(t)\ dt\right)$. What is $q'(x)$?

  Well, $q(x) = (\cos \circ m \circ F \circ s)(x)$ where $s(x) = \sqrt{x}$, $F(x) = \int_a^x g(t)\ dt$, and $m(x) = -x$, so

  $$\begin{aligned} q'(x) &= \cos'(m \circ F \circ s(x))m'((F \circ s)(x))F'(s(x))s'(x) \\ &= \left[-\sin\left(\int_{\sqrt{x}}^a g(t)\ dt\right)\right] \times [-1] \times \left[g(\sqrt{x})\right] \times \left[\frac{1}{2x^{1/2}}\right]. \end{aligned}$$

These and other similar examples are essentially all exercises in the chain rule.

10.7 Improper integrals

The aim of this section is to make sense of expressions like

- $\int_0^\infty f$,

and

- $\int_a^b f$ when f is not necessarily bounded on $[a, b]$.

We will refer to integrals like this – whose interpretations are derived from the definition of the integral, but do not perfectly fit the conditions of the definition — as *improper integrals*.

We start with integrals like $\int_0^\infty f$, where there is an obvious definition:

Definition of improper integral (I) The expression $\int_a^\infty f$ is defined to mean

$$\int_a^\infty f := \lim_{N\to\infty} \int_a^N f,$$

as long as this limit exists (and so, in particular, as long as f is bounded on $[a, N]$ and $\int_a^N f$ exists, for every large enough N[176]

As an example, we consider $\int_1^\infty \frac{dx}{x^r}$, where $r \geq 2$ is an integer.[177]. Certainly $\int_1^N \frac{dx}{x^r}$ exists for all $N \geq 1$, since $f(x) = 1/x^r$ is continuous on any such interval. Moreover the integral is easy to calculate since f has a simple antiderivative: $g(x) = -1/((r-1)x^{r-1})$. So

$$\int_1^N \frac{dx}{x^r} = g(N) - g(1) = \frac{1}{r-1} - \frac{1}{(r-1)N^{r-1}}.$$

Now for $r \geq 2$ we have $r - 1 \geq 1$ and $1/N^{r-1} \to 0$ as $N \to \infty$, and so

$$\int_1^\infty \frac{dx}{x^r} = \lim_{N\to\infty} \left(\frac{1}{r-1} - \frac{1}{(r-1)N^{r-1}} \right) = \frac{1}{r-1}.$$

This clearly doesn't work when $r = 1$; and in fact the above computation hints that $\int_1^\infty \frac{dx}{x}$ probably does not exist. To see this formally we want to show that

$$\lim_{N\to\infty} \int_1^N \frac{dx}{x}$$

does not exist. We will do this by considering $\int_1^{2^n} \frac{dx}{x}$, for integers $n \geq 0$. We cannot evaluate this integral directly (we do not yet know anything about an antiderivative for $1/x$), but we can put a lower bound on the integral by observing that on the interval $[1, 2]$, the function f

[176] But of course that immediately means, via Lemma 10.7, that f is bounded on $[a, N]$, and $\int_a^N f$ exists, for all $N \geq a$.

[177] We haven't yet properly defined x^r for r *not* an integer — that will come soon.

has $1/2$ as a lower bound; on $[2,4]$ it has $1/4$, and in general on $[2^{k-1}, 2^k]$ there is a lower bound of $1/2^k$. So the partition $P = (1, 2, 4, \ldots, 2^i, \ldots, 2^n)$ has

$$L(f,P) = \sum_{i=1}^{n} m_i(t_i - t_{i-1}) = \sum_{i=1}^{n} \frac{2^{i-1}}{2^i} = \frac{n}{2}.$$

Since $n/2 \to \infty$ as $n \to \infty$, this shows that $L(f)$ does not exist.

Definition of improper integral (II) The expression $\int_{-\infty}^{a} f$ is defined to mean

$$\int_{-\infty}^{a} f := \lim_{N \to -\infty} \int_{N}^{a} f,$$

as long as this limit exists.

What about an expression like $\int_{-\infty}^{\infty} f$? It is tempting to consider $\lim_{N\to\infty} \int_{-N}^{N} f$, and take this limit (if it exists) to be the value of the improper integral. But this approaches leads to some unfortunate oddities. For example, we would have

$$\int_{-\infty}^{\infty} x \, dx \neq \int_{-\infty}^{a} x \, dx + \int_{a}^{\infty} x \, dx$$

since the integral on the left would equal 0, and neither of the two integrals on the right exist.[178]

A much better approach is to notice that the range of integration $(-\infty, \infty)$ has two "problem points" ($+\infty$ and $-\infty$). We can break up the range of integration into two pieces, in such a way that each piece has only one problem point, and then use the previous definitions.

Definition of improper integral (III) The expression $\int_{-\infty}^{\infty} f$ is defined to mean

$$\int_{-\infty}^{\infty} f := \int_{-\infty}^{0} f + \int_{0}^{\infty} f,$$

as long as both of the integrals on the right exist.

The was nothing special about the choice of 0 in this definition. The following easy lemma is left as an exercise.

Lemma 10.23. *If $\int_{-\infty}^{\infty} f$ exists, then for* every *real a, both of $\int_{-\infty}^{a} f, \int_{a}^{\infty} f$ exist, and $\int_{-\infty}^{\infty} f = \int_{-\infty}^{a} f + \int_{a}^{\infty} f$. Conversely, if $\int_{-\infty}^{a} f, \int_{a}^{\infty} f$ both exist for* some *$a \in \mathbb{R}$, then also both $\int_{-\infty}^{0} f, \int_{0}^{\infty} f$ exist, and so $\int_{-\infty}^{\infty} f$ exists.*

[178] One fix might be: since $\int_{-\infty}^{a} x \, dx = -\infty$, and $\int_{a}^{\infty} x = \infty$, we could *declare* "$0 = -\infty + \infty$". But this leads to serious problems; for example, we would have

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

Before giving an example of an integral of this kind, we present a useful "comparison" lemma.

Lemma 10.24. *Suppose*

- $0 \leq g(x) \leq f(x)$ *for all* $x \in [a, \infty)$*;*
- $\int_a^\infty f$ *exists; and*
- $\int_a^N g$ *exists for all* $N \geq a$*.*

Then $\int_a^\infty g$ *exists, and* $\int_a^\infty g \leq \int_a^\infty f$*.*

Proof: Consider the set $A = \{\int_a^N g : N \geq a\}$. This certainly non-empty, and it is bounded above by $\int_a^\infty f$ (since, for all $N \geq a$,

$$\int_a^N g \leq \int_a^N f \leq \int_a^\infty f).$$

It follows that A has a supremum $\alpha \leq \int_a^\infty f$. We claim that $\lim_{N\to\infty} \int_a^N g$ exists and equals α, from which the claim follows.

Fix $\varepsilon > 0$. There is an N_ε such that $\alpha - \varepsilon \leq \int_a^{N_\varepsilon} g \leq \alpha$. Now for all $N > N_\varepsilon$ we have $\int_a^{N_\varepsilon} g \leq \int_a^N g \leq \alpha$ (the latter inequality because $\alpha = \sup A$, the former because $g(x) \geq 0$ so $\int_a^N g$ increases as N increases). So for all $N > N_\varepsilon$ we have $\alpha - \varepsilon \leq \int_a^N g \leq \alpha$. Since this is true for arbitrary $\varepsilon > 0$, it follows that $\lim_{N\to\infty} \int_a^N g = \alpha$, as claimed. □

We illustrate both the definition of $\int_{-\infty}^\infty f$ and the utility of Lemma 10.24 with the example of $\int_{-\infty}^\infty \frac{dt}{1+t^2}$. To see if this integral exists, we need to examine $\int_0^\infty \frac{dt}{1+t^2}$ and $\int_{-\infty}^0 \frac{dt}{1+t^2}$. For the first of these, note that on $(0, \infty)$ we have

$$0 \leq \frac{1}{1+t^2} \leq \frac{1}{t^2}.$$

Unfortunately, this inequality does not extend to $t = 0$. However, we can use it on $[1, \infty)$ to conclude, via Lemma 10.24 (and the fact proved earlier that $\int_1^\infty \frac{dt}{t^2}$ exists), to conclude that $\int_1^\infty \frac{dt}{1+t^2}$ exists; and since $\int_0^1 \frac{dt}{1+t^2}$ clearly exists, it quickly follows that $\int_0^\infty \frac{dt}{1+t^2}$ exists.

For $\int_{-\infty}^0 \frac{dt}{1+t^2}$, we could develop an analog of Lemma 10.24 for the interval $(-\infty, a]$, or we could appeal to symmetry. One half of the following lemma appears in the homework; the other half (the half we'll use here) is left as an exercise.

Lemma 10.25. *Let* $f : [-b, b] \to \mathbb{R}$ *be a function that is integrable on the interval* $[0, b]$*.*

- *If* f *is an odd function* ($f(-x) = -f(x)$) *then* $\int_{-b}^b f$ *exists, and equals* 0*.*
- *If* f *is an even function* ($f(-x) = f(x)$) *then* $\int_{-b}^b f$ *exists, and equals* $2\int_0^b f$*.*

From this it quickly follows that $\int_{-\infty}^{0} \frac{dt}{1+t^2}$ exists and equals $\int_{0}^{\infty} \frac{dt}{1+t^2}$, and so $\int_{-\infty}^{\infty} \frac{dt}{1+t^2}$ exists.[179]

We discuss one more kind of improper integral.

Definition of improper integral (IV) Suppose that the function $f : (a, b] \to \mathbb{R}$ is unbounded, but that it is bounded on every interval $[a + \varepsilon, b]$ for $\varepsilon > 0$. Define $\int_a^b f$ by

$$\int_a^b f := \lim_{\varepsilon \to 0^+} \int_{a+\varepsilon}^b f,$$

as long as this limit exists.

As an example, we consider $\int_0^a \frac{dx}{\sqrt{x}}$. We can compute $\int_\varepsilon^a \frac{dx}{\sqrt{x}}$ using the fundamental theorem of calculus (part 2): since $g(x) = 2\sqrt{x}$ has $g'(x) = 1/\sqrt{x}$, we have

$$\int_\varepsilon^a \frac{dx}{\sqrt{x}} = g(a) - g(\varepsilon) = 2\sqrt{a} - 2\sqrt{\varepsilon}.$$

Since $2\sqrt{a} - 2\sqrt{\varepsilon} \to 2\sqrt{a}$ as $\varepsilon \to 0$, we conclude that

$$\int_0^a \frac{dx}{\sqrt{x}} = 2\sqrt{a}.$$

There are many other variants of improper integrals. For example, what would $\int_0^\infty f$ mean if f is bounded on $[\varepsilon, \infty)$ for all $\varepsilon > 0$, but f is unbounded on $(0, \infty)$? The paradigm is to break up range of integration into pieces, in such a way that each piece has only one problem point. So to check if $\int_0^\infty f$ exists in this scenario, we would check each of $\int_0^a f$, $\int_a^\infty f$, using the definitions we have given earlier.

[179] Notice that Lemma 10.24 gives no hint as to what the *value* of the integral is; we will return to this later.

11 Inverse functions

In this short chapter we digress back to the material of the fall semester. There, we defined functions, and showed how new functions could be formed from old, by addition, subtraction, multiplication by a constant, multiplication, division, and, most importantly, composition. In the chapter on integration, we introduced another technique to obtain a new function F from an old function f: setting $F(x) = \int_a^x f$. Before exploiting the full power of that technique, we need one more way of forming new functions from old: inverting.

11.1 Definition and basic properties

First, some background: recall the notation

$$f : A \to B,$$

standing for "f is a function with domain A, co-domain B". Precisely, this means that f is a set of pairs, with each element of A — the *domain* of f — occurring *exactly once* as a first entry of a pair, and with the set of second entries being a subset of B — a *co-domain* of f.

Here are some special kinds of functions, that come up frequently:

Injective functions $f : A \to B$ is *injective* or *one-to-one* if: no element of B appears more than once as a second entry; or, equivalently,

$$\text{if } x, y \in A \text{ are different, then } f(x), f(y) \in B \text{ are different.}$$

Such an f is also called *an injection* or *an injective map*.

Surjective functions $f : A \to B$ is *surjective* or *onto* if: every element of B appears at least once as a second entry; or, equivalently, if

$$\text{for every } y \in B \text{ there is a (not necessarily unique) } x \in A \text{ with } f(x) = y.$$

Another way to say this is that B is not just a co-domain for f, it is in fact the *range* — the exact set of second entries of the pairs that comprise f. Such an f is also called *a surjection* or *a surjective map*.

Bijective functions $f : A \to B$ is *bijective* if: f is both injective and surjective; or, equivalently, if

$$\text{for every } y \in B \text{ there is a } \textit{unique } x \in A \text{ with } f(x) = y.$$

Such an f is also called *a bijection* or *a bijective map*.

Note that if

$$f : A \to B \tag{10}$$

is an injective function, then there is naturally associated with f a bijective function, namely

$$f : A \to R \tag{11}$$

where $R \subseteq B$ is the range of f. The f's in (10) and (11) are *the same function* — they are comprised of the same set of pairs. The only difference between them is in the notion with which they are presented.[180]

For an injective, or a bijective, function f, we can form a new function g that we can think of as "undoing the action" of f, by simply reversing all the pairs that make up f. For example, if $f(1) = 2$, i.e., $(1, 2) \in f$, then we put the pair $(2, 1)$ in g, i.e., set $g(2) = 1$. Is this really a function? It is a set of ordered pairs, certainly. Suppose that some number b appears twice in the set of ordered pairs, say as (b, a_1) and (b, a_2), with $a_1 \neq a_2$. Then that means that (a_1, b) and (a_2, b) are both in f (that's how (b, a_1) and (b, a_2) got into g). The presence of (a_1, b) and (a_2, b) in f then contradicts that f is injective.

We've just argued that if f is injective, then g is a function. On the other hand, if f is *not* injective, then the process we have described does *not* produce a new function g. Indeed, let $(a_1, b), (a_2, b) \in f$ with $a_1 \neq a_2$ witness the failure of injectivity of f. We have $(b, a_1), (b, a_2) \in g$, witnessing the failure of g to be a function.

This all shows that the process of reversing all the pairs that make up a function f, to form a new function g, makes sense if and only f is injective. We now note some further properties.

- if f is injective, then so is g. Indeed, suppose g is not injective, and let $(b_1, a), (b_2, a) \in g$ with $b_1 \neq b_2$ witness failure of injectivity of g. Then $(a, b_1), (a, b_2) \in f$ witness the failure of f to be a function, a contradiction;

- the range of f is the domain of g; the domain of f is the range of g (this is obvious);

- $f \circ g = \mathrm{id}$, where $\mathrm{id} : \mathrm{Domain}(g) \to \mathrm{Domaing}$ is the identity function, consisting of pairs whose first and second coordinates are the same; and $g \circ f = \mathrm{id}$, where $\mathrm{id} : \mathrm{Domain}(f) \to \mathrm{Domainf}$ is the identity function[181] (this should also be obvious); and

- if the operation that is used to produce g from f is applied to g (this makes sense, since g is injective), then the result is f (this should also be obvious).

[180] Which begs the question: why not just present all functions in the form $f : A \to R$, where A is the domain and R the range? The issue is that it is often very difficult to pin down the exact range of a function, so it is often convenient to simply present a definitely valid co-domain, such as $\mathbb{R}$. Try, for example, finding the exact range of $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^6 + x^5 + 1$.

[181] But note that the two identity functions here are not necessarily the same — there is no reason why $\mathrm{Domain}(g)$ should equal $\mathrm{Domain}(f)$.

We formalize all this, in a definition and a theorem. The theorem has already been proven; it is just the above discussion, reframed in the language of the definition.

Definition of inverse function Let $f = \{(a, b) : a \in \text{Domain} f\}$ be an injective function. The *inverse* of f, denoted f^{-1}, is defined by

$$f^{-1} = \{(b, a) : (a, b \in f)\}.$$

If f has an inverse, it is said to be *invertible*[182].

Theorem 11.1. *Let f be an injective function, with domain D and range R.*

- *f^{-1} is a function, with domain R and range D.*
- *f^{-1} is injective, and $(f^{-1})^{-1} = f$.*
- *$f \circ f^{-1}$ is the identity function on R (that is, for all $x \in R$, $f(f^{-1}(x)) = x$).*
- *$f^{-1} \circ f$ is the identity function on D.*

There's a very easy way to construct the graph of f^{-1} from the graph of f: the set of points of the form (b, a) (that comprises the graph of f^{-1}) is the reflection across the line $x = y$ of the set of points of the form (a, b) (that comprises the graph of f). Because vertical lines in the plane are mapped to horizontal lines by reflection across the line $x = y$, this leads to an easy visual test for when a function is invertible: f is invertible if it's graph passes the

Horizontal line test : every horizontal line in the plane crosses the graph of f at most once.

Which functions are invertible?

- Certainly, if f is increasing[183] on it's domain, then it is invertible (and it is an easy check that the inverse f^{-1} in this case is also increasing). On the other hand, if f is only weakly increasing, then it may not necessarily be invertible (think of the constant function).
- Similarly if f is decreasing, it's invertible, and f^{-1} is decreasing too.
- There are certainly examples of invertible functions that are *not* monotone (increasing or decreasing); consider, for example, $f : [0, 1] \to [0, 1]$ given by

$$f(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } x = 1 \\ x & \text{if } 0 < x < 1. \end{cases}$$

[182] For practical purposes, we can think of "invertible" and "injective" as synonymous.

[183] Recall: $f : A \to \mathbb{R}$ is *increasing* on A if $x < y \in A$ implies $f(x) < f(y)$; we sometimes say *strictly* increasing, but our convention is that without any qualification, "increasing" is the same as "strictly increasing"; we use *weakly increasing* to indicate $x < y \in A$ implies $f(x) \leq f(y)$.

- Even adding the assumption of continuity, there are still non-monotone invertible functions; consider, for example, $f : (0,1) \cup (1,2) \to \mathbb{R}$ given by

$$f(x) = \begin{cases} x & \text{if } 0 < x < 1 \\ 3 - x & \text{if } 1 < x < 2. \end{cases}$$

If f is continuous, however, *and* defined on a single interval, then it seems reasonable to expect that invertibility forces monotonicity. This is the content of our first significant theorem on invertibility.

Theorem 11.2. *Suppose that $f : I \to \mathbb{R}$ is continuous on the interval I. If f is invertible, then it is monotone (either increasing or decreasing).*

Proof: We prove the contrapositive. Suppose that f is *not* monotone on I. That means that

- there is $x_1 < x_2$ with $f(x_1) \leq f(x_2)$ (witnessing that f is not decreasing), and
- there is $y_1 < y_2$ with $f(y_1) \geq f(y_2)$ (witnessing that f is not increasing).

If either $f(x_1) = f(x_2)$ or $f(y_1) = f(y_2)$ then f is not invertible. So we may assume that in fact $f(x_1) < f(x_2)$ and $f(y_1) > f(y_2)$.

There are twelve possibilities for the relative order of x_1, x_2, y_1, y_2:

- $y_1 < y_2 < x_1 < x_2$
- $x_1 < y_1 < y_2 < x_2$
- $x_1 < x_2 < y_1 < y_2$
- $y_1 < x_1 < y_2 < x_2$
- $y_1 < x_1 < x_2 < y_2$
- $x_1 < y_1 < x_2 < y_2$
- $y_1 < x_1 < x_2 = y_2$
- $x_1 < y_1 < x_2 = y_2$
- $y_1 < x_1 = y_2 < x_2$
- $y_1 = x_1 < y_2 < x_2$
- $y_1 = x_1 < x_2 < y_2$
- $x_1 < x_2 = y_1 < y_2$

In each of these twelve cases, it is possible to find $x < y < z$ with either

- $f(y) > f(x), f(z)$

or

- $f(y) < f(x), f(z)$

For example, if $y_1 = x_1 < y_2 < x_2$, then we may take $y = y_2$, $x = x_1 = y_1$ and $z = x_2$ to get $f(y) < f(x), f(z)$. For a more involved example, consider $y_1 < y_2 < x_1 < x_2$. We have $f(y_2) < f(y_1)$. If $f(x_2) < f(y_2)$ then we may take $y = x_1$, $x = y_2$ or y_1 and $z = x_2$ to get $f(y) < f(x), f(z)$, while if $f(x_2) > f(y_2)$ we may take $y = y_2$, $x = y_1$ and $z = x_2$ to again get $f(y) < f(x), f(z)$ (note that we won't have $f(x_2) = f(y_2)$, as this automatically implies non-invertibility).

Suppose $f(y) > f(x), f(z)$. Let $m = \max\{f(x), f(z)\}$. By the intermediate value theorem applied to the interval $[x, y]$, f takes on the value $(f(y) + m)/2$ in (x, y). But by the intermediate value theorem applied to the interval $[y, z]$, f takes on the value $(f(y) + m)/2$ in (y, z). Since (x, y) and (y, z) don't overlap, this shows that f takes on the same value at least two different times, so is not invertible.

Here's an alternate, direct, proof, that uses a shorter case analysis. Suppose that f is invertible. Then it is injective, so $x \neq y$ implies $f(x) \neq f(y)$. Fix $y \in I$ that is not an endpoint; let I_1 be $\{x \in I : X < y\}$ and I_2 be $\{z \in I : z > y\}$.

It cannot be the case that there is some $x \in I_1$ with $f(x) > f(y)$, and some $x' \in I_1$ with $f(x') < f(y)$; for then we could easily find $x' < y' < z'$ with either $f(y') > f(x'), f(z')$ or $f(y') < f(x'), f(z')$, and the IVT argument from above gives a contradiction. So either $f(x) > f(y)$ for all $x \in I_1$, or $f(x) < f(y)$ for all $x \in I_1$. Similarly, either $f(z) > f(y)$ for all $z \in I_2$, or $f(z) < f(y)$ for all $z \in I_2$.

If either

- $f(x) > f(y)$ for all $x \in I_1$ and $f(z) > f(y)$ for all $z \in I_2$

or

- $f(x) < f(y)$ for all $x \in I_1$ and $f(z) < f(y)$ for all $z \in I_2$

then we could easily find $x' < y' < z'$ with either $f(y') > f(x'), f(z')$ or $f(y') < f(x'), f(z')$, for a contradiction.

If $f(x) > f(y)$ for all $x \in I_1$ and $f(z) < f(y)$ for all $z \in I_2$, then we claim that f is monotone decreasing. Indeed, consider $a < b \in I$. If one of a, b is y, we immediately have $f(a) > f(b)$. If $a < y < b$, we immediately have $f(a) > f(y) > f(b)$, so $f(a) > f(b)$. If $a, b < y$, and $f(a) < f(b)$, then we $x' < y' < z'$ with $f(y') > f(x'), f(z')$, a contradiction, so $f(a) > f(b)$ in this case. Similarly, if $a, b > y$ we also get $f(a) > f(b)$.

Finally, in the case If $f(x) < f(y)$ for all $x \in I_1$ and $f(z) > f(y)$ for all $z \in I_2$, a similar agument gives that f is monotone increasing. □

A consequence of the first proof above is the following useful fact:

If $f : I \to \mathbb{R}$ is not monotone, then there $x < y < z \in I$ with either $f(y) > f(x), f(z)$ or $f(y) < f(x), f(z)$.

Note that this fact does not need continuity of f.

The proof of Theorem 11.2 uses IVT, which raises a nice challenge: in $\mathbb{Q}$-world, find an example of a function f defined on an interval, that is continuous and invertible, but not monotone. (Such an example would show that the completeness axiom is necessary to prove Theorem 11.2).

Suppose that $f : I \to \mathbb{R}$ is continuous and invertible, and so, by Theorem 11.2, monotone. We can easily determine the range of f, and so the domain of f^{-1}. The verification of all of these are left as exercises.

- if $I = [a, b]$ and f is increasing, then $\text{Range}(f) = [f(a), f(b)]$, while if f decreasing, then $\text{Range}(f) = [f(b), f(a)]$;
- if $I = (a, b)$ or $(-\infty, b)$ or (a, ∞) or $(-\infty, \infty)$, then, whether f is increasing or decreasing, we have $\text{Range}(f) = (\inf\{f(x) : x \in I\}, \sup\{f(x) : x \in I\})$ (where here we allow $\inf\{f(x) : x \in I\}$ to possibly take the value $-\infty$, and $\sup\{f(x) : x \in I\}$ to possibly take the value $-\infty$);
- and if I is a mixed interval (open at one end, closed at the other end), then we do the obvious thing: for example, if $I = [a, b)$ and f is decreasing, then $\text{Range}(f) = (\inf\{f(x) : x \in I\}, f(a)]$.

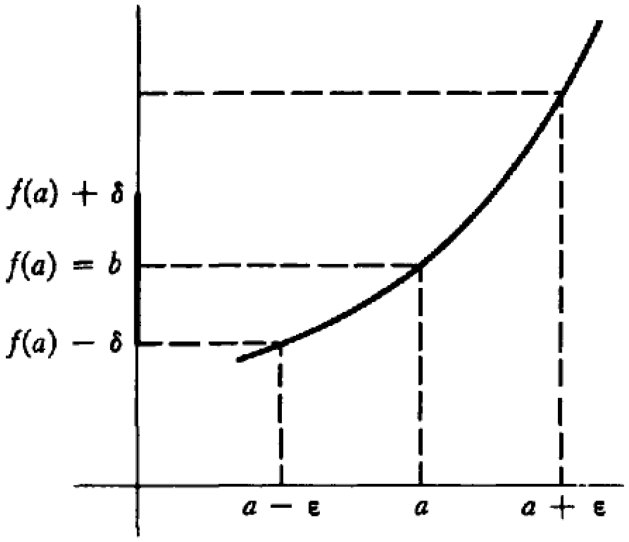
11.2 The inverse, continuity and differentiability

The inverse function behaves well with respect to continuity and differentiability, as we now show.

Theorem 11.3. *If I is an interval, and $f : I \to \mathbb{R}$ is continuous on I, and invertible, then f^{-1} is also continuous on its whole domain.*

Proof: By Theorem 11.2, we know that f is either increasing or decreasing on I. We can assume that f is increasing; if it is decreasing, we obtain the result by considering (increasing) $-f$.

Given $b \in \text{Domain}(f^{-1})$, we want to show that $\lim_{x\to b} f^{-1}(x) = f^{-1}(b)$. Now there is $a \in \text{Domain}(f)$ with $f(a) = b$, so $f^{-1}(b) = a$. Given $\varepsilon > 0$, we want to find a $\delta > 0$ such that $f(a) - \delta < x < f(a) + \delta$ implies $a - \varepsilon < f^{-1}(x) < a + \varepsilon$. The picture below, taken from Spivak, should both make the choice of notation clear, and suggest how to proceed:



Let δ be small enough that

$$f(a-\varepsilon) < f(a) - \delta < b = f(a) < f(a) + \delta < f(a+\varepsilon)$$

(since $f(a+\varepsilon) - f(a) > 0$ and $f(a) - f(a-\varepsilon) > 0$, such a δ can be found — just take δ to be anything smaller than the minimum of $f(a+\varepsilon) - f(a)$ and $f(a) - f(a-\varepsilon)$).

For $f(a) - \delta < x < f(a) + \delta$ we have $f(a-\varepsilon) < x < f(a+\varepsilon)$ and so (using that f^{-1} is increasing) we that get $a - \varepsilon < f^{-1}(x) < a + \varepsilon$, as required. □

What about differentiability and the inverse? By considering the graph of an increasing, continuous, f, at a point $(a, f(a))$, where f is differentiable, and by then considering the reflection of the graph across $x = y$, it is fairly easy to form the hypothesis that $(f^{-1})'(f(a))$ is well-defined — unless $f'(a) = 0$, when it appears that the tangent line through $(f(a), f^{-1}(f(a)))$ is vertical.[184] In other words, it appears that for b in the domain of f^{-1}, we have that $(f^{-1})'(b)$ is well defined unless $f'(f^{-1}(b)) = 0$.

Differentiating both sides of $(f \circ f^{-1})(x) = x$ we get $f'(f^{-1}(x))(f^{-1})'(x) = 1$, that is,

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

suggesting what the derivative of f^{-1} should be at b (as long as $f'(f^{-1}(b)) \neq 0$) ("suggesting" because we don't a priori know that f^{-1} is differentiable at b).

All this can be made precise.

Theorem 11.4. *Suppose I is an interval, and that $f : I \to \mathbb{R}$ is continuous on I, and invertible. Suppose further that for some b in the domain of f^{-1}, f is differentiable at $f^{-1}(b)$.*

- *If $f'(f^{-1}(b)) = 0$ then f^{-1} is* not *differentiable at b.*

[184] Draw some graphs! Convince yourself.

- *If* $f'(f^{-1}(b)) \neq 0$ *then* f^{-1} is *differentiable at* b*, with derivative* $1/(f'(f^{-1}(b)))$.

Proof: We consider the first bullet point first: Suppose (for a contradiction) that f^{-1} *is* differentiable at b. We apply the chain rule to conclude

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

but this is impossible since $f'(f^{-1}(b)) = 0$.

We now move on to the second bullet point. Let a be such that $f(a) = b$. We have

$$\frac{f^{-1}(b+h) - f^{-1}(b)}{h} = \frac{f^{-1}(b+h) - a}{h} = \frac{k}{h}$$

where $k = k(h)$ is such that $f^{-1}(b+h) = a + k$. We also have $f(a) + h = b + h = f(a+k)$, so the we have

$$\frac{k}{h} = \frac{k}{f(a+k) - f(a)} = \frac{1}{\left(\frac{f(a+k)-f(a)}{k}\right)}.$$

Because f^{-1} is continuous, we have $\lim_{h\to 0} f^{-1}(b+h) = f^{-1}(b) = a$, $\operatorname{solim}_{h\to 0} k = 0$. So as h approaches 0, so does k, and $(f(a+k) - f(a))/k$ approaches $f'(a) = f'(f^{-1}(b))$ (which exists by hypothesis). Since (also be hypothesis) $f'(f^{-1}(b))$ is non-zero, we can put everything together to get

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

□

What about the interaction between invertibility and integrability? Certainly, if $f : [a, b] \to \mathbb{R}$ is *continuous* and invertible (say, for simplicity, increasing), then since f^{-1} is continuous on $[f^{-1}(a), f^{-1}(b)]$, it is integrable over that range. In all the applications that are coming up, that is all that we will need. What happens to f^{-1} vis a vis integrability, when f is only assumed to be *integrable* and invertible, will be explored in an exercise.

12 The logarithm, exponential, and trigonometric functions

We know what x^n means for any real x and natural number n. But what does, for example, $x^{\sqrt{2}}$ mean? Even when x is a natural number, there is no obvious intuitive interpretation, and the situation is far more unclear when we ask about, say, $\sqrt{3}^{\sqrt{2}}$.

The goal of this section is to introduce the logarithm and exponential functions, which allow us to interpret sensible and unambiguous expressions of the form x^y for arbitrary reals x, y.

12.1 Informal introduction

For natural number n, define $f_n : (0, \infty) \to \mathbb{R}$ by $f_n(x) = x^n$.[185] We already know alot about f_n: it has domain $(0, \infty)$, range $(0, \infty)$, is increasing on its domain, is continuous everywhere and differentiable everywhere, and has derivative $f_n'(x) = nx^{n-1}$.

By the results of the section on inverse functions, we know that f_n has an inverse, call it g_n, which has domain $(0, \infty)$, range $(0, \infty)$, is increasing on its domain, is continuous everywhere. We denote $g_n(x)$ by $x^{1/n}$, and also refer to g_n as $f_{1/n}$. Since $f_n'(x) \neq 0$ for any x, $f_{1/n}$ is differentiable everywhere, with derivative

$$f_{1/n}'(x) = \frac{1}{f_n'(x^{1/n})} = \frac{1}{n} x^{\frac{1}{n}-1}$$[186].

For any positive rational $r = m/n$ (with $m, n \in \mathbb{N}$) we can define f_r by $f_r(x) = (x^{1/n})^m$. It is a straightforward check that this is in fact well defined, that is, that the value of $f_r(x)$ doesn't depend on the particular choice of m, n. This amounts to using the axioms of the real numbers to check that if $m/n = p/q$ with $m, n, p, q \in \mathbb{N}$ then $(x^{1/n})^m = (x^{1/q})^p$. Again, f_r has domain $(0, \infty)$, range $(0, \infty)$, is increasing on its domain, and is continuous everywhere. By basic properties of the derivative, it is differentiable everywhere, and an application of the chain rule yields

$$f_r'(x) = m(x^{1/n})^{m-1} \times \frac{1}{n} x^{\frac{1}{n}-1} = rx^{r-1}.$$

Denote by x^r the value $f_r(x)$.

For negative rational r, we can define a function f_r by $f_r(x) = 1/f_{-r}(x)$, and again denote by x^r the value $f_r(x)$. Again, f_r has domain $(0, \infty)$, range $(0, \infty)$, but now is decreasing on

[185] The natural domain of $f(x) = x^n$ for $n \in \mathbb{N}$ is $\mathbb{R}$, but we choose to restrict to the domain on $(0, \infty)$. The reason for this is that f is increasing on this domain, for *every* n (whereas on its natural domain, f is only increasing for odd n); we will avoid a lot of annoyance by focussing exclusively on non-negative inputs to the power function, and we wouldn't gain much by trying to extend to negative inputs.

[186] This is jumping the gun a little bit. Really, $f_{1/n}'(x) = 1/(n(x^{1/n})^{n-1})$. With the definition we will give in a moment for x^r for positive rational r, this becomes $1/(n(x^{(n-1)/n}))$, and with the definition we will give a moment later for x^r for negative rational r, this becomes $(1/n)x^{(1/n)-1}$. But technically we need those later definitions to jump to the final answer.

its domain. It is continuous everywhere, and by the chain rule (or the reciprocal rule, or the quotient rule), it is easily seen to satisfy $f_r'(x) = rx^{r-1}$.

The conclusion of all this is we can define, for *every* rational r, a function f_r that acts as a "raising to the power r" function, i.e., $f_r(x) = x^r$, that can be applied to any positive x. This function has the properties that

- f_r has domain $(0, \infty)$ and range $(0, \infty)$ (unless $r = 0$, in which case it has range $\{1\}$; recall that $x^0 = 1$ for all $x \neq 0$);
- f_r is increasing if $r > 0$, decreasing if $r < 0$, and constant if $r = 0$;
- f_r is continuous everywhere; and
- f_r is differentiable everywhere, with derivative $f_r'(x) = rx^{r-1}$.

Moreover, f_r agrees with our usual notion of "raising to the power r" where r is a natural number.

But what can we do to make sense of x^a when a is *not* rational? One approach is through the completeness axiom: for $x > 1$ and rational numbers $0 < r_1 < r_2$ it can be checked that $x^{r_1} < x^{r_2}$. It follows that for $x > 1$ and real $a > 0$, the set $A = \{x^r : 0 < r < a\}$ is bounded above (by $x^{r'}$ for any $r' > a$), It's also non-empty (obviously), so by completeness, $\sup A$ exists. We could then *declare* x^a to be $\sup A$. Similarly, for $0 < x < 1$, we could declare x^a to be $\inf\{x^r : r < a\}$, and (of course) declare 1^a to be 1. Then, for $a < 0$, we could declare x^a to be $1/x^{-a}$.All this certainly defines, for each real a, a function $f_a : (0, \infty) \to (0, \infty)$.

It's an easy check that this agrees with the previous definition when a is rational. It is far from easy to check that when a is irrational, f_a is still continuous and differentiable, with derivative $f_a'(x) = ax^{a-1}$, that is increasing when $a > 0$ and decreasing when $a < 0$, and that it has range $(0, \infty)$.

It is also far from easy to check that f_a satisfies all the properties that we would expect of the "raising to the power a" function, properties such as

- $x^{a+b} = x^a x^b$

and

- $(x^a)^b = x^{ab}$;

its not even all that straightforward to verify these properties for rational a, b.

So, we'll take an alternate approach to defining the power function for general exponents. Instead of constructing a function that seems like it should work, and then verifying that the properties we want to hold do actually hold, we'll list those properties, considering them as axioms, and then try to argue that there exists a unique function that satisfies those properties.

Actually, we will address a related, but slightly different, question:

Fix a real number $a > 0$. For real x, what does a^x mean?

(The difference here is that we now are considering the base to be fixed, and we are varying the exponent, whereas before we were considering the exponent to be fixed, and we were varying the base.[187])

One approach follows the lines we described earlier:

- Set $a^0 = 1$ and set $a^n = aa^{n-1}$ for $n \in \mathbb{N}$.
- For $n \in \mathbb{N}$ define $a^{1/n}$ via the intermediate value theorem, as we did last semester.
- For positive rational $r = m/n$, set $a^r = (a^{1/m})^n$ (after checking that this is well-defined, i.e., doesn't depend on the choice of representation of $r = m/n$ with $m, n \in \mathbb{N}$).
- For negative rational r, set $a^r = 1/(a^{-r})$.

This defines a^x for rational x, and a series of tedious inductions, together with lots of algebraic manipulation, verifies the relation

$$a^{x+y} = a^x a^y \quad \text{for all } x, y \in \mathbb{Q}. \quad (\star)$$

Then, for general real x, we can define

$$a^x = \begin{cases} \sup\{a^r : r \in \mathbb{Q}, r < x\} & \text{if } a > 1 \\ \inf\{a^r : r \in \mathbb{Q}, r < x\} & \text{if } a < 1 \\ 1 & \text{if } a = 1. \end{cases}$$

It is a long and intricate exercise that for each $a > 0$, this yields a continuous function that satisfies $(\star)$ (in fact, this gives the unique such continuous function that extends the given definition of a^r for rational r, as we discuss in a moment).

Instead of taking this approach, we'll take an axiomatic approach. Fix $a > 0$, with $a \neq 1$[188]. Let $\exp_a : \mathbb{R} \to \mathbb{R}$ be a function (an as-yet unknown function) that captures the notion of "raising a base a to a power"; that is, $\exp_a(x)$ is a sensible interpretation of a^x for all real x. Here are the properties that we want $\exp_a$ to satisfy:

- $\exp_a(1) = a$ and $e_a(0) = 1$ (a normalizing property);
- for all real x, y, $\exp_a(x + y) = \exp_a(x) \exp_a(x)$ (this property, together with the normalizing property, and a lot of induction, is what's need to ensure that $\exp_a(r) = a^r$ for rational r, where a^r is defined in the natural way that we described above);

[187] Of course, it will amount to the same thing in the end — we'll end up define u^v for any $u > 0$ and any real v.

[188] If $a = 1$, there is an obvious choice for $\exp_a$, namely $\exp_a(x) = 1$ for all x.

- $\exp_a$ is continuous on its domain (this condition ensures that $\exp_a(x) = a^x$ for all real x, where a^x is defined via the completeness axiom as described above. Indeed, let $f : \mathbb{Q} \to \mathbb{R}$ be the function $f(r) = a^r$, with the natural definition, and let $g : \mathbb{Q} \to \mathbb{R}$ be the function $g(r) = \exp_a^r$. Suppose both f and g extend to continuous functions on the whole real line. By the first two properties, $f - g$ is identically 0 on the rationals; but it is also continuous on the reals. A simple argument based on the density of the rationals quickly gives that $f - g$ is identically 0 on the reals, i.e., that $f = g$[189]);
- $\exp_a$ is differentiable; and
- $\exp_a$ is monotone (these last two properties will allow us to derive non-obvious properties that $\exp_a$ must also satisfy, if it satisfies the ones listed above; from these we will get our actual explicit expression for $\exp_a$).

Assuming such a function exists, here is what its derivative must look like:

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

And here is what the derivative of its inverse must look like:

$$\begin{aligned} (\exp_a^{-1})'(x) &= \frac{1}{\exp_a'(\exp_a^{-1}(x))} \\ &= \frac{1}{\exp_a'(0) \exp_a(\exp_a^{-1}(x))} \\ &= \frac{1}{\exp_a'(0)x}. \end{aligned}$$

Now the fundamental theorem of calculus (part 1) tells us that if we define (for any specific constant c)

$$\exp_a^{-1}(x) = \int_c^x \frac{dt}{\exp_a'(0)t}$$

then we indeed have $(\exp_a^{-1})'(x) = 1/(\exp_a'(0)x)$. We should probably choose $c = 1$, because with the above definition, we have $\exp_a^{-1}(c) = 0$ so $\exp_a(0) = c$, and we want $\exp_a(0) = 1$.

So we have been led to defining not $\exp_a$, but rather $\exp_a^{-1}$; and we have been led to the definition

$$\exp_a^{-1}(x) = \int_1^x \frac{dt}{\exp_a'(0)t}.$$

A problem with this definition is that we don't know what $\exp_a'(0)$ is. So as it stands, the definition is somewhat circular.

[189] This is an example of the general result: if $f, g : \mathbb{R} \to \mathbb{R}$ are continuous, and agree on a dense set, then they agree everywhere.

Here's a solution: presumably, there is a base a for which $\exp_a'(0) = 1$. For that special base, we have a completely explicit candidate for $\exp_a^{-1}$, namely

$$\exp_a^{-1}(x) = \int_1^x \frac{dt}{t}.$$

All this was hypothetical — *if* there is a function $\exp_a$ satisfying all of the required properties, then (at least for one special, undetermined as-yet a), its inverse has the simple explicit expression given above. In the next section we start the whole process over, and put it on firm foundations. It will go quickly, because we now know where to start from — with the integral $\int_1^x dt/t$.

12.2 Defining the logarithm and exponential functions

Motivated by the discussion in the previous section, we now define the (natural) logarithm function.

Definition of logarithm The function log[190] is defined by

$$\log(x) = \int_1^x \frac{dt}{t}.$$

Because the function $f(t) = 1/t$ is continuous and bounded on every interval of the form $[\varepsilon, N]$ (for arbitrarily small $\varepsilon > 0$ and arbitrarily large $N > 0$), it follows that the natural domain of log is (at least) $(0, \infty)$. In fact, this is the full natural domain, because it is easy to check (in a manner similar to how we checked that $\int_1^\infty dt/t$ diverges) that $\int_0^1 dt/t$ diverges.[191]

Moreover, since $1/t$ is non-negative, log is increasing on its domain. It takes the value 0 once, at $x = 1$. Because $\int_1^N dt/t$ can be made arbitrarily large by choosing N large enough, and $\int_1^\varepsilon$ can be made arbitrarily small (large and negative) by choosing $\varepsilon > 0$ close enough to zero, it follows that the range of log is $(-\infty, \infty)$. Specifically,

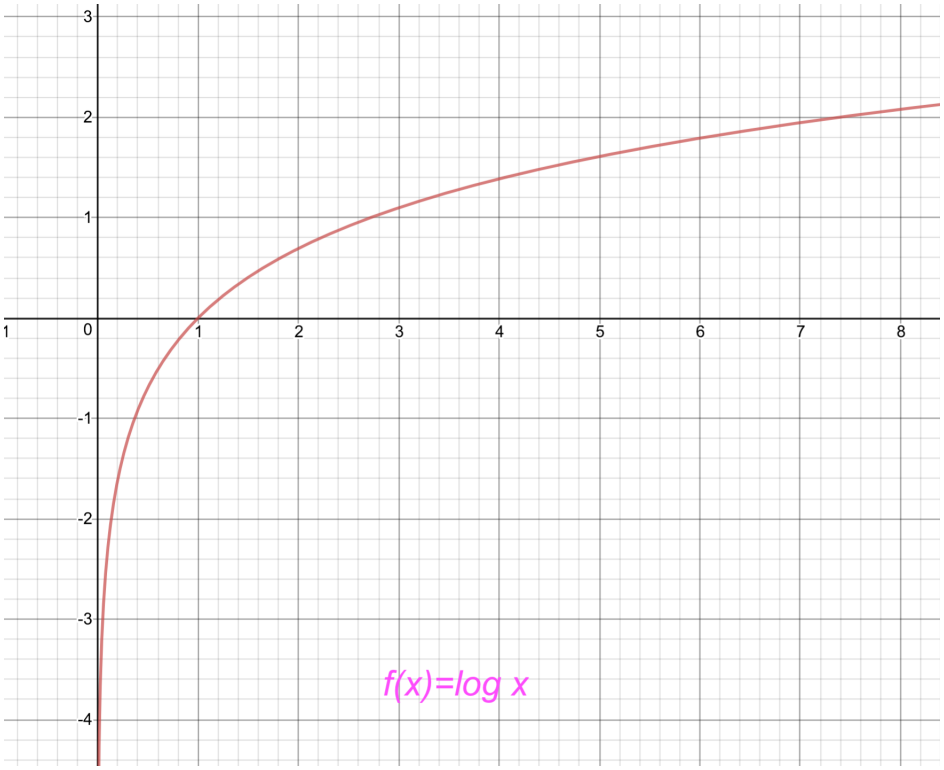
- $\lim_{x \to 0^+} \log(x) = -\infty$
- $\lim_{x \to \infty} \log(x) = +\infty$.

Because it is defined as the integral of an integrable function, log is continuous on its whole domain, and because $f(t) = 1/t$ is itself continuous, log is moreover differentiable everywhere, with derivative $\log'(x) = 1/x$ and second derivative $\log''(x) = -1/x^2$. Since this latter is always negative, log is concave on its whole domain.

We are now in a good position to sketch the graph of log:

[190] The name ln is also sometimes used for this function, but this notation is far more commonly seen in calculus textbook that in scientific papers.

[191] Consider the integral on the intervals $[1/2, 1]$, $[1/4, 1/2]$, $[1/8, 1/4$, et cetera.



Because log is increasing, it has an inverse, which we denote by exp (for "exponential"). From our general discussion of inverse functions, together with the properties we have just established about log, we immediately get that exp is continuous and increasing, has domain $\mathbb{R}$ and range $(0, \infty)$, and satisfies

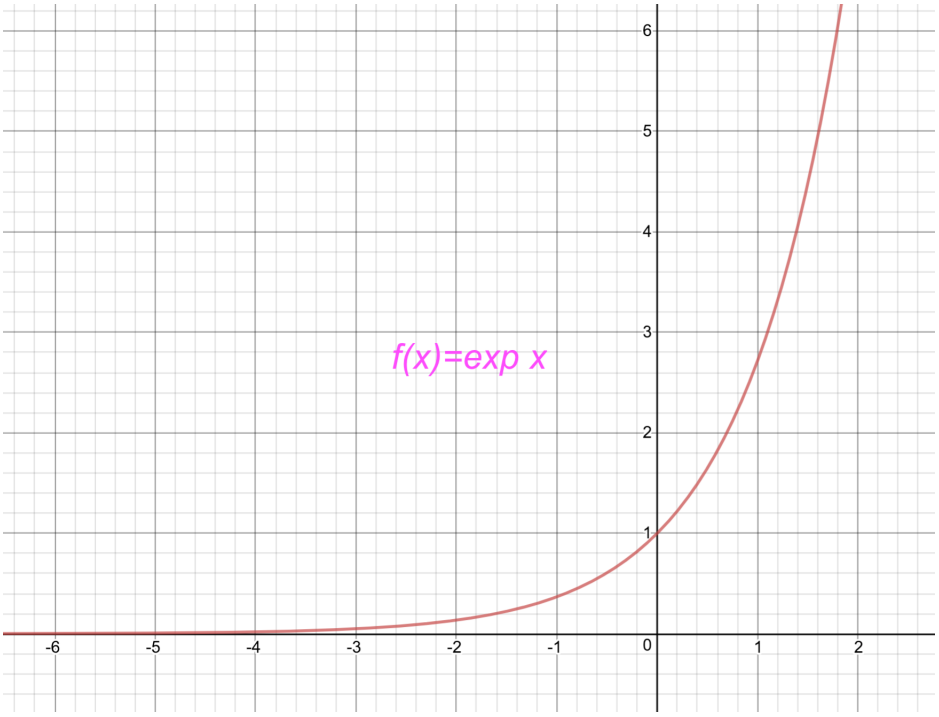
- $\lim_{x \to -\infty} \exp(x) = 0$
- $\lim_{x \to \infty} \exp(x) = +\infty$.

Because the derivative of *log* is never 0, the derivative of exp exists at all points in its domain, and we have

$$\exp'(x) = (\log^{-1})'(x) = \frac{1}{\log'(\log^{-1}(x))} = \log^{-1}(x) = \exp(x).$$

So $\exp''(x) = \exp(x) > 0$, and exp is convex. Since $\log(1) = 0$ we get $\exp(0) = 1$. We are now in a good position to sketch the graph of exp[192]:

[192] Of course, we could have also obtained the graph of exp by reflecting the graph of log across $x = y$



There is some unique number $\alpha > 1$ that has the property

$$\int_1^{\alpha} \frac{dt}{t} = 1.$$

We call this number e. The two basic properties of e, the first of which is the defining relation (reframed in the language of the log function), and the second of which is an immediate consequence of the first:

$$\log(e) = 1 \quad \text{and} \quad \exp(1) = e.$$

The number e is ubiquitous in mathematics. It has numerous "equivalent definitions", such as

- $\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n$
- $\lim_{n\to\infty} \sum_{k=1}^{n} \frac{1}{k!}$
- $\lim_{n\to\infty} \frac{n^n}{\sqrt[n]{n!}}$.

Hopefully our approach to defining e shows why it is natural: it attempting to define a function e_a that could sensibly serve as an interpretation for a^x, we discovered that e_a should satisfy

$$\exp_a^{-1}(x) = \int_1^{x} \frac{dt}{\exp_a'(0)t}.$$

The number e turns out to be the unique choice of a for which $\exp_a'(0) = 1$, leading to a particularly clean definition.

We can use the definition of e to give a numerical estimate. First, we show that $e > 2.7182$. To do this, we need to show that $\int_1^{2.7182} dt/t < 1$. Dividing the interval $[1, 2.7182]$ into n equal subintervals, we use that $1/t$ is decreasing to get

$$\int_1^{2.7182} \frac{dt}{t} \leq \sum_{i=1}^{n} M_i(t_i - t_{i-1}) = \frac{1.7182}{n} \sum_{i=1}^{n} \frac{1}{1 + \frac{1.7182(i-1)}{n}}.$$

A `Mathematica` calculation shows that when $n = 100,000$ the right-hand side above is $0.99997\cdots$. Next, we show that $e < 2.7183$. To do this, we need to show that $\int_1^{2.7183} dt/t > 1$. Using the same approach as before, we have

$$\int_1^{2.7183} \frac{dt}{t} \geq \sum_{i=1}^{n} m_i(t_i - t_{i-1}) = \frac{1.7183}{n} \sum_{i=1}^{n} \frac{1}{1 + \frac{1.7183i}{n}}.$$

A `Mathematica` calculation shows that when $n = 100,000$ the right-hand side above is $1.000001\cdots$. So we have the bounds

$$2.7182 < e < 2.7183.$$

We now derive the key property of the log function.

Theorem 12.1. *For all a, b in the domain of* log*,*

$$\log(ab) = \log(a) + \log(b)$$

and

$$\log(a/b) = \log(a) - \log(b).$$

Proof: To prove that $\log(ab) = \log(a) + \log(b)$ we want to show that for all $a, b \in (0, \infty)$,

$$\int_1^a \frac{dt}{t} + \int_1^b \frac{dt}{t} = \int_1^{ab} \frac{dt}{t}.$$

By the basic properties of integration, this is equivalent to

$$\int_1^a \frac{dt}{t} = \int_b^{ab} \frac{dt}{t}.$$

Define $G(x) = \int_1^x \frac{dt}{t}$, so $G'(x) = 1/x$, and $H(x) = \int_b^{xb} \frac{dt}{t}$, so $H'(x) = (1/xb)b = 1/x$. Since $G'(x) = H'(x)$ for all x, we have that $G - H$ is constant. Since $G(1) = H(1) = 0$, that constant is 0, so $G = H$, and in particular $G(a) = H(a)$, which is what we wanted to show.

For the second identity we have

$$\log a = \log(a/b)b = \log(a/b) + \log b$$

(using the result we have just proven), so

$$\log(a/b) = \log a - \log b.$$

□

Translating this to the exponential function, we obtain the following corollary.

Corollary 12.2. *For all $a, b \in \mathbb{R}$,*

$$\exp(a+b) = \exp(a)\exp(b)$$

and

$$\exp(a-b) = \exp(a)/\exp(b).$$

Proof: Let a', b' be such that $\log(a') = a$, $\log(b') = b$. We have

$$\exp(a+b) = \exp(\log(a') + \log(b')) = \exp(\log(a'b')) = a'b' = \exp(a)\exp(b)$$

(since $a = \exp(x)$, $b = \exp(y)$). Similarly

$$\exp(a-b) = \exp(\log(a') - \log(b')) = \exp(\log(a'/b')) = a'/b' = \exp(a)/\exp(b).$$

□

Both Theorem 12.1 and Corollary 12.2 can be extended by induction: for $a_1, \ldots, a_n \in (0, \infty)$,

$$\log(a_1 \cdot a_2 \cdot \cdots \cdot a_n) = \log a_1 + \log a_2 + \cdots + \log a_n$$

and for $a_1, \ldots, a_n \in \mathbb{R}$,

$$\exp(a_1 + a_2 + \cdots + a_n) = \exp a_1 \cdot \exp a_2 \cdot \cdots \cdot \exp a_n.$$

Recall that we set out to find, for each $a > 0$, a function $\exp_a : \mathbb{R} \to \mathbb{R}$ that captures the notion of "base a raised to a power", and we decided that such a function should be continuous, differentiable, invertible, and satisfy $\exp_a(0) = 1$, $\exp_a(1) = a$, and $\exp_a(x+y) = \exp_a(x)\exp_a(y)$ for all real x, y. Looking back on what we have done so far, we see that that function exp satisfies these conditions for the specific value $a = e \approx 2.7128\ldots$. It therefore makes sense to make the following definition.

Definition of e raised to the power x For real x, e^x means $\exp x$.

This agrees with the natural definition of e^x, for rational x, given earlier. Recall that specified that

- first $e^0 = 1$ ($\exp(0) = 1$), and for $n \in \mathbb{N}$, $e^n = e \cdot e^{n-1}$ ($\exp n = \exp(1 + (n-1)) = \exp(1)\exp(n-1)$).
- It then specified that for $n \in \mathbb{N}$, $e^{1/n}$ is that unique positive number such that $(e^{1/n})^n = e$; but $(\exp(1/n))^n = \exp(1/n) \cdot \cdots \cdot \exp(1/n) = \exp(1/n + \cdots + 1/n) = \exp(1) = e$, and $\exp(1/n) > 0$, so $\exp(1/n)$ is indeed that unique positive number that $e^{1/n}$ was defined to be.
- It then specified that $e^{m/n} = (e^{1/n})^m$; but also $\exp(m/n) = \exp((1/n) + \cdots + (1/n))$ (where there are m summands, all $1/n$), and this equals $\exp(1/n) \cdot \cdots \exp(1/n) = (\exp(1/n))^m$.

- Finally, it specified that $e^r = 1/e^{-r}$ for $r < 0$ and rational; but $1/\exp(-r) = \exp(0)/\exp(-r) = \exp(0 - -(r)) = \exp(r)$.

So $\exp(x)$ agrees with the natural definition of e^x for all rational x; and since there is at most one continuous function on the reals that agrees with e^x on the rationals, that fact that exp is a continuous function on the reals makes it the only sensible choice for an interpretation of e^x.

What about defining a^x, for arbitrary $a > 0$? To start the process, we need the identity

$$\log(a^x) = x \log a$$

for *rational* x, which we can verify from previously established properties. For $x > 0$ with $x = m/n$, $m, n \in \mathbb{N}$ we have

$$\log(a^x) = \log(a^{m/n}) = \log((a^{1/n})^m) = m \log(a^{1/n}),$$

with the last equality following from $\log(a_1 \cdot \cdots \cdot a_n) = \log a_1 + \cdots + \log a_n$, applied with $n = m$ and $a_i = a^{1/n}$. Also, from the same identity we get

$$\log(a) = \log\left((a^{1/n})^n\right) = n \log a^{1/n},$$

so $\log a^{1/n} = n \log a$, and $m \log(a^{1/n}) = (m/n) \log a$. From this we get $\log a^x = x \log a$.

If $a < 0$ then

$$\log a^x = \log(1/a^{-x}) = \log 1 - \log a^{-x} = 0 - (-x) \log a = x \log a.$$

So, for *rational* x we have $\log a^x = x \log a$ or

$$a^x = e^{x \log a}.$$

This suggests a very obvious choice for a^x, for arbitrary real x.

Definition of a raised to the power x For $a > 0$, and real x, a^x means $\exp(x \log a)$ (or $e^{x \log a}$).

Clearly the function that sends x to $\exp(x \log a)$ is continuous — it is the composition of the continuous function "multiply by $\log a$" with the continuous function exp — and by the discussion above it agrees with the natural definition of a^x for $x \in \mathbb{Q}$; so it is the only sensible choice for a^x for arbitrary real x.

Notice that the fundamental relation for logarithms,

$$\log(a^x) = x \log a \quad \text{for all } a > 0 \text{ and all real } x$$

follows now *by definition* of a^x.

Here are some of the basic properties of the a^x function, all of which follow from just unravelling the definition, and using properties of exp and log:

- $a^0 = e^{0 \log a} = e^0 = 1$ and $a^1 = e^{1 \log a} = e^{\log a} = a$;
- for any reals x, y, $a^{x+y} = e^{(x+y)\log a} = e^{x \log a + y \log a} = e^{x \log a} e^{y \log a} = a^x a^y$ (these two properties are the ones we desired of a^x, along with the continuity we have already established);
- $(a^b)^c = e^{c \log(a^b)} = e^{(bc) \log a} = a^{bc}$; and
- if $\exp_a(x) = a^x = e^{x \log a}$ then $\exp_a'(x) = (\log a) a^x$ and so if $a > 1$ then $\exp_a$ is increasing (from 0 to ∞), while if $a < 1$ it is decreasing (from ∞ to 0), and in either case it is invertible; and furthermore, since $\exp_a''(x) = (\log a)^2 a^x > 0$, $\exp_a$ is concave.

We denote the inverse of $\exp_a$ by $\log_a$, so $\log_a(x) = y$ means $a^y = x$. There is an easy translation between $\log_a$ and log, namely:

$$\log_a x = \frac{\log x}{\log a}$$

(if $\log_a x = y$, then $a^y = x$, so $e^{y \log a} = x$, so $y \log a = \log x$). This allows us, for example, to quickly deduce that

$$\log_a'(x) = \frac{1}{x \log a}.$$

There are some basic algebraic identities that $\log_a$ satisfies, but we won't bother mentioning them here; in general when working with $\log_a$ or $\exp_a$ it is best to translate back to log and exp to do algebraic manipulations.

As an example of this, consider, for fixed $a \in \mathbb{R}$, the power function $f_a : (0, \infty) \to \mathbb{R}$ given by $f_a(x) = x^a$. (We have previously only considered the power function for rational a). We have $f_a(x) = e^{a \log x}$, and so

$$f_a'(x) = e^{a \log x} \cdot a \cdot \frac{1}{x} = a e^{a \log x} \cdot x^{-1} = a e^{a \log x} e^{-\log x} = a e^{(a-1) \log x} = a x^{a-1};$$

and so the usual rule for differentiating the power function holds, even when the exponent is an arbitrary real.

We end this discussion of the exponential and logarithm functions by presenting two theorems, one of which is a nice result with a short proof, that will not get used again, and the other of which is a slightly more technical result with a longer proof, that will be very useful to us later.

Theorem 12.3. *Suppose $f : \mathbb{R} \to \mathbb{R}$ is differentiable at all x, and $f' = f$. Then there is a constant c such that $f(x) = ce^x$ for all x. In particular, if $f(0) = 1$ then $f(x) = e^x$.*

Proof: Consider the function $g : \mathbb{R} \to \mathbb{R}$ defined by $g(x) = f(x)/e^x$. This is differentiable everywhere, with derivative

$$g'(x) = \frac{f(x)e^x - f'(x)e^x}{e^{2x}} = 0$$

(since $f'(x) = f(x)$ for all x). So $g(x) = c$ for all x, for some constant c. In other words, $f(x) = ce^x$. And if $f(0) = 1$, then $1 = ce^0$ so $c = 1$ and $f(x) = e^x$. □

Theorem 12.4. *For each fixed $n \in \mathbb{N}$,*

$$\lim_{x\to\infty} \frac{e^x}{x^n} = \infty.$$

Essentially this says that the exponential function grows faster than any polynomial. Spivak gives a fairly delicate proof of this, but there are lots of proofs; one, using L'Hôpital's rule, appears in homework. The proof we give uses derivatives, and needs the following lemma.

Lemma 12.5. *If $f, g : [a, \infty)$ are both differentiable, with $f(a) = g(a)$ and $f'(x) \geq g'(x)$ for all x, then $f(x) \geq g(x)$ for all x. In particular, if $g(x) \to \infty$ as $x \to \infty$ then also $f(x) \to \infty$ as $x \to \infty$.*

Proof: We use the mean value theorem. Suppose there was some $b > a$ with $f(b) < g(b)$. Then, applying the mean value theorem to the function $h = f - g$ on $[a, b]$, we have that there is some $c \in (a, b)$ with

$$f'(c) - g'(c) = h'(c) = \frac{h(b) - h(a)}{b - a} = \frac{f(b) - g(b)}{b - a} < 0,$$

so $f'(c) < g'(c)$, a contradiction. □

Proof (of Theorem 12.4): Set $f(x) = e^x/x^n$. We have

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

This is positive for $x > n$, so f is increasing on $[n, \infty)$, and in particular that means that $f(x) \geq f(n) = e^n/n^n$ for $x \geq n$. It follows that for $x \in [n, \infty)$ we have

$$f'(x) \geq \frac{e^n}{n^n}\left(1 - \frac{n}{x}\right).$$

In particular, for $x \geq 2n$ we have

$$f'(x) \geq \frac{e^n}{2n^n} = c_n,$$

where c_n is some positive constant.

Now let $g(x)$ be the linear function with slope c_n that passes through the point $(2n, f(2n))$, that is,

$$g(x) = c_n(x - 2n) + f(2n).$$

On the interval $[2n, \infty)$ the conditions of Lemma 12.5 are satisfied by f and g, so, since $g(x) \to \infty$ as $x \to \infty$, we conclude that $f(x) \to \infty$ as $x \to \infty$, as claimed. □

Almost the same proof can be used to show that every exponential function (with base at least one) grows faster than every power function:

$$\text{for } a > 1 \text{ and } b < \infty, \lim_{x\to\infty} \frac{a^x}{x^b} = \infty.$$

For example, 1.00000001^x grows faster than $x^{1{,}000{,}000{,}000}$ (though you have to look at very large values of x to see this!)

12.3 The trigonometric functions sin and cos

In this section, we use the integral to formally define the trigonometric functions sin (*sine*) and cos (*cosine*), and establish all their properties.

Recall first our provisional definition of sin and cos:

Provisional definition of sin and cos The points reached on unit circle centered at the origin, starting from $(1, 0)$, after traveling a distance θ, measured counter-clockwise, is $(\cos\theta, \sin\theta)$.

Knowing that the area of the unit circle is π, and that the circumference is 2π, we would get exactly the same thing if we said that the point P on the unit circle $x^2 + y^2 = 1$ has coordinates $(\cos\theta, \sin\theta)$ when $\theta/2$ is *area* of circle sector between $(1, 0)$ and P — at least, as long as P is in upper half plane (so $0 \le \theta \le \pi$).

To start the formal definition, we *define* π to be the "area" of the unit circle, or more specifically to be twice the area of that part of the unit circle $x^2 + y^2 = 1$ that lies in the upper half plane.

Definition of π

$$\pi = 2\int_{-1}^{1} \sqrt{1-x^2}\, dx.$$

Using upper and lower Darboux sums for the partition $(-1, -4/5, -3/5, 0, 3/5, 4/5, 1)$ we get the very rough estimates

$$2.4 \le \pi \le 3.52.$$

Next, we set up a function $A(x)$ that captures the notion of the area of circle sector between $(1, 0)$ and $P = (x, \sqrt{1-x^2})$, where P is in the upper half plane, that is, $-1 \le x \le 1$. For $0 \le x \le 1$ we have

$$A(x) = \frac{x\sqrt{1-x^2}}{2} + \int_x^1 \sqrt{1-t^2}\, dt$$

and for $-1 \le x \le 0$,

$$A(x) = \int_x^1 \sqrt{1-t^2}\, dt - \frac{(-x)\sqrt{1-x^2}}{2}$$

So in fact for every $x \in [-1, 1]$ we have

$$A(x) = \frac{x\sqrt{1-x^2}}{2} + \int_x^1 \sqrt{1-t^2}\, dt.$$

A is a continuous function on $[-1,1]$, and it is differentiable on $(-1,1)$, with derivative

$$A'(x) = \frac{-1}{2\sqrt{1-x^2}}.$$

This derivative is never 0, and in fact is always negative on $(-1,1)$, so A is decreasing on $[-1,1]$. It follows that the range of A is $[A(1), A(-1)] = [0, \pi/2]$. All this says that A has an inverse $A^{-1} : [0,\pi/2] \to [-1,1]$ which is decreasing.

Following our informal definition of sin and cos, we want that for $0 \le \theta \le \pi$, $(\cos(\theta), \sin(\theta)$ is the point P on the circle $x^2 + y^2 = 1$ for which the area of the circle sector between $(1,0)$ and P equals $\theta/2$. That is, we want $A(\cos\theta) = \theta/2$, or $\cos\theta = A^{-1}(\theta/2)$ (note that this seems to be sensible: as θ goes from 0 to π, $\theta/2$ goes from 0 to $\pi/2$, exactly the domain of A^{-1}).

Initial definition of cos **and** sin Define $\cos : [0,\pi] \to [-1,1]$ by

$$\cos\theta = A^{-1}(\theta/2).$$

Define $\sin : [0,\pi] \to [0,1]$ by

$$\sin\theta = \sqrt{1-\cos^2\theta}.$$

Observe that, since A is differentiable on $(-1,1)$ with derivative never 0, A^{-1} is differentiable on $(0,\pi)$, and for $\theta \in (0,\pi)$ cos is differentiable, with Have

$$\begin{aligned}
\cos'\theta &= (A^{-1})'(\theta/2) \\
&= \frac{1}{2A'(A^{-1}(\theta/2))} \\
&= -\sqrt{1-A^{-1}(\theta/2))^2} \\
&= -\sin\theta.
\end{aligned}$$

Now differentiating the equation $\sin^2\theta + \cos^2\theta = 1$ get, on $(0,\pi)$,

$$\sin'\theta = \cos\theta.$$

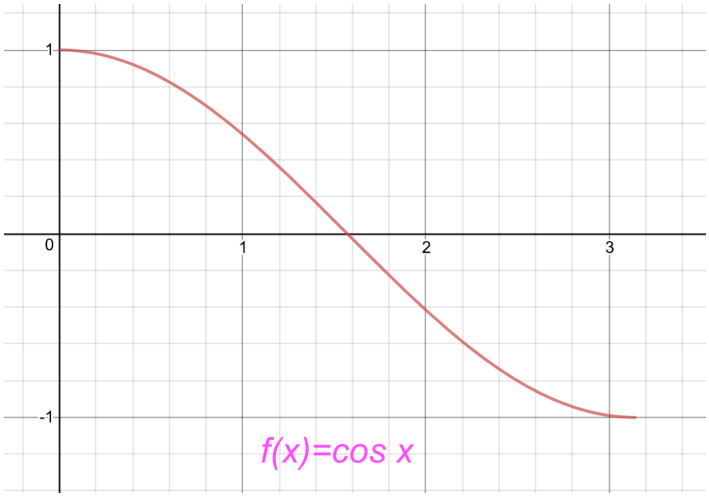
We are now in a position to sketch a reasonable graph of cos on the interval $[0,\pi]$. We have

- $\cos 0 = A^{-1}(0) = 1$
- $\cos\pi = A^{-1}(\pi/2) = -1$
- $\cos' = -\sin < 0$ on $(0,\pi)$, so cos decreasing
- cos is continuous, so by the intermediate value theorem there is an $m \in (0,\pi)$ with $\cos m = 0$. We have $A^{-1}(m/2) = 0$, so

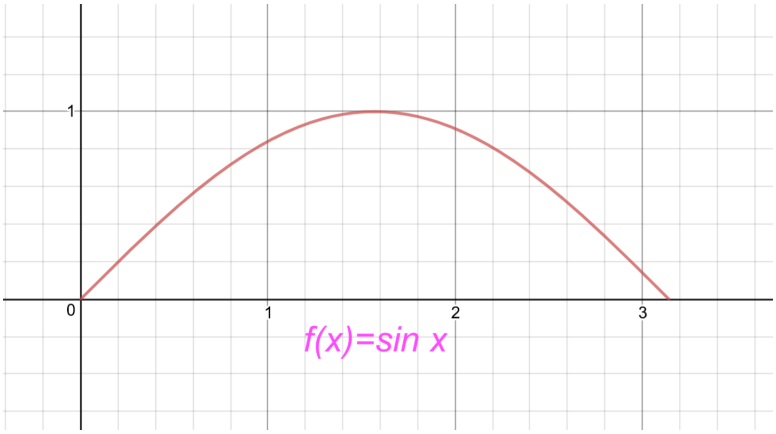
$$m = 2A(0) = \int_0^1 \sqrt{1-t^2}\ dt = 2\int_0^1 \sqrt{1-t^2}\ dt = \pi/2.$$

- $\cos'' = -\cos$, so cos is concave on $[0, \pi/2]$ and convex on $[\pi/2, \pi]$.
- $\cos'(\pi/2) = -\sin(\pi/2) = -\sqrt{1 - \cos^2(\pi/2)} = -1$, so as the graph crosses the x-axis it has slope -1.
- As $\theta \to 0^+$ we have $\cos'(\theta) = -\sin(\theta) = -\sqrt{1 - \cos^2(\theta)} \to 0$, and similarly as $\theta \to \pi^-$ we have $\cos'(\theta) \to 0$, so the graph is flat near 0 and π.

We sketch the graph of cos below, on the interval $[0, \pi]$.



The same reasoning can be used to sketch a graph of sin (again on $[0, \pi]$); this is left as an exercise.



We now extend sin and cos to the whole real line.

- We begin with the interval $[\pi, 2\pi]$. For θ in this range, the x-coordinate of the point that is distance θ from $(1, 0)$, is the same as the x-coordinate of the point that is distance θ' from $(1, 0)$, where $\theta' = 2\pi - \theta$. This motivates the definition that for $\theta \in [\pi, 2\pi]$,

$$\cos\theta = \cos(2\pi - \theta).$$

- Similarly for $\theta \in [\pi, 2\pi]$ we set

$$\sin\theta = -\sin(2\pi - \theta).$$

Observe that since cos is continuous on $[0, \pi]$, with $\lim_{x\to\pi^-} \cos x = \cos\pi = -1$, it follows that cos is continuous on $[\pi, 2\pi]$, with

$$\lim_{x\to\pi^+} \cos x = \lim_{x\to\pi^+} \cos(2\pi - x) = \lim_{x\to\pi^-} \cos x = \cos\pi = -1.$$

But from this it follows that actually cos is continuous on $[0, 2\pi]$. Similarly it can be argued that sin is continuous on $[0, 2\pi]$.

The relation $\cos^2\theta + \sin^2\theta = 1$ for $\theta \in [\pi, 2\pi]$ follows immediately from the same relation for $\theta \in [0, \pi]$, so in fact it too holds for all $\theta \in [0, 2\pi]$.

Finally, we turn to differentiability. Since $\cos' = -\sin$ on $(0, \pi)$, we have for $\theta \in (\pi, 2\pi)$ that

$$\cos'(\theta) = \cos'(2\pi - \theta) = -\sin(2\pi - \theta) \times (-1) = \sin(2\pi - \theta) = -\sin\theta.$$

What about $\cos'\pi$? We utilize the following lemma.

Lemma 12.6. *Suppose*

- *f is continuous at a,*
- *f' exists near a, and*
- *$\lim_{x\to a} f'(x) = L$ exists.*

Then $f'(a) = L$.

Proof: For $h > 0$, by the mean value theorem there's $\alpha_h \in (a, a + h)$ with

$$\frac{f(a+h) - f(a)}{h} = f'(\alpha_h).$$

As $h \to 0^+$ we have $\alpha_h \to a$, and so $f'(\alpha_h) \to L$. This shows that f is differentiable from above at a, with derivative L.

A similar argument applies for $h < 0$. □

We apply this lemma with $f = \cos$ and $a = \pi$. We've shown that cos is continuous at π, and it's differentiable near π. It's derivative near π is $-\sin$, which approaches 0 near π, so we conclude that cos is differentiable at π, with derivative 0 — which is $-\sin\pi$, so in fact $\cos' = -\sin$ on all of $(0, 2\pi)$. Similarly we can argue $\sin' = \cos$ on $[0, 2\pi]$.

- Finally, we extend both $\cos, \sin$ periodically to $\mathbb{R}$, via

$$\cos\theta = \cos\theta', \quad \sin\theta = \sin\theta'$$

for $\theta = 2k\pi + \theta'$, $0 \le \theta' \le 2\pi$, $k \in \mathbb{Z}$.

That

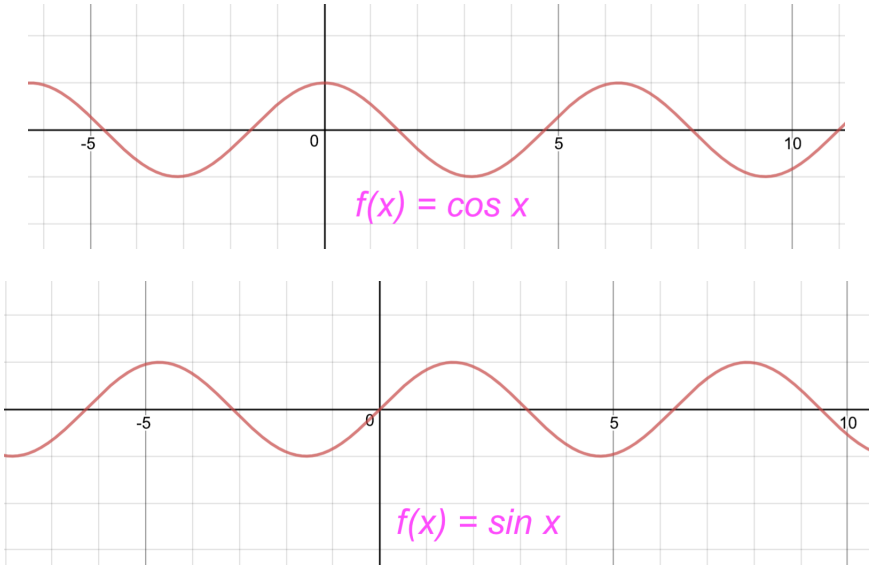
$$\sin^2\theta + \cos^2\theta = 1$$

for all θ, follows almost immediately from the same relation for $\theta \in [0, 2\pi]$. That

$$\cos'(\theta) = -\sin(\theta), \quad \sin'(\theta) = \cos(\theta)$$

for all θ, follows exactly as this relation extended from $(0, \pi)$ to $(0, 2\pi)$ (via Lemma 12.6).

Here are the graphs of cos and sin on their full domains:



We make a digression here, to give another application of Lemma 12.6, that will be useful later. Consider the function f defined by

$$f(x) = \begin{cases} e^{-1/x^2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

We claim that f is continuous and differentiable, and that $f'(0) = 0$. Away from 0, the function is clearly continuous and differentiable arbitrarily many times (also known as *infinitely differentiable*).

To show continuity at 0, we need to establish $\lim_{x\to 0} e^{-1/x^2} = 0$. Recall that we have proven that

$$\lim_{x\to 0^+} g(x) = \lim_{y\to\infty} g(1/y) \quad \text{and} \quad \lim_{x\to 0^-} g(x) = \lim_{y\to -\infty} g(1/y).$$

So to show $\lim_{x\to 0} e^{-1/x^2} = 0$, it suffices to show that

$$\lim_{y\to\infty} e^{-y^2} = 0 \quad \text{and} \quad \lim_{y\to -\infty} e^{-y^2}.$$

Since $e^y \to \infty$ as $y \to \infty$, and $y^2 > y$ for all large y, it follows that $e^{y^2} \to \infty$ as $y \to \infty$, so that indeed $\lim_{y\to\infty} e^{-y^2} = 0$; that $\lim_{y\to -\infty} e^{-y^2}$ is established similarly (see below for details in a similar case). This shows that f is continuous at 0.

For differentiability, we use Lemma 12.6. Here, f is continuous at 0, and differentiable near 0. So to establish that f is differentiable at 0, with derivate 0, if suffices to show that

$$\lim_{x\to 0} \frac{2e^{-1/x^2}}{x^3} = 0$$

(note that $f'(x) = 2e^{-1/x^2}/x^3$ if $x \neq 0$).

As before, this limit is implied by

$$\lim_{y\to\infty} \frac{y^3}{e^{y^2}} = 0 \quad \text{and} \quad \lim_{y\to-\infty} \frac{y^3}{e^{y^2}} = 0.$$

Since $e^{y^2} > e^y$ for all large positive y, and since $y^3/e^y \to 0$ as $y \to \infty$ (a basic estimate that we proved in class), it follows that $\lim_{y\to\infty} \frac{y^3}{e^{y^2}} = 0$. For the negative limit, notice that

$$\lim_{y\to-\infty} \frac{y^3}{e^{y^2}} = \lim_{z\to\infty} \frac{(-z)^3}{e^{(-z)^2}} = -\lim_{z\to\infty} \frac{z^3}{e^{z^2}} = -0 = 0.$$

So we conclude that f is differentiable at 0, with derivative 0.

In fact, we can do more: f is k times differentiable for every natural number k, and $f^{(k)}(0) = 0$. To see this we will make use of the fact that away from 0, the kth derivative of f has the following form:

$$f^{(k)}(x) = P_k(1/x)e^{-1/x^2}$$

where P_k is a polynomial. This fact can be proven by induction on k. Indeed, for $k = 1$ we have already shown it (in part (a)), with specifically the polynomial being $P_1(z) = 2z^3$.

For $k > 1$, suppose that $f^{(k-1)}(x) = P_{k-1}(1/x)e^{-1/x^2}$ where P_{k-1} is a polynomial. We then have

$$\begin{aligned} f^{(k)}(x) &= P_{k-1}(1/x)e^{-1/x^2}\left(2/x^3\right) + e^{-1/x^2}P'_{k-1}(1/x)\left(-1/x^2\right) \\ &= \left(\frac{2}{x^3}P_{k-1}(1/x) - \frac{P'_{k-1}(1/x)}{x^2}\right)e^{-1/x^2}, \end{aligned}$$

so that indeed $f^{(k)}(x) = P_k(1/x)e^{-1/x^2}$ with P_k the polynomial given by $P_k(z) = 2z^3P_{k-1} - z^2P'_{k-1}(z)$. This completes the induction.

We will also make use of the fact that if P is a polynomial, then $\lim_{x\to 0} P(1/x)e^{-1/x^2} = 0$. Indeed, to show this it suffices to show

$$\lim_{y\to\infty} P(y)/e^{y^2} = 0 \quad \text{and} \quad \lim_{y\to-\infty} P(y)/e^{y^2} = 0.$$

The first of these follows immediately from $e^{y^2} > e^y$ for large (positive y) and the fact that y^k/e^y goes to 0 as $y \to \infty$ for any natural number k; then the second follows from the first on observing that

$$\lim_{y\to-\infty} P(y)/e^{y^2} = \lim_{z\to\infty} P(-z)/e^{z^2} = 0.$$

We now prove that the predicate $p(k)$: "f is k times differentiable and $f^{(k)}(0) = 0$" is true for all natural numbers k, by induction on k, with the base case $k = 1$ having been proven earlier.

For the induction step, suppose that f is k times differentiable and $f^{(k)}(0) = 0$. Let g be the kth derivative of f. By induction $g(0) = 0$, and by the calculations done above g approaches 0 near 0, so that g is continuous at 0. But now again by the calculations done above, g' approaches 0 near 0 (away from zero, g' is the $(k+1)$st derivative of f calculated above), so by Lemma 12.6, g' exists at 0 and takes value 0 there. This completes the induction.

The function f is as flat as it possible can be at 0 — its value, and the values of all its derivatives, are 0. And yet the function is *not* the zero function everywhere. We will return to this when we talk about Taylor polynomials.

Returning to trigonometric functions: it is now immediate that $\sin''(\theta) = -\sin(\theta)$ and that $\cos''(\theta) = -\cos(\theta)$ for all θ, that is, that $\sin, \cos$ are both solutions to the differential equation $f'' + f = 0$. Just as exp was (essentially) characterized by the differential equation $f' = f$ (Theorem 12.3), it turns out that sin and cos are (essentially) characterized by the differential equation $f'' + f = 0$. Unlike Theorem 12.3, which is nice but will not get used again this year, the following analogous theorem will have an immediate and important pay-off.

Theorem 12.7. *Suppose $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable at all x, that $f'' + f = 0$, and that $f(0) = a$ and $f'(0) = b$. Then $f(x) = a\cos x + b\sin x$ for all x.*

Proof: We begin with the special case $a = b = 0$. Since $f'' + f = 0$, we have $f'f'' + f'f = 0$,[193] and so $((f')^2 + f^2)' = 0$ and so $(f')^2 + f^2 = C$ for some constant C. Evaluating the left-hand side at 0, we find that $C = 0$, so f^2, $(f')^2$ are both zero, and in particular $f = 0$, as claimed.

Now for general a, b, set $g = f - a\cos - b\sin$. We have $g'' + g = 0$, $g(0) = 0$, $g'(0) = 0$, and so $g = 0$. This says that $f = a\cos x - b\sin x$, as claimed. □

The immediate pay-off is that the addition formulae for sin and cos are now almost immediate.

Theorem 12.8. *For all x, y,*

- $\sin(x + y) = \sin x \cos y + \sin y \cos x$ *and*
- $\cos(x + y) = \cos x \cos y - \sin x \sin y$.

Proof: We just prove the first identity; the second is similar. For each *fixed* y, the function $f(x) = \sin(x + y)$ (a function of x only) satisfies $f'' + f = 0$, $f(0) = \sin y$, and $f'(0) = \cos y$, so by Theorem 12.7 we have $\sin(x + y) = f(x) = \sin y \cos x + \cos y \sin x$. □

This allows us to calculate some particular values of the functions sin and cos. The first we will use fairly soon, so we derive that fully; the others we will never use, so are left as exercises.

[193] So ... this is a rabbit-out-of-a-hat proof.

- $\sin \pi/4 = \cos \pi/4 = \sqrt{2}/2$. We have $0 = \cos \pi/2 = \cos(\pi/4 + \pi/4) = \cos^2(\pi/4) - \sin^2(\pi/4)$, so $\cos^2(\pi/4) = \sin^2(\pi/4)$. Since both are positive, we have $\cos(\pi/4) = \sin(\pi/4) > 0$. Now from $\cos^2(\pi/4) + \sin^2(\pi/4) = 1$ we get $2\cos^2(\pi/4) = 1$ or $\cos(\pi/4) = \sqrt{2}/2$.
- $\sin \pi/6 = \cos \pi/3 = 1/2$.
- $\sin \pi/3 = \cos \pi/6 = \sqrt{3}/2$.

Another consequence of Theorems 12.7 and 12.8 is that we can use it to verify some properties of sin and cos that appears obvious from the graphs of the two functions, but up until now would have been quite hard to prove.

Theorem 12.9. *1. The graph of* sin *is a shift of the graph of* cos*; specifically, for all* x, $\sin(x + \pi/2) = \cos(x)$.

2. sin *is an odd function; that is, for all* x, $\sin(-x) = -\sin(x)$.

3. cos *is an even function; that is, for all* x, $\cos(-x) = \cos(x)$.

Proof: For item 1 we have, using Theorem 12.8 and some special values of sin and cos that come out of the definition,

$$\sin(x + \pi/2) = \sin(x)\cos(\pi/2) + \cos(x)\sin(\pi/2) = \cos(x).$$

For item 2, consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = \sin(-x)$. We have $f'(x) = -\cos(-x)$ and $f''(x) = -\sin(-x)$, so f satisfies the equation $f'' + f = 0$. It follows from Theorem 12.8 that for all x,

$$f(x) = f(0)\cos x + f'(0)\sin x = 0 \cdot \cos x - 1 \cdot \sin x = -\sin x.$$

But since $f(x) = \sin(-x)$ for all x, we immediately get that $\sin(-x) = -\sin(x)$ for all x.

The proof of item 3 is similar to that of item 2, and is omitted. □

12.4 The other trigonometric functions

Having defined sin and cos, we can now define some auxiliary trigonometric functions. The most important of these is the tangent function.

Definition of tan The *tangent* function $\tan : \mathbb{R} \setminus \{(n + 1/2)\pi : n \in \mathbb{Z}\}$[194] is defined by

$$\tan \theta = \frac{\sin \theta}{\cos \theta}.$$

[194]Note that the domain is precisely those points where $\cos \neq 0$.

Since sin and cos are periodic with period 2π ($\sin(x+2\pi) = \sin x$ for all x), it is clear that tan is also periodic with period 2π. But in fact, tan has period π: using the angle sum formulae for $\sin, \cos$, along with $\sin \pi = 0$, $\cos \pi = -1$ we get

$$\tan(x+\pi) = \frac{\sin(x+\pi)}{\cos(x+\pi)} = \frac{-\sin x}{-\cos x} = \tan x.$$

To understand the tan function, then, it suffices to examine it on the interval $(-\pi/2, \pi/2)$. On this interval it is continuous and differentiable, with (by the quotient rule)

$$\tan'(x) = \frac{(\cos x)(\cos x) - (\sin x)(-\sin x)}{\cos^2 x} = \frac{\cos^2 x + \sin^2 x}{\cos^2 x} = \frac{1}{\cos^2 x}.$$

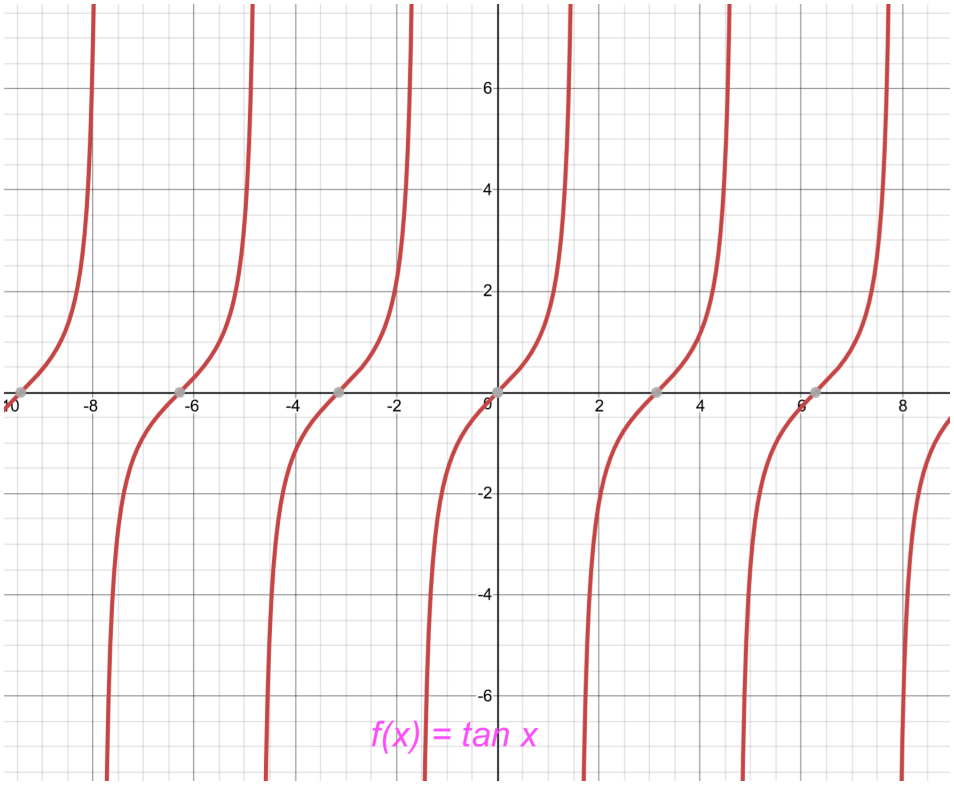
Since this is positive, tan is increasing on $(-\pi/2, \pi/2)$. From our knowledge of sin and cos, we have

$$\lim_{x \to \pi/2^-} \tan x = +\infty, \qquad \lim_{x \to -\pi/2^+} \tan x = -\infty.$$

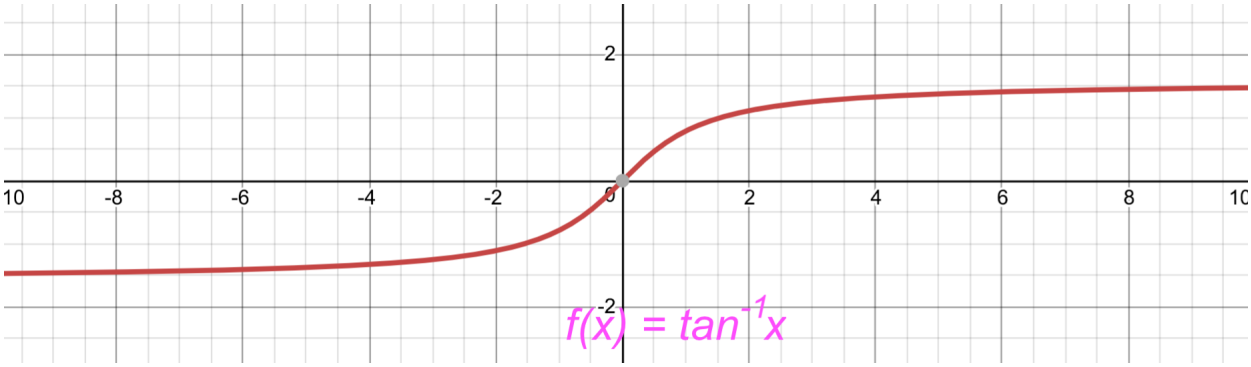
Also,

$$\tan''(x) = \frac{2 \sin x}{\cos^3 x},$$

which is positive for $x \in [0, \pi/2)$ (so tan is convex on that interval) and negative for $x \in (-\pi/2, 0]$ (so tan is concave on that interval). Finally noting that $\tan 0 = 0$, we have enough information to produce an accurate graph tan:



tan is clearly not invertible, but it becomes invertible if it is restricted to the domain $(\pi/2, \pi/2)$ (on that interval it is monotone increasing from $-\infty$ to ∞). We define the function $\tan^{-1} : \mathbb{R} \to (\pi/2, \pi/2)$[195] to be the inverse of the function $\tan : (\pi/2, \pi/2) \to \mathbb{R}$ (the restriction of tan to the domain $(\pi/2, \pi/2)$). That is, for each real x, $\tan^{-1}(x)$ is defined to be the unique $\theta \in (\pi/2, \pi/2)$ such that $\tan\theta = x$. From the graph of tan we easily get the graph of $\tan^{-1}$:



Notice that it is monotone increasing on the whole real line, but bounded.

We now compute the derivative of $\tan^{-1}$, which turns out to be $1/(1+x^2)$[196]. We will use the identity

$$\tan^2 x + 1 = \frac{1}{\cos^2 x},$$

valid on $(-\pi/2, \pi/2)$, which follows immediately from $\sin^2 x + \cos^2 x = 1$. We have

$$\begin{aligned}
(\tan^{-1})'(x) &= \frac{1}{\tan'(\tan^{-1}(x))} \\
&= \cos^2(\tan^{-1}(x)) \\
&= \frac{1}{1+\tan^2(\tan^{-1}(x))} \\
&= \frac{1}{1+x^2}.
\end{aligned}$$

Note that this tallies with the graph of $\tan^{-1}$:

- $(\tan^{-1})'$ is positive, the function is increasing;
- $(\tan^{-1})'' = -2x/(1+x^2)^2$, which is negative for negative x (where the function is concave), and positive for positive x (where the function is convex);
- $\lim_{x\to\pm\infty}(\tan^{-1})'(x) = 0$, and the graph is flat at $\pm\infty$.

[195] Sometimes called "arctan".

[196] The appearance of such a simple, rational function, as the derivative of $\tan^{-1}$, should not be surprising; recall that cos was defined as the inverse of a function that very clearly has a rational function as its derivative.

The relation $(\tan^{-1})'(x) = 1/(1+x^2)$ leads to an integral relation. Define $F : \mathbb{R} \to (-\pi/2, \pi/2)$ by $F(x) = \int_0^x dt/(1+t^2)$. Since by the fundamental theorem of calculus $F'(x) = 1/(1+x^2)$, we have that $F(x) = \tan^{-1}(x) + C$ for some constant C; and setting $x = 0$ we get $C = 0$. So:

$$\tan^{-1}(x) = \int_0^x \frac{dt}{1+t^2}.$$

Recall that previously we showed, by comparison with $1/t^2$, that $\int_0^\infty dt/(1+t^2)$ exists; now we give a value to that integral:

$$\int_0^\infty \frac{dt}{1+t^2} = \lim_{x\to\infty} \int_0^\infty \frac{dt}{1+t^2} = \lim_{x\to\infty} \left(\tan^{-1}(x) - \tan^{-1}(0)\right) = \frac{\pi}{2}.$$

The main point of all of this discussion of $\tan^{-1}$, though, is the following. We have shown $\cos(\pi/4) = \sin(\pi/4) = \sqrt{2}/2$, so $\tan(\pi/4) = 1$. In other words,

$$\int_0^1 \frac{dt}{1+t^2} = \frac{\pi}{4}.$$

There is a way to estimate $\int_0^1 dt/(1+t^2)$ that does not require trigonometric functions. We have, for each natural number n that is divisible by 4, at for each $t \geq 0$,

$$1 - t^2 + t^4 - t^6 + t^8 - \cdots - t^{n-2} \leq \frac{1}{1+t^2} \leq 1 - t^2 + t^4 - t^6 + t^8 - \cdots + t^n.$$

Indeed, if we multiply across by $1+t^2$, this becomes

$$1 - t^n \leq 1 \leq 1 + t^{n+2},$$

which is clearly true. It follows that

$$\int_0^1 \left(1 - t^2 + t^4 - t^6 + t^8 - \cdots - t^{n-2}\right)\, dt \leq \int_0^1 \frac{dt}{1+t^2} \leq \int_0^1 \left(1 - t^2 + t^4 - t^6 + t^8 - \cdots + t^n\right)\, dt.$$

We know that the integral in the middle is $\pi/4$. The integral on the right can easily be evaluated by the fundamental theorem of calculus:

$$\int_0^1 \left(1 - t^2 + t^4 - t^6 + t^8 - \cdots + t^n\right)\, dt = 1 - \frac{1}{3} + \frac{1}{5} - \cdots + \frac{1}{n+1},$$

while

$$\int_0^1 \left(1 - t^2 + t^4 - t^6 + t^8 - \cdots - t^{n-2}\right) = 1 - \frac{1}{3} + \frac{1}{5} - \cdots - \frac{1}{n-1}.$$

Combining, we get that for any n that is a multiple of 4,

$$1 - \frac{1}{3} + \frac{1}{5} - \cdots - \frac{1}{n-1} \leq \frac{\pi}{4} \leq 1 - \frac{1}{3} + \frac{1}{5} - \cdots + \frac{1}{n+1}.$$ [197]

[197] When we come to learn about infinite series, we will see that this translates to the sum

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots.$$

This is known variously as *Leibniz formula for* π, or as *Gregory's series*.

The difference between the right- and left-hand sides of this series of inequalities is $1/(n+1)$, which goes to 0 as n goes to infinity. This says that for all $\varepsilon > 0$ we can find an interval $[a, b]$, with rational endpoints and of length at most ε, inside which $\pi/4$ must lie; and of course, by using

$$4 - \frac{4}{3} + \frac{4}{5} - \cdots - \frac{4}{n-1} \leq \pi \leq 4 - \frac{4}{3} + \frac{4}{5} - \cdots + \frac{4}{n+1}$$

we can pin down π itself into a window of arbitrarily small width. For example, taking $n = 10,000$ we find that $\pi \in [3.14154, 3.14165]$.

This is nice, though not very efficient. But we can do better. From $\tan^{-1}(x) = \int_0^x dt/(1+t^2)$ we can use exactly the same argument to conclude that for positive x and n a multiple of 4, we have

$$x - \frac{x^3}{3} + \cdots - \frac{x^{n-1}}{n-1} \leq \tan^{-1}(x) \leq x - \frac{x^3}{3} + \cdots + \frac{x^{n+1}}{n+1}.$$

The difference between the right- and left-hand sides here is $x^{n+1}/(n+1)$, which goes very quickly to 0 as n grows, as long as $0 < x < 1$ (in contrast to $1/(n+1)$, which goes to 0 very slowly). So if we had some expression for π involving $\tan^{-1}(x)$ for small x, could get more accurate estimates for π more quickly.

Many such expressions are known. The most famous of them[198] is

$$\pi = 16 \tan^{-1} \frac{1}{5} - 4 \tan^{-1} \frac{1}{239}$$

(whose proof is left as an exercise).

This leads to the following bounds for π: with $x = 1/5$ and $y = 1/239$,

$$\pi \leq 16 \left(x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + \frac{x^{4n+1}}{4n+1} \right) - 4 \left(y - \frac{y^3}{3} + \frac{y^5}{5} - \cdots - \frac{y^{4n-1}}{4n-1} \right)$$

and

$$\pi \geq 16 \left(x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + \frac{x^{4n-1}}{4n-1} \right) - 4 \left(y - \frac{y^3}{3} + \frac{y^5}{5} - \cdots - \frac{y^{4n+1}}{4n+1} \right).$$

At $n = 5$ this already leads to

$$3.14159265358979 \leq \pi \leq 3.14159265358980,$$

accurate to 12 decimal places!

There are three other (somewhat) commonly encountered trigonometric functions, that are (essentially) the reciprocals of sin, cos and tan[199]. We mention them here for completeness, without really delving too deeply into them.

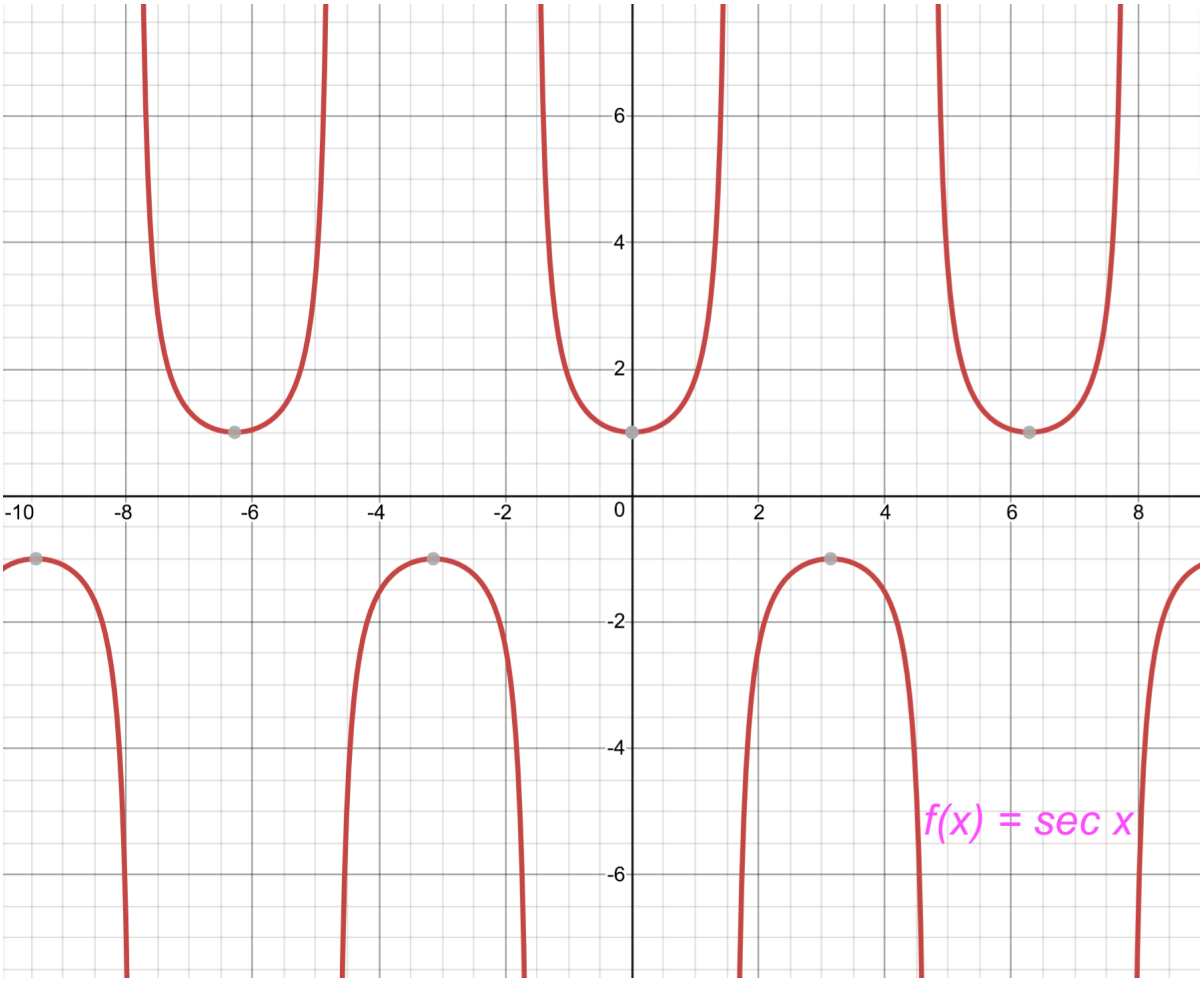
[198] Discovered in 1706 by J. Machin, and hence referred to as *Machin* or *Machin-like formulae*. Machin used his formula to calculate π to 100 decimal places in 1706. Today, much more elaborate Machin-like formulae are known, that allow π to be rapidly calculated to unfathomably many decimal places — see for example https://en.wikipedia.org/wiki/Machin-like_formula for many examples.

[199] Not exactly. tan is defined as sin / cos, while cot is defined as cos / sin. But cot is *not* the reciprocal of tan. Why not?

Definition of sec**, the secant function** $\sec : \mathbb{R}\setminus\{(n+1/2)\pi : n \in \mathbb{Z}\} \to (-\infty, -1]\cup[1, \infty)$ is defined by

$$\sec x = \frac{1}{\cos x}.$$

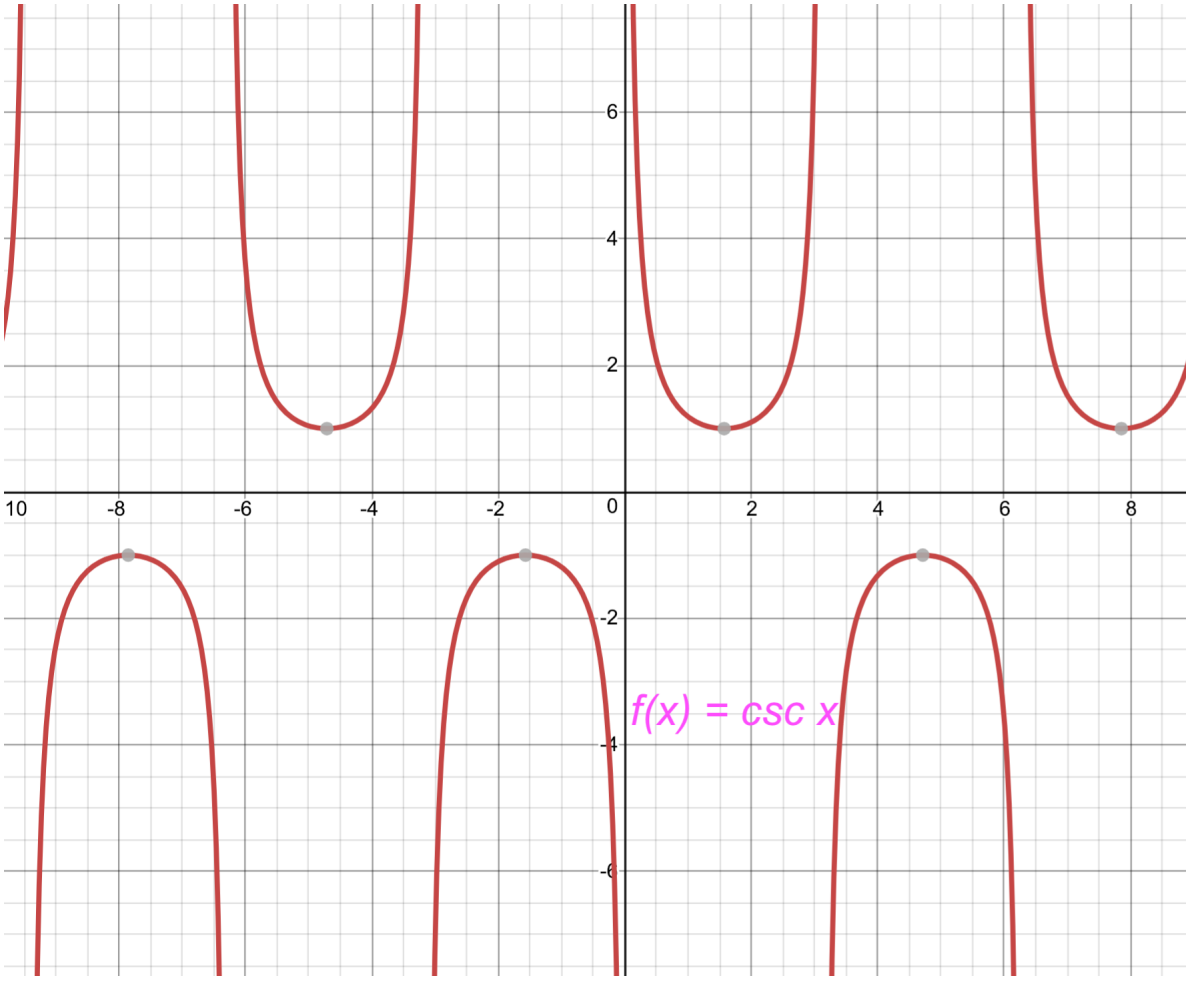
Here is a graph of sec:



Definition of csc**, the cosecant function** $\csc : \mathbb{R} \setminus \{n\pi : n \in \mathbb{Z}\} \to (-\infty, -1] \cup [1, \infty)$ is defined by

$$\csc x = \frac{1}{\sin x}.$$

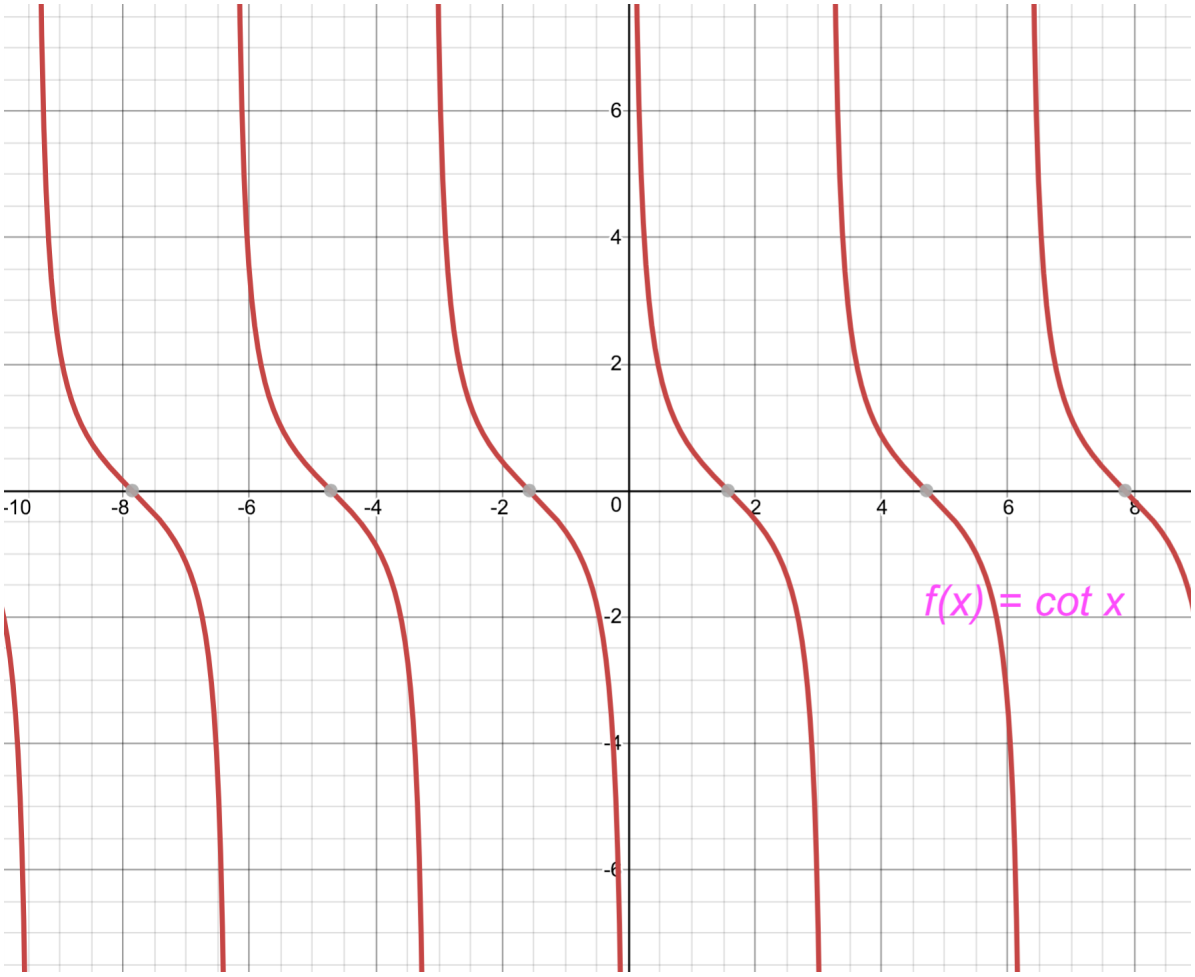
Here is a graph of csc:



Definition of cot**, the cotangent function** $\cot : \mathbb{R} \setminus \{n\pi : n \in \mathbb{Z}\} \to (-\infty, -1] \cup [1, \infty)$ is defined by

$$\cot x = \frac{\cos x}{\sin x}.$$

Here is a graph of cot:



All these functions are easily differentiated. It's not worth writing down any of the derivatives (at least at the moment); but it is worth noting that since $\tan' = 1/\cos^2$, we have in this language that

$$\tan' = \sec^2,$$

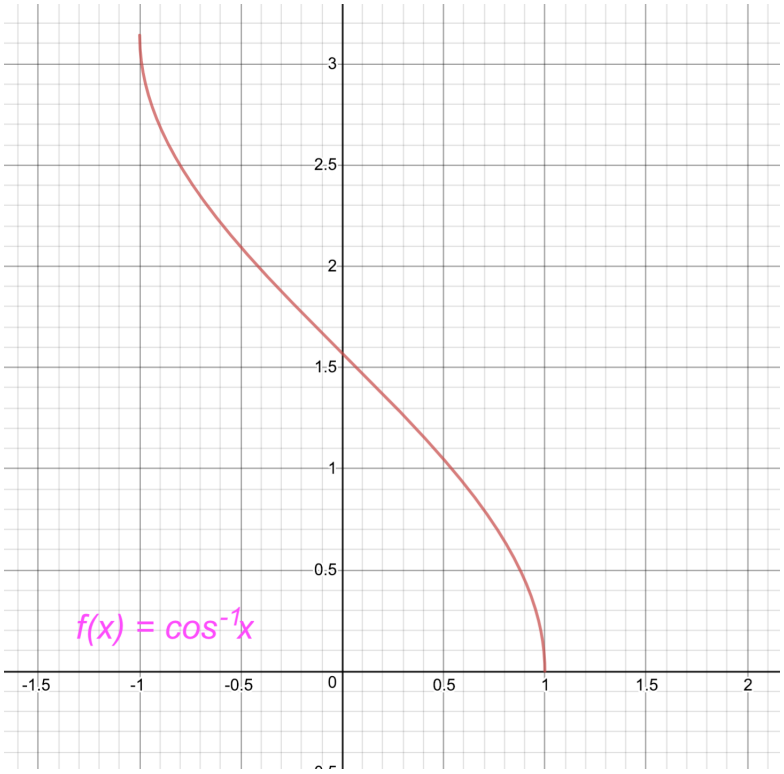
and also that since $\tan^2 + 1 = 1/\cos^2$, we have in this language that

$$\tan^2 + 1 = \sec^2 .$$

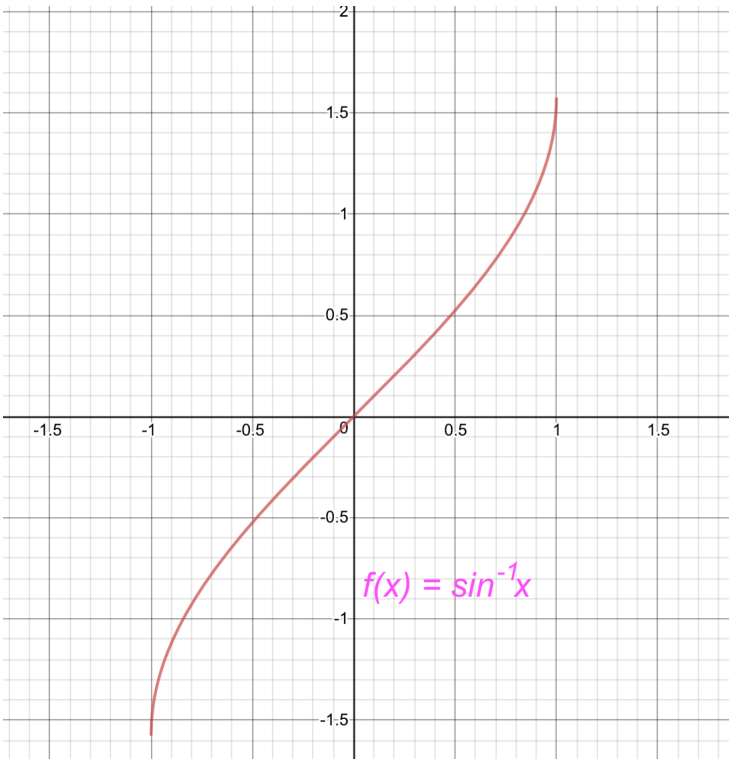
We have already discussed the inverse of the tangent function. Just like tan, none of the other trigonometric functions are invertible on their full domains, but they become invertible if suitably restricted. We only discuss here the inverses of cos, sin and sec.

It is standard to restrict cos to the interval $[0, \pi]$, and sin to the interval $[-\pi/2, \pi/2]$, to define their inverses. Formally we define the function $\cos^{-1} : [-1, 1] \to [0, \pi]$[200] to be the inverse of the function $\cos : [0, \pi] \to [-1, 1]$ (the restriction of cos to the domain $[0, \pi]$). That is, for each $x \in [-1, 1]$, $\cos^{-1}(x)$ is defined to be the unique $\theta \in [0, \pi]$ such that $\cos\theta = x$. From the graph of cos we easily get the graph of $\cos^{-1}$:

[200] Sometimes called "arccos".



And we define the function $\sin^{-1} : [-1, 1] \to [-\pi/2, \pi/2]$[201] to be the inverse of the function $\sin : [-\pi/2, \pi/2] \to [-1, 1]$. Here is the graph of $\sin^{-1}$:



[201] Sometimes called "arcsin".

These two functions have nice derivatives. For $x \in [-1, 1]$,

$$(\cos^{-1})'(x) = \frac{1}{\cos'(\cos^{-1}(x))} = \frac{-1}{\sin(\cos^{-1}(x))}.$$

Now

$$1 = \sin^2(\cos^{-1}(x)) + \cos^2(\cos^{-1}(x)) = \sin^2(\cos^{-1}(x)) + x^2,$$

, so

$$\sin^2(\cos^{-1}(x)) = 1 - x^2 \quad \text{and} \quad \sin(\cos^{-1}(x)) = \sqrt{1 - x^2}$$

(we take the positive square root since $\arccos(x) \in [0, \pi]$, on which interval sin is non-negative). It follows that

$$(\cos^{-1})'(x) = \frac{-1}{\sqrt{1 - x^2}}.$$

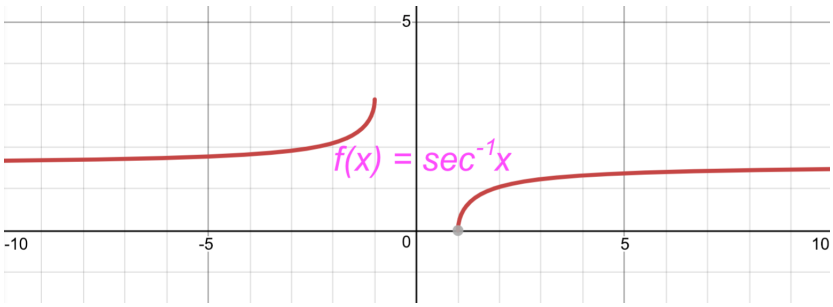
Similarly,

$$(\sin^{-1})'(x) = \frac{1}{\sqrt{1 - x^2}}.$$

Later, when we talk about trigonometric substitutions, it will be useful to know about the inverse of the secant function. It is conventional here to restrict sec to the domain $[0, \pi/2) \cup (\pi/2, \pi]$. As x ranges over $[0, \pi/2)$, $\cos x$ ranges over $[1, \infty)$ (and is increasing), and as x ranges over $(\pi/2, \pi]$, $\cos x$ ranges over $(-\infty, -1]$ (and is also increasing). So

$$\sec^{-1} : (-\infty, -1] \cup [1, \infty) \to [0, \pi/2) \cup (\pi/2, \pi]$$

and has the following graph:

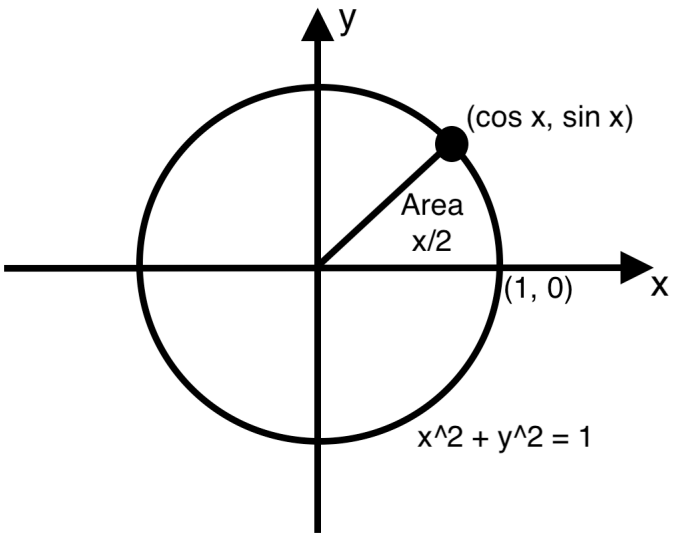


This is probably the first natural instance of a function whose domain is a union of intervals.

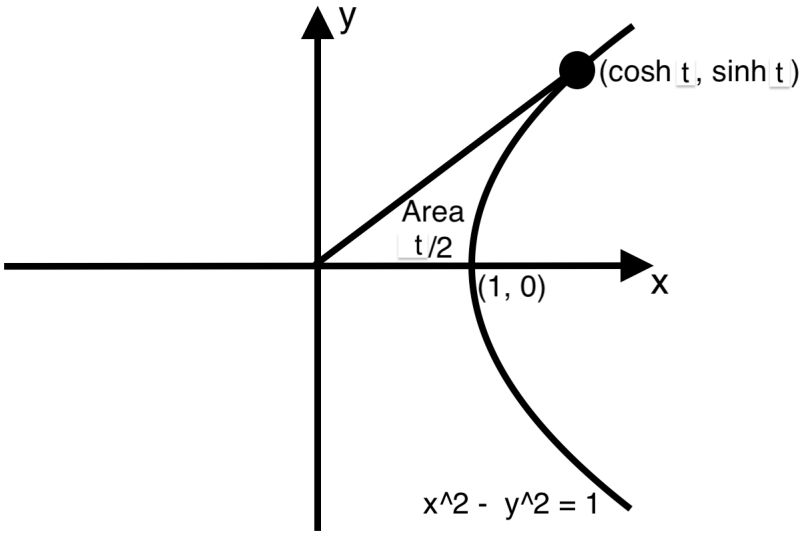
12.5 The hyperbolic trigonometric functions

This section discusses the definitions of, and derivation of the basic properties of, the so-called *hyperbolic* functions.

Just as the trigonometric functions were defined via the following picture:



the hyperbolic functions are defined via the following picture:



That is: let $P = (a, b)$ be a point on the curve $x^2 - y^2 = 1$, with $a \geq 1$ and $b \geq 0$. If the area A bounded by

- the x-axis between $(1, 0)$ and $(0, 0)$,
- the line segment from $(0, 0)$ to P, and
- the curve $x^2 - y^2 = 1$ between P and $(1, 0)$

is $t/2$, then $a = \cosh t$ and $b = \sinh t$. The curve $x^2 - y^2 = 1$ is a hyperbola, hence the adjective "hyperbolic".

It's obvious that this defines the functions cosh, sinh, both on some domain that starts at 0 (and includes 0). It's not entirely obvious just what that domain is — that depends on what is the area of the slanted needle bounded by the line $x = y$ (that the hyperbola is approaching, for large x), the hyperbola, and the the x-axis. If this needle has infinite area, then cosh, sinh have just been defined on $[0, \infty)$, whereas if it has finite area, L say, then cosh, sinh have just been defined on $[0, L)$.

We will discover the answer to the question, "what is the domain on cosh, sinh" in a quite direct way. Unlike with the trigonometric functions, it is possible to come up an explicit expression for $\cosh x$ and $\sinh x$ in terms of functions we have previously defined; specifically, in terms of the exponential function:

$$\cosh t = \frac{e^t + e^{-t}}{2}, \qquad \sinh t = \frac{e^t - e^{-t}}{2}.$$

To see this, first consider the function $f : [0, \infty) \to \mathbb{R}$ given by $f(t) = (e^t + e^{-t})/2$. It is an easy check that this is a is monotone increasing function, with range $[1, \infty)$. This says that for any point $P = (a, b)$ on $x^2 - y^2 = 1$ with $a \geq 1$ and $b \geq 0$, there is a unique $t \in [0, \infty)$ with $a = (e^t + e^{-t})/2$.

Next, it is easy the verify that if $a = (e^t + e^{-t})/2$, then $b = (e^t - e^{-t})/2$. Indeed, given $a^2 - b^2 = 1$ and $a = (e^t + e^{-t})/2$, simple algebra gives that b is one of $\pm(e^t - e^{-t})/2$; and the correct choice to make $b \geq 0$ is easily seen to be $(e^t - e^{-t})/2$. This, together with the observation of the last paragraph, shows that we can parameterize the points of the hyperbola in the first quadrant by $\{(e^t + e^{-t})/2, (e^t - e^{-t})/2 : t \in (0, \infty)\}$. It also says that if we can show that $\cosh t = (e^t + e^{-t})/2$, then we automatically get that $\sinh t = (e^t - e^{-x})/2$.

Let $P = ((e^t + e^{-t})/2, (e^t - e^{-t}/2)$ be a parameterized point on the hyperbola, with $t \geq 0$. The function $A(t)$ that calculates the area of the region A bounded by

- the x-axis between $(1, 0)$ and $(0, 0)$,
- the line segment from $(0, 0)$ to P, and
- the curve $x^2 - y^2 = 1$ between P and $(1, 0)$

is

$$A(t) = \frac{1}{2}\left(\frac{e^t + e^{-t}}{2}\right)\left(\frac{e^t - e^{-t}}{2}\right) - \int_1^{\frac{e^t+e^{-t}}{2}} \sqrt{x^2 - 1}\, dx.$$

This does not look like a very pleasant function to work with! But in fact, it has a very simple re-formulation. Using the fundamental theorem of calculus to differentiate $A(t)$, after a lot of algebra one gets to $A'(t) = 1/2$. Since $A(0) = 0$, it follows that $A(t) = t/2$ for all $t \geq 0$. We conclude that indeed

$$\cosh t = \frac{e^t + e^{-t}}{2}, \qquad \sinh t = \frac{e^t - e^{-t}}{2}, \tag{12}$$

for $t \geq 0$. We extend cosh and sinh to all t by simply taking (12) as the defining relation for all $t \in \mathbb{R}$. This makes $\cosh : \mathbb{R} \to [1, \infty)$ an even function, and $\sinh : \mathbb{R} \to \mathbb{R}$ an odd function.

Here are some basic facts about the hyperbolic functions, including the function tanh defined by $\tanh x = \sinh x / \cosh x$, which can all be verified very easily from (12):

- sinh has domain and range $\mathbb{R}$, and is increasing on its domain. This says that there is a function $\sinh^{-1}$, domain and range $\mathbb{R}$, also increasing, that is the inverse of sinh.
- cosh has domain $\mathbb{R}$ and range $[1, \infty)$. It is not monotone, so not invertible. However, it is increasing on $[0, \infty)$, and on this restricted domain its range is still $[1, \infty)$. This says that there is a function $\cosh^{-1}$, domain $[1, \infty)$ and range $[0, \infty)$, also increasing, that is the inverse of cosh.
- tanh has domain $\mathbb{R}$ and range $(-1, 1)$ (this last is the most non-obvious fact to verify). It is increasing on its domain. This says that there is a function $\tanh^{-1}$, domain $(-1, 1)$ and range $\mathbb{R}$, also increasing, that is the inverse of tanh.

It is worthwhile to look at the graphs of the curves of sinh, cosh, tanh, $\sinh^{-1}$, $\cosh^{-1}$ and $\tanh^{-1}$. The graph of cosh looks like that of a parabola, but it is not (as we will see below, the second derivative of cosh is not zero, but the second derivative of a parabolic function is zero). This graph (cosh) has physical significance — it is the shape formed by a hanging chain, acted on only by the force of gravity.[202]

The hyperbolic functions satisfy many identities that are similar to familiar trigonometric identities. It's easy to verify the following.

- $\cosh^2 - \sinh^2 = 1$.
- $\tanh^2 + 1/\cosh^2 = 1$.
- $\sinh(x + y) = \sinh x \cosh y + \sinh y \cosh x$
- $\cosh(x + y) = \cosh x \cosh y + \sinh y \sinh x$.
- $\sinh' = \cosh$.
- $\cosh' = \sinh$.
- $\tanh' = 1/\cosh^2$.

Just as the inverse trigonometric functions have derivatives that are either rational functions or square roots of rational functions, so too are the derivatives of the inverse hyperbolic functions quite simple. This is one reason why the hyperbolic functions will be important for us: they provide information about the integrals (primitives, antiderivatives) of some very simple functions. Following the approach we took to computing the derivatives of the inverse trigonometric functions, the following are all fairly straightforward to verify:

[202] Google *catenary*. The most famous catenary in the world is upside-down, and is located in St. Louis, Missouri.

- $(\sinh^{-1})'(x) = \frac{1}{\sqrt{x^2+1}}$.
- $(\cosh^{-1})'(x) = \frac{1}{\sqrt{x^2-1}}$, for $x > 1$.
- $(\tanh^{-1})'(x) = \frac{1}{1-x^2}$, for $-1 < x < 1$.

Just as the hyperbolic functions can be explicitly expressed in terms of functions we have defined earlier, so too can the inverse hyperbolic functions. For example, if $y = \cosh x$ (with $x \in [0, \infty)$ and $y \in [1, \infty)$) then $x = \cosh^{-1} y$. So to get an expression for $\cosh^{-1} y$, we can solve

$$y = \frac{e^x + e^{-x}}{2}$$

for x in terms of y. One way to do this is to say that since $y = (e^x + e^{-x})/2$ we have $e^{2x} - 2ye^x + 1 = 0$. This is a quadratic equation in e^x, with solutions

$$y + \sqrt{y^2 - 1} \quad \text{and} y - \sqrt{y^2 - 1}.$$

The expression $y - \sqrt{y^2 - 1}$ is decreasing from 1 on $[1, \infty)$, so taking this solution would give $x\ (= \log y) \leq 0$. On the other hand the expression $y + \sqrt{y^2 - 1}$ is increasing from 1 on $[1, \infty)$, so this is the right expression to take. We conclude that $x = \log(y + \sqrt{y^2 - 1})$, so that

$$\cosh^{-1}(x) = \log(x + \sqrt{x^2 - 1}).$$

We will use this formula in the next section, when we discuss antiderivatives.

12.6 The length of a curve

How long is a piece of string? If it is stretched straight, we can measure it with a ruler, but if it is curved (and we don't have a possibility to straighten it), it is less clear what to do.

Here's a mathematical formulation of this question:

> Let $f : [a, b] \to \mathbb{R}$ be a bounded function. What is the length, $\ell(f)$, of the graph of f from the points $(a, f(a))$ to $(b, f(b))$?[203][204][205]

[203] This doesn't cover all possible ways in which a piece of string could be curved. For example, it doesn't cover any string that intersects itself (and so can't be modeled as the graph of a *function*). We could get over this by introducing *parameterized curves*, but we won't (yet). The treatment we give here already covers many practically important cases.

[204] On the other hand, it also covers many more possibilities than the arrangement of a piece of string. Presumably any sensible model for the position of a piece of string on the plane would use a *continuous* function; but we are just assuming bounded. As we'll soon see, not a whole lot can be said if we don't add the assumption of continuity.

[205] The notation really should mention a and b, to allow us to talk about lengths of different portions of the graph of the same function. But that would lead to a pretty cumbersome expression (maybe something like $\ell_{[a,b]}(f)$), so we don't bother with the extra information. Usually the interval we are working over is going to be clear from the context.

If f is a linear function $x \mapsto mx + d$ then we can use the Pythagorean formula, since the graph of f is the straight line joining $(a, ma + d)$ to $(b, mb + d)$, to get

$$\ell(f) = \sqrt{(b-a)^2 + (mb - ma)^2} = (b-a)\sqrt{1+m^2}.$$

If f is piecewise linear, we could just compute the lengths of each linear segment, and add them all up. But what if f is nowhere linear?

We have seen how the Darboux integral arises as an answer to a question about understanding area. That suggests an approach to understanding length: let $P = (t_0, t_1, \ldots, t_n)$ be a partition of $[a, b]$ (so $a = t_0 < t_1 < \cdots < t_n = b$). The piecewise linear curve that joins $(t_{i-1}, f(t_{i-1}))$ to $(t_i, f(t_i))$ by a straight line, for each $i = 1, \ldots, n$ is a piecewise linear approximation to the graph of f between $(a, f(a))$ to $(b, f(b))$, and it has length

$$\ell(f, P) = \sum_{i=1}^{n} \sqrt{(t_i - t_{i-1})^2 + (f(t_i) - f(t_{i-1}))^2}.$$

As we add points to P, the above expression gets larger (or at least doesn't get smaller). Indeed, if $P = (t_0, t_1, \ldots, t_k, t_{k+1}, \ldots, t_n)$ and $P' = (t_0, t_1, \ldots, t_k, u, t_{k+1}, \ldots, t_n)$, then where $\ell(f, P)$ has a term that measures the straight-line distance between $(t_k, f(t_k))$ and $(t_{k+1}, f(t_{k+1}))$, $\ell(f, P')$ has *two* terms that measures the straight-line distance between $(t_k, f(t_k))$ and $(t_u, f(t_u))$ *plus* the straight-line distance between $(t_u, f(t_u))$ and $(t_{k+1}, f(t_{k+1}))$. Although we haven't actually proven a "triangle inequality" in two dimensions[206], it is both true and intuitively clear that if A, B, C are three distinct points in the plane, then the straight line distance from A to B is never more than the sum of the straight line distances from A to C and from C to B — in other words, it is never shorter to travel between A and B by going via a third point C. So we get that $\ell(f, P') \geq \ell(f, P)$, and more generally we get that if P, Q are two unrelated partitions of $[a, b]$, then the common refinement partition $P \cup Q$ satisfies that $\ell(f, P \cup Q)$ is at least as large as both $\ell(f, P)$ and $\ell(f, Q)$.[207]

[206] and don't need to, for the purposes of defining $\ell(f)$ — it just helps motivate the eventual definition.

[207] As mentioned in an earlier footnote, we don't actually need the triangle inequality in two dimensions to define length. But it is going to be useful, for some examples. So here's a fairly precise statement, and a proof:

Theorem: Let A, B, C be three points in the plane. Denote by $d(X, Y)$ the straight line distance between points X and Y in the plane. Then

$$d(A, B) \leq d(A, C) + d(C, B),$$

and in fact the inequality is strict ($d(A, B) < d(A, C) + d(C, B)$) unless C lies on the line segment joining A and B (in which case there is equality — $d(A, B) = d(A, C) + d(C, B)$).

Proof: If A and B are the same point, the result is obvious. So we from now on assume that A and B are different points. By translating, rotating and scaling we may put A at $(0, 0)$ and B at $(1, 0)$ (so $d(A, B) = 1$). If C lies along the x-axis (i.e., C is at $(x, 0)$ for some $x \in \mathbb{R}$) then the result is easy — if $x > 1$ then $d(A, C) + d(C, B) = x + (x - 1) > 1 = d(A, B)$; if $x < 0$ then $d(A, C) + d(C, B) = -x + (-x + 1) > 1 = d(A, B)$; while if $0 \leq x \leq 1$ (the only case where C lies on the line segment joining A and B) then $d(A, C) + d(C, B) = x + (1 - x) = 1 = d(A, B)$. So what is left to consider is the case where C is at coordinates

It may be that there is no bound to the possible values that $\ell(f, P)$ can take on as P varies over all partitions[208]. If this happens, then we cannot use this idea of piecewise linear approximations to make sense of the length of the graph. But if there *is* an absolute upper bound on the possible values that $\ell(f, P)$ can take on as P varies over all partitions, then there is a *least* such upper bound, and that number seems like a very good candidate for the length of the curve: it is a number that we can approach arbitrarily closely by a piecewise linear approximation, and there is no larger number with that property.

Definition: Suppose that $f : [a, b] \to \mathbb{R}$ is bounded function. The *length* $\ell(f)$ of the graph of f (from $(a, f(a))$ to $(b, f(b))$) is

$$\begin{aligned} \ell(f) &= \sup\{\ell(f, P) : P \text{ a partition of } [a, b]\} \\ &= \sup\left\{\sum_{i=1}^{n} \sqrt{(t_i - t_{i-1})^2 + (f(t_i) - f(t_{i-1}))^2} : P \text{ a partition of } [a, b]\right\}, \end{aligned}$$

if this supremum exists. If the supremum does not exist, then the curve graph does not have a length.

Let's check that this definition gives the answer we would expect, for a linear function, say the function $f : [a, b] \to \mathbb{R}$ from earlier (given by $x \mapsto mx + d$). For any partition

---

(x, y) with $y \neq 0$. In this case $d(A, C) = \sqrt{x^2 + y^2}$ and $d(C, B) = \sqrt{(x-1)^2 + y^2}$, and our goal is to show that

$$\sqrt{x^2 + y^2} + \sqrt{(x-1)^2 + y^2} > 1. \quad (\star)$$

This is equivalent to $\sqrt{(x-1)^2 + y^2} \geq 1 - \sqrt{x^2 + y^2}$. Now we may assume that $\sqrt{x^2 + y^2} \leq 1$ (C is on or inside the circle of radius 1 around A), because otherwise $(\star)$ is trivially true. So $\sqrt{(x-1)^2 + y^2} \geq 1 - \sqrt{x^2 + y^2}$ is equivalent to $(\sqrt{(x-1)^2 + y^2})^2 \geq (1 - \sqrt{x^2 + y^2})^2$ (both sides are non-negative), which (after squaring out, doing some canceling and rearranging, and dividing both sides by -2) is equivalent to $x < \sqrt{x^2 + y^2}$. If $x < 0$ this is trivially true (a negative is less than a positive, which $\sqrt{x^2 + y^2}$ is, since $y \neq 0$ — notice that $x < 0$ puts C outside the circle of radius 1 centered at B, so of course in this case $1 = d(A, B) < d(A, C) + d(C, B)$). If $x > 0$ it is also true, because (again using $y \neq 0$) $\sqrt{x^2 + y^2} > \sqrt{x^2} = x$. So we have established $(\star)$, and finished the proof of the two-dimensional triangle inequality.

[208] It is a good exercise to try to come up with a bounded continuous function on domain $[0, 1]$ for which $\ell(f, P)$ can take on arbitrarily large values as P varies over all partitions of $[0, 1]$. If you think about it in terms of how the function is built, and don't get overly hung up on giving an exact formula for $f(x)$ at each x, it should be quite easy. Hint: there's a piecewise linear example.

$P = (t_0, \ldots, t_n)$ of $[a, b]$ we have

$$\begin{aligned}
\ell(f, P) &= \sum_{i=1}^{n} \sqrt{(t_i - t_{i-1})^2 + ((mt_i + d) - (mt_{i-1} + d))^2} \\
&= \sum_{i=1}^{n} \sqrt{(t_i - t_{i-1})^2 + m^2(t_i - t_{i-1})^2} \\
&= \sum_{i=1}^{n} (t_i - t_{i-1})\sqrt{1 + m^2} \\
&= \sqrt{1 + m^2} \sum_{i=1}^{n} (t_i - t_{i-1}) \\
&= (b - a)\sqrt{1 + m^2} \quad \text{(via a telescoping sum).}
\end{aligned}$$

This last expression is a constant (not depending on P), so

$$\ell(f) = \sup\{(b - a)\sqrt{1 + m^2} : P \text{ a partition}\} = (b - a)\sqrt{1 + m^2},$$

exactly as we calculated previously using the Pythagorean formula.

On the other hand, suppose that $g : [a, b] \to \mathbb{R}$ is a bounded function satisfying $g(a) = ma + d$ and $g(b) = mb + d$ (i.e., whose graph starts and finishes at the same points as the linear function f above), but which is *not* the linear function f. Since g is different from f (but agrees with f at a and b) there must be a $c \in (a, b)$ with $f(c) \neq g(c)$. Consider the partition $P = (a, c, b)$. It is intuitively clear (and can be made precise — triangle inequality in two dimensions, again) that $\ell(g, P)$ is strictly greater than $\ell(f)$ — indeed, $\ell(f)$ is the straight-line distance between $(a, ma + d)$ and $(b, mb + d)$, while $\ell(g, P)$ is the sum of the straight line distance between $(a, ma + d)$ and $(c, g(c))$ and the straight line distance between $(c, g(c))$ and $(b, mb + d)$, where $(c, g(c))$ is a point *not* on the straight line between $(a, ma + d)$ and $(b, mb + d)$.[209] Since $\ell(g, P) > \ell(f)$ it follows immediately that $\ell(g) > \ell(f)$ (if $\ell(g)$ exists).

What we have just established is the famous

> **Dictum**: "The shortest distance between two points is a straight line". That is, among all graphs of bounded functions on domain $[a, b]$ that start at $(a, ma + d)$ and end at $(b, mb + d)$, and that have a well defined length, the unique function with the shortest length is the linear function $x \mapsto mx + d$.

There is a close connection between length and the integral, and between the length of a piecewise linear approximation of a graph coming from a partition P, and Darboux sums, coming also from P, of a certain function. We explore that connection now.

The expression $\sum_{i=1}^{n} \sqrt{(t_i - t_{i-1})^2 + (f(t_i) - f(t_{i-1}))^2}$ (i.e., $\ell(f, P)$) doesn't much look like a Darboux sum, but it can be made to look more like one by pulling out a factor of

[209] If you read the earlier long footnote on the two dimensional triangle inequality, you will know that this intuition can be made precise.

$t_i - t_{i-1}$ (a.k.a. Δ_i) from each summand:

$$\ell(f,P) = \sum_{i=1}^n (t_i - t_{i-1}) \sqrt{1 + \left(\frac{f(t_i) - f(t_{i-1})}{t_i - t_{i-1}} \right)^2} = \sum_{i=1}^n \Delta_i \sqrt{1 + \left(\frac{f(t_i) - f(t_{i-1})}{t_i - t_{i-1}} \right)^2}.$$

Now it looks more like a Darboux sum. It's not quite one yet, because the expression

$$\sqrt{1 + \left(\frac{f(t_i) - f(t_{i-1})}{t_i - t_{i-1}} \right)^2}$$

isn't obviously either the infimum or the supremum of a function. But, the above expression does strongly suggest thinking about the function $\sqrt{1 + (f')^2}$ (if this exists); after all, for $[t_{i-1}, t_i]$ short,

$$\frac{f(t_i) - f(t_{i-1})}{t_i - t_{i-1}} \approx f'(t_i), f'(t_{i-1}).$$

So: let's assume that f is continuous and differentiable on $[a,b]$ and also that f' is bounded on $[a,b]$. Applying the mean value theorem to the interval $[t_{i-1}, t_i]$, we find that there is $c \in (t_{i-1}, t_i)$ with

$$f'(c) = \frac{f(t_i) - f(t_{i-1})}{t_i - t_{i-1}}$$

so

$$\sqrt{1 + (f'(c))^2} = \sqrt{1 + \left(\frac{f(t_i) - f(t_{i-1})}{t_i - t_{i-1}} \right)^2}.$$

Since

$$\inf \left\{ \sqrt{1 + (f'(x))^2} : x \in [t_{i-1}, t_i] \right\} \leq \sqrt{1 + (f'(c))^2} \leq \sup \left\{ \sqrt{1 + (f'(x))^2} : x \in [t_{i-1}, t_i] \right\}$$

we get (summing over i)

$$\begin{aligned} \sum_{i=1}^n \Delta_i \inf \left\{ \sqrt{1 + (f'(x))^2} : x \in [t_{i-1}, t_i] \right\} &\leq \sum_{i=1}^n \Delta_i \sqrt{1 + \left(\frac{f(t_i) - f(t_{i-1})}{t_i - t_{i-1}} \right)^2} \\ &= \ell(f,P) \\ &\leq \sum_{i=1}^n \Delta_i \sup \left\{ \sqrt{1 + (f'(x))^2} : x \in [t_{i-1}, t_i] \right\}. \end{aligned}$$

The first and last expressions above are *exactly* Darboux sums — lower and upper Darboux sums, respectively, for the (bounded) function $\sqrt{1 + (f')^2}$ with respect to the partition P. In other words: for any partition P of $[a,b]$

$$L(\sqrt{1 + (f')^2}, P) \leq \ell(f,P) \leq U(\sqrt{1 + (f')^2}, P). \quad (\star)$$

From $(\star)$, in particular from the second inequality therein, we can read off a very useful fact:

Proposition: If bounded $f : [a, b] \to \mathbb{R}$ is continuous and differentiable, and if also f' is bounded, then $\ell(f)$ exists.

Here is the proof: let P and Q be any two (unrelated) partitions of $[a, b]$. We have

$$\begin{aligned} \ell(f, P) &\leq \ell(f, P \cup Q) \quad \text{(by the earlier triangle inequality observation)} \\ &\leq U(\sqrt{1 + (f')^2}, P \cup Q) \quad \text{(by the second inequality in } (\star)) \\ &\leq U(\sqrt{1 + (f')^2}, Q) \quad \text{(by one of our earliest observations about Darboux sums).} \end{aligned}$$

So em every upper Darboux sum of $\sqrt{1 + (f')^2}$ is at least as large as *every* piecewise linear approximation to the length of the graph of f. It follows that $\sup\{\ell(f, P) : P \text{ a partition}\}$ exists, as claimed (any upper Darboux sum for $\sqrt{1 + (f')^2}$ provides an upper bound for the $\ell(f, P)$'s).

The argument above gives more, since it says that

$$\sup\{\ell(f, P) : P \text{ a partition}\} \leq \inf\{U(\sqrt{1 + (f')^2}, P) : P \text{ a partition}\}. \quad (\star\star)$$

But also, from the first inequality of $(\star)$ (that $L(\sqrt{1 + (f')^2}, P) \leq \ell(f, P)$), and the new-found knowledge that $\sup\{\ell(f, P) : P \text{ a partition}\}$ exists, we immediately get[210]

$$\sup\{L(\sqrt{1 + (f')^2}, P) : P \text{ a partition}\} \leq \sup\{\ell(f, P) : P \text{ a partition}\}. \quad (\star\star\star)$$

Combining $(\star\star)$ and $(\star\star\star)$ we get upper and lower bounds on $\ell(f)$:

$$L(\sqrt{1 + (f')^2}) \leq \ell(f) \leq U(\sqrt{1 + (f')^2}).$$

If we know that $\sqrt{1 + (f')^2}$ is not just bounded on $[a, b]$, but is also integrable, then the right and left sides of the above string of inequalities are *equal*, and we get the following theorem (the main point of this section):

Theorem: Suppose that bounded $f : [a, b] \to \mathbb{R}$ is continuous and differentiable, with bounded derivative, and that also $\sqrt{1 + (f')^2}$ is integrable on $[a, b]$. Then the length of the graph of f (from $(a, f(a))$ to $(b, f(b))$) exists and is

$$\ell(f) = \int_a^b \sqrt{1 + (f')^2}.$$

In particular this formula is valid is f is *continuously differentiable* — continuous and differentiable, with continuous derivative.

[210] This is one of those things that you either see instantly, or don't. If you don't, that's fine, because you can *prove* it! What you have to proof is this: if f, g are two functions on domain A, and $f(x) \leq g(x)$ for all $x \in A$, and $\sup\{g(x) : x \in A\}$ exists, then so also does $\sup\{f(x) : x \in A\}$, and moreover $\sup\{f(x) : x \in A\} \leq \sup\{g(x) : x \in A\}$.

It's hard right now to give many applications of this formula for the length of a curve, because for most functions f (even quite reasonable ones) the function $\sqrt{1+(f')^2}$ is hard to integrate (antidifferentiate). Here is one very relevant example, though. We've defined the number π by saying that $\pi/2$ is the area of half of a unit circle. It is much more usual to see π defined by the relation that 2π is the circumference of a unit circle. It is natural to ask

> do these two definitions actually lead to the *same* π?

Another way to put this question is:

> if we *define* the area of the unit circle $x^2+y^2=1$ to be π, then can we *prove* that circumference of the circle is 2π?

Because we have developed a theory of lengths of graphs, we are now in a position to answer this question. The circumference of the circle $x^2+y^2=1$ is twice the length of the graph of the function $f:[-1,1]\to\mathbb{R}$ given by $f(x)=\sqrt{1-x^2}$. This is a bounded, continuous function, with continuous derivative

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

It follows (after some algebra) that

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

So, by the formula we have just developed, the circumference of the circle $x^2+y^2=1$ is[211]

$$2\int_{-1}^{1}\frac{dx}{\sqrt{1-x^2}}.$$

But we know that the derivative of $\sin^{-1}(x)$ is $1/\sqrt{1-x^2}$. So

$$2\int_{-1}^{1}\frac{dx}{\sqrt{1-x^2}}=2\left[\sin^{-1}(x)\right]_{x=-1}^{1}=2\left(\frac{\pi}{2}-\frac{-\pi}{2}\right)=2\pi.$$

So (thankfully) our definition of π is consistent with (for example) Archimedes' definition.

[211] Not quite. The function $\sqrt{1+(f'(x))^2}$ is not bounded on $[-1,1]$, because of that pesky $1-x^2$ in the denominator. It is, however, bounded on $[-1+\varepsilon_1,1-\varepsilon_2]$ for every $\varepsilon_1,\varepsilon_2>0$. So we can treat the integral we have to compute as an improper integral. Essentially we are saying that the length of a proportion α of the circle approaches π as α approaches one half (from below).

13 Primitives and techniques of integration

This section is concerned with *integration* or *antidifferentiation* — the process of finding a function whose derivative is some given function.

Definition of primitive A function F is a *primitive* of a function f, or an *antiderivative* of f, if $F' = f$. The notation we use to denote this relationship is either

$$F = \int f \quad (\text{or } \int f = F)$$

(when working with some generic function f), or

$$\int f(x) = F(x) \quad (\text{or } \int f(x) = F(x))$$

(when working with a specific, named function, given by a certain rule).

Here is why primitives are useful:

if F is a primitive of f, on an interval that includes $[a, b]$, *and if* f *is integrable on* $[a, b]$[212], then by the fundamental theorem of calculus (part 2) we have

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

The expression $F(b) - F(a)$ comes up so frequently, it has a few different notations:

$$F(b) - F(a) = F|_a^b = F(x)|_{x=a}^b.$$

A number of important comments are in order about the definition of a primitive. We give a few examples, and make the comments along the way.

Example 1 $x^3 + 3x + \pi$ is a primitive of $3x^2 + 3$ (obviously!) and so

$$\int (3x^2 + 3)\, dx = x^3 + 3x + \pi.$$

But equally obviously

$$\int (3x^2 + 3)\, dx = x^3 + 3x + e.$$

The critical comment to be made here is that

[212] This is a subtle but important point. FTOC (part 2) says that if F satisfies $F' = f$ on $[a, b]$ *and if* f *is integrable on* $[a, b]$ then $\int_a^b f = F(b) - F(a)$. If we don't add the assumption that f is integrable, then we cannot draw this conclusion. There are examples of differentiable functions whose derivatives are not integrable. If V is such a function (I use V here because the first example of such a function was discovered by Volterra, and is called *Volterra's function*) then while it is true to say that $V = \int V'$ (V is a primitive of V'), it is *not* true to say that $\int_a^b V' = V(b) - V(a)$, since the left-hand side exists but the right-hand side doesn't. Remember that the FTOC part 2 says that if there's a function g with $g' = f$ *and* f is integrable then $\int_a^b f = g(b) - g(a)$.

the "=" in $\int f = F$ is **not** a true equality!

If it was, from $\int (3x^2+3)\,dx = x^3+3x+\pi$ and $\int (3x^2+3)\,dx = x^3+3x+e$ we would conclude the patently absurd

$$x^3+3x+\pi = x^3+3x+e.$$

It is very important to remember that "$\int f = F$" is actually shorthand for "$F' = f$", and *not* an assertion that two functions are identical. This is a clear abuse of the "=" sign, but hopefully one you can live with. We'll see some odd paradoxes that can arise when we forget this.[213]

Example 2 $-2/(1+\tan(x/2))$ is a primitive of $1/(1+\sin x)$. This is obvious, no? Probably not; but once it has been asserted, it can be easily checked, by differentiating $F(x) = -2/(1+\tan(x/2))$. After *a lot* of algebra, the derivative can be massaged into the form $1/(1+\sin x)$.

The comment to be made here is that

> unlike finding derivatives, which is a mechanical process, finding antiderivatives is often hard, always requires ingenuity and usually (see a later example) is practically impossible.

See https://xkcd.com/2117/ for a could summary of the situation!

Example 3 If F is a primitive of f, so is $F+c$ for any constant c.

The comment to be made here is that

> A function with an antiderivative, has infinitely many antiderivatives.

It is tempting to at this point try to prove a theorem, along the lines of: if F is a primitive of f, then *all* primitives of f are of the form $F+c$ for some constant c. We won't try to prove this, because it is false. (See later examples).

Another comment is in order here:

> There is no great value in writing "$\int f = F + C$".

The "$+C$" adds nothing — since "$\int f = F$" is shorthand for "$F' = f$", adding the "$+C$" (to convey "$(F+C)' = f$") is just saying "by the way, the derivative of a constant function is 0".

In the next example, we'll see that not only is there no value in writing "$+C$", it can sometimes be misleading.

[213] Spivak gets over this issue by defining $\int f$ to be the *set of all primitives* of f; so $F' = f$ translates to $F \in \int f$.

Example 4 $\int dx/x \neq \log x$. This seems strange. Of course $(\log x)' = 1/x$. The (somewhat subtle) issue here is that in an equation like "$F' = f$", asserting that two functions are identical, if there is no specific statement of domains, then our convention is to assume that both F' and f are each defined on their natural domain — the largest subset of the the reals for which the rule defining the function makes sense. In the equation "$(\log x)' = 1/x$", the domain of $\log x$ is $(0, \infty)$, and since log is differentiable at all points in its domain, the domain of $(\log x)'$ is $(0, \infty)$. On the other hand, the domain of $1/x$ is $\mathbb{R} \setminus \{0\}$. So, without qualification on the domains on which the two sides are being considered, it is incorrect to say $(\log x)' = 1/x$.

It is, on the other hand, perfectly correct to say

> on $(0, \infty)$, $(\log x)' = 1/x$, so $\int dx/x = \log x$.

What about on $(-\infty, 0)$? This is the domain of $\log(-x)$, and (by the chain rule) the derivative of $\log(-x)$ is $1/x$. So it is correct to say

> on $(-\infty, 0)$, $(\log(-x))' = 1/x$, so $\int dx/x = \log(-x)$.

This leads to some examples of primitives of $1/x$:

- the function that maps x to $\log x$ if $x > 0$ and $\log(-x)$ if $x < 0$; this can be more compactly expressed as $x \mapsto \log|x|$;
- for any real constant C, the function that maps x to $\log|x| + C$;
- the function that maps x to $3 + \log x$ if $x > 0$ and $12\pi^2 + \log(-x)$ if $x < 0$.

The comment that relates to this example is that

> if the domain of f is not an interval, then
>
> - one has to be careful about $\int f$, and
> - it's not true that any two antiderivatives of f differ by a constant.

We mentioned earlier that it it is not true that if F is a primitive of f, then all primitives of f are of the form $F + c$ for some constant c. In light of the current example, there is a natural modification to this statement, that is indeed true, and can easily be shown to be true:

> if f is continuous on its domain, and that domain is an interval, and if F is a primitive of f, then all primitives of f are of the form $F + c$ for some constant c. If the domain of f is a union of intervals, and if F and G are two primitives of f, then $F - G$ is constant on each of the intervals.

Example 5 It seems hard to find an antiderivative of e^{-x^2} (for more on this, see the discussion of elementary functions below). However, this function has a very simple antiderivative:

$$\int e^{-x^2}\ dx = \int_c^x e^{-t^2}\ dt$$

where c is any constant (since, by the fundamental theorem of calculus, the derivative of $\int_c^x e^{-t^2}\ dt$ with respect to x is e^{-x^2}).

The comment to be made on this example is:

every continuous function $x \mapsto f(x)$ has a primitive, namely $\int_c^x f(t)\ dt$.

Of course, this is not a particularly *useful* primitive: if we try to use it to calculate a definite integral like $\int_a^b f(x)\ dx$, we get

$$\int_a^b f(x)\ dx = \int_c^b f(x)\ dx - \int_c^a f(t)\ dt,$$

which really doesn't help.

The last example above shows that while finding primitives is easy, what we really want to know about is finding simple, compact expressions for primitives. Computer algebra systems can do this very well: for example, entering

"antiderivative of $1/(1+\sin x)$"

into Wolfram Alpha yields the answer

$$\text{“}\frac{2\sin(x/2)}{\sin(x/2)+\cos(x/2)} + \text{constant}''.$$

(This is not quite the same as $-2/(1+\tan(x/2))$ that we mentioned earlier; but a little algebra shows that the two expressions $-2/(1+\tan(x/2))$ and $2\sin(x/2)/(\sin(x/2)+\cos(x/2))$ differ from each other by a universal constant).

Given that computers are very good at finding compact expressions for primitives, it's natural to ask why it's useful to spend time, as we will do, developing techniques to find primitives by hand. Here are three reasons why being able to **find** primitives is useful:

1. knowing something of the theory of finding compact expressions for primitives, allows one to troubleshoot when things go wrong using a computer algebra system (as it inevitably will);

2. underlying some of the techniques we describe (in particular integration by parts and integration by partial fractions) are valuable theorems, that are useful to know; and

3. questions about finding primitives they come up on exams, like the GRE.

So, out goal for a while will be to develop techniques to find compact expressions for primitives. "Compact" here means that we are looking for *elementary* functions as primitives:

- rational functions;
- exponential, log, trigonometric (and so hyperbolic) functions and their inverses;
- algebraic functions: functions g satisfying a polynomial equation with rational functions as coefficients (so, for example, functions that extract roots); and
- any function obtained from the previous functions by finitely many additions, subtractions, multiplications, divisions, and compositions.

Essentially, elementary functions are those that can be described in finite time using any combination of the functions $1, x, \sin, \cos, \tan, \arcsin, \arccos, \arctan, \exp$ and $\log$. It is a theorem (though a very hard one) that the function $x \mapsto e^{-x^2}$ does not have an elementary primitive; nor does $\sin x^2$, nor $\sqrt{1+x^3}$. In fact, "most" elementary functions do not have elementary primitives. But still, it will prove very worthwhile to think about those functions that *do* have elementary primitives; and that will be the topic of the next few sections.

13.1 Techniques of integration

There are five basic techniques of integration:

- Know lots of integrals!
- Linearity
- Integration by parts
- Integration by substitution
- Integration by partial fractions

The first two can be discussed quickly. First, know lots of integrals! Every differentiation, when turned on its head, leads to an integration formula, and the more of these you can recognize quickly, the better you will be at integration. Here are some of the integrals we have seen so far:

- $\int x^n dx = x^{n+1}/(n+1)$ for $n \in \mathbb{N}$, as long as $n \neq -1$.
- $\int dx/x = \log|x|$ (as long as $x \neq 0$).
- $\int x^a dx = x^{a+1}/(a+1)$ for real $a \neq -1$ (as long as $x \in (0, \infty)$).
- $\int \sin x dx = -\cos x$.

- $\int \cos x dx = \sin x$.
- $\int \sec^2 x dx = \tan x$.
- $\int e^x dx = e^x$.
- $\int \frac{dx}{\sqrt{1-x^2}} = \sin^{-1}(x)$ (as long as $-1 < x < 1$).
- $\int \frac{dx}{1+x^2} = \tan^{-1}(x)$.
- $\int \frac{dx}{\sqrt{x^2+1}} = \sinh^{-1}(x) = \log\left(x + \sqrt{x^2+1}\right)$.
- $\int \frac{dx}{\sqrt{x^2-1}} = \cosh^{-1}(x) = \log\left(x + \sqrt{x^2-1}\right)$ (as long as $x \geq 1$). But for this last example, it is easy to see that if $x < -1$ then, since $x + \sqrt{x^2-1} < 0$, we get $\log - \left(x + \sqrt{x^2-1}\right)$ as an antiderivative of $1/(\sqrt{x^2-1})$; so in fact

$$\int \frac{dx}{\sqrt{x^2-1}} = \log|\left(x + \sqrt{x^2-1}\right)| \quad \text{(as long as } |x| > 1\text{)}.$$

Second, linearity: if $F = \int f$ and $G = \int g$ then it is an easy check that $aF + bG = \int(af + bg)$.

The other three techniques, integration by parts, substitution and partial fractions, require significantly more discussion.

13.2 Integration by parts

Suppose f', g' are both continuous (so all integrals below exist). We have

$$(fg)' = f'g + fg' \quad \text{or} \quad fg' = (fg)' - f'g.$$

An antiderivative of $(fg)'$ is fg. Suppose $A = \int f'g$ is an antiderivative of $f'g$. Then

$$(fg - A)' = (fg)' - f'g = fg'.$$

In other words, $fg - A$ is an antiderivative of fg'. The traditional way to write this is

$$\int fg' = fg - \int f'g$$

or

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

This identity is referred to as *integration by parts*, and allows the calculation of one integral ($\int fg'$) to be reduced to the calculation of another integral ($\int f'g$).

Integration by parts has a definite integral form: since

$$(fg)' = f'g + fg',$$

from the fundamental theorem of calculus (part 2) we get that, as long as $[a, b]$ is fully contained in the domains of both f and g,

$$\int_a^b (f'g + fg') = (fg)_a^b$$

or

$$\int_a^b f(x)g'(x)dx = f(x)g(x)|_{x=a}^b - \int_a^b f'(x)g(x)dx.$$

The key to applying integration by parts is to identify that the function to be integrated can be decomposed into the product of two functions, one of which is easy to differentiate (this will play the role of f), and the other of which has an obvious antiderivative (this will play the role of g').

As an example, consider $\int x \log x \; dx$. Here we take $f(x) = \log x$ (so $f'(x) = 1/x$) and $g'(x) = x$ (so one valid choice for g is $g(x) = x^2/2$). We have

$$\int x \log x \; dx = \frac{x^2 \log x}{2} - \int \frac{x}{2} \; dx = \frac{x^2 \log x}{2} - \frac{x^2}{4},$$

a result which can easily be checked by differentiating.

More generally, consider $\int x^a \log x \; dx$ with $a \neq -1$. Here we again take $f(x) = \log x$ (so $f'(x) = 1/x$) and $g'(x) = x^a$, so one valid choice for g is $g(x) = x^{a+1}/(a+1)$). We have

$$\int x \log x \; dx = \frac{x^{a+1} \log x}{a+1} - \int \frac{x^a}{a+1} \; dx = \frac{x^{a+1} \log x}{a+1} - \frac{x^{a+1}}{(a+1)^2}.$$

What about $a = -1$? Again taking $f(x) = \log x$, $f'(x) = 1/x$, $g'(x) = 1/x$, $g(x) = \log x$ (Note: we don't need $\log |x|$ here, since the domain of $(\log x)/x$ is $(0, \infty)$), we get

$$\int \frac{\log x}{x} \; dx = \log^2 x - \int \frac{\log x}{x} \; dx.$$

It appears that we have gone in a circle! But no: we have an (easy) equation which we can solve for $\int (\log x)/x \; dx$, that yields

$$\int \frac{\log x}{x} \; dx = \frac{\log^2 x}{2},$$

again a result which can easily be checked by differentiating.

We need to be a little careful in justifying the above, because of the previous observation that the "=" in $F = \int f$ has to be treated with care. Formally what we are doing is saying: "if A is an antiderivative of $(\log x)/x$, then from integration by parts, $\log^2 x - A$ is also an antiderivative of $\int \frac{\log x}{x} \; dx$. But now, since $(\log x)/x$ is a continuous function defined on an interval, that says that A and $\log^2 x - A$ differ by a constant, or in other words, $A = (\log^2 x)/2 + C$ for some constant C".

Other manipulations that we do with integral equalities can be just as easily be justified formally; we won't do so any more, unless there is an extra subtlety that needs to be pointed out.

Hidden inside the last example was the special case $a = 0$, where we considered $\int \log x \; dx$ — which isn't obviously the product of two functions — and applied integration by parts by "introducing" the function $g'(x) = 1$ into the picture. More generally, for any f with f' continuous, we have

$$\int f = \int f \cdot 1 = xf - \int xf'.$$

For example

$$\int \tan^{-1} x \; dx = x \tan^{-1} x - \int \frac{x dx}{1+x^2} = x \tan^{-1} x - \frac{1}{2} \log(1+x^2).$$

Integration by parts sometimes reduces a more complicated integral to a less complicated one, that still needs some non-trivial works to solve; sometimes even another iteration of integration by parts.

Example: For $n \geq 0$, $n \in \mathbb{N}$, set $I_n = \int x^n e^x \; dx$. We have $I_0 = e^x$ rather easily. For $n > 0$, we use integration by parts with $f(x) = x^n$, $g'(x) = e^x$ to get

$$I_n = x^n e^x - n \int x^{n-1} e^x \; dx = x^n e^x - n I_{n-1}.$$

This is a *reduction formula* that allows us to calculate I_n recursively:

- $I_0 = e^x$
- $I_1 = xe^x - 1 \cdot I_0 = xe^x - e^x = e^x(x-1)$
- $I_2 = x^2 e^x - 2e^x(x-1) = e^x(x^2 - 2x + 2)$
- $I_3 = x^3 e^x - 3e^x(x^2 - 2x + 2) = e^x(x^3 - 3x^2 + 6x - 6)$
- $I_4 = x^4 e^x - 4e^x(x^3 - 3x^2 + 6x - 6) = e^x(x^4 - 4x^3 + 12x^2 - 24x + 24)$

and in general (this is an easy induction) $I_n = e^x P_n(x)$ where $P_n(x)$ is a polynomial of degree n defined recursively by $P_0(x) = 1$ and $P_n = x^n - nP_{n-1}(x)$.

We'll see plenty more reduction formulae.

Notice that in this example, we had a choice: both x^n and e^x are easy both to integrate and differentiate. There's no golden rule for what to do in this case. Sometimes one choice works and the other doesn't, sometimes both do, and sometimes neither work. With lots of practice you should start to develop an intuition; but for the moment, a good rule-of-thumb is:

> if one of the functions involved is a polynomial, try to make the choice that reduces the degree of the polynomial.

This doesn't always work, but often does.

Integration by parts is sometimes symbolically written

$$\int udv = uv - \int vdu$$

Here "u" can be thought of as the part of the function that's easy to differentiate (so, f; its derivative appears as "du"), while "dv" can be thought of as the part of the function that has an easy antiderivative (so g'; its antiderivative appears as "v").

For an example in this language, consider $A_n = \int \frac{dx}{(x^2+1)^n}\, dx$, $n \geq 0$. We have $A_0 = 1$ and $A_1 = \arctan x$. For general $n \geq 2$, set $u = 1/(1+x^2)^n$, $dv = dx$, so $v = x$ and $du = -2nxdx/(1+x^2)^{n+1}$. We get

$$\begin{aligned} A_n &= \frac{x}{(1+x^2)^n} + 2n \int \frac{x^2}{(1+x^2)^{n+1}}\, dx \\ &= \frac{x}{(1+x^2)^n} + 2n \int \frac{(1+x^2)-1}{(1+x^2)^{n+1}}\, dx \\ &= \frac{x}{(1+x^2)^n} + 2nA_n - 2nA_{n+1}. \end{aligned}$$

So

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

(valid for $n \geq 1$). For example, at $n = 1$ we get

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

For the rest of this section we'll use an integration by parts reduction formula to derive Wallis' formula for π, and see the connection between Wallis' formula and the binomial coefficients.

We begin by defining, for integers $n \geq 0$, $S_n := \int_0^{\pi/2} \sin^n x dx$. We have

$$S_0 = \frac{\pi}{2}, \quad S_1 = \int_0^{\pi/2} \sin x dx = 1,$$

and for $n \geq 2$ we get from integration by parts (taking $u = \sin^{n-1} x$ and $dv = \sin x dx$, so that $du = (n-1)\sin^{n-2} x \cos x dx$ and $v = -\cos x$) that

$$\begin{aligned} S_n &= (\sin^{n-1} x)(-\cos x)|_{x=0}^{\pi/2} - \int_0^{\pi/2} -(n-1)\cos x \sin^{n-2} x \cos x dx \\ &= (n-1)\int_0^{\pi/2} \cos^2 x \sin^{n-2} x dx \\ &= (n-1)\int_0^{\pi/2} (1-\sin^2 x)\sin^{n-2} x dx \\ &= (n-1)S_{n-2} - (n-1)S_n, \end{aligned}$$

which leads to the recurrence relation

$$S_n = \frac{n-1}{n} S_{n-2} \quad \text{for } n \geq 2.$$

Iterating the recurrence relation until the initial conditions are reached, we get that

$$S_{2n} = \left(\frac{2n-1}{2n}\right)\left(\frac{2n-3}{2n-2}\right)\cdots\left(\frac{3}{4}\right)\left(\frac{1}{2}\right)\frac{\pi}{2}$$

and

$$S_{2n+1} = \left(\frac{2n}{2n+1}\right)\left(\frac{2n-2}{2n-1}\right)\cdots\left(\frac{4}{5}\right)\left(\frac{2}{3}\right)1.$$

Taking the ratio of these two identities and rearranging yields

$$\frac{\pi}{2} = \left(\frac{2}{1}\right)\left(\frac{2}{3}\right)\left(\frac{4}{3}\right)\left(\frac{4}{5}\right)\cdots\left(\frac{2n}{2n-1}\right)\left(\frac{2n}{2n+1}\right)\frac{S_{2n}}{S_{2n+1}}.$$

Now since $0 \leq \sin x \leq 1$ on $[0, \pi/2]$ we have also

$$0 \leq \sin^{2n+1} x \leq \sin^{2n} x \leq \sin^{2n-1} x,$$

and so, integrating and using the recurrence relation, we get

$$0 \leq S_{2n+1} \leq S_{2n} \leq S_{2n-1} = \frac{2n+1}{2n} S_{2n+1}$$

and so

$$1 \leq \frac{S_{2n}}{S_{2n+1}} \leq 1 + \frac{1}{2n}.$$

This says that by choosing n large enough, the ratio S_{2n}/S_{2n+1} can be made arbitrarily close to 1, and so the product

$$\left(\frac{2}{1}\right)\left(\frac{2}{3}\right)\left(\frac{4}{3}\right)\left(\frac{4}{5}\right)\cdots\left(\frac{2n}{2n-1}\right)\left(\frac{2n}{2n+1}\right)$$

can be made arbitrarily close to $\pi/2$ by choosing n large enough. This fact as usually expressed by saying that $\pi/2$ can be described by an "infinite product":

$$\frac{\pi}{2} = \left(\frac{2}{1}\right)\left(\frac{2}{3}\right)\left(\frac{4}{3}\right)\left(\frac{4}{5}\right)\left(\frac{6}{5}\right)\left(\frac{6}{7}\right)\cdots.$$

This infinite product was probably first written down by John Wallis in 1655. Wallis' other claim to fame is that he was probably the first mathematician to use the symbol "∞" for infinity.

Note that Wallis' formula is not a particularly good way to actually estimate π; because we have $1 \leq S_{2n}/S_{2n+1} \leq 1 + 1/2n$, it turns out that to get an estimate of π correct to k decimal places, we need to take $n \approx 10^k$. This is similar to the rate of convergence of the approximation based on $\arctan 1$.

Wallis' formula can be used to estimate the binomial coefficient $\binom{2n}{n}$. Indeed,

$$\begin{aligned}\binom{2n}{n} &= \frac{(2n)(2n-1)(2n-2)\cdots(3)(2)(1)}{(n)(n-1)\cdots(2)(1)(n)(n-1)\cdots(2)(1)} \\ &= 2^n \frac{(2n-1)(2n-3)\cdots(3)(1)}{(n)(n-1)\cdots(2)(1)} \\ &= 2^{2n} \frac{(2n-1)(2n-3)\cdots(3)(1)}{(2n)(2n-2)\cdots(4)(2)} \\ &= \frac{2^{2n}}{\sqrt{2n+1}} \sqrt{\frac{(2n+1)(2n-1)(2n-1)(2n-3)(2n-3)\cdots(3)(3)(1)}{(2n)(2n)(2n-2)(2n-2)\cdots(4)(4)(2)(2)}}\end{aligned}$$

and so

$$\frac{\sqrt{n}\binom{2n}{n}}{2^{2n}} = \sqrt{\frac{n}{2n+1}} \sqrt{\frac{(2n+1)(2n-1)(2n-1)(2n-3)(2n-3)\cdots(3)(3)(1)}{(2n)(2n)(2n-2)(2n-2)\cdots(4)(4)(2)(2)}}$$

For large enough n, $\sqrt{n/(2n+1)}$ can be made arbitrarily close to $1/\sqrt{2}$, and the other term on the right-hand side above can (by Wallis' formula) be made arbitrarily close to $\sqrt{2/\pi}$, so the whole right-hand side can be made arbitrarily close to $\sqrt{1/\pi}$. In other words,

$$\lim_{n\to\infty} \frac{\sqrt{n}\binom{2n}{n}}{2^{2n}} = \frac{1}{\sqrt{\pi}}.$$

(Note that this is not a very helpful limit: at $n = 10,000$ the expression $\sqrt{n}\binom{2n}{n}/2^{2n}$ evaluates to around 0.564183, whereas $1/\sqrt{\pi} \approx 0.564189$).

This limit is usually written[214]

$$\frac{\binom{2n}{n}}{2^{2n}} \sim \frac{1}{\sqrt{n\pi}} \quad \text{as } n \to \infty;$$

here I am introducing the symbol "$\sim$", read as "asymptotic to", which is defined as follows:

$$f(n) \sim g(n) \quad \text{as } n \to \infty$$

if $\lim_{n\to\infty} f(n)/g(n) = 1$. The sense is that f and g grow at essentially the same rate as n grows. Note that this does *not* say that f and g get closer to one another absolutely as n grows; for example $n^2 \sim n^2 + n$ as $n \to \infty$, but the difference between the two sides goes to infinity too. It's the *relative* (or proportional) difference that gets smaller.

This estimate for $\binom{2n}{n}$ has a connection to probability. If a fair coin is tossed $2n$ times, then the probability that it comes up heads exactly k times is $\binom{2n}{k}/2^{2n}$. This quantity is at its largest when $n = k$ (some easy algebra), at which point it takes value very close to $1/\sqrt{n\pi}$ (as we have just discovered).

[214] Another way to write this is: $\pi = \lim_{n\to\infty} \frac{16^n}{n\binom{2n}{n}^2}$.

Some easy algebra also suggests that we should expect $\binom{2n}{k}/2^{2n}$ to be quite close to $\binom{2n}{n}/2^{2n}$ for k fairly close to n. If this is the case, then we might expect that the probability of getting some number of heads between $n - n_0$ and $n + n_0$ to be somewhat close to $2n_0$ times the probability of getting n heads, or somewhat close to $2n_0/\sqrt{n\pi}$. If this is true, then by the time n_0 gets up to somewhere around $\sqrt{n}$, the probability of getting some number of heads between $n - n_0$ and $n + n_0$ should be somewhat close to 1.

This intuition can be made precise, in a result called the *central limit theorem*, one of the most important results in probability. One very specific corollary of the central limit theorem is that if a coin is tossed $2n$ times, then for any constant C the probability of getting between $n - C\sqrt{n}$ and $n + C\sqrt{n}$ heads is at least $1 - e^{-C^2/3}$. For example, with $n = 1,000,000$ and $C = 5$, on tossing a coin $2,000,000$ times, the probability of getting between $995,000$ and $1,005,000$ heads is at least $1 - e^{-25/3} \approx .99976$.

13.3 Integration by substitution

Just as the product rule led to integration by parts, the chain rule also leads to an integration principle, integration by substitution.

Integration by substitution, easy case If f, g' are both continuous (so all integrals in question exist), and if F is a primitive of f, then (since $(F \circ g)'(x) = F'(g(x))g'(x) = f(g(x))g'(x)$) we have

$$\int f(g(x))g'(x)\,dx = (F \circ g)(x) = F(g(x)).$$

The key to applying this form of integration by substitution is to recognize that the integrand $f(g(x))g'(x)$ can be written as the product of two things — one, $f(g(x))$, is a function f of some basic building block g, and the other, g', is the derivative of the basic building block. If f has a known antiderivative F, then the integral can be expressed in terms of this.

Consider, for example, $\int \tan x\,dx$. Re-expressing as

$$\int \frac{\sin x}{\cos x}\,dx = -\int \frac{-\sin x}{\cos x},$$

there are obvious choices for f ($f(x) = 1/x$, so $F(x) = \log|x|$), and $g(x) = \cos x$, with $g'(x) = -\sin x$. This leads to

$$\int \frac{\sin x}{\cos x}\,dx = -\int \frac{-\sin x}{\cos x} = -\log|\cos x| = \log|\sec x|.$$

Notice that we multiplied the integrand by -1 to "massage" it into the correct form: initially, the integrand was almost of the form $f(g(x))g'(x)$, but not quite (we will return to this issue later).

As another example, consider

$$\int \frac{x\,dx}{(1+x^2)^n}$$

for $n = 1, 2, \ldots$. Recognizing that the integrand is mostly a function of $1 + x^2$, and that the rest of the integral is almost the derivative of $1 + x^2$, we take

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

Noting that $g'(x) = 2x$ and

$$F(x) = \int f(x)\ dx = \begin{cases} \log|x| & \text{if } n = 1 \\ \frac{-1}{(n+1)x^{n+1}} & \text{if } n \geq 2, \end{cases}$$

we get

$$\int \frac{x dx}{(1+x^2)^n} = \frac{1}{2}\int \frac{2x dx}{(1+x^2)^n} = \begin{cases} \frac{\log(1+x^2)}{2} & \text{if } n = 1 \\ \frac{-1}{(n+1)(1+x^2)^{n+1}} & \text{if } n \geq 2. \end{cases}$$

There is a shorthand for this process. If we make a *change of variables* — a *substitution* — $u = g(x)$, and (formally) write $du = g'(x)\ dx$, then, re-expressing $\int f(g(x))g'(x)\ dx$ entirely in terms of the new variable u, the integral becomes

$$\int f(u)\ du,$$

which is $F(u)$, or, going back to expressing in terms of variable x, $F(g(x))$. The message here is: if we simply make the substitution $u = g(x)$, re-express the integral in terms of u solve (as a function of u), then go back to expressing in terms of variable x, we get the correct answer — this, even though the expression "$du = g'(x)\ dx$" doesn't (yet) have any official meaning.

This shorthand allows integration by substitution to be done quite quickly, without having to explicitly identify f, g, et cetera.

As an example: via the substitution $u = \log x$ (so $du = dx/x$) we have

$$\int \frac{(\log x)^2}{x} dx = \int u^2 du = \frac{u^3}{3} = \frac{(\log x)^3}{x}.$$

In all examples so far we were extremely fortunate that either (as in the last example) one part of the integrand was easily recognizable as the derivative of another, or (as in the other examples), that became the case after a simple manipulation. We now present a more general substitution method that is far more versatile, in that in can (in principle) be used for *any* integrand. The idea is that we for (essentially) any function f that is given, and (essentially) any function g (of our choosing), we can (usually) express $f(x)$ in the form $(H \circ g)(x)g'(x)$ for a suitably constructed function H, and then use "easy" integration by substitution to complete the integration.

Indeed,

$$\begin{aligned} f(x) &= \frac{f(x)}{g'(x)} g'(x) \\ &= \frac{f(g^{-1}(g(x)))}{g'(g^{-1}(g(x)))} g'(x) \\ &= \left(\frac{f \circ g^{-1}}{g' \circ g^{-1}}\right)(g(x))g'(x), \end{aligned}$$

all steps valid as long as:

- g is differentiable (so continuous)
- g' is never 0, and
- g^{-1} exists (so, since g continuous, g must be monotone, assuming that we are working on some interval).

From this we get:

Integration by substitution, general case If f and g' are both continuous (so all integrals in question are certain to exist), and if g is invertible, with non-zero derivative, and if H is a primitive of

$$\frac{f \circ g^{-1}}{g' \circ g^{-1}},$$

then (since

$$\begin{aligned}(H \circ g)'(x) &= H'(g(x))g'(x) \\ &= \frac{f(g^{-1}(g(x)))}{g'(g^{-1}(g(x)))}g'(x) \\ &= \frac{f(x)}{g'(x)}g'(x) \\ &= f(x)\end{aligned}$$

we have

$$\int f(x)\ dx = (H \circ g)(x) = H(g(x)).$$

As a simple example, suppose that we are considering $\int x/(1+x^2)\ dx$, and we do not recognize that after a simple manipulation the integrand becomes of the form $f(g(x))g'(x)$. We have at our disposal the general method of integration by substitution, which allows us to re-express the integrand, by "substituting out" any expression that we may choose. A general rule-of-thumb to keep in mind for integration by substitution is:

> identify any awkward/prominent/annoying part of the integrand, and try to substitute that out.

Here, the awkward/prominent/annoying part of the integrand is the $1 + x^2$, so we set $g(x) = 1 + x^2$. We have to be a little careful now, since g is *not* an invertible function either on its domain, or on the domain of the integrand. It becomes invertible if we restrict it to either non-negative reals, or non-positive reals, so let us do that.

First, consider the problem on non-negative reals. We have $g(x) = 1 + x^2$, $g'(x) = 2x$[215] and $g^{-1}(x) = \sqrt{x-1}$[216]. So

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

which has primitive $(\log x)/2$ (note x non-negative here, so we don't need the absolute value). So we may take $H(x) = (\log x)/2$, and get that

$$H(g(x)) = \frac{\log(1+x^2)}{2}$$

is an antiderivative of the original function f, at least when we restrict to the domain of positive reals. A similar calculation gives that an antiderivative of f is $(1/2)\log(1+x^2)$ on negative reals (now $g^{-1} = -\sqrt{x-1}$, but the negative sign disappears in the calculation of $(f \circ g^{-1})/(g' \circ g^{-1})$, since it appears in numerator and denominator).

As with the easy substitution method, there is a shorthand way to proceed. Start with the substitution $u = g(x)$, and then re-express everything in the integrand in terms of u:

- $u = g(x)$ so $x = g^{-1}(u)$ (requiring g to be invertible)
- $du = g'(x)dx$ so $dx = du/g'(x) = du/g'(g^{-1}(x))$ (requiring g' not to ever be 0)
- $f(x) = f(g^{-1}(u))$.

The integral becomes

$$\int \frac{(f \circ g^{-1})(u)}{(g' \circ g^{-1})(u)}\, du,$$

so if H is a primitive of $(f \circ g^{-1})/(g' \circ g^{-1})$, then the integral is $H(u)$, or, in terms of x, $H(g(x))$.

As a simple example, consider $\int f(ax+b)\, dx$, where a, b are constants and where F is a known primitive of f. Via the substitution $u = ax + b$ (so $du = adx$, $dx = du/a$), we get

$$\begin{aligned}
\int f(ax+b)\, dx &= \int \frac{f(u)}{a} du \\
&= \int \frac{1}{a} \int f(u)\, du \\
&= \frac{1}{a} F(u) \\
&= \frac{F(ax+b)}{a}.
\end{aligned}$$

[215] There will clearly be a problem at 0, since g' is 0 there. So let's restrict the domain of g a little further, to *positive* reals.

[216] Note that restricting g to positive reals, it has range $(1, \infty)$, so that is the domain of g^{-1}, while the range of g^{-1} is positive reals. That is why we take the positive square root when computing g^{-1}.

Notice that we didn't have to think about the involved expression $(f \circ g^{-1})/(g' \circ g^{-1})$ here; simply by formally re-expressing the whole integrand in terms of the new variable u, we inevitably reach $(f \circ g^{-1})(u))/((g' \circ g^{-1})(u)\ du$.

The process of integration by substitution in the general case can be thought of as being similar to the process of integration by parts, in that it can be used to replace one integration problem with another, hopefully simpler, one. It can be done entirely mechanically. As stated earlier in more general terms, a rough guiding principle should be:

> identify a "prominent" part of the integrand, call it $g(x)$, and substitute for it by setting $u = g(x)$ and then completely re-expressing the integrand in terms of u.

This leads to a new integral that, since it makes no reference to g, is hopefully simpler to evaluate that the original. As the next few examples show, this new, simpler integral may require the application of some other integration techniques, (maybe another application of integration by substitution) to crack; or, as we will see in at least one example, the new integral may be just as hopeless as the old.

Example 1 $\int 1/(1+\sqrt{1+x})\ dx$. Here an obvious substitution is $u = \sqrt{1+x}$, which gives $du = dx/(2\sqrt{1+x}) = dx/2u$, so $dx = 2udu$. We get

$$\begin{aligned}\int \frac{dx}{1+\sqrt{1+x}} &= \int \frac{2udu}{1+u} \\ &= 2\int \frac{udu}{1+u}.\end{aligned}$$

We now do another substitution, $w = 1+u$, so $dw = du$, and $u = w-1$, leading to

$$\begin{aligned}\int \frac{udu}{1+u} &= \int \frac{(w-1)dw}{w} \\ &= \int dw - \int \frac{dw}{w} \\ &= w - \log|w|.\end{aligned}$$

Reversing the substitutions,

$$\begin{aligned}\int \frac{dx}{1+\sqrt{1+x}} &= 2\int \frac{udu}{1+u} \\ &= 2(w - \log|w|) \\ &= 2\left((1+u) - \log|1+u|\right) \\ &= 2\left(1+\sqrt{1+x} - \log\left(1+\sqrt{1+x}\right)\right)\end{aligned}$$

(with the absolute value sign removed in the last log, since $1+\sqrt{1+x} > 0$ always).[217]

[217] The substitution $u = 1+\sqrt{1+x}$ would also have worked here; as would the trick of writing $u/(1+u) = ((u+1)-1)/(u+1) = 1 - 1/(u+1)$, with obvious antiderivative $u - \log|1+u|$, instead of using a second substitution; note that this would have lead to the final answer $2\left(\sqrt{1+x} - \log\left(1+\sqrt{1+x}\right)\right)$, differing from the answer we got by a constant.

Example 2 $\int e^{\sqrt{x}}\ dx$. An obvious substitution is $u = \sqrt{x}$, with $du = dx/2\sqrt{x}$, so $dx = 2\sqrt{x}du = 2udu$, leading to

$$\begin{aligned}\int e^{\sqrt{x}}\ dx &= 2\int ue^u\ du \\ &= 2\left(ue^u - \int e^u\right) \quad \text{(integration by parts)} \\ &= 2\left(ue^u - e^u\right) \\ &= 2\left(\sqrt{x}e^{\sqrt{x}} - e^{\sqrt{x}}\right).\end{aligned}$$

We could have also tried the substitution $w = e^{\sqrt{x}}$, so $dw = ((e^{\sqrt{x}})/2\sqrt{x})dx$, or $dx = 2(\log w)/w$, which leads to

$$\int e^{\sqrt{x}}\ dx 2\int \log w\ dw,$$

and again an application of integration by parts finishes things.

Example 3 $\int e^{x^2}\ dx$. Here the obvious substitution is $u = x^2$, [218] so $du = 2xdx$, and $dx = du/(2x) = du/(2\sqrt{u})$, leading to

$$\int e^{x^2}\ dx = \frac{1}{2}\int \frac{e^u du}{\sqrt{u}}.$$

The obvious substitution here, $w = \sqrt{u}$, just returns us to $\int e^{w^2}\ dw$, and no other substitution or clever integration by parts helps matters — as mentioned earlier, e^{x^2} is a function with no elementary antiderivative.

There is a definite integral version of integration by substitution. With the notation as in the indefinite version, we have

$$\int_a^b f(x)\ dx = (H \circ g)(x)|_{x=a}^b = H(u)|_{u=g(a)}^{g(b)} = \int_{g(a)}^{g(b)} \frac{(f \circ g^{-1})(u)}{(g' \circ g^{-1})(u)}\ du.$$

So, the only change between definite and indefinite integration is that after the substitution $u = g(x)$, as well as re-expressing the integrand in terms of u, we also re-express the limits of integration in terms of u; and then there is no need to re-express things in terms of x before evaluating the integral.

We illustrate with some examples. Consider first $\int_{\pi/4}^{\pi/2} \cot x\ dx = \int_{\pi/4}^{\pi/2} \frac{\cos x\ dx}{\sin x}$. Set $u = \sin x$, so $du = \cos x dx$. At $x = \pi/4$ we have $u = \sin(\pi/4) = \sqrt{2}/2$, and at $x = \pi/2$ we have $u = 1$, so

$$\int_{\pi/4}^{\pi/2} \frac{\cos x\ dx}{\sin x} = \int_{\sqrt{2}/2}^{1} \frac{du}{u} = \log 1 - \log(\sqrt{2}/2) = (\log 2)/2.$$

[218] As with the example of $x/(1+x^2)$, we formally should split the domain of x^2 into two invertible parts to do this example correctly.

As another simple example, consider $\int_0^1 \frac{xdx}{1+x^2} = \frac{1}{2}\int_0^1 \frac{2xdx}{1+x^2}$. We set[219] $u = 1+x^2$, so $du = 2xdx$. At $x = 0$ we have $u = 1$, and at $x = 1$ we have $u = 2$. So the integral is

$$\frac{1}{2}\int_1^2 \frac{du}{u} = (\log 2)/2.$$

It's tempting here to conjecture:

> all definite integrals calculated by substitution evaluate to $(\log 2)/2$.

This is false, however.[220]

13.4 Some special (trigonometric) substitutions

To illustrate how the rough guiding principle of integration by substitution might sometimes break down, consider $\int \sqrt{1-x^2}dx$. It's tempting to try the substitution $u = x^2$, so $du = 2xdx$, $dx = du/2x = \frac{1}{2}\frac{du}{\sqrt{u}}$, making the integral

$$\frac{1}{2}\int \sqrt{\frac{1-u}{u}}du,$$

and any obvious substitution gets you right back where you started.

Alternately, one could try the completely non-obvious substitution $u = \arcsin x$ (note that the domain of the integrand is $[-1,1]$, which is exactly the domain of arcsin), so $x = \sin u$, $1-x^2 = \cos^2 u$, $\sqrt{1-x^2} = \cos u$ (as x ranges over $[-1,1]$, u ranges over $[-\pi/2, \pi/2]$, where cos is positive), $dx = \cos u\ du$, and the integral becomes

$$\int \cos^2 u\ du,$$

a completely different kettle of fish, and possibly amenable to a more direct attack than $\int \sqrt{1-x^2}dx$.

There is a general principle here.

Trigonometric substitutions, 1 A function involving a square root of a quadratic expression can often be reduced to an integral involving trigonometric functions, via the following substitutions (all motivated by the identity $\sin^2 + \cos^2 = 1$ and its relatives). Note that throughout we may assume $a, b > 0$.

- If the integrand involves $\sqrt{a^2 - b^2x^2}$, try the substitution $u = \arcsin\frac{bx}{a}$, or $x = \frac{a}{b}\sin u$. With this substitution,

$$\sqrt{a^2-b^2x^2} = a\sqrt{1-b^2x^2/a^2} = a\sqrt{1-\sin^2 u} = a\sqrt{\cos^2 u} = a\cos u,$$

[219] Note that $1+x^2$ *is* invertible on the domain $[0,1]$.

[220] This was a joke.

(so here we are motivated by $\cos^2 u = 1 - \sin^2 u$), while $dx = \frac{a}{b}\cos u\ du$. The domain of $\sqrt{a^2 - b^2x^2}$ is $[-a/b, a/b]$. For x on this domain, bx/a ranges over $[-1, 1]$ (the domain of arcsin), so the substitution makes sense. Since arcsin has range $[-\pi/2, \pi/2]$, and on this range cos is positive, we can justify the line $a\sqrt{\cos^2 u} = a\cos u$ above.

- If the integrand involves $\sqrt{a^2 + b^2x^2}$, try the substitution $u = \arctan\frac{bx}{a}$, or $x = \frac{a}{b}\tan u$. With this substitution,

$$\sqrt{a^2 + b^2x^2} = a\sqrt{1 + b^2x^2/a^2} = a\sqrt{1 + \tan^2 u} = a\sqrt{\sec^2 u} = a\sec u,$$

(so here we are motivated by $\sec^2 u = 1 + \tan^2 u$), while $dx = \frac{a}{b}\sec^2 u\ du$. The domain of $\sqrt{a^2 + b^2x^2}$ is $\mathbb{R}$. For x on this domain, bx/a ranges over $\mathbb{R}$ (the domain of arctan), so the substitution makes sense. Since arctan has range $[-\pi/2, \pi/2]$, and on this range sec is positive, we can justify the line $a\sqrt{\sec^2 u} = a\sec u$ above.

- If the integrand involves $\sqrt{b^2x^2 - a^2}$, try the substitution[221] $u = \operatorname{arcsec}\frac{bx}{a}$, or $x = \frac{a}{b}\sec u$. With this substitution,

$$\sqrt{b^2x^2 - a^2} = a\sqrt{b^2x^2/a^2 - 1} = a\sqrt{\sec^2 u - 1} = a\sqrt{\tan^2 u} = a|\tan u|,$$

(so here we are motivated by $\tan^2 u = \sec^2 u - 1$), while $dx = \frac{a}{b}\sec u\tan u\ du$. In the last two cases we wrote (and justified) $a\sqrt{\cos^2 u} = a\cos u$ and $a\sqrt{\sec^2 u} = a\sec u$; here we have to be a little more careful, and really need to write $a\sqrt{\tan^2 u} = a|\tan u|$. Indeed, the domain of $\sqrt{b^2x^2 - a^2}$ is $(-\infty, -a/b] \cup [a/b, \infty)$. If we are on the negative part of this domain then bx/a ranges over $(-\infty, -1]$, which is the negative part of the domain of arcsec. On this domain, arcsec ranges over the values $(\pi/2, \pi]$, and the tangent function is negative here. If we are on the positive part of the domain of $\sqrt{b^2x^2 - a^2}$ then bx/a ranges over $[1, \infty)$, which is the positive part of the domain of arcsec. On this domain, arcsec ranges over the values $[0, \pi/2)$, and the tangent function is positive here. So we get

$$\sqrt{b^2x^2 - a^2} = \begin{cases} -a\tan u & \text{if } x < -a/b, \\ a\tan u & \text{if } x > a/b. \end{cases}$$

We've already seen the example of $\int \sqrt{1 - x^2}dx$ transforming into $\int \cos^2 u\ du$ via the substitution $x = \sin u$. Here is another example, that involves the arcsec function, and so requires some care.

Example $\int \frac{\sqrt{25x^2-4}}{x}\ dx$. Following the discussion above, the sensible substitution is $x = (2/5)\sec u$, so $dx = (2/5)\sec u\tan u\ du$, and

$$\sqrt{25x^2 - 4} = 5\sqrt{x^2 - (2/5)^2} = 5\sqrt{(2/5)^2\sec^2 u - (2/5)^2} = 2\sqrt{\sec^2 u - 1} = 2\sqrt{\tan^2 u}.$$

Following the discussion above, we know that we have to treat separately the cases $x \geq 2/5$ and $x \leq -2/5$.

[221] The arcsec function, which somewhat weird, is discussed at the very end of Section 12.4.

- **Case of $x \geq 2/5$:** Here $2\sqrt{\tan^2 u} = 2\tan u$, so

$$\sqrt{25x^2 - 4} = 2\tan u,$$

and the integral becomes

$$2\int \tan^2 u du.$$

We'll discuss trigonometric integrals like this in general in a short while, but this one can be dealt with fairly easily: using $\sec^2 - 1 = \tan^2$ we get

$$2\int \tan^2 u du = 2\int (\sec^2 u - 1) du = 2(\tan u - u).$$

We would like to re-express this in terms of x, recalling $x = (2/5)\sec u$. One way is to simply write

$$2(\tan u - u) = 2\left(\tan\left(\sec^{-1}\left(\frac{5x}{2}\right)\right) - \sec^{-1}\left(\frac{5x}{2}\right)\right).$$

This can be considerable cleaned up.

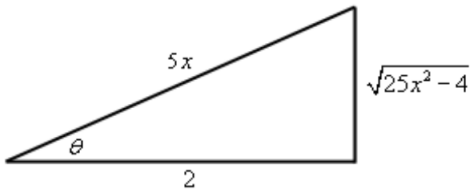
Since $x \geq 2/5$ we have $5x/2 \geq 1$, and so $\sec^{-1}(5x/2)$ is between 0 and $\pi/2$. Now we use $\sec^2 = 1 + \tan^2$ to get that $(5x/2)^2 = 1 + \tan^2(\sec^{-1}(5x/2))$, so $\tan(\sec^{-1}(5x/2)) = \pm\sqrt{(5x/2)^2 - 1} = \pm\sqrt{25x^2 - 4}/2$. But which is it, plus or minus? Well, since $\sec^{-1}(5x/2)$ is between 0 and $\pi/2$, and tan is positive in that domain, we must take the positive square root.

As it happens, we can also re-express $\sec^{-1}(5x/2)$ in terms of simpler (more fundamental) trigonometric functions. We have that $\sec^{-1}$ is the inverse of the composition $(r \circ \cos)$, where r is the reciprocal function $x \mapsto 1/x$. We know[222] that $(r \circ \cos)^{-1} = \cos^{-1} \circ r^{-1} = \cos^{-1} \circ r$ (because $r = r^{-1}$). So $\sec^{-1}(5x/2) = \cos^{-1}(2/5x)$ (and this calculation does not depend on $x \geq 2/5$; it works for all x in the domain). So we get, in this case,

$$\int \frac{\sqrt{25x^2 - 4}}{x}\, dx = \sqrt{25x^2 - 4} - 2\cos^{-1}\left(\frac{2}{5x}\right).$$

There is another way to do the simplifying calculation, that may be more intuitive. Recall that in the present case $\sec^{-1}(5x/2)$ is between 0 and $\pi/2$. Remembering that in a right-angled triangle with one angle θ, the secant of θ is the length of the side opposite the angle, divided by the length of the hypothenuse, we are led to the following right-angled triangle:

[222] This was a homework problem.



The angle θ is $\sec^{-1}(5x/2)$, which can clearly be expressed in terms of more fundamental inverse trigonometric functions: for example, it is $\cos^{-1}(2/5x)$; and $\tan\theta$ is $\sqrt{25x^2-4}/2$.

An issue with the "right-angled triangle" approach is that it is not so easy to implement it when the angle one is working with is negative, or greater than $\pi/2$; in this case the earlier method is probably more reliable.

- **Case of** $x \leq -2/5$: Here, following the discussion earlier, $2\sqrt{\tan^2 u} = -2\tan u$, so the integral becomes

$$\begin{aligned} -2\int \tan^2 u du &= -2\int (\sec^2 u - 1)du \\ &= -2\tan u + 2u \\ &= -2\tan\left(\sec^{-1}\left(\frac{5x}{2}\right)\right) + 2\sec^{-1}\left(\frac{5x}{2}\right) \\ &= -2\tan\left(\sec^{-1}\left(\frac{5x}{2}\right)\right) + 2\cos^{-1}\left(\frac{2}{5x}\right). \end{aligned}$$

Since $x \leq -2/5$ we have $5x/2 \leq -1$, and so $\sec^{-1}(5x/2)$ is between $\pi/2$ and π. Now we use $\sec^2 = 1 + \tan^2$ to get that $(5x/2)^2 = 1 + \tan^2(\sec^{-1}(5x/2))$, so $\tan(\sec^{-1}(5x/2)) = \pm\sqrt{(5x/2)^2 - 1} = \pm\sqrt{25x^2-4}/2$. Since $\sec^{-1}(5x/2)$ is between $\pi/2$ and π, and tan is *negative* in that domain, we must take the *negative* square root — $\tan(\sec^{-1}(5x/2)) = -\sqrt{25x^2-4}/2$ and

$$\int \frac{\sqrt{25x^2-4}}{x}\,dx = \sqrt{25x^2-4} + 2\cos^{-1}\left(\frac{2}{5x}\right)$$

in this case. Note the subtle difference between the two regimes: when x is positive we subtract $2\cos^{-1}(2/5x)$, while when x is negative we add that term.[223]

[223] If you ask Wolfram Alpha to evaluate the integral in this example, you get the answer

$$\int \frac{\sqrt{25x^2-4}}{x}\,dx = \sqrt{25x^2-4} + 2\tan^{-1}\left(\frac{2}{\sqrt{25x^2-4}}\right),$$

valid for all x with $|x| \geq 2/5$. Amusingly, this answer *does not* differ from our answer by a universal constant. On $[2/5, \infty)$, $\tan^{-1}(2/\sqrt{25x^2-4})$ and $-\cos^{-1}(2/5x)$ differ by a constant, while on $(-\infty, 2/5]$, $\tan^{-1}(2/\sqrt{25x^2-4})$ and $+\cos^{-1}(2/5x)$ differ by a *different* constant; see the graph below:

The above discussion covers all cases where there is a quadratic expression under a square root, *when the quadratic has no linear term.*[224] But every quadratic with a linear term can be massaged into a quadratic without a linear term, by a simple linear substitution, a process known as "completing the square". This goes as follows: first assuming that $p > 0$,

$$\begin{aligned} px^2 + qx + r &= \left(\sqrt{p}x + \frac{q}{2\sqrt{p}}\right)^2 + r - \frac{q^2}{4p} \\ &= y^2 + r - \frac{q^2}{4p}. \end{aligned}$$

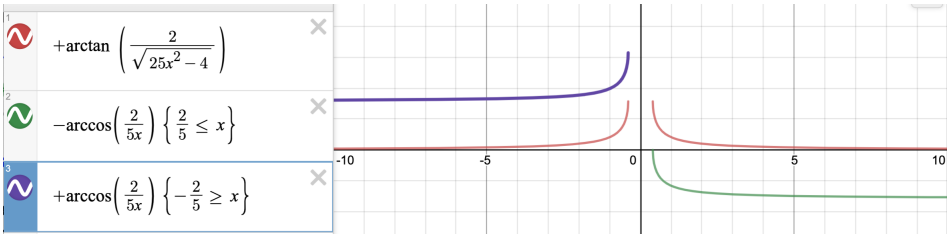
If $r > q^2/4p$ then the substitution $y = \sqrt{p}x + q/(2\sqrt{p})$ reduces $px^2 + qx + r$ to the form $y^2 + a^2$; if $r < q^2/4p$ then it reduces it to the form $y^2 - a^2$. If, on the other hand, $p < 0$, then

$$\begin{aligned} px^2 + qx + r &= -((-p)x^2 - qx) + r \\ &= -\left(\sqrt{-p}x - \frac{q}{2\sqrt{-p}}\right)^2 + r - \frac{q^2}{4p} \\ &= \left(r - \frac{q^2}{4p}\right) - y^2. \end{aligned}$$

If $r > q^2/4p$ then the substitution $y = \sqrt{-p}x - q/(2\sqrt{-p})$ reduces $px^2 + qx + r$ to the form $y^2 - a^2$. If, on the other hand, $r < q^2/4p$, then from the quadratic formula we find that the (real) range of $px^2 + qx + r$ is empty, and we are in the one case where there is no point to considering the integration.

Here's an alternate approach to completing the square, using the quadratic formula:

$$\begin{aligned} px^2 + qx + r &= p\left(x^2 + \frac{q}{p}x + \frac{r}{p}\right) \\ &= p\left(x - \left(\frac{-\frac{q}{p} + \sqrt{\frac{q^2}{p^2} - 4\frac{r}{p}}}{2}\right)\right)\left(x - \left(\frac{-\frac{q}{p} - \sqrt{\frac{q^2}{p^2} - 4\frac{r}{p}}}{2}\right)\right) \\ &= p\left(\left(x + \frac{q}{2p}\right) - \frac{1}{2}\sqrt{\frac{q^2}{p^2} - \frac{4r}{p}}\right)\left(\left(x + \frac{q}{2p}\right) + \frac{1}{2}\sqrt{\frac{q^2}{p^2} - \frac{4r}{p}}\right) \\ &= p\left(\left(x + \frac{q}{2p}\right)^2 - \left(\frac{1}{2}\sqrt{\frac{q^2}{p^2} - \frac{4r}{p}}\right)^2\right) \end{aligned}$$



This is cautionary example that shows that you have to keep your wits about you when dealing with integrals of functions that are defined on unions of intervals.

[224] We don't consider the fourth case, $\sqrt{-b^2x^2 - a^2}$, since this has empty domain.

If $p > 0$ we have reduced to one of the the forms $a^2 + b^2x^2$ or $b^2x^2 - a^2$ (which one depending on whether $(q^2/p^2) - (4r)(p)$ is positive or negative). If $p < 0$ then $(q^2/p^2) - (4r)(p)$ must be non-negative (otherwise there are no reals in the range of $px^2 + qx + r$), and we have reduced to the form $a^2 + b^2x^2$.

Example $\int \frac{x}{\sqrt{2x^2-4x-7}}\,dx$. To make things easier, we pull out a factor of $\sqrt{2}$:

$$\int \frac{x}{\sqrt{2x^2 - 4x - 7}}\,dx = \frac{1}{\sqrt{2}} \int \frac{x}{\sqrt{x^2 - 2x - (7/2)}}\,dx.$$

We now complete the square:

$$x^2 - 2x - (7/2) = (x-1)^2 - 1 - (7/2) = (x-1)^2 - (3/\sqrt{2})^2.$$

We make the substitution $u = x - 1$, so $du = dx$, and $x = u + 1$, so the integral becomes

$$\begin{aligned} \int \frac{x}{\sqrt{x^2 - 2x - (7/2)}}\,dx &= \int \frac{u+1}{\sqrt{u^2 - (3/\sqrt{2})^2}}\,dx \\ &= \int \frac{u}{\sqrt{u^2 - (3/\sqrt{2})^2}}\,dx + \int \frac{1}{\sqrt{u^2 - (3/\sqrt{2})^2}}\,dx. \end{aligned}$$

The first of these integrals can be handled by a simple substitution $v = u^2 - (3/\sqrt{2})^2$. For the second, the earlier discussion suggests the substitution $u = 3/\sqrt{2}\sec v$. The details are left as an exercise.

The examples given above indicate that it is important to understand the integrals of functions of sin, cos, et cetera, as these pop up naturally in the study functions involving square roots of quadratics. All functions of trigonometric functions can be re-expressed purely in terms of sin and cos, so we concentrate our attention only on functions of sin and cos. In particular, we are going to focus attention on functions of the form $\cos^p \sin^q$, where p and q are integers; using linearity of the integral, virtually every function of trigonometric functions that we will need to be able to deal with, can be reduced to this form.

Our examination of this integrals will break into cases, according to the parity (oddness or evenness) of p, q.

- **Case 1**: p is odd. Here we write

$$\begin{aligned} \int \cos^p x \sin^q x\ dx &= \int \cos^{p-1} x \sin^q x\ \cos x\ dx \\ &= \int \left(\cos^2 x\right)^{\frac{p-1}{2}} \sin^q x\ \cos x\ dx \\ &= \int \left(1 - \sin^2 x\right)^{\frac{p-1}{2}} \sin^q x\ \cos x\ dx \end{aligned}$$

(i.e., we "peel off" one copy of $\cos x$, to join the dx; note that we can do this whether p is positive or negative; note also that since p is odd, $(p-1)/2$ is an integer). We now make the substitution $u = \sin x$, so $du = \cos x dx$, to get

$$\int \cos^p x \sin^q x \; dx = \int \left(1-u^2\right)^{\frac{p-1}{2}} u^q \; du.$$

If $(p-1)/2 \geq 0$ then expanding out the polynomial $(1-u^2)^{\frac{p-1}{2}}$ and multiplying through by u^q, we have reduced to integrating a linear combination of functions of the form u^k, $k \in \mathbb{Z}$ (very easy); if $(p-1)/2 < 0$, then at least we have reduced to integrating a rational function in u, a topic that we will shortly address.

Example: $\int \cos^3 x \sin^4 x \; dx$. We make the substitution $u = \sin x$, $du = \cos x dx$, so

$$\int \cos^3 x \sin^4 x \, dx = \int \cos^2 x \sin^4 \cos x \, dx = \int (1-\sin^2) x \sin^4 \cos x \, dx = \int (1-u^2) u^4 \, du.$$

- **Case 2**: q is odd. This is almost identical to Case 1. Here we "peel off" one copy of $\sin x$, to join the dx, and make the substitution $u = \cos x$, so $du = -\sin x dx$, and we reduce to a rational function in u via $(\sin^2 x)^{(p-1)/2} = (1-\cos^2 x)^{(p-1)/2} = (1-u^2)^{(p-1)/2}$.

  Example: $\int \frac{\cos^2 x}{\sin^5 x} \; dx$. After the substitution $u = \cos x$,

$$\int \frac{\cos^2 x}{\sin^5 x} \; dx = \int \frac{\cos^2 x}{\sin^6 x} \sin x \; dx = -\int \frac{u^2}{(1-u^2)^3} \; du.$$

- **Case 3**: p, q even, both non-negative, at least one positive[225]. From

$$\begin{aligned} \cos^2 x + \sin^2 x &= 1 \\ \cos^2 x - \sin^2 x &= \cos 2x, \end{aligned}$$

  we get the identities

$$\cos^2 x = \frac{1+\cos 2x}{2}, \quad \text{and} \quad \sin^2 x = \frac{1-\cos 2x}{2}$$

  which leads to

$$\cos^p x \sin^q x = \left(\frac{1+\cos 2x}{2}\right)^{\frac{p}{2}} \left(\frac{1+\cos 2x}{2}\right)^{\frac{q}{2}}.$$

  Expanding this out, and separating out the monomials in the polynomial, we get a collection of integrands of the form $\cos^{p'} 2x$ with p' non-negative, and with p' smaller than p. Any such terms with p' odd can be dealt with by an application of Case 1;

[225] Things are rather trivial if both $p, q = 0$...

any such terms with both p' even can be dealt with by *another* application of Case 3. Because the highest powers involved are strictly decreasing, this process terminates after a finite number of iterations.

Example: $\int \frac{\cos^6 x}{\csc^2 x}\, dx$. We write

$$\begin{aligned}\frac{\cos^6 x}{\csc^2 x} &= (\cos^2 x)^3(\sin^2 x)\\ &= \left(\frac{1+\cos 2x}{2}\right)^3\left(\frac{1-\cos 2x}{2}\right)\\ &= \frac{1}{16}\left(1+2\cos 2x-2\cos^3 2x-\cos^4 2x\right).\end{aligned}$$

So (ignoring the constants) we've reduced to four integrals:

- the 1 is trivial;
- the $\cos 2x$ is easy (after the substitution $u = 2x$, which is the sort of easy substitution one should just do in one's head);
- the $\cos^3 2x$ is an instance of Case 1, and can be dealt with by the substitution $u = \sin 2x$.
- the $\cos^4 2x$ is an instance of Case 3, but with smaller powers than the original instance. We write

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

  We have three simpler integrals, the first trivial, the second an instance of Case 1, and the third and even simpler instance of Case 3 (clearly, the last one that will be encountered in this particular example).

This doesn't cover every expression of the form $\cos^p x \sin^q x$ with $p, q \in \mathbb{Z}$; for example, it omits the case where p, q are both even and non-positive, with at least one of them negative.[226] This case, and a whole many more trigonometric integrals, can be dealt with by the following "magic bullet".

The last-resort trigonometric substitution Consider the substitution $t = \tan x/2$. We have

$$dt = \frac{\sec^2 x/2}{2}dx = \frac{1+\tan^2 x/2}{2}dx = \frac{1+t^2}{2}dx,$$

so

$$dx = \frac{2dt}{1+t^2}.$$

[226] I think that this is the only non-trivial omitted case.

Also,

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

with everything valid exactly as long as $\tan(x/2)$ is defined. And since $\cos x = \cos^2(x/2) - \sin^2(x/2)$ and $1 = \cos^2(x/2) + \sin^2(x/2)$, we have

$$\begin{aligned}
\cos x &= 1 - 2\sin^2(x/2) \\
&= 1 - 2\frac{\sin^2(x/2)\cos^2(x/2)}{\cos^2(x/2)} \\
&= 1 - 2\tan^2(x/2)\cos^2(x/2) \\
&= 1 - 2t^2\cos^2(x/2) \\
&= 1 - 2\left(\frac{t^2}{\sec^2(x/2)}\right) \\
&= 1 - 2\left(\frac{t^2}{1+\tan^2(x/2)}\right) \\
&= 1 - 2\left(\frac{t^2}{1+t^2}\right) \\
&= \frac{1-t^2}{1+t^2},
\end{aligned}$$

again with everything valid exactly as long as $\tan(x/2)$ is defined.

The upshot of this is

> *any* integrand in the variable x that is a function of $\sin x$, $\cos x$ (and the other trigonometric functions) (not necessarily a rational function — it could involve roots, and exponentials, too) can be converted into an integrand in the variable t that does *not* mention any trigonometric functions, by the substitution $t = \tan x/2$ (though this substitution does not do away with roots or exponentials). In particular, if an integrand is a *rational* function of trigonometric functions, it can be converted into a rational function of t by this substitution.

This is a "last resort" substitution, because in general if there is *any* other way to approach the integration problem, the path is almost always going to be easier that way![227]

We give some examples:

Example 1 $\int \cos^2 x\ dx$. The obvious thing to do here is to write

$$\int \cos^2 x\ dx = \int \left(\frac{1+\cos 2x}{2}\right)\ dx = \frac{x}{2} + \frac{\sin 2x}{4}.$$

Using the "last resort" substitution $t = \tan(x/2)$ we get

$$\begin{aligned}\int \cos^2 x\ dx &= \int \left(\frac{1-t^2}{1+t^2}\right)^2 \frac{2}{1+t^2}\ dt \\ &= 2\int \frac{(1-t^2)^2}{(1+t^2)^3}\ dt.\end{aligned}$$

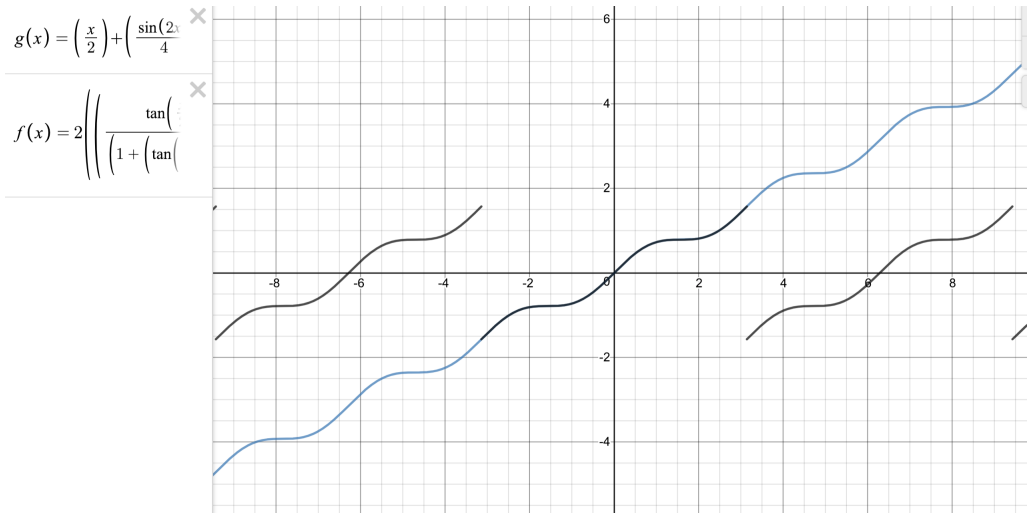
As we will shortly see, this kind of integral can be handled quite mechanically, but the result is quite hideous. `Mathemaica` gives the integral as

$$2\left(\frac{t}{(1+t^2)^2} - \frac{t}{2(1+t^2)} + \frac{\arctan(t)}{2}\right),$$

so that

$$\int \cos^2 x\ dx = 2\left(\frac{\tan(x/2)}{(1+(\tan(x/2))^2)^2} - \frac{\tan(x/2)}{2(1+(\tan(x/2))^2)} + \frac{\arctan(\tan(x/2))}{2}\right) := f(x).$$

This obviously equals $x/2+(\sin 2x)/4 := g(x)$, right, at least up to an additive constant? Not exactly ... the Desmos screenshot below shows the two functions:



[227] Also, there is a slight issue with this method — see example 1, or (exercise) see if you can spot the issue before looking at example 1.

Notice that the domain of f is *not* all reals; it is all reals other than $\pm\pi, \pm 3\pi, \pm 5\pi, \ldots$. On the interval $(-\pi, \pi)$, f and g agree; on all other intervals they agree up to a constant. The issue here is that in making the substitution $t = \tan(x/2)$, one has to give up all values of x of the form $\pm\pi, \pm 3\pi, \pm 5\pi, \ldots$; $\tan(x/2)$ is not defined for these values. On all other values, the substitution works fine (modulo dealing with the hideous expressions that come out of it).

Example 2 $\int \frac{dx}{1+\sin x}$. Here, the only course of action seems to be to apply the "last resort" substitution, and in fact it works beautifully, leading to

$$\int \frac{dx}{1+\sin x} = 2\int \frac{1}{(1+t)^2}\, dt = \frac{-2}{1+t} = \frac{-2}{1+\tan(x/2)}.$$

13.5 Integration by partial fractions

In the last section we saw that many integrals can be reduced to integrals of rational functions, via appropriate substitutions. There is a method that, in principle at leasts, can find a primitive of any rational function. The method is, on the whole, fairly okay to understand at a theoretical level, but (unfortunately) rather difficult to implement practically except in some simple cases.

Setup

A *rational function* is a function f given by $f(x) = \frac{P(x)}{Q(x)}$ where P and Q are both polynomials, and Q is not zero.

Because our concern is with finding antiderivatives of rational functions, and this is easy when Q is a constant (in which case the rational function is just a polynomial), we will throughout assume that the degree of Q is at least 1. Recall that the *degree* $\deg(Q)$ of the polynomial $Q(x)$ is the highest power of x in the polynomial that has a non-zero coefficient.

Also, since it's easy to find an antiderivative of the zero function, we will assume that P is not zero.

By scaling Q by a constant, if necessary, we can assume that

$$Q(x) = x^n + q_1 x^{n-1} + \cdots + q_{n-1}x + q_n$$

where $n \geq 1$, and that

$$P(x) = p_0 x^m + p_1 x^{m-1} + \cdots + p_{m-1}x + p_m$$

where $m \geq 0$ and $p_0 \neq 0$.

Three key facts

To find an antiderivative of $P(x)/Q(x)$ we will need to use three facts from algebra, that we will not prove.

Fact 1 (the division algorithm for polynomials): If $\deg(P) \geq \deg(Q)$ then there are polynomials $A(x)$ (the *quotient*) and $B(x)$ (the *remainder*) with $\deg(B) < \deg(Q)$ such that $P(x) = A(x)Q(x) + B(x)$, or

$$\frac{P(x)}{Q(x)} = A(x) + \frac{B(x)}{Q(x)}.$$

A and B can be found by polynomial long division.

Running example: Consider the rational function

$$\frac{3x^6 - 7x^5 + 9x^4 - 9x^3 + 7x^2 - 4x - 1}{x^4 - 2x^3 + 2x^2 - 2x + 1}.$$

When we start long division, the first term will definitely be $3x^2$ (that's what's needed to get the leading x^4 in the denominator up to $3x^6$). Now

$$(x^4 - 2x^3 + 2x^2 - 2x + 1)(3x^2) = 3x^6 - 6x^5 + 6x^4 - 6x^3 + 3x^2,$$

and when this is subtracted from $3x^6 - 7x^5 + 9x^4 - 9x^3 + 7x^2 - 4x - 1$ we get

$$-x^5 + 3x^4 - 3x^3 + 4x^2 - 4x - 1.$$

So we continue the long division with $-x$ (that's what's needed to get the leading x^4 in the denominator up to $-x^5$). Now

$$(x^4 - 2x^3 + 2x^2 - 2x + 1)(-x) = -x^5 + 2x^4 - 2x^3 + 2x^2 - x,$$

and when this is subtracted from $-x^5 + 3x^4 - 3x^3 + 4x^2 - 4x - 1$ we get

$$x^4 - x^3 + 2x^2 - 3x - 1.$$

So the next term in the long division is 1; and when $x^4 - 2x^3 + 2x^2 - 2x + 1$ is subtracted from $x^4 - x^3 + 2x^2 - 3x - 1$ we get $x^3 - x - 2$. This has degree smaller than 4, so the remainder term has been reached, and the long division is finished:

$$\frac{3x^6 - 7x^5 + 9x^4 - 9x^3 + 7x^2 - 4x - 1}{x^4 - 2x^3 + 2x^2 - 2x + 1} = 3x^2 - x + 1 + \frac{x^3 - x - 2}{x^4 - 2x^3 + 2x^2 - 2x + 1}.$$

Examples to work out: Find A and B for the following rational functions.

1. $\frac{x^5}{x^4+x+1}$
2. $\frac{2x^7+3x^6-x^5+4x^4+5x^2-1}{x^4+x^2-x}$
3. $\frac{(1-x)^4}{(1+x)^4}$
4. $\frac{x^4+2x^3+3x^2+2x+1}{x^2+x+1}$

The point of Fact 1 is that since

$$\int \frac{P}{Q} = \int A + \int \frac{B}{Q},$$

and $\int A$ is easy to find (A being a polynomial), from here on in the method of partial fractions we need only concentrate on rational functions of the form $B(x)/Q(x)$, i.e., those where the degree of the numerator is less than the degree of the denominator.

Fact 2 (a corollary of the fundamental theorem of algebra): The polynomial $\deg(Q)$ can be factored into linear and quadratic terms:

$$Q(x) = (x - r_1)^{\alpha_1} \cdots (x - r_k)^{\alpha_k} (x^2 - 2s_1 x + t_1)^{\beta_1} \cdots (x^2 - 2s_\ell x + t_\ell)^{\beta_\ell}$$

where the r_i's, s_i's and t_i's are reals, the α_i's and β_i's are natural numbers, the r_i's are distinct from each other, the pairs (s_i, t_i) are distinct from each other (so there is no co-incidence between any pairs of factors), $s_i^2 < t_i$ for each i (so none of the quadratic terms can be factored further into linear terms), and $\deg(Q) = \sum_i \alpha_i + 2\sum_j \beta_j$.

Moreover, each quadratic term $x^2 - 2s_i x + t_i$ can be written in the form $(x - a_i)^2 + b_i^2$ with a_i and b_i real and b_i positive (this comes straight from $s_i^2 < t_i$: we have $x^2 - 2s_i x + t_i = (x - s_i)^2 + t_i - s_i^2 = (x - a_i)^2 + b_i^2$ where $a_i = s_i$ and $b_i = \sqrt{t_i - s_i^2}$).

Lurking behind Fact 2 is the *fundamental theorem of algebra*, which says that every polynomial with complex coefficients has a root in the complex numbers. Given a complex polynomial $C(z)$ with root c, using the division algorithm it is possible to write $C(z) = (z - c)\tilde{C}(z)$, where $\tilde{C}(z)$ is a complex polynomial whose degree is one less than that of C, and repeating this process we get that C factors fully into linear terms, as $C(z) = (z - c_1) \cdots (z - c_n)$ where $n = \deg(C)$. Here the c_i are complex numbers; but Q, having only real coefficients, possibly has some of these roots being real (these are the r_i above). It turns out that for a polynomial with all real coefficients, the complex roots appear in what are called *complex conjugate pairs*: pairs of the form $a + b\sqrt{-1}$ and $a - b\sqrt{-1}$. Noting that

$$(z - (a + b\sqrt{-1}))(z - (a - b\sqrt{-1})) = (z - a)^2 + b^2,$$

this "explains" the form of the quadratic factors above.

In general, it is *very difficult* to fully factor a real polynomial into linear and quadratic factors.

Running example: Consider $x^4 - 2x^3 + 2x^2 - 2x + 1$. After some trial-and error, we find that 1 must be a root, since $(1)^4 - 2(1)^3 + 2(1)^2 - 2(1) + 1 = 0$. So $x - 1$ is a factor, and long division gives

$$x^4 - 2x^3 + 2x^2 - 2x + 1 = (x - 1)(x^3 - x^2 + x - 1).$$

Again 1 is a root of $x^3 - x^2 + x - 1$, and

$$x^4 - 2x^3 + 2x^2 - 2x + 1 = (x - 1)(x - 1)(x^2 + 1).$$

The quadratic formula tells us that we cannot factor any further.

Examples to work out: Fully factor the polynomials below into linear factors of the form $x - r$ and quadratic factors of the form $(x - a)^2 + b^2$. Start by trying a few small values of r (positive and negative) to find one with the polynomial evaluating to 0 at r; then divide by $x - r$ and repeat.

1. Factorize $x^4 - x^3 - 7x^2 + x + 6$
2. Factorize $x^4 - x^3 - 7x^2 + x + 6$
3. Factorize $x^3 - 3x^2 + 3x - 1$
4. Factorize $x^6 + 3x^4 + 3x^2 + 1$
5. Factorize $x^4 + 1$ (tricky)

The point of Fact 2 is that it feeds nicely into Fact 3.

Fact 3 (partial fractions decomposition): Let Q and B be polynomials as described above (Q has degree at least 1, and leading coefficient 1, and B has degree less than that of Q). Let Q be factored into linear and quadratic terms, exactly as outlined in Fact 2:

$$Q(x) = (x - r_1)^{\alpha_1} \cdots (x - r_k)^{\alpha_k}((x - a_1)^2 + b_1^2)^{\beta_1} \cdots ((x - a_\ell)^2 + b_\ell^2)^{\beta_\ell}.$$

Then there are real constants

$$\begin{gathered} A_{11}, \ldots, A_{1\alpha_1}, \\ A_{21}, \ldots, A_{2\alpha_2}, \\ \cdots, \\ A_{k1}, \ldots, A_{k\alpha_k}, \end{gathered}$$

$$\begin{gathered} B_{11}, \ldots, B_{1\beta_1}, \\ B_{21}, \ldots, B_{2\beta_2}, \\ \cdots, \\ B_{k1}, \ldots, B_{k\beta_k}, \end{gathered}$$

$$\begin{gathered} C_{11}, \ldots, C_{1\beta_1}, \\ C_{21}, \ldots, C_{2\beta_2}, \\ \cdots, \\ C_{k1}, \ldots, C_{k\beta_k}, \end{gathered}$$

such that

$$\begin{array}{rl} & \frac{A_{11}}{(x-r_1)} + \frac{A_{12}}{(x-r_1)^2} + \cdots + \frac{A_{1\alpha_1}}{(x-r_1)^{\alpha_1}} + \\ & \frac{A_{21}}{(x-r_2)} + \frac{A_{22}}{(x-r_2)^2} + \cdots + \frac{A_{2\alpha_2}}{(x-r_2)^{\alpha_2}} + \\ & \cdots + \\ & \frac{A_{k1}}{(x-r_k)} + \frac{A_{k2}}{(x-r_k)^2} + \cdots + \frac{A_{k\alpha_k}}{(x-r_k)^{\alpha_k}} + \\ \frac{B(x)}{Q(x)} = & \frac{B_{11}x+C_{11}}{((x-a_1)^2+b_1^2)} + \frac{B_{12}x+C_{12}}{((x-a_1)^2+b_1^2)^2} + \cdots + \frac{B_{1\beta_1}x+C_{1\beta_1}}{((x-a_1)^2+b_1^2)^{\beta_1}} + \\ & \frac{B_{21}x+C_{21}}{((x-a_2)^2+b_2^2)} + \frac{B_{22}x+C_{22}}{((x-a_2)^2+b_2^2)^2} + \cdots + \frac{B_{2\beta_2}x+C_{2\beta_2}}{((x-a_2)^2+b_2^2)^{\beta_2}} + \\ & \cdots + \\ & \frac{B_{\ell 1}x+C_{\ell 1}}{((x-a_\ell)^2+b_\ell^2)} + \frac{B_{\ell 2}x+C_{\ell 2}}{((x-a_\ell)^2+b_\ell^2)^2} + \cdots + \frac{B_{\ell\beta_\ell}x+C_{\ell\beta_\ell}}{((x-a_\ell)^2+b_\ell^2)^{\beta_\ell}}. \end{array}$$

The proof of Fact 3 is not very difficult, but it requires too much familiarity with linear algebra to describe here.

It is somewhat straightforward to locate the values of the constants asserted in Fact 3. Start with the equation given in Fact 3 (with all the constants unknown). Multiply both sides by $Q(x)$. The right-hand side becomes a polynomial of degree $\deg(Q) - 1$, so with $\deg(Q)$ coefficients, expressed in terms of a number of unknowns — $\deg(Q)$ unknowns, to be precise. The left-hand side becomes a polynomial with known coefficients with degree at most $\deg(Q) - 1$. Equating the constant terms on both sides, the linear terms, the quadratic terms, et cetera, one gets a collection of $\deg(Q)$ equations in $\deg(Q)$ unknowns. Using techniques from linear algebra, such a system can be solved relatively quickly to find the (unique, as it turns out) values for the constants (the A's, B's and C's).

Even without knowing linear algebra, it is fairly straightforward to perform this task, if the degrees of the polynomials involved are all reasonably small.

Running example: We seek to find the partial fractions decomposition of $\frac{x^3-x-2}{(x-1)^2(x^2+1)}$. We start with

$$\frac{x^3-x-2}{(x-1)^2(x^2+1)} = \frac{A}{x-1} + \frac{B}{(x-1)^2} + \frac{Cx+D}{x^2+1}.$$

Multiplying through by $(x-1)^2(x^2+1)$ yields

$$\begin{aligned} x^3 - x - 2 &= A(x-1)(x^2+1) + B(x^2+1) + (Cx+D)(x-1)^2 \\ &= (A+C)x^3 + (-A+B-2C+D)x^2 + (A+C-2D)x + (-A+B+D). \end{aligned}$$

Equating coefficients gives

$$A + C = 1, \quad -A + B - 2C + D = 0, \quad A + C - 2D = -1, \quad -A + B + D = -2.$$

One can solve this system of four equations in four unknowns by, for example, using the first equation to write $A = C - 1$, then substituting this into the remaining three to get three equations in three unknowns, then substitute again to get two equations in two unknowns, then again to get one equation in one unknown, which is easy to solve. Plugging in that one known value, the whole system now becomes one of three equations in three unknowns; rinse

and repeat. (There are systematic ways to do this process, which are very efficient, and are explored in linear algebra).

Solving this system of equations in this way gives $A = 2$, $B = -1$, $C = -1$ and $D = 1$, so that

$$\frac{x^3 - x - 2}{(x-1)^2(x^2+1)} = \frac{2}{x-1} - \frac{1}{(x-1)^2} - \frac{(x-1)}{x^2+1}.$$

Examples to work out: Find the partial fractions decompositions of the following expressions.

1. $\frac{2x^2+7x-1}{x^3+x^2-x-1}$
2. $\int \frac{2x+1}{x^3-3x^2+3x-1}$
3. $\int \frac{3x}{(x^2+x+1)^3}$
4. $\frac{1}{x^4+1}$

Finding antiderivatives of rational functions

Using the three facts above, we can reduce the task of finding an antiderivative of a rational function to that of finding antiderivatives of functions of the following types:

- polynomials — these are easy
- functions of the form $\frac{A}{(x-r)^\alpha}$ where A and r are constants, and α is a natural number. These are straightforward:

$$\int \frac{A}{(x-r)^\alpha} dx = \begin{cases} \frac{A}{(1-\alpha)(x-r)^{\alpha-1}} & \text{if } \alpha \neq 1 \\ A \log(x-r) & \text{if } \alpha = 1. \end{cases}$$

- functions of the form $\frac{Cx+D}{((x-a)^2+b^2)^\beta}$ where C, D, a and b are real constants, with b positive, and β is a natural number. To deal with these, we first write

$$\frac{Cx+D}{((x-a)^2+b^2)^\beta} = \frac{(C/2)2(x-a)}{((x-a)^2+b^2)^\beta} + \frac{Ca+D}{((x-a)^2+b^2)^\beta}.$$

  Using the substitution $u = (x-a)^2 + b^2$ we get

$$\int \frac{(C/2)2(x-a)}{((x-a)^2+b^2)^\beta} dx = \frac{C}{2} \int \frac{du}{u^\beta} = \begin{cases} \frac{C/2}{(1-\beta)u^{\beta-1}} = \frac{C/2}{(1-\beta)((x-a)^2+b^2)^{\beta-1}} & \text{if } \beta \neq 1 \\ (C/2) \log u = (C/2) \log((x-a)^2+b^2) & \text{if } \beta = 1. \end{cases}$$

  To deal with the $(Ca+D)/(((x-a)^2+b^2)^\beta)$ terms, we have

$$\frac{Ca+D}{((x-a)^2+b^2)^\beta} = \frac{Ca+D}{b^{2\beta}} \left(\frac{1}{\left(\frac{x-a}{b}\right)^2 + 1} \right)$$

so using the substitution $u = (x-a)/b$ we get

$$\int \frac{Ca+D}{((x-a)^2+b^2)^\beta}dx = \frac{Ca+D}{b^{2\beta-1}}\int \frac{du}{(u^2+1)^\beta}$$

In class we studied the integral $\int \frac{du}{(u^2+1)^\beta}$. If we set

$$A_\beta = \int \frac{du}{(u^2+1)^\beta}$$

then we saw that we have $A_0 = 1$, $A_1 = \arctan u = \arctan((x-a)/b)$, and for $\beta \geq 1$,

$$A_{\beta+1} = \frac{u}{2\beta(1+u^2)^\beta} + \frac{(2\beta-1)}{2\beta}A_\beta = \frac{(x-a)/b}{2\beta(1+((x-a)/b)^2)^\beta} + \frac{(2\beta-1)}{2\beta}A_\beta.$$

So we can recursively figure out an antiderivative.

Running example: We seek an antiderivative of

$$\frac{3x^6-7x^5+9x^4-9x^3+7x^2-4x-1}{x^4-2x^3+2x^2-2x+1}.$$

As we have seen, this function can be expressed as

$$3x^2 - x + 1 + \frac{2}{x-1} - \frac{1}{(x-1)^2} - \frac{(x-1)}{x^2+1}.$$

Only the last of these terms requires effort. We have

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

and so the desired antiderivative of our rational function is

$$x^3 - \frac{x^2}{2} + x + 2\log(x-1) + \frac{1}{x-1} - \frac{1}{2}\log(x^2+1) + \arctan x.$$

Examples to work out

1. Find $\int \frac{2x^2+7x-1}{x^3+x^2-x-1}dx$
2. Find $\int \frac{2x+1}{x^3-3x^2+3x-1}dx$
3. Find $\int \frac{3x}{(x^2+x+1)^3}dx$
4. Find $\int \frac{dx}{x^4+1}$
5. Use the $t = \tan(x/2)$ substitution, and the method of partial fractions, to find antiderivatives for each of the following trigonometric functions:

- $\sec x$ (the answer you get is unlikely to be $\log(\sec x + \tan x)$, which is the expression that you are most likely to see if you look up a table of antiderivatives. Check that $\log(\sec x + \tan x)$ differentiates to $\sec x$, and also that the expression that you get is equal to $\log(\sec x + \tan x)$, perhaps up to an additive constant).
- $\sec^3 x$.

6. Using the "magic" substitution $t = \tan(x/2)$, and partial fractions, we see that every rational function of sin and cos has an elementary antiderivative. Show that also every rational function of e^x has an elementary antiderivative.

14 Taylor polynomials and Taylor's theorem

14.1 Definition of the Taylor polynomial

Suppose we know a lot about a function f at a point a (the function value, in some exact form, the value of the derivative, et cetera), but don't have such easy access to values away from a. Can we use the information we have at a to say *something* about the function away from a?

Examples of this type of situation include:

- $f(x) = \sin x$ at 0 (we know *everything* about the function at 0, in a quite exact way, but very little about it away from 0)
- $f(x) = \sin x$ at π, or $-\pi$, or $3\pi/2$, ...
- $f(x) = \log x$ at 1
- $f(x) = \sqrt{x^2+9}$ at 4, or 0, or -4.

There's an obvious, but next-to-useless, way to approximate f near a, using data at a — just use the constant function $f(a)$. A less obvious, and much more useful, way, is to use the linearization of f at a to approximate f near a, that is, to use the function

$$f(a) + f'(a)(x-a),$$

which has the property that it agrees with f at a, and also agrees with f' at a, so its graph agrees with the graph of f at a, and is also "traveling in the same direction" as the graph of f at a.

We can push this further: the function

$$f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2$$

is easily seen to agree with each of f, f' and f'' at a, and more generally the function

$$f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n$$

is easily seen to agree with each of $f, f', f'', \ldots, f^{(n)}$ at a.[228]

This example leads to the definition of the *Taylor polynomial*.

[228] One way to prove this formally is to prove by induction on k that for $n \geq k \geq 0$, the kth derivative of

$$f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n$$

is

$$\sum_{j=k}^{n} \frac{(j)_k}{j!}(x-a)^{j-k},$$

where $(j)_k$ is defined to be $j(j-1)(j-2)\cdots(j-k+1)$ ("j to the power k falling"). Evaluating at $x = a$ then gives that the kth derivative at a is $f^{(k)}(a)$.

Taylor polynomial of f at a of order n Suppose f is a function defined at and near a.

The *Taylor polynomial of f at a of order n* is

$$P_{n,a,f}(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n$$

We give some examples now, the details of which are left as exercises:

- $P_{n,0,\exp}(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!}$.
- $P_{2n+1,0,\sin}(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!}$.
- $P_{2n,0,\cos}(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots + (-1)^n \frac{x^{2n}}{(2n)!}$.
- $P_{n,1,\log}(x) = (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \frac{(x-1)^4}{4} + \cdots + (-1)^n \frac{(x-1)^n}{n}$.

The Taylor polynomial is not always easy to calculate. For example, consider $P_{n,0,\tan}$. We have

- $\tan 0 = 0$,
- $\tan' = \sec^2$, so $\tan' 0 = 1$,
- $\tan'' = (\sec^2)' = 2\sec^2 \tan$, so $\tan'' 0 = 0$,
- $\tan''' = (2\sec^2 \tan)' = 2\sec^4 + 4\sec^2 \tan^2$, so $\tan''' 0 = 2$,

and so $P_{3,0,\tan} = x + x^3/3$, but it does not seem very easy to continue.

14.2 Properties of the Taylor polynomial

An important property of the linearization $P_{1,a,f}$ of a function f at a (formerly denoted $L_{f,a}$) is that not only does $P_{1,a,f}(x) - f(x) \to 0$ as $x \to a$, but also

$$\frac{P_{1,a,f}(x) - f(x)}{x-a} = f'(a) - \frac{f(x)-f(a)}{x-a} \to 0 \quad \text{as} \quad x \to a$$

So the error in using $P_{1,a,f}$ to approximate f not only gets smaller as x gets closer to a, but gets smaller relative to $x-a$. But it is not necessarily the case that $(P_{1,a,f}(x)-f(x))/((x-a)^2)$ goes to zero as x approaches a. Consider, for example, the function $f(x) = x^2$ at $a = 0$, for which

$$\frac{P_{1,0,f}(x) - f(x)}{(x-a)^2} = -1 \not\to 0 \quad \text{as} \quad x \to a.$$

What about limit as $x \to a$ of

$$\frac{P_{2,a,f}(x) - f(x)}{(x-a)^2} = \frac{f(a) + f'(a)(x-a) + f''(a)(x-a)^2/2 - f(x)}{(x-a)^2}?$$

By L'Hôpital's rule, this limit is same as

$$\lim_{x\to a}\frac{f'(a)+f''(a)(x-a)-f'(x)}{2(x-a)},$$

if this limit exists; but

$$\frac{f'(a)+f''(a)(x-a)-f'(x)}{2(x-a)}=\frac{1}{2}\left(f''(a)-\frac{f'(x)-f'(a)}{x-a}\right),$$

which approaches 0 as $x\to a$, by the definition of the second derivative at a; so the original limit is 0. Note that $f(x)=x^3$, $a=0$, shows $(P_{2,a,f}(x)-f(x))/((x-a)^3)$ does not necessarily tend to 0.

This example leads to the following definition.

Definition of functions agreeing to order n A function g *agrees* with a function f *to order* n ($n\geq 0$ an integer) *at* a if both g and f are defined near a and if

$$\lim_{x\to a}\frac{g(x)-f(x)}{(x-a)^n}$$

exists and equals 0.

We use the shorthand $g\sim_{n,a} f$ to denote that g agrees with f to order n at a.

Note that if $g\sim_{n,a} f$ then automatically $f\sim_{n,a} g$, so it is legitimate to say "f and g agree to order n at a". Note also that if $f\sim_{n,a} g$ then $f\sim_{m,a} g$ for all $0\leq m<n$, since

$$\lim_{x\to a}\frac{g(x)-f(x)}{(x-a)^m}=\lim_{x\to a}(x-a)^{n-m}\frac{g(x)-f(x)}{(x-a)^n}=0,$$

although it is not necessarily the case that $f\sim_{m,a} g$ for any $m>n$ (as some earlier examples show). Finally, note that if $f\sim_{n,a} g$ and if $g\sim_{n,a} h$ then it follows that $f\sim_{n,a} g$. Indeed:

$$\frac{f(x)-h(x)}{(x-a)^n}=\frac{f(x)-g(x)}{(x-a)^n}+\frac{g(x)-h(x)}{(x-a)^n},$$

the right-hand side above tends at 0 as $x\to a$, so the left-hand side does also.

The example that lead to this definition strongly suggests that the Taylor polynomial $P_{n,a,f}$ agrees with f to order n. That's the content of the next theorem.

Theorem 14.1. *Suppose f is a function such that each of $f, f', f'', \ldots, f^{(n)}$ exist at a. Then $P_{n,a,f}\sim_{n,a} f$ (but is is not necessarily the case that $P_{n,a,f}\sim_{n+1,a} f$).*

Proof: Consider $f(x)=x^{n+1}$ at $a=0$ to see that we may not have agreement to order $n+1$.

For first part we want to show that

$$\lim_{x\to a}\frac{f(a)+f'(a)(x-a)+\cdots+\frac{f^{(n)}(a)}{n!}(x-a)^n-f(x)}{(x-a)^n}=0.\ (\star)$$

Set

$$P_n(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n - f(x)$$

and $Q_n(x) = (x-a)^n$. It is an easy check that of the following limits exist and equal 0:

- $\lim_{x\to a} P_n(x), \lim_{x\to a} P_n'(x), \ldots, \lim_{x\to a} P_n^{(n-2)}(x)$ and
- $\lim_{x\to a} Q_n(x), \lim_{x\to a} Q_n'(x), \ldots, \lim_{x\to a} Q_n^{(n-2)}(x)$.

So, applying L'Hôpital's rule $n-1$ times, we get that the limit in $(\star)$ exists if

$$\lim_{x\to a} \frac{P_n^{(n-1)}(x)}{Q_n^{(n-1)}(x)} = \lim_{x\to a} \frac{f^{(n-1)}(a) + f^{(n)}(a)(x-a) - f^{(n-1)}(x)}{n!(x-a)}$$

exists; but

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

which exists and equals 0 by the definition of the nth derivative; so the limit in $(\star)$ exists and equals 0. □

So, the Taylor polynomial of f of degree n at a agrees with f to order n at a. Does this property *characterize* the Taylor polynomial, among polynomials of degree n? Essentially, "yes", as we now see. First we need some notation. Say that Q *is a polynomial of degree at most* n *in* $x-a$ if

$$Q(x) = a_0 + a_1(x-a) + a_2(x-a)^2 + \cdots + a_n(x-a)^n$$

(where a_n is not necessarily non-zero).[229]

Theorem 14.2. *Suppose that f is a function that is n times differentiable at a[230], and that Q is a polynomial of degree at most n in $x-a$ that agrees with f at a to order n. Then $Q = P_{n,a,f}$ (so the degree n Taylor polynomial of f at a is the* unique *polynomial of degree at most n that agrees with f to order n at a).*

[229] For every polynomial of degree m, and for every real a, the polynomial can be expressed as a polynomial of degree m in $x-a$. This is a future-fact — it comes from Linear Algebra. It's easy to see why it is true, though. Here's an example, of expressing an ordinary quadratic polynomial a quadratic polynomial in x) as a quadratic polynomial in $x-1$:

$$x^2 - 4x + 5 = (x-1)^2 - 2x + 4 = (x-1)^2 - 2(x-1) + 2.$$

The trick is to work from the higher powers down.

[230] This hypothesis is necessary. It is *not* true that if f is a function defined at at near a, and if Q is a polynomial of degree at most n in $x-a$ that agrees with f to order n at a, then $Q = P_{n,a,f}$. The issue is that although Q may agree with f to order n at a, it may not be the case that f has the necessary derivatives existing to have a Taylor polynomial. (Spivak gives a specific example in his text.)

This will be a corollary of the following lemma.

Lemma 14.3. *If P and Q are polynomials of degree at most n, and $P \sim_{n,a} Q$, then $P = Q$.*

To see that Theorem 14.2 follows from this, note that

- Q (in the statement of Theorem 14.2) agrees with f to order n at a (by hypothesis of Theorem 14.2),
- $P_{n,a,f}$ agrees with f to order n at a (by Theorem 14.1), so
- Q agrees with $P_{n,a,f}$ to order n at a (as discussed earlier), and so
- $Q = P_{n,a,f}$ (by Lemma 14.3).

Proof (of Lemma 14.3): Set $R = P - Q$, so

$$\frac{R(x)}{(x-a)^n} \to 0 \quad (\star)$$

as $x \to a$. Write $R(x) = r_0 + r_1(x-a) + \cdots + r_n(x-a)^n$.

From $(\star)$ it follows that

$$\frac{R(x)}{(x-a)^i} \to 0 \quad (\star\star)$$

as $x \to a$, for each $i = 0, \ldots, n$.

We have $R(x) \to r_0$ as $x \to a$; but considering $(\star\star)$ at $i = 0$, we get also $R(x) \to 0$ as $x \to a$. So $r_0 = 0$.

From this it follows that $R(x)/(x-a) \to r_1$ as $x \to a$; but considering $(\star\star)$ at $i = 1$, we get $R(x)/(x-a) \to 0$ as $x \to a$. So $r_1 = 0$.

Continuing in this many, we get that $r_i = 0$ for all i, and so $R = 0$ and $P = Q$. □

This theorem suggests an alternate approach to finding Taylor polynomials: if f is n times differentiable at a, and we can someone guess or intuit a polynomial of degree n around a that agrees with f to order n at a, then that polynomial must by the Taylor polynomial of order n at a of f.

Here's an example. Consider $\tanh^{-1} x$ (recall that $\tanh x = (e^x - e^{-x})/(e^x + e^{-x})$), a function with domain $\mathbb{R}$ and range $(-1, 1)$. We know (or can derive) that

$$(\tanh^{-1})'(x) = \frac{1}{1-x^2},$$

so

$$(\tanh^{-1})''(x) = \frac{2x}{(1-x^2)^2} \quad \text{and} \quad (\tanh^{-1})'''(x) = \frac{6x^2+2}{(1-x^2)^3},$$

so $P_{3,0,\tanh^{-1}}(x) = x + x^3/3$. It does not seem like it will be a very pleasant task to continue calculating Taylor polynomials via derivatives!

But we also have

$$\begin{aligned}\tanh^{-1} x &= \int_0^x \frac{dt}{1-t^2} \\ &= \int_0^x \left(1 + t^2 + t^4 + \cdots + t^{2n} + \frac{t^{2n+2}}{1-t^2}\right)\, dt \\ &= x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots + \frac{x^{2n+1}}{2n+1} + \int_0^x \frac{t^{2n+2}}{1-t^2}\, dx. \quad (\star)\end{aligned}$$

If we can show

$$\lim_{x\to 0} \frac{1}{x^{2n+1}} \int_0^x \frac{t^{2n+2}}{1-t^2} dx = 0 \quad (\star\star)$$

then that would exactly say (via $(\star)$) that the polynomial $\sum_{k=0}^n x^{2k+1}/(2k+1)$ agrees with $\tanh^{-1}(x)$ to order $2n+1$ at 0, and so is the Taylor polynomial $P_{2n+1,0,\tanh^{-1}}(x)$.

Using the evenness of the integrand, we have that for $|x| < 1/2$

$$\begin{aligned}\left|\int_0^x \frac{t^{2n+2}}{1-t^2}\, dx\right| &= \int_0^{|x|} \frac{t^{2n+2}}{1-t^2}\, dx \\ &\leq \frac{4}{3}\int_0^{|x|} t^{2n+2} dx \\ &= \frac{4|x|^{2n+3}}{3(2n+3)}\end{aligned}$$

and so indeed $(\star\star)$ holds[231] and

$$P_{2n+1,0,\tanh^{-1}}(x) = x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots + \frac{x^{2n+1}}{2n+1}.$$

We can do a little better than this. Let $x \in (-1,1)$ be fixed (note that $(-1,1)$ is the domain of $\tanh^{-1}$). Arguing as above we have

$$\left|\int_0^x \frac{t^{2n+2}}{1-t^2}\, dx\right| \leq \frac{|x|^{2n+3}}{(2n+3)(1-x^2)} \to 0 \quad \text{as } n \to \infty.$$

A corollary of this calculation is that for all $x \in (-1,1)$ we have

$$\left|\tanh^{-1}(x) - P_{2n+1,0,\tanh^{-1}}(x)\right| \leq \frac{4|x|^{2n+3}}{3(2n+3)}.$$

So we can estimate $\tanh^{-1}(x)$ for any particular $x \in (-1,1)$, to any accuracy, by using $P_{2n+1,0,\tanh^{-1}}(x)$ for large enough n. This hints at the main point of what we are about to do: if we can estimate the *difference* between $f(x)$ and $P_{n,a,f}(x)$ (perhaps just for x in some little interval around a), then we can have the potential to use the Taylor polynomial as a reliable way to estimate the function that its the Taylor polynomial of.

[231] Here's another approach: applying L'Hôpital's rule (and the fundamental theorem of calculus) to $(\star\star)$ we find that the limit exists and equals

$$\lim_{x\to 0}\left(\frac{1}{(2n+1)x^{2n}}\right)\left(\frac{x^{2n+2}}{1-x^2}\right) = \lim_{x\to 0}\frac{x^2}{(2n+1)(1-x^2)},$$

as long as these limits exist. But the last limit evidently exists and equals 0.

14.3 Taylor's theorem and remainder terms

Definition of remainder term If f is a function and $P_{n,a,f}$ exists, then the *remainder term* $R_{n,a,f}(x)$ is defined by

$$f(x) = P_{n,a,f}(x) + R_{n,a,f}(x).$$

Our goal for the next while is to find good estimates for $R_{n,a,f}(x)$, that allow us to say that for x sufficiently close to a, $R_{n,a,f}(x) \to 0$ as $n \to \infty$ (so the Taylor polynomial at a is a good approximation for f near a). For example, as we have already discussed, if $x \in [-1/2, 1/2]$ is fixed then

$$R_{2n+1,0,\tanh^{-1}}(x) = \int_0^x \frac{t^{2n+2}}{1-t^2}\,dx \le \frac{4|x|^{2n+3}}{3(2n+3)} \to 0$$

as $n \to \infty$.[232]

In what follows, we slowly derive Taylor's Theorem with integral remainder term (Theorem 14.4 below). We won't both to mention explicitly the assumptions we are making; those will be stated explicitly in the the theorem, and will easily be seen to be exactly the hypothesis needed to make the argument we are about to describe work.

Let f be a function, with a and x two *fixed* points of the domain of f.[233] We assume $x \neq a$, since it is rather trivial to understand $R_{n,a,f}(a)$. From the fundamental theorem of calculus we have

$$f(x) - f(a) = \int_a^x f'(t)\,dt \quad \text{or} \quad f(x) = f(a) + \int_a^x f'(t)\,dt \text{or} \quad f(x) = P_{0,a,f}(x) + \int_a^x f'(t)\,dt$$

which says that $R_{0,a,f}(x)$ can be expressed as $\int_a^x f'(t)\,dt$.

Now we apply integration by parts to $\int_a^x f'(t)\,dt$, taking

$$\begin{aligned} u = f'(t) \quad &\text{so} \quad du = f''(t)\,dt \\ dv = dt \quad &\text{so} \quad v = t - x. \end{aligned}$$

Notice here that we are *not* taking $v = t$, the obvious choice for an antiderivative of 1. We could, but it would lead us nowhere. Instead we are taking another, non-obvious but equally correct (because x is just some fixed constant), antiderivative; as we will see in a moment, it

[232] Note that we have made a subtle change in viewpoint: we are thinking now of x as being fixed (some number close to a), and thinking about n growing, rather than thinking about n as being fixed with x approaching a.

[233] It is critical that a and x are both consider to be fixed here. Think of a as a point at which we know a lot about f, and of x as some other point, perhaps close to a.

is this non-obvious choice that drives the proof of Taylor's theorem. We get

$$\begin{aligned} f(x) &= f(a) + \int_a^x f'(t)\ dt \\ &= f(a) + [(t-x)f'(t)]_{t=a}^x - \int_a^x (t-x)f''(t)\ dt \\ &= f(a) + (x-a)f'(a) + \int_a^x (x-t)f''(t)\ dt \\ &= P_{1,a,f}(x) + \int_a^x (x-t)f''(t)\ dt. \end{aligned}$$

This says that $R_{1,a,f}(x)$ can be expressed as $\int_a^x (x-t)f''(t)\ dt$.

Now we apply integration by parts to $\int_a^x (x-t)f''(t)\ dt$, taking

$$\begin{aligned} u = f''(t) \quad \text{so} \quad du = f'''(t)\ dt \\ dv = (x-t)dt \quad \text{so} \quad v = \frac{-(x-t)^2}{2}. \end{aligned}$$

Notice here that we *are* taking the obvious choice for antiderivative of 1; as we will in all subsequent applications in this proof. We get

$$\begin{aligned} f(x) &= f(a) + (x-a)f'(t) + \int_a^x (x-t)f''(t)\ dt \\ &= f(a) + (x-a)f'(t) + \left[\frac{-(x-t)^2}{2}f''(t)\right]_{t=a}^x + \int_a^x \frac{(x-t)^2}{2}f'''(t)\ dt \\ &= f(a) + (x-a)f'(t) + \frac{(x-a)^2 f''(a)}{2} + \int_a^x \frac{(x-t)^2}{2}f'''(t)\ dt \\ &= P_{2,a,f}(x) + \int_a^x \frac{(x-t)^2}{2}f'''(t)\ dt. \end{aligned}$$

This says that $R_{2,a,f}(x)$ can be expressed as $\int_a^x \frac{(x-t)^2}{2}f'''(t)\ dt$.

We try this one more time. We apply integration by parts to $\int_a^x \frac{(x-t)^2}{2}f'''(t)\ dt$, taking

$$\begin{aligned} u = f'''(t) \quad \text{so} \quad du = f''''(t)\ dt \\ dv = \frac{(x-t)^2}{2}\ dt \quad \text{so} \quad v = \frac{-(x-t)^3}{3!}. \end{aligned}$$

We get

$$\begin{aligned} f(x) &= +\int_a^x \frac{(x-t)^2}{2}f'''(t)\ dt \\ &= f(a) + (x-a)f'(t) + \frac{(x-a)^2 f''(a)}{2} + \left[\frac{-(x-t)^3}{3!}f'''(t)\right]_{t=a}^x + \int_a^x \frac{(x-t)^3}{3!}f'''(t)\ dt \\ &= f(a) + (x-a)f'(t) + \frac{(x-a)^2 f''(a)}{2} + \frac{(x-a)^3 f'''(a)}{3!} + \int_a^x \frac{(x-t)^3}{3!}f'''(t)\ dt \\ &= P_{3,a,f}(x) + \int_a^x \frac{(x-t)^3}{3!}f'''(t)\ dt. \end{aligned}$$

This says that $R_{3,a,f}(x)$ can be expressed as $\int_a^x \frac{(x-t)^3}{3!} f''''(t)\ dt$.

An obvious pattern is emerging, and we can verify it by induction. Suppose, for some k, we have shown that

$$f(x) = P_{k,a,f}(x) + \int_a^x \frac{(x-t)^k}{k!} f^{(k+1)}(t)\ dt.$$

We apply integration by parts to $\int_a^x \frac{(x-t)^k}{k!} f^{(k+1)}(t)\ dt$, taking

$$\begin{aligned} u = f^{(k+1)}(t) \quad &\text{so} \quad du = f^{(k+2)}(t)\ dt \\ dv = \frac{(x-t)^k}{k!}\ dt \quad &\text{so} \quad v = -\frac{(x-t)^{k+1}}{(k+1)!}. \end{aligned}$$

We get

$$\begin{aligned} f(x) &= P_{k,a,f}(x) + \int_a^x \frac{(x-t)^k}{k!} f^{(k+1)}(t)\ dt \\ &= P_{k,a,f}(x) + \left[\frac{-(x-t)^{k+1}}{(k+1)!} f^{(k+1)}(t)\right]_{t=a}^{x} + \int_a^x \frac{(x-t)^{(k+1)}}{(k+1)!} f^{(k+2)}(t)\ dt \\ &= P_{k,a,f}(x) + \frac{(x-a)^{k+1}}{(k+1)!} f^{(k+1)}(t) + \int_a^x \frac{(x-t)^{(k+1)}}{(k+1)!} f^{(k+2)}(t)\ dt \\ &= P_{k+1,a,f}(x) + \int_a^x \frac{(x-t)^{(k+1)}}{(k+1)!} f^{(k+2)}(t)\ dt. \end{aligned}$$

This says that $R_{k+1,a,f}(x)$ can be expressed as $\int_a^x \frac{(x-t)^{(k+1)}}{(k+1)!} f^{(k+2)}(t)\ dt$.

We have proven the following important theorem:

Theorem 14.4. (Taylor's theorem with integral remainder term) *Suppose $f, f', \ldots, f^{(n+1)}$ are all defined on an interval that includes a and x, and that $f^{(n+1)}$ is integrable on that interval. Then*

$$f(x) = P_{n,a,f}(x) + \frac{1}{n!} \int_a^x (x-t)^n f^{(n+1)}(t)\ dt.$$

That is,

$$R_{n,a,f}(x) = \frac{1}{n!} \int_a^x (x-t)^n f^{(n+1)}(t)\ dt.$$

There is another form of the remainder term that is usually much easier to work with. Suppose that $f^{(n+1)}(t)$ is continuous on the closed interval I that has a and x as endpoints[234]. Then, by the extreme value theorem, there are numbers $m < M$ such that

$$m \leq f^{(n+1)}(t) \leq M$$

[234]We write this, rather than the more natural "on the interval $[a, x]$", to allow for the possibility that $x < a$.

for $t \in I$, and moreover there are numbers $t_1, t_2 \in I$ with $f^{(n+1)}(t_1) = m$ and $f^{(n+1)}(t_2) = M$. If $x > a$ we have that on I

$$m(x-t)^n \leq (x-t)^n f^{(n+1)}(t) \leq M(x-t)^n$$

so, integrating,

$$\frac{m(x-a)^{n+1}}{(n+1)!} \leq R_{n,a,f}(x) \leq \frac{M(x-a)^{n+1}}{(n+1)!}$$

or

$$m \leq \frac{(n+1)!R_{n,a,f}(x)}{(x-a)^{n+1}} \leq M.$$

By the intermediate value theorem, there is some c between t_1 and t_2 (and so between a and x) with

$$f^{(n+1)}(c) = \frac{(n+1)!R_{n,a,f}(x)}{(x-a)^{n+1}}$$

or

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

We can use a similar argument to reach the same conclusion, when $x < a$. We summarize in the following theorem, as important as Theorem 14.4.

Theorem 14.5. (Taylor's theorem with Largrange remainder term, weak form[235]) *Suppose $f, f', \ldots, f^{(n+1)}$ are all defined on an interval that includes a and x, and that $f^{(n+1)}$ is continuous on that interval. Then there is some number c (strictly) between a and x such that*

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

That is,

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

14.4 Examples

Example 1, sin at 0 We illustrate the use of Theorem 14.5 with the example of the function $f(x) = \sin x$, at $a = 0$. Fix $x \in \mathbb{R}$. Recall that we have

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + R_{2n+1,0,\sin}(x).$$

The Lagrange form of the remainder term is

$$|R_{2n+1,0,\sin}(x)| = \left| \frac{\sin^{(2n+2)}(c)x^{n+1}}{(n+1)!} \right| \leq \frac{|x|^{2n+2}}{(2n+2)!},$$

[235] Why is this the *weak form*? Because this theorem is also true, without the hypothesis that $f^{(n+1)}$ is continuous. However, since in every example that we will see, we will have continuity of the $(n+1)$st derivative, we will not discuss the stronger form here.

where c is some number between 0 and x. The inequality above follows from the fact that $|\sin^{(2n+2)}(c)| = |\sin(c)| \leq 1$, regardless of the values of n and c.

The continue the analysis, we need the following lemma, that will be extremely useful for many other applications, that says that the factorial function grows faster than any power function.

Lemma 14.6. *For each $x > 0$ and $\varepsilon > 0$, for all sufficiently large n we have*

$$\frac{x^n}{n!} < \varepsilon.$$

Proof: Pick any integer $n_0 > 2x$. We have that for $n > n_0$,

$$\frac{x^n}{n!} \leq \frac{(n_0/2)^n}{n_0^{n-n_0}(n_0-1)!} = \frac{n_0^{n_0}}{(n_0-1)!}\frac{1}{2^n}.$$

Noting that $(n_0^{n_0}/(n_0-1)!)$ is just a constant, we can make $1/(2^n) < \varepsilon/(n_0^{n_0}/(n_0-1)!)$, so $x^n/n! < \varepsilon$, for all sufficiently large n.

Alternately: if n is even, then the largest $n/2$ terms in the product $n!$ are all bigger than $n/2$, so $n! > (n/2)^{n/2}$ while if n is odd, then the largest $(n+1)/2$ terms in $n!$ are all bigger than $n/2$, so $n! > (n/2)^{(n+1)/2}$. Either way, for all n

$$n! > \left(\frac{n}{2}\right)^{n/2} = \left(\frac{\sqrt{n}}{\sqrt{2}}\right)^n,$$

so

$$\frac{x^n}{n!} < \left(\frac{x\sqrt{2}}{\sqrt{n}}\right)^n.$$

For all $n \geq 8x^2$ we therefore have

$$\frac{x^n}{n!} < \left(\frac{1}{2}\right)^n.$$

Since $(1/2)^n$ can be made smaller than ε by choosing n sufficiently large, so too can $x^n/n!$.

□

An immediate corollary is that for each real x,

$$R_{2n+1,0,\sin}(x) \to 0 \quad \text{as } n \to \infty,$$

and so for each real x

$$P_{2n+1,0,\sin}(x) \to \sin x \quad \text{as} \quad n \to \infty.$$

For example, to estimate $\sin 355$ to within $\pm.001$, we simply choose n large enough that $355^{2n+2}/((2n+2)!) < .001$, and then calculate $P_{2n+1,0,\sin}(355)$. A `Mathematica` calculation tells us that $n = 483$ is sufficiently large, and that $P_{2(483)+1,0,\sin}(355) = -0.000233397\ldots$, so we conclude

$$\sin 355 = -0.000233397\ldots \pm 0.001.$$

(In fact $\sin 355 = -0.0000301444\cdots$[236]).

Given that $\sin 355$ is so close to zero, it is rather remarkable that the sequence $(P_{2n+1,0,\sin}(355))_{n\geq 0}$ starts $(355, -7\times 10^6, 4\times 10^{10}, \ldots)$, and that along the way to the term $P_{967,0,\sin}(355)$, the sequence rises up to as large as 1.6×10^{152}!

Example 2, cos **at 0** By an almost identical argument to the one used for sin, we find that for all real x,

$$P_{2n,0,\cos}(x) \to \cos x \quad \text{as} \quad n\to\infty.$$

Example 3, exp We have, for each fixed x,

$$P_{n,0,\exp}(x) = 1 + x + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!}$$

with (Lagrange form of the remainder, c between 0 and x)

$$R_{n,0,\exp}(x) = \frac{\exp(c)}{(n+1)!}x^{n+1}$$

so that, using the fact that exp is an increasing function.

$$|R_{n,0,\exp}(x)| \leq e^{\max\{0,x\}}\frac{|x|^{n+1}}{(n+1)!}$$

From Lemma 14.6 we get that

$$R_{n,0,\exp}(x) =\to 0 \quad \text{as } n\to\infty$$

and, as with sin, this is valid for *all* real x, so the Taylor polynomial of exp at 0 can be used to estimate $\exp^x$ to arbitrary precision for all x; that is,

$$P_{n,0,\exp}(x) \to \exp x \quad \text{as} \quad n\to\infty.$$

As an illustrative example we estimate e^{-1}. Setting $x=-1$ we have

$$|R_{n,0,\exp}(-1)| \leq \frac{1}{(n+1)!}.$$

[236]Why is this so close to zero? It's because 355 is almost an integer multiple of *pi*; in fact, $355 \approx 113\pi = 354.9999698556467\ldots$. That begs the question, "why is 355/113 such a good approximation to π? That gets into the theory of continued fractions.

Since $1/11! = 0.000000025\ldots$ we get that

$$P_{10,0,\exp}(-1) = \frac{16481}{44800} = 0.367879464\ldots$$

is an approximation of $1/e$ accurate to ± 0.000000025. (In fact $1/e = 0.367879441\ldots$.)

Example 4, $f(x) = e^{-1/x^2}$ **at** 0 Consider the function f defined by

$$f(x) = \begin{cases} e^{-1/x^2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

We studied this function in some detail, as an example of the application of Lemma 12.6. We proved there that f is differentiable arbitrarily many times at 0, and that all derivatives at 0 are 0. It follows that

$$P_{n,0,f}(x) = 0$$

for all n (for all x; $P_{n,0,f}$ is the identically 0 polynomial). It follows that

$$R_{n,0,f}(x) = f(x)$$

for all n and x. So: for $x = 0$ we have $R_{n,0,f}(x) \to 0$ as $n \to \infty$ (trivially), but for $x \neq 0$ we have

$$R_{n,0,f}(x) = e^{-x^2/2} \not\to 0 \quad \text{as } n \to \infty.$$

In this example, the Taylor polynomial is useless as an approximation tool.

Example 5, $\tan^{-1}$ **at** 0 Consider $\tan^{-1} : \mathbb{R} \to (-\pi/2, \pi/2)$. Proceeding as we did for $\tanh^{-1}$, we have

$$\begin{aligned} \tan^{-1}(x) &= \int_0^x \frac{1}{1+t^2} dt \\ &= x - \frac{x^3}{3} + \cdots + \frac{(-1)^n x^{2n+1}}{2n+1} + \int_0^x \frac{(-1)^{n+1} t^{2n+2}}{1+t^2} dt. \end{aligned}$$

Is $P_{2n+1,0,\tan^{-1}}(x) = \sum_{k=0}^{n} \frac{(-1)^k x^{2k+1}}{2k+1}$? Yes, because for all x,

$$\begin{aligned} \left| \frac{1}{x^{2n+1}} \int_0^x \frac{(-1)^{n+1} t^{2n+2}}{1+t^2} dt \right| &= \frac{1}{|x|^{2n+1}} \int_0^{|x|} \frac{t^{2n+2}}{1+t^2}\, dt \\ &\leq \frac{1}{|x|^{2n+1}} \int_0^{|x|} t^{2n+2}\, dt \\ &= \frac{x^2}{2n+3} \end{aligned}$$

which goes to 0 as x goes to 0, and so the degree $2n+1$ polynomial we have found agrees with $\tan^{-1}$ to order $2n+1$ at 0.

The calculation we have just done shows that

$$|R_{2n+1,0,\tan^{-1}}(x)| \leq \frac{|x|^{2n+3}}{2n+3}.$$

As long as $x \in [-1,1]$, we therefore have $R_{2n+1,0,\tan^{-1}}(x) \to 0$ as $n \to \infty$, and so in this range $P_{2n+1,0,\tan^{-1}}(x) \to \tan^{-1}(x)$.

What about when $|x| > 1$? We claim that here, $|R_{2n+1,0,\tan^{-1}}(x)| \not\to 0$. We first consider positive x. If $|R_{2n+1,0,\tan^{-1}}(x)| \to 0$ in this case, then, since

$$\begin{aligned}\left|\int_1^x \frac{(-1)^{n+1}t^{2n+2}}{1+t^2}\,dt\right| &= \left|\int_0^x \frac{(-1)^{n+1}t^{2n+2}}{1+t^2}\,dt - \int_0^1 \frac{(-1)^{n+1}t^{2n+2}}{1+t^2}\,dt\right| \\ &\leq \left|\int_0^x \frac{(-1)^{n+1}t^{2n+2}}{1+t^2}\,dt\right| + \left|\int_0^1 \frac{(-1)^{n+1}t^{2n+2}}{1+t^2}\,dt\right|\end{aligned}$$

and each of $\int_0^x \frac{(-1)^{n+1}t^{2n+2}}{1+t^2}\,dt$, $\int_0^1 \frac{(-1)^{n+1}t^{2n+2}}{1+t^2}\,dt \to 0$ as $n \to \infty$, we would have

$$\int_1^x \frac{(-1)^{n+1}t^{2n+2}}{1+t^2}\,dt \to 0.$$

But now, on the interval $[1,x]$ we have

$$\frac{(-1)^{n+1}t^{2n+2}}{1+t^2} \geq \frac{1}{1+x^2}$$

so that

$$\text{either} \quad \int_1^x \frac{(-1)^{n+1}t^{2n+2}}{1+t^2}\,dt \geq \frac{1}{1+x^2} \quad \text{or} \quad \int_1^x \frac{(-1)^{n+1}t^{2n+2}}{1+t^2}\,dt \leq \frac{-1}{1+x^2}$$

(depending on whether n is odd or even), and so it cannot possibly be that the integral tends to 0 as n grows. We conclude that

$$P_{2n+1,0,\tan^{-1}}(x) \to \tan^{-1}(x) \quad \textit{only} \text{ on the interval } [-1,1].$$

The upshot of these examples, is that it seems that for each f and $a \in \text{Domain}(f)$, there is a range of x around a for which, for each fixed x in that range, $R_{n,a,f}(x)$ goes to 0 as n goes to infinity, and so for which $P_{n,a,f}(x)$ approaches $f(x)$ as n gets large.

- In the case of $\sin, \cos, \exp$ at 0, that range is all of $\mathbb{R}$.
- In the case of $\tanh^{-1}$ at 0, that range is $(-1,1)$, which coincides with the domain of $\tanh^{-1}$.
- In the case of $\tan^{-1}$ at 0, that range is goes from -1 to 1 which again includes only a portion of the domain.

- In the case of e^{-1/x^2} at 0, that range includes *only* the point 0.

That $P_{n,0,\exp}(x)$ approaches e^x as n gets large, for all real x, suggests that we can meaningfully write something like

$$\begin{aligned} e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} \cdots \\ &= \sum_{n=0}^{\infty} \frac{x^n}{n!} \end{aligned}$$

for all $x \in \mathbb{R}$, and that $P_{n,0,\tan^{-1}}(x)$ approaches $\tan^{-1} x$ as n gets large, for all $x \in (-1, 1)$ suggests that the equation

$$\tan^{-1}(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots$$

is meaningful for all $x \in [-1, 1]$.

The goal of rest of these notes is to study infinite sequences and series, to make what we have just discussed precise.

15 Sequences

15.1 Introduction to sequences

Formally an *infinite sequence* is a function $a : \mathbb{N} \to \mathbb{R}$, Informally, a sequence[237] is an ordered list of real numbers:

$$(a_1, a_2, a_3, \ldots) \quad \text{or} \quad (a_k)_{k=1}^{\infty}.$$

We sometimes even just write (a_n), if it clear from the context that this is representing a sequence. Some remarks:

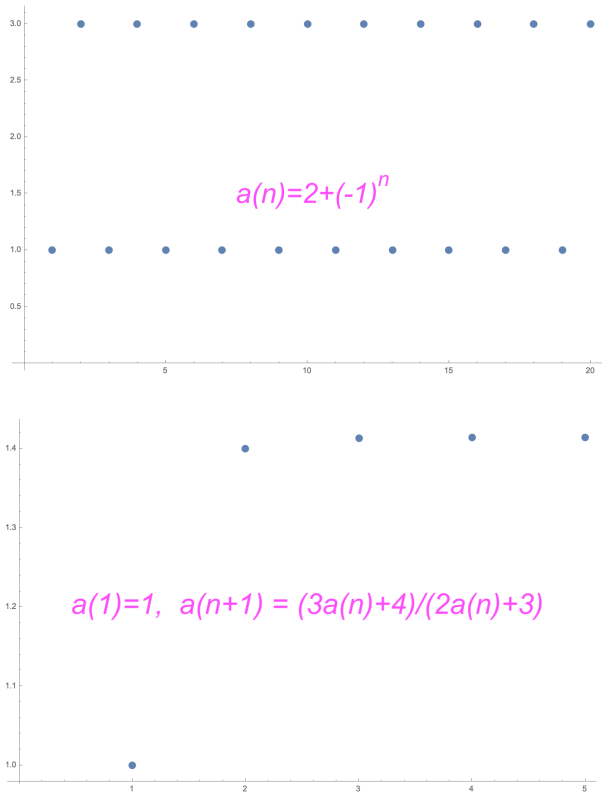
- The number a_k (formally, the image of k under the map a), is called the kth *term* of the sequence. Notice that we write a_k rather than $a(k)$; this is a tradition, but not a requirement.
- Spivak writes $\{a_1, a_2, a_3, \ldots\}$ or $\{a_k\}_{k=1}^{\infty}$. I much prefer "$(\cdots)$" to "$\{\cdots\}$"; because we use "$\{\cdots\}$" for a *set* of elements, this notation might incorrectly convey (at a subconscious level) that the order of the elements in a sequence doesn't matter.
- A sequence doesn't necessarily have to start at element a_1; it will be useful to allow sequences of the form $(a_k, a_{k+1}, a_{k+2}, \ldots)$ (denoted also $(a_j)_{j=k}^{\infty}$) for arbitrary integers k. In particular we will very frequently work with sequences of the form $(a_0, a_1, a_2, \ldots)$.

We give some examples, mostly to indicate different ways that sequences might be presented:

- $a : \mathbb{N} \to \mathbb{R}$, $a(k) = 2 + (-1)^k$.
- $a_n = n^2 + 1$, $n = 0, 1, 2, \ldots$.
- $(2, 3, 5, 7, 11, 13, 17, \ldots)$. (This is a very typical way to present a sequence — list only a few terms, and let the pattern speak for itself. But it should be used with caution. Is the pattern *really* obvious? In this case, I think that the answer is *no*, since the sequence I'm thinking of dies *not* have 19 as its next element.)
- $a_n = \sum_{k=1}^{n} x^k/k!$. (Notice that this is a *family* of sequences, one for each $x \in \mathbb{R}$.)
- $a_1 = 1$, and, for $n \geq 1$, $a_{n+1} = \frac{3a_n+4}{2a_n+3}$. (This is a *recursive* (or *recursively defined*) sequence.)

A sequence can be graphically represented: here are some examples:

[237] From here on we will typically say "sequence" rather than "infinite sequence".



As illustrated by the last two examples, a sequence can exhibit very different behaviors as the terms get larger.

15.2 Convergence

Definition of a sequence converging A sequence $(a_n)_{n=1}^{\infty}$ *converges to a limit* L *as* n *approaches infinity*, written

- $(a_n) \to L$ as $n \to \infty$
- $a_n \to L$ as $n \to \infty$
- $(a_1, a_2, \ldots) \to L$ as $n \to \infty$
- $\lim_{n\to\infty} a_n = L$,

if for all $\varepsilon > 0$ there exists n_0 such that

$$n > n_0 \quad \text{implies} |a_n - L| < \varepsilon.$$

If a sequence converges to a limit L as n approaches infinity then it is said to be a *convergent* sequence. If a sequence does not converge to a limit, then it is said to *diverge*, or be a *divergent* sequence.

Definition of a sequence converging to ∞ A divergent sequence $(a_n)_{n=1}^{\infty}$ *converges to* ∞ *as* n *approaches infinity*, written

- $(a_n) \to \infty$ as $n \to \infty$
- $a_n \to \infty$ as $n \to \infty$
- $(a_1, a_2, \ldots) \to \infty$ as $n \to \infty$
- $\lim_{n\to\infty} a_n = \infty$,

if for all $N > 0$ there exists n_0 such that

$$n > n_0 \quad \text{implies} \quad a_n > N.$$

The definition of a sequence converging to $-\infty$ is analogous.

A note on notation: any way of notating the concept that the function f approaches a limit near a, it is necessary to include some reference to the parameter a; but when notating the concept that the sequence (a_n) converges to a limit as n approaches infinity, it is usually unnecessary to include reference to the fact that n is approaching infinity; usually it will be perfectly clear from the context that the only place that n can go is to infinity. So we will often write simply:

- $(a_n) \to \infty$
- $a_n \to \infty$,
- $(a_1, a_2, \ldots) \to \infty$.

We illustrate the concept of convergence with three examples:

- $\left(\frac{n-1}{n+1}\right)_{n\geq 1}$. Evidently this converges to 1 as $n \to \infty$. To prove this formally, note that for each $\varepsilon > 0$ we require an n_0 such that $n > n_0$ implies

$$\left|\frac{n-1}{n+1} - 1\right| < \varepsilon.$$

This is equivalent to

$$\left|\frac{(n+1)-2}{n+1} - 1\right| < \varepsilon$$

or

$$\frac{2}{n+1} < \varepsilon$$

or

$$n > \frac{2}{\varepsilon} - 1.$$

So taking $n_0 = (2/\varepsilon) - 1$ we get that the sequence converges to 1 as $n \to \infty$.

- $(n^2+1)_{n=1}^{\infty}$. This evidently diverges, and tends to ∞. To verify this formally, we need to show that for each N, there is n_0 such that $n > n_0$ implies $n^2 + 1 > N$. If $N \leq 1$, simply take $n_0 = 1$; if $N > 1$ take n_0 to be anything greater than $\sqrt{N-1}$.

- $(2+(-1)^k)_{k\geq 1}$. This evidently diverges. To see this formally, suppose that it converges to a limit L. Whatever L is, it must be at distance at least 1 from at least one of $1, 3$. Suppose it is distance at least 1 from 3. Take $\varepsilon = 1/2$. By the assumption that the limit is L, there is n_0 such that for all $n > n_0$, $2+(-1)^n$ is within $1/2$ of L. This implies that for all $n > n_0$, $2+(-1)^n$ cannot take the value 3 (since 3 is not within $1/2$ of L, by assumption). But this is a contradiction, since $2+(-1)^n$ takes the value 3 for all even n. We get a similar contradiction under the assumption that the distance from L to 1 is at least 1. So the assumption that the sequence converges to a limit L is untenable, regardless of the choice of L, and $(2+(-1)^k)_{k\geq 1}$ diverges. Note that because $2+(-1)^k$ always lies between 1 and 3, we easily see that $(2+(-1)^k)_{k\geq 1}$ does not converge to either of $\infty, -\infty$ either.

Just as with limits of functions, it is quite annoying to compute limits of sequences directly from the definition. Fortunately, just as in the functions case, there are some basic facts about limits of sequences that allow for relatively straightforward calculation of limits without employing ε-n_0 formalism. These facts mostly mirror those concerning limits of functions.

Theorem 15.1. *We have the following facts.*

- *If a sequence (a'_n) is obtained from the sequence (a_n) by changing finitely many of the a_n (where "changing" includes "making undefined"), then the behavior of both sequence as $n \to \infty$ is the same.*
- *For any natural number k, the sequences (a_{n+k}) and (a_n) have the same limiting behavior as $n \to \infty$.*
- *If (a_n) converges to a limit (including possibly $\pm\infty$), then that limit is unique.*
- *If $(a_n) \to L_a$ and $(b_n) \to L_b$ ($L_a, L_b \in \mathbb{R}$) then*
  - *$(ca_n + db_n) \to cL_a + dL_b$ for any real constants c, d, and*
  - *$(a_n b_n) \to L_a L_b$.*
  - *Moreover, if there is some n_0 such that for all $n > n_0$ we have $a_n = b_n$ then $L_a = L_b$.*
- *If $(a_n) \to L_a$ and $(b_n) \to \infty$ (respectively, $-\infty$) then $(a_n + b_n) \to \infty$ (respectively, $-\infty$).*
- *If (a_n) converges to a limit $L \neq 0$ then*
  - *there is some n_0 such that for all $n > n_0$, a_n is within $L/2$ of L (and so in particular is either always positive or always negative, depending on whether L is positive or negative), and*
  - *$(1/a_n)$ converges to $1/L$.*

- *If (a_n) converges to ∞ (respectively, $-\infty$), then*
  - *for every positive constant C (respectively, negative constant C) there is some n_0 such that for all $n > n_0$, $a_n > C$ (respectively, $a_n < C$) (and so in particular a_n is eventually either always positive or always negative, depending on whether the limit is $+\infty$ or $-\infty$), and*
  - *$(1/a_n)$ converges to 0.*
- *$(1) \to 1$ and $(n) \to \infty$.*
- *If $p(n)$ is polynomial of degree r, with the coefficient n^r being c_r, and $q(n)$ a polynomial of degree s, with the coefficient of n^s being 1, then*

$$\lim_{n\to\infty} \frac{p(n)}{q(n)} = \begin{cases} c_r & \text{if } r = s \\ 0 & \text{if } s > r \\ +\infty & \text{if } r > s,\ c_r > 0 \\ -\infty & \text{if } r > s,\ c_r < 0 \end{cases}$$

Proof: We leave all of these as exercises! The proofs here are very similar to the proofs of similar statements for limits of functions, and this theorem is a good exercise in reviewing those proofs. □

Armed with this theorem we can easily say, for example, that

$$\left(\frac{2n^4 - n + 1}{n^3 + 1}\right) \to \infty, \quad \left(\frac{2n^4 - n + 1}{n^4 + 1}\right) \to 2 \quad \text{and} \quad \left(\frac{2n^4 - n + 1}{n^5 + 1}\right) \to 0.$$

15.3 Sequences and functions

There is a natural (collection of) connections between limits of functions and limits of sequences. All three of the following facts are left as (easy) exercises:

1. Given a function $f : [1, \infty) \to \mathbb{R}$ with $\lim_{x\to\infty} f(x) = L$ (or ∞, or $-\infty$), define a_n by $a_n = f(n)$. Then $(a_n) \to \ell$ (or ∞, or $-\infty$).

2. The converse of point 1 above is *not* true: if $(a_n) \to \ell$ (or ∞, or $-\infty$) and $f : [1, \infty) \to \mathbb{R}$ satisfies $f(n) = a_n$ for all n, it is not necessarily the case that $\lim_{x\to\infty} f(x) = \ell$ (or ∞, or $-\infty$).

3. However, point 1 above has a partial converse: if $(a_n) \to \ell$ (or ∞, or $-\infty$) and $f : [1, \infty) \to \mathbb{R}$ satisfies $f(n) = a_n$ for all n, and furthermore $\lim_{x\to\infty} f(x)$ exists, then $\lim_{x\to\infty} f(x) = \ell$ (or ∞, or $-\infty$). (And note that there is always such an f. For example, define $f : [1, \infty) \to \mathbb{R}$ by $f(n) = a_n$ for all $n \in \mathbb{N}$, and then extend f to all of $[1, \infty)$ by linear interpolation (for $x \in (n, n + 1)$, $f(x) = (x - n)f(n + 1) + (n + 1 - x)f(n)$).

For example, consider $\lim_{n\to\infty} a^n$.

- For $a > 0$, we have $a^n = e^{n \log a}$.
  - For $a > 1$ we have $\log a > 0$ and so $\lim_{x\to\infty} e^{x \log a} = \infty$. By point 1 above $\lim_{n\to\infty} a^n = \infty$.
  - For $a < 1$ we have $\log a < 0$ and so $\lim_{x\to\infty} e^{x \log a} = 0$. By point 1 above, $\lim_{n\to\infty} a^n = 0$.
  - Rather trivially, for $a = 1$ we have $\lim_{n\to\infty} a^n = 1$.
- For $a < 0$, we write $a^n = (-1)^n(-a)^n$.
  - For $a > -1$ we have $\lim_{n\to\infty}(-a)^n = 0$ (from earlier), and it is an easy exercise that this implies that $\lim_{n\to\infty} a^n = \lim_{n\to\infty}(-1)^n(-a)^n = 0$.
  - for $a \leq -1$ it is an easy exercise to directly verify that $\lim_{n\to\infty} a^n$ does not exist.

In summary

$$\lim_{n\to\infty} a^n = \begin{cases} \infty & \text{if } a > 1 \\ 1 & \text{if } a = 1 \\ 0 & \text{if } -1 < a < 1 \\ \text{does not exist} & \text{if } a < -1. \end{cases}$$

The most important connection between limits of sequences and limits of functions is conveyed in the following result:

Theorem 15.2. *Suppose that f is continuous at c and that $(a_n) \to c$. Then $\lim_{n\to\infty} f(a_n) = f(c)$*[238]*.*

Conversely, suppose f is defined at and near c, and that $\lim_{n\to\infty} f(a_n) = f(c)$ for all sequences (a_n) that tend to c. Then f is continuous at c.

Before proving this, we give some examples.

Example 1 For the first, consider the sequence defined recursively by $a_1 = 1$ and $a_{n+1} = (3a_n + 4)/(2a_n + 3)$. A little computation shows that it is highly plausible that this sequence converges to the limit $\sqrt{2}$. We are not yet in a position to prove this. But, *suppose* we know that (a_n) converges to some limit, say c. We can use Theorem 15.2 to prove that the limit must be $\sqrt{2}$. Indeed, consider the function

$$f(x) = \frac{3x + 4}{2x + 3},$$

[238] This expression may not make sense — not all the a_n may be in the domain of f. However, since f is continuous at c, its domain includes $(c - \Delta, c + \Delta)$ for some $\Delta > 0$; and by definition of limit, there is some n_0 such that for all $n > n_0$, $a_n \in (c - \Delta, c + \Delta)$. So *eventually* the sequence $(f(a_n))_{n=1}^{\infty}$ makes sense. And, as we have seen, this is all that is necessary for the expression $\lim_{n\to\infty} f(a_n)$ to make sense.

which is continuous on all its domain ($\mathbb{R} \setminus \{-3/2\}$), and in particular is continuous at c ($a_n \geq 0$ for all n, so $c \geq 0$). From $(a_n) \to c$ we conclude $(f(a_n)) \to f(c)$. But $f(a_n) = a_{n+1}$, and $(a_{n+1}) \to c$. We conclude that $f(c) = c$, or

$$\frac{3c+4}{2c+3} = c.$$

After some easy algebra, we get that the only non-negative solution to this equation is $c = \sqrt{2}$. We conclude that *if* (a_n) converges, then it must converge to $\sqrt{2}$.

Note that the *if* is important here. Consider the recursively defined sequence $a_1 = 2$ and $a_{n+1} = a_n^2$ for $n \geq 1$. If the limit exists and equals c (clearly positive), then by the continuity of $f(x) = x^2$ at c we get by the same argument as above that $c = c^2$ so $c = 1$. But the limit is clearly not 1; the sequence diverges.

Example 2 As a second example, consider $\lim_{n\to\infty} \sqrt{n+a\sqrt{n}} - \sqrt{n+b\sqrt{n}}$. We have

$$\begin{aligned}
\sqrt{n+a\sqrt{n}} - \sqrt{n+b\sqrt{n}} &= \sqrt{n+a\sqrt{n}} - \sqrt{n+b\sqrt{n}} \left(\frac{\sqrt{n+a\sqrt{n}} + \sqrt{n+b\sqrt{n}}}{\sqrt{n+a\sqrt{n}} + \sqrt{n+b\sqrt{n}}} \right) \\
&= \frac{(a-b)\sqrt{n}}{\sqrt{n+a\sqrt{n}} + \sqrt{n+b\sqrt{n}}} \\
&= \frac{(a-b)}{\sqrt{1+\frac{a}{\sqrt{n}}} + \sqrt{1+\frac{b}{\sqrt{n}}}}.
\end{aligned}$$

Now the function

$$f(x) = \frac{a-b}{\sqrt{1+a\sqrt{x}} + \sqrt{1+b\sqrt{x}}}$$

is continuous at 0, with $f(0) = (a-b)/2$, so from Theorem 15.2 and the fact that $(1/n) \to 0$ we conclude that $f(1/n) \to (a-b)/2$, and so

$$\lim_{n\to\infty} \sqrt{n+a\sqrt{n}} - \sqrt{n+b\sqrt{n}} = \frac{a-b}{2}.$$

Example 3 Fix $a > 0$. What is $\lim_{n\to\infty} a^{1/n}$? Write $a^{1/n}$ as $e^{(\log a)/n}$. We have $(\log a)/n \to 0$ as $n \to \infty$, and the function $f(x) = e^x$ is continuous at 0, so by Theorem 15.2 we get $\lim_{n\to\infty} a^{1/n} = \lim_{n\to\infty} f((\log a)/n) = f(0) = 1$.

Proof (of Theorem 15.2): First suppose that f is continuous at c and that $(a_n) \to c$. Fix $\varepsilon > 0$. There is $\delta > 0$ such that $|x - c| < \delta$ implies $|f(x) - f(c)| < \varepsilon$. Also, there is n_0 such that $n > n_0$ implies $|a_n - c| < \delta$, so $|f(a_n) - f(c)| < \varepsilon$. Since ε was arbitrary this shows that $(f(a_n)) \to f(c)$.

For other direction, suppose $\lim_{n\to\infty} f(a_n) = f(c)$ for all sequences (a_n) that tend to c, but that $\lim_{x\to c} f(x) \neq f(c)$. So there is an $\varepsilon > 0$, such that for all $\delta > 0$, there is x_δ with

$|x_\delta - c| < \delta$ but $|f(x) - f(c)| \geq \varepsilon$. Applying this with $\delta = 1/n$ for each $n \in \mathbb{N}$, we get a sequence (x_n) with $|x_n - c| < 1/n$ but $|f(x_n) - f(c)| \geq \varepsilon$, so $(f(x_n)) \not\to f(c)$. But evidently $(x_n) \to c$, contradicting our hypotheses. □

There is a slight modification of this result, that has an almost identical proof:

> Suppose f is defined near c (but not necessarily *at* c) and that $\lim_{x\to c} f(x) = \ell$. If $(a_n) \to c$, and for all large enough n we have $a_n \neq c$, then $\lim_{n\to\infty} f(a_n) = \ell$.
>
> Conversely if f is defined near (but not necessarily at) c, and $\lim_{n\to\infty} f(a_n) = \ell$ for all sequences (a_n) that tend to c and that eventually (for all sufficiently large n) avoid c. Then $\lim_{x\to c} f(x) = \ell$.

We will have no need to use this strengthening, so will say no more about it.

One more useful result concerning sequences and convergence is a very natural "squeeze theorem".

Theorem 15.3. *Let $(a_n), (b_n)$ and (c_n) be sequences with $(a_n), (c_n) \to L$. If eventually (for all $n > n_0$, for some finite n_0) we have $a_n \leq b_n \leq c_n$, then $(b_n) \to L$ also.*

Proof: Fix $\varepsilon > 0$. There is n_1, n_2 such that $n > n_1$ implies $a_n \in (L - \varepsilon, L + \varepsilon)$, and $n > n_2$ implies $c_n \in (L - \varepsilon, L + \varepsilon)$. For $n > \max\{n_0, n_1, n_2\}$ (n_0 as in the statement of the theorem), we have

$$L - \varepsilon < a_n \leq b_n \leq c_n < L + \varepsilon$$

so $b_n \in (L - \varepsilon, L + \varepsilon)$. □

Consider, for example,

$$\lim_{n\to\infty} \left(\frac{2n^2 - 1}{3n^2 + n + 2} \right)^{\frac{1}{n}}.$$

We have $(2n^2 - 1)/(3n^2 + n + 2) \to 2/3$ as $n \to \infty$, so for all sufficiently large n

$$0.6^{\frac{1}{n}} \leq \left(\frac{2n^2 - 1}{3n^2 + n + 2} \right)^{\frac{1}{n}} \leq 0.7^{\frac{1}{n}}.$$

Since, as we have seen previously, both $0.6^{\frac{1}{n}}, 0.7^{\frac{1}{n}} \to 1$ as $n \to \infty$, the squeeze theorem allows us to conclude

$$\lim_{n\to\infty} \left(\frac{2n^2 - 1}{3n^2 + n + 2} \right)^{\frac{1}{n}} = 1.$$

There is an "infinite" variant of the squeeze theorem, whose simple proof we omit.

> If $(a_n) \to \infty$ and (b_n) is such that eventually $b_n \geq a_n$, then $(b_n) \to \infty$ also.

15.4 Monotonicity, subsequences and Bolzano-Weierstrass

Definition of a sequence increasing/decreasing A sequence (a_n) is said to be *increasing* (a.k.a. *strictly increasing*) if $a_n > a_m$ whenever $n > m$. It is said to be *weakly increasing* (a.k.a. *non-decreasing*) if $a_n \geq a_m$ whenever $n > m$. The analogous definitions of *decreasing* (a.k.a. *strictly decreasing*) and *weakly decreasing* (a.k.a. *non-increasing*) are omitted. The sequence is said to be *monotone* (a.k.a. *strictly monotone*) if it is either increasing or decreasing, and *weakly monotone* if it is either non-decreasing or non-increasing.

Definition of a sequence being bounded A sequence (a_n) is said to be *bounded above* if there is M such that $a_n \leq M$ for all n, and *bounded below* if there is m such that $m \leq a_n$ for all n. It is said to be bounded if it is both bounded above and bounded below.

Note that

- (a_n) is bounded if and only if there is M such that for all n, $|a_n| \leq M$, and
- If there is a number M' such that for all $n > n_0$ we have $a_n < M'$, then (a_n) is bounded, for example by $\max\{a_1, \ldots, a_n, M'\}$. So, as with converging to a limit, the property of being bounded is one that is not compromised by changing a sequence at finitely many values.

If a sequence (a_n) is bounded above, then $\{a_n : n \in \mathbb{N}\}$ is non-empty and bounded above, so $\alpha = \sup\{a_n : n \in \mathbb{N}\}$ exists. It is certainly not necessarily the case, though, that (a_n) converges under these circumstances, nor the limit, if it exists, has to be α. If, however, the sequence is also non-decreasing, the story is different.

Lemma 15.4. *If (a_n) is non-decreasing and bounded above then $(a_n) \to \alpha := \sup\{a_n : n \in \mathbb{N}\}$.*

Proof: Let $\varepsilon > 0$ be given. There is some n_0 with $a_{n_0} \in (\alpha - \varepsilon, \alpha]$ (otherwise, $\alpha - \varepsilon$ would be an upper bound for (a_n), contradicting that α is the least upper bound. Since (a_n) is non-decreasing we have that $a_n \in (\alpha - \varepsilon, \alpha]$ for all $n > n_0$, so $|a_n - \alpha| < \varepsilon$ for all such n. $\square$

The analogous result, that a non-increasing sequence that is bounded below tends to a limit, and that that limit is $\inf\{a_n : n \in \mathbb{N}\}$, is proven almost identically.

As an example, consider the recursively defined sequence $a_1 = 1$, $a_{n+1} = (3a_n+4)/(2a_n+3)$ for $n \geq 1$. We showed previously that *if* this sequence converges to a limit, that limit must be $\sqrt{2}$. We now show that it does converge to a limit, by showing that it is non-decreasing and bounded above.

We first note that obviously $a_n > 0$ for all n. We have

$$\begin{aligned} a_{n+1} \geq a_n \quad &\text{if and only if} \quad \frac{3a_n + 4}{2a_n + 3} \geq a_n \\ &\text{if and only if} \quad 4 \geq 2a_n^2 \\ &\text{if and only if} \quad a_n \leq \sqrt{2}. \end{aligned}$$

We now show by induction on n that $a_n \leq \sqrt{2}$ for all n; as well as this showing that (a_n) is non-decreasing, it also shows that (a_n) is bounded above, so by the lemma converges.

The base case of the induction is trivial. For the induction step, we assume $a_n \leq \sqrt{2}$ for some $n \geq 1$. We have

$$\begin{aligned} a_{n+1} \leq \sqrt{2} \quad &\text{if and only if} \quad \frac{3a_n + 4}{2a_n + 3} \leq \sqrt{2} \\ &\text{if and only if} \quad 3a_n + 4 \leq \sqrt{2}(2a_n + 3) \\ &\text{if and only if} \quad 9a_n^2 + 24a_n + 16 \leq 8a_n^2 + 24a_n + 18 \\ &\text{if and only if} \quad a_n \leq \sqrt{2}. \end{aligned}$$

This completes the induction, and the verification that $(a_n) \to \sqrt{2}$. Notice that $(a_1, a_2, a_3, \ldots)$ is a list of every-better rational approximations to $\sqrt{2}$.

In general, determining whether a sequence is bounded above or not is not easy! Consider, for example:

- (a_n) where $a_n = 1 + 1/2 + 1/3 + \cdots + 1/n$,
- (b_n) where $b_n = \sum_{p \leq n,\ p \text{ a prime number}} 1/p$, and
- (c_n) where $c_n = \sum_{k \leq n,\ k \text{ has no 7 in its decimal representation}} 1/k$.

We will shortly develop techniques to a test sequences of this form — sequences who generic terms are sums of other sequences — for boundedness.

We now turn to considering *subsequences*. Informally, a subsequence of a sequence

$$(a_1, a_2, a_3, \ldots)$$

is a sequence of the form

$$(a_{n_1}, a_{n_2}, a_{n_3}, \ldots)$$

with $n_1 < n_2 < n_3 \cdots$. In other words, it is a sequence obtained from another sequence by extracting an infinite subset of the elements of the original sequence, *keeping the elements in the same order as they were in the original sequence.*

Formally, a subsequence is a restriction of a sequence $a : \mathbb{N} \to \mathbb{R}$ to an infinite subset S of $\mathbb{N}$, that is, a function $a|_S : S \to \mathbb{N}$ defined by $a|_S(n) = a(n)$ for $n \in S$.

Here is the fundamental lemma concerning subsequences of a sequence.

Lemma 15.5. *Every sequence has a subsequence which is either non-decreasing or non-increasing. In fact, every sequence has a subsequence which is either weakly increasing or strictly decreasing. Also, every sequence has a subsequence which is either strictly increasing or weakly decreasing.*

Proof: Call a term a_n in a sequence a *horizon point* if $a_n > a_m$ for all $m > n$. If a sequence has infinitely many horizon points, say $a_{n_1}, a_{n_2}, \ldots$, then we get a strictly decreasing subsequence. If there are only finitely many horizon points, then pick a_{n_1} after the last horizon point. Since a_{n_1} is not a horizon point, there is $n_2 > n_1$ with $a_{n_2} \geq a_{n_1}$. Since a_{n_2} is not a horizon point, there is $n_3 > n_2$ with $a_{n_3} \geq a_{n_2}$. Repeating, we get a weakly increasing subsequence.

This shows that every sequence has a subsequence which is either weakly increasing or strictly decreasing. Applying this result to the sequence $(-a_n)$ shows that (a_n) also has a subsequence which is either strictly increasing or weakly decreasing. □

Combining Lemmas 15.4 and 15.5 we get the following, one of the cornerstone theorems of analysis.

Theorem 15.6. *(Bolzano-Weierstrass) If (a_n) is bounded then it has a convergent subsequence.*

Here is a consequence of the Bolzano-Weierstrass theorem. Suppose (a_n) is a bounded sequence, bounded, say, by M (so $-M \leq a_n \leq M$ for all n). Let S be the set of all numbers s such that (a_n) has a subsequence which converges to s. We have that S is non-empty (by Bolzano-Weierstrass). Also, by the squeeze theorem every element in S lies between $-M$ and M. So S is bounded, both from above and from below. By the completeness axiom, then, S has both a supremum and an infimum.

Defintion of $\limsup$ **and** $\liminf$ With the notation as above, the *limit superior* of the sequence (a_n), or *lim sup*, denoted $\limsup a_n$, is the supremum of S, and the *limit inferior*, or *lim inf*, denoted $\liminf a_n$, is the infimum of S.

Think of $\limsup a_n$ as the "largest" of all subsequential limits of (a_n), and $\liminf a_n$ as the "smallest" of all subsequential limits. The quotes around "largest" and "smallest" are there since, as usual when working with infima and suprema, there may not actually be subsequences that converge to the lim sup or lim inf. But in fact they are unnecessary in this case.

Lemma 15.7. *Suppose that (a_n) is a bounded sequence, with $\limsup a_n = \alpha$ and $\liminf a_n = \beta$. Then (a_n) has a subsequence that converges to α, and one that converges to β.*

Proof: We'll just show that there is a sequence that converges to α; the proof of the existence of a sequence converging to β is similar.

If $\alpha \in S$, we are immediately done. If not, for each $n \in \mathbb{N}$ there is $\alpha_n \in S$ with $\alpha - 1/n < \alpha_n < \alpha$ (since $\alpha = \sup S$), and there is a subsequence that converges to α_n.

Build a new subsequence as follows: chose a_{n_1} to be any term of the subsequence converging to α_1, that is distance less than 1 from α_1; chose a_{n_2} to be any term of the subsequence converging to α_2, that is distance less than $1/2$ from α_2; and in general chose a_{n_k} to be any term of the subsequence converging to α_k, that is distance less than $1/k$ from α_k.

Notice that a_{n_1} is distance less than $1+1 = 2$ from α; a_{n_1} is distance less than $1/2+1/2 = 1$ from α; and in general a_{n_k} is distance less than $1/k + 1/k = 2/k$ from α. Since $2/k \to 0$ as $n \to \infty$, it follows that for any $\varepsilon > 0$ the terms of the sequence $(a_{n_1}, a_{n_2}, \ldots)$ eventually are all within ε of α. □

So:

> The lim sup of a bounded sequence is the *largest* subsequential limit, and the lim inf is the *smallest* subsequential limit.

A bounded sequence may not have a limit, but it always has a lim inf and a lim sup, and it is for this reason that these parameters are introduced. As an example, the sequence whose nth term a_n is 0 if n is odd, and $1 - (1/n)$ if n is even, has no limit, but it has $\liminf a_n = 0$ and $\limsup a_n = 1$.

The notions of lim sup and lim inf can be thought of as generalizations of the notion of the limit of a sequence, because if a sequence (a_n) has a limit, then it is fairly easy to prove (exercise!) that $\lim a_n = \limsup a_n = \liminf a_n$ (and conversely, if (a_n) is a sequence with $\limsup a_n = \liminf a_n$ then the sequence converges to the common value).

lim sup and lim inf can also be thought of as capturing the "eventual" behavior of a sequence. Suppose (a_n) is a sequence with $\limsup a_n = \alpha$. For each $\varepsilon > 0$, it must be the case that only finitely many terms of the sequence are larger than $\alpha + \varepsilon$ (if not, there would be a subsequence consisting only of terms larger than $\alpha + \varepsilon$, and by the Bolzano-Weierstrass theorem this subsequence would have a subsequence converging to a limit that lies at or above $\alpha + \varepsilon$, contradicting that α is the lim sup). On the other hand, for each $\varepsilon > 0$, it must be the case that infinitely many terms of the sequence are larger than $\alpha - \varepsilon$ (by definition of α there must be a subsequence converging to some value between $\alpha - \varepsilon/2$ and α, and that subsequence eventually always has terms greater than $\alpha - \varepsilon$). This leads to an alternate characterization of lim sup and lim inf (we skip the nitty gritty details of verifying this):

> If (a_n) is a bounded sequence, then $\limsup a_n$ is the unique real number α such that for each $\varepsilon > 0$ only finitely many terms of the sequence are larger than $\alpha + \varepsilon$, while infinitely many terms of the sequence are larger than $\alpha - \varepsilon$. Also $\liminf a_n$ is the unique real number β such that for each $\varepsilon > 0$ only finitely many terms of the sequence are smaller than $\beta - \varepsilon$, while infinitely many terms of the sequence are smaller than $\beta + \varepsilon$.

There are other characterizations of lim sup and lim inf, and extensions to unbounded sequences, which we do not address.

We end our discussion of convergence of sequences by introducing one last test for convergence.

Definition of a Cauchy sequence A sequence (a_n) is *Cauchy*, or *a Cauchy sequence*, if for all $\varepsilon > 0$ there is n_0 such that $n, m > n_0$ implies $|a_n - a_m| < \varepsilon$.

In other words, a sequence is Cauchy not necessarily if the terms eventually get close to a particular limit, but if they eventually get close to *one another*.

For example, every sequence (a_n) that converges to a limit, is Cauchy. Indeed, suppose that the limit is L. Fix $\varepsilon > 0$. There is n_0 such that $n, m > n_0$ implies both $|a_n - L| < \varepsilon/2$ and $|a_m - L| = L - a_m| < \varepsilon/2$. But then,

$$|a_n - a_m| = |a_n - L + L - a_m| \leq |a_n - L| + |L - a_m| < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

In fact, convergent sequences are the *only* Cauchy sequences:

Lemma 15.8. *If (a_n) is a Cauchy sequence, then (a_n) converges.*

Proof: Let (a_n) be a Cauchy sequence. The proof that (a_n) converges goes in three steps.

- (a_n) **is bounded**: There is n_0 such that $n, m > n_0$ implies $|a_n - a_m| < 1$. In particular, for every $m > n_0$, $|a_{n_0+1} - a_m| < 1$, so $a_m \in (a_{n_0+1} - 1, a_{n_0+1} + 1)$. So (a_n) is bounded above by $\max\{a_1, \ldots, a_{n_0}, a_{n_0+1} + 1\}$, and below by $\min\{a_1, \ldots, a_{n_0}, a_{n_0+1} - 1\}$.

- (a_n) **has a convergent subsequence**: Directly from the Bolzano-Weierstrass theorem, there is a subsequence of (a_n), $(a_{n_1}, a_{n_2}, \ldots)$ say, that converges to a limit, L say.

- (a_n) **converges to** L: Suppose not. Then there is $\varepsilon > 0$ for which it is not the case that eventually all terms of the sequence are within ε of L. In other words, there is a subsequence $(a_{n'_1}, a_{n'_2}, \ldots)$ with $|a_{n'_k} - L| \geq \varepsilon$. Now pick any n_0. There is $n'_k > n_0$; there is also $n_k > n_0$ such that $|a_{n_k} - L| < \varepsilon/10$. It follows that $|a_{n_k} - a_{n'_k}| > \varepsilon/2$,[239] and so it is not possible to find an n_0 such that for all $n, m > n_0$, we have $|a_n - a_m| < \varepsilon/2$. This contradicts that (a_n) is Cauchy.

□

Showing that a sequence is Cauchy allows us to show that it is convergent, without

[239] Draw a picture!

actually finding the limit. Consider, for example, $a_n = \sum_{k=1}^{n} 1/k^2$. We have, for $n > m$,

$$\begin{aligned}
|a_n - a_m| &= \frac{1}{(m+1)^2} + \frac{1}{(m+2)^2} + \cdots + \frac{1}{n^2} \\
&\leq \frac{1}{(m+1)^2 - (m+1)} + \frac{1}{(m+2)^2 - (m+2)} + \cdots + \frac{1}{n^2 - n} \\
&= \left(\frac{1}{m} - \frac{1}{m+1}\right) + \left(\frac{1}{m+1} - \frac{1}{m+2}\right) + \cdots + \left(\frac{1}{n-1} - \frac{1}{n}\right) \\
&= \frac{1}{m} - \frac{1}{n} \\
&\leq \frac{1}{m} \\
&< \frac{1}{n_0}.
\end{aligned}$$

So, given $\varepsilon > 0$, if we choose n_0 such that $1/n_0 \leq \varepsilon$, then for all $n, m > n_0$ we have $|a_n - a_m| < \varepsilon$, showing that (a_n) is Cauchy, and so converges to a limit.

Notice that we were able here to establish that (a_n) converges, without actually identifying the limit. This illustrates the value of the concept of Cauchy sequences.[240]

[240] The problem of evaluating the limit of (a_n) in this case is known as the *Basel problem*, and was famously solved by Euler, who showed the remarkable formula

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

See e.g. https://en.wikipedia.org/wiki/Basel_problem.

16 Series

16.1 Introduction to series

Informally a *series* or *infinite series* is an expression of the form

$$a_1 + a_2 + a_3 + \dots$$

or

$$\sum_{k=1}^{\infty} a_k.$$

We clearly have to take care with such an expression, as it is far from clear that the operation of adding infinitely many things is well defined. For example, we could argue

$$\sum_{k=1}^{\infty} (-1)^k = 0$$

by writing

$$\begin{aligned} \sum_{k=1}^{\infty} (-1)^k &= (-1+1) + (-1+1) + (-1+1) + \cdots \\ &= 0 + 0 + 0 + \cdots \\ &= 0; \end{aligned}$$

but we could equally well argue

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

by writing

$$\begin{aligned} \sum_{k=1}^{\infty} (-1)^k &= -1 + (1-1) + (1-1) + \cdots \\ &= -1 + 0 + 0 + \cdots \\ &= -1. \end{aligned}$$

We will see more startling paradoxes later.

Formally, given a sequence (a_n), define the *nth partial sum* of (a_n) by

$$s_n = a_1 + \ldots + a_n = \sum_{k=1}^{n} a_k$$

Definition of summability Say that (a_n) is *summable* if (s_n) converges to some limit ℓ.

If (a_n) is summable we write $\sum_{k=1}^{\infty} a_k$ or $a_1 + a_2 + \ldots$ for ℓ. Informally, we say "the series $\sum_{k=1}^{\infty} a_k$ *converges* (to ℓ)"[241], or

$$\sum_{k=1}^{\infty} a_k = \ell.$$

We give some examples here:

- if (a_n) is eventually (for all sufficiently large n) 0, then it is summable.
- $((-1)^n)$ is *not* summable: the sequence of partial sums is

  $$(-1, 0, -1, 0, -1, 0 \ldots),$$

  which does *not* converge to a limit.
- $(1/n)$ is *not* summable. The nth partial sum s_n is

  $$s_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}.$$

  There are many ways to see that the sequence (s_n) does not tend to a limit. Perhaps the simplest is to consider the subsequence (s_{2^n}). We have

  $$\begin{aligned} s_{2^n} &= 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \cdots + \frac{1}{8}\right) + \cdots + \left(\frac{1}{2^{n-1}+1} + \frac{1}{2^n}\right) \\ &> 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \cdots + \frac{1}{8}\right) + \cdots + \left(\frac{1}{2^n} + \frac{1}{2^n}\right) \\ &= 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots + \frac{1}{2} \\ &= 1 + \frac{n}{2}. \end{aligned}$$

  Since $1 + n/2$ can be made arbitrarily large by choosing n large enough, we see that (s_n) cannot possibly tend to a finite limit.[242]
- Consider the sequence $(r^n)_{n=0}^{\infty}$ with $|r| < 1$. We have

  $$s_n = 1 + r + r^2 + \cdots + r^n = \frac{1 - r^{n+1}}{1 - r} = \frac{1}{1 - r} - \frac{r^{n+1}}{1 - r}$$

  For $|r| < 1$, $\lim_{n\to\infty} r^{n+1} = 0$, so $(s_n) \to 1/(1 - r)$ as $n \to \infty$. We conclude that

  for $|r| < 1$, $(r^n)_{n=0}^{\infty}$ is summable, and

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

  This is the incredibly useful *geometric series sum*.

[241] Note that this is quite informal; the expression "$\sum_{k=1}^{\infty} a_k$" is just a single expression, that is not varying, so is not really in any sense converging to anything

[242] The partial sum $s_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$ is called the *nth Harmonic number*, usually denoted H_n. It's properties will be explored in a homework problem.

16.2 Tests for summability

We now develop a collection of tests/criteria for summability.

Basic closure properties If (a_n), (b_n) are both summable, then so are

- $(a_n + b_n)$ — with
$$\sum_{n=1}^{\infty}(a_n + b_n) = \sum_{n=1}^{\infty} a_n + \sum_{n=1}^{\infty} b_n$$
- and (ca_n) — with
$$\sum_{n=1}^{\infty} ca_n = c\sum_{n=1}^{\infty} a_n.$$

The proofs of these facts are left as exercises.

For all $k \geq 0$, the sequences $(a_1, a_2, a_3, \ldots)$ and $(a_k, a_{k+1}, a_{k+2}, \ldots)$ are either both summable or both not, and if they are both summable, then they have the same sum; from this it follows that if two sequences can be made equal by shifting and changing finitely many terms, then they are either both summable or both not. Again, the proof is left as an exercise.

Cauchy criterion (a_n) is summable if and only if (s_n) is Cauchy. This follows from Lemma 15.8, and from the observation before that lemma that convergent sequences are Cauchy. This means that (a_n) is summable if and only if for all $\varepsilon > 0$ there's n_0 such that $n, m > n_0$ implies $|s_m - s_n| < \varepsilon$, that is (assuming without loss of generality that $m > n$)
$$|a_{n+1} + a_{n+2} + \ldots + a_m| < \varepsilon.$$
The intuition here is that a sum $\sum_{n=1}^{\infty} a_n$ converges if and only if its "tail" $a_n + a_{n+1} + a_{n+2} + \cdots$ can be made arbitrarily small. But this is just an intuition; the tail of an infinite sum is itself a sum of infinitely many things, and so to properly understand it we need the theory that we are in the process of developing. The Cauchy criterion expresses the idea that the tail of the sequence can be made arbitrarily small, while only ever referring to the sum of finitely many terms.

Vanishing condition If (a_n) is summable then, from the Cauchy condition, for all $\varepsilon > 0$ there is n_0 such that $n, m > n_0$ implies $|s_n - s_m| < \varepsilon$. Applying this with $m = n - 1$ we get that for sufficiently large n, $|a_n| < \varepsilon$. It follows that $\lim_{n\to\infty} |a_n| = 0$, so $\lim_{n\to\infty} a_n = 0$. The contrapositive of this is what is usually used:

> if $\lim_{n\to\infty} a_n \neq 0$ then (a_n) is *not* summable.

For example, for $|r| \geq 1$ we have $\lim_{n\to\infty} r^n \neq 0$, so in this range (r^n) *not* summable (we have already seem that it is summable, with sum $1/(1-r)$ for all other r).

Note: the converse of vanishing condition is **not** true — $(a_n) \to 0$ does *not* imply that (a_n) is summable. An example to consider is $(1/n)$.

Boundedness criterion This is more a theoretical than a practical criterion. If $a_n \geq 0$ for all n, then (s_n) is increasing. So in this situation

> (a_n) is summable iff (s_n) is bounded above.

As an example, we earlier showed that $\left(\sum_{k=1}^{n} \frac{1}{k^2}\right)$ is bounded above, and so now we know that $(1/n^2)$ is summable, i.e., that the expression $\sum_{n=1}^{\infty} \frac{1}{n^2}$ makes sense (is finite).

Comparison test Suppose $b_n \geq a_n \geq 0$ for all n. If (b_n) is summable, then so is (a_n) — this is because the partial sums of (a_n) are bounded above by $\sum_{n=1}^{\infty} b_n$, so we can apply the boundedness criterion. Contrapositively, if (a_n) is not summable, then neither is (b_n).

We give some examples.

- $(1/n^\alpha)$, $\alpha < 1$. We have $n^\alpha < n$, so $1/n^\alpha > 1/n \geq 0$. By comparison with $(1/n)$, $(1/n^\alpha)$ is not summable.
- $(n^3/3^n)$. We know that $(1/3^n)$ is summable, so we would like to say that $(\frac{n^3}{3^n})$ is too, by comparison. But the inequality goes the wrong way — we have $1/3^n \leq \frac{n^3}{3^n}$. We can, however, compare with $(1/2^n)$. We have $n^3/3^n \leq 1/2^n$ for all large enough n, so by comparison $(\frac{n^3}{3^n})$ is summable.[243]
- $(n^2/(n^3+1))$. This looks a lot like $1/n$, so we suspect that it is not summable. But it is not true that $n^2/(n^3+1) > 1/n$. However, for all large enough n, we have $n^2/(n^3+1) > 1/2n$ (actually for $n > 1$), so $(n^2/(n^3+1))$ is not summable, by comparison with $(1/2n)$.

Limit comparison test As the last few examples show, sometimes the comparison test can be awkward to apply. A much more convenient version is the limit comparison test:

> Suppose $a_n, b_n > 0$ and $\lim_{n\to\infty} a_n/b_n = c > 0$. Then (a_n) is summable if and only if (b_n) is.

To prove this, first suppose that (b_n) is summable. There is n_0 such that $n > n_0$ implies $a_n < 2cb_n$. Since (b_n) is summable, so is $(2cb_n)$. We can now conclude that (a_n) is summable, by comparison with $(2cb_n)$. (As usual, we are ignoring finitely many terms of (a_n) that might be larger than their companion terms in $(2cb_n)$).

In the other direction, if (a_n) is summable, then since $b_n/a_n \to 1/c > 0$ we get that (b_n) is summable by the above argument.

We give a few examples:

[243] We don't care what happens for finitely many n; that changes the sum, but not whether the sequence is summable.

- $(n^2/(n^3+1))$. We have
$$\lim_{n\to\infty} \frac{n^2/(n^3+1)}{1/n} = 1,$$
so by limit comparison with $(1/n)$, we get that $(n^2/(n^3+1))$ is not summable (and note that this went more smoothly than the comparison test).
- $(2/\sqrt[3]{n^2+1})$. Since $1/n^{2/3}$ diverges, and
$$\lim_{n\to\infty} \frac{2/\sqrt[3]{n^2+1}}{1/n^{2/3}} = 2,$$
we get that $(2/\sqrt[3]{n^2+1})$ is not summable.

Ratio test This is possibly the most useful criterion for summability. If $a_n > 0$ for all (sufficiently large) n, and
$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = r,$$
then:

- if $r < 1$, the series $\sum_{n=1}^{\infty} a_n$ converges (that is, (a_n) is summable),
- if $r > 1$, the series does not converge, and
- if $r = 1$, no conclusion can be reached.

Before proving this, we give some examples.

- If $a_n = x^{2n}/(2n)!$ (positive for all n and all $x \in \mathbb{R}$), then $\lim_{n\to\infty} a_n = 0$, and so
$$\sum_{n\geq 0} \frac{x^{2n}}{(2n)!}$$
converges to a limit for all real x. By the same argument, so does
$$\sum_{n\geq 0} \frac{x^n}{n!}.$$
We should strongly suspect that the sum is e^x (and in fact we can easily prove this at this point); we will return to this later. Notice that from the summability of $(x^n/n!)$ and the vanishing criterion, we recover a previous result, that
$$\lim_{n\to\infty} \frac{x^n}{n!} = 0$$
for all real x.

- $a_n = n^k/c^n$, $c > 1$, $k > 0$. We apply the ratio test, and see that

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \lim_{n\to\infty} \frac{(n+1)^k}{cn^k} = \frac{1}{c} < 1$$

  and so (n^k/c^n) summable, that is,

$$\sum_{n=0}^{\infty} \frac{n^k}{c^n}$$

  is a finite number. Note that this tells us (via the vanishing criterion) that $\lim_{n\to\infty} n^k/c^n = 0$ (something we have seen before).

We now turn to the proof of ratio test. Suppose $r < 1$. Fix s with $r < s < 1$. There is n_0 such that $a_{n+1}/a_n < s$ for all $n > n_0$. We have

$$a_{n+1} < sa_n,$$

$$a_{n+2} < sa_{n+1} < s^2 a_n,$$

and in general

$$a_{n+k} < s^k a_n,$$

so applying at $n = n_0 + 1$ we get

$$a_{(n_0+1)+k} < s^k a_{n_0+1}$$

for all $k \geq 1$. Since $(s^k a_{n_0+1})_{k=1}^{\infty}$ is summable (it is a geometric series), so is $(a_{(n_0+1)+k})_{k=1}^{\infty}$ (by comparison), and so so also is $(a_n)_{n\geq 1}$.

On the other hand, suppose $r > 1$. Fix s with $1 < s < r$. There n_0 such that for all $n > n_0$, $a_{n+1}/a_n > s$ so that (by the same reasoning as above) $a_{(n_0+1)+k} > s^k a_{n_0+1}$, so $\lim_{n\to 0} a_n \neq 0$ and so (a_n) not summable.

Finally, the sequences $(1/n)$ (which is not summable) and $(1/n^2)$ (which is) both have

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = 1,$$

which verifies that no conclusion can be reached if $r = 1$.

Integral test Suppose that $f : [1, \infty) \to (0, \infty)$ is non-increasing, and $f(n) = a_n$. Then (a_n) is summable if and only if $\int_1^{\infty} f$ exists.

As an example, consider the *p-series* $\sum_{n=1}^{\infty} 1/n^p$.

- If $p \leq 0$, the sum diverges, by the vanishing criterion.
- $p > 0$, the sum converges if and only if $\int_1^{\infty} dx/x^p$ exists, which is the case exactly if $p > 1$.

At $p = 1$ we recover the divergence of the Harmonic series

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots .$$

At $p = 2$ we get that $\sum_{n=1}^{\infty} \frac{1}{n^2}$ is some finite number (as we have seen before, via ad-hoc methods); that number happens to be $\pi^2/6$.

Here is a proof of the integral test. Consider the sequence whose nth term is $\int_n^{n+1} f$. We have that $\int_1^{\infty} f$ exists if and only if $(\int_n^{n+1} f)_{n\geq 1}$ is summable.

Because f is decreasing we have

$$a_{n+1} \leq \int_n^{n+1} f \leq a_n.$$

Suppose $(\int_n^{n+1} f)_{n\geq 1}$ is summable. Then the first inequality, together with comparison, says that (a_{n+1}) is summable, and so (a_n) is summable.

Suppose on the other hand that (a_n) is summable. Then the second inequality, together with comparison, says that $(\int_n^{n+1} f)_{n\geq 1}$ is summable.

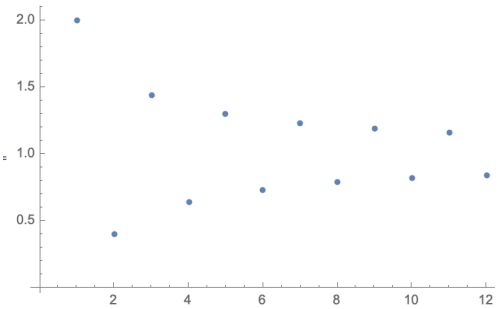
Leibniz' theorem on alternating series Most of the powerful tests for summability presented so far have concerned non-negative sequence. We now present a test which considers sequences that have some negative terms. Suppose (a_n) is non-increasing and tends to 0 (and so necessarily $a_n \geq 0$ for all n). Then Leibniz' theorem is the assertion that the *alternating* series $((-1)^{n-1}a_n)$ is summable, i.e.,

$$a_1 - a_2 + a_3 - a_4 \cdots$$

converges to a (finite) limit.

For example, $(1/n)$ is not summable, but $((-1)^n/n)$ is; and if p_n is the nth prime, then $(1/p_n)$ not summable, but $((-1)^n/p_n)$ is.

To prove Leibniz' theorem, it helps to draw a picture of the sequence of partial sums, which strongly suggests that $s_1 \geq s_3 \geq \cdots \geq \ell$ and $s_2 \leq s_4 \leq s_6 \cdots \leq L$, for some limit L, and with every odd partial sum exceeding every even one:



Inspired by this picture, we prove Leibniz' theorem in small steps:

- First, $s_1 \geq s_3 \geq s_5 \geq \cdots$ (odd partial sums form a non-increasing sequence). Indeed, for $n \geq 1$ we have

$$s_{2n-1} - s_{2n+1} = a_{2n} - a_{2n+1} \geq 0$$

  since (a_n) is non-increasing.

- Next, $s_2 \leq s_4 \leq s_6 \leq \cdots$ (even partial sums form a non-decreasing sequence). Indeed, for $n \geq 2$ we have

$$s_{2n} - s_{2n-2} = a_{2n-1} - a_{2n} \geq 0$$

  again since (a_n) is non-increasing.

- Next, if k is even and ℓ is odd, then $s_k \leq s_\ell$ (all odd partial sums are at least as large as all even partial sums). Indeed, for every n we have $s_{2n-1} - s_{2n} = a_n \geq 0$, so $s_{2n} \leq s_{2n-1}$. So, choosing n large enough that $2n > k$ and $2n - 1 > \ell$, we have from the first two observations that

$$s_k \leq s_{2n} \leq s_{2n-1} \leq s_\ell.$$

- Next, the sequence $(s_{2n})_{n=1}^{\infty}$ converges to a limit, say α. Indeed, it is non-decreasing and bounded above, say by s_1, by previous observations, so converges.

- Next, the sequence $(s_{2n-1})_{n=1}^{\infty}$ converges to a limit, say β. Indeed, it is non-increasing and bounded below, say by s_2, by previous observations, so converges.

- Next, $\alpha = \beta$. Indeed, as previously observed we have $s_{2n-1} - s_{2n} = a_n$. Taking limits of both sides as n goes to infinity, and using the last unused hypothesis (that $(a_n) \to 0$) we get $\beta - \alpha = 0$ so $\alpha = \beta$.

- Finally, letting L be the common value of α, β, we have $(s_n) \to L$. Indeed, fix $\varepsilon > 0$. Because (s_{2n}) increases to limit L there is n_1 such that if $n > n_1$ and n is even, then $|s_n - L| < \varepsilon$, and because (s_{2n-1}) decreases to limit L there is n_2 such that if $n > n_2$ and n is odd, then $|s_n - L| < \varepsilon$. So for $n > \max\{n_1, n_2\}$ we have $|s_n - L| < \varepsilon$.

In fact, this proof gives something more: for each n we have $s_{2n} \leq L \leq s_{2n+1}$, so

$$L - s_{2n} \leq s_{2n+1} - s_{2n} = a_{2n+1},$$

and also $s_{2n+2} \leq L \leq s_{2n+1}$, so

$$s_{2n+1} - L \leq s_{2n+1} - s_{2n+2} = a_{2n+2}.$$

In other words:

Let $S = \sum_{n=1}^{\infty}(-1)^{n-1}a_n$ where $a_n \geq 0$, (a_n) is non-increasing and $(a_n) \to 0$ (so S is finite, by Leibniz' theorem). The sum of the first $2n$ terms (truncating at a subtracted term) underestimates S, but by at most a_{2n+1}, while the sum of the first $2n+1$ terms (truncating at an added term) overestimates S, but by at most a_{2n+2}.

Consider, for example,

$$S = \frac{1}{2} - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{11} - \frac{1}{13} + \cdots = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{p_n}$$

where p_n is the nth prime number. The number S is finite by Leibniz' theorem. If we want to estimate S to within ± 0.001, we note that the 168th prime number is 997, while the 169th is 1009. So the sum of the first 168 terms of the series is within $1/1009$ of the limit, and moreover this partial sum underestimates the limit. We conclude that

$$\sum_{n=1}^{168} \frac{(-1)^{n-1}}{p_n} \leq S \leq \sum_{n=1}^{168} \frac{(-1)^{n-1}}{p_n} + \frac{1}{1009} \leq \sum_{n=1}^{168} \frac{(-1)^{n-1}}{p_n} + 0.001.$$

A tedious calculation shows

$$\sum_{n=1}^{168} \frac{(-1)^{n-1}}{p_n} = 0.269086\ldots$$

while

$$S = 0.269606351\ldots.$$

16.3 Absolute convergence

Leibniz' theorem allows us to deal with the question of convergence of *alternating* series. For series which have more arbitrary patterns of signs, we need the concept of *absolute convergence*.

Definition of absolute convergence If (a_n) is a sequence of real numbers, we say that (a_n) is *absolutely summable* (or, $\sum_{n=1}^{\infty} a_n$ is *absolutely convergent*) if $(|a_n|)$ is summable (that is, if $\sum_{n=1}^{\infty} |a_n|$ converges). If (a_n) is summable but not absolutely summable we say that is *conditionally summable* (or that $\sum_{n=1}^{\infty} a_n$ is *conditionally convergent*).

For example, $((-1)^{n-1}/n)$ is summable (by Leibniz' theorem) but not absolutely summable (the Harmonic series diverges), and so is an example of a conditionally summable sequence; while if $f : \mathbb{N} \to \{+1, -1\}$ is any function then $(f(n)/2^n)$ is absolutely convergent.

The example of $((-1)^{n-1}/n)$ shows that conditional convergence does not imply convergence. On the other hand, we have the following theorem that makes the notion of absolute convergence a very useful one.

Theorem 16.1. *(Absolute convergence implies convergence) If (a_n) is absolutely summable, it is summable.*

Proof: Fix $\varepsilon > 0$. Since $\sum_{n=1}^{\infty} |a_n|$ converges, we get from the Cauchy criterion that there is n_0 such that

$$|a_{n+1}| + \cdots + |a_m| < \varepsilon$$

for all $m > n > n_0$. But the triangle inequality says

$$|a_{n+1} + \cdots + a_m| < |a_{n+1}| + \cdots + |a_m|$$

and so we have

$$|a_{n+1} + \cdots + a_m| < \varepsilon$$

for all $m > n > n_0$, which says (again by the Cauchy criterion) that $\sum_{n=1}^{\infty} a_n$ converges. □

As a corollary we get that if (a_n) is a non-negative summable sequence, and $f : \mathbb{N} \to \{+1, -1\}$ is *any* function, $(f(n)a_n)$ is summable; this allows us to deal with many "irregularly alternating" series. As a specific example, consider

$$\left(\frac{\sin(n^2+1)}{n\sqrt{n}}\right)_{n=1}^{\infty}.$$

It is very difficult to keep track of the way that the sign of $\sin(n^2+1)$ changes as n changes (it does so quite chaotically, starting out

$$1, -1, -1, -1, 1, -1, -1, 1, 1, 1, 1, 1, 1, 1, -1, -1, 1, -1, -1, -1, 1, 1, 1, -1, -1, -1, 1, -1, 1, 1$$

according to `Mathematica`). But, the sequence is easy seen to be absolutely convergent:

$$0 \leq \left|\frac{\sin(n^2+1)}{n\sqrt{n}}\right| \leq \frac{1}{n\sqrt{n}},$$

so we get absolute convergence by comparison with the p-series $(1/n^{3/2})$ (which converges since $p > 1$); and so we get convergence of the original sequence.

We have previously seen that *finite* addition is commutative: for any three reals a_1, a_2, a_3 we have

$$a_1 + a_2 + a_3 = a_1 + a_3 + a_2 = a_2 + a_1 + a_3 = a_2 + a_3 + a_1 = a_3 + a_1 + a_2 = a_3 + a_2 + a_1$$

and more generally, no matter the order that n reals are arranged, their sum remains unchanged.

What about *infinite* addition? If $(a_n)_{n=1}^{\infty}$ is a summable sequence, does the sum depend on the order in which the a_i are written? Unfortunately[244], the answer is *yes*.

[244]Or maybe **fortunately** — odd results like the one described here make the mathematical landscape richer.

Example: By Leibniz' theorem we have that

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \frac{1}{9} - \cdots = L \quad (\star)$$

for some finite number L. Whatever L is, by Leibniz' theorem we have $L > 1 - (1/2) = 1/2 > 0$ (on truncating a Leibniz alternating series after a subtracted term, the partial sum underestimates the limit).

Exercise: Consider the sum

$$1 - \frac{1}{2} - \frac{1}{4} + \frac{1}{3} - \frac{1}{6} - \frac{1}{8} + \frac{1}{5} - \frac{1}{10} \cdots \quad (\star\star)$$

obtained from the left-hand side of $(\star)$ by rearranging the terms as follows: take the first positive term first, then the first two negative terms, then the next positive term, then the next two negative terms, and so on. By combining the 1 with the 1/2, the 1/3 with the 1/6, the 1/5 with the 1/10, and so on (combine two, skip one, repeat), argue that $(\star\star)$ converges to $L/2$ — a different sum to $(\star)$, even though the terms are the same, just written in a different order!

Worse[245] is true:

Theorem 16.2. *If* (a_n) *is* conditionally *summable, then for* any *real number* α *there is a rearrangement*[246] (b_n) *of* (a_n) *with* $\sum_{n=1}^{\infty} b_n = \alpha$*; and there are also rearrangements with* $\sum_{n=1}^{\infty} b_n = \infty$ *and* $\sum_{n=1}^{\infty} b_n = -\infty$.

Proof: We proceed in steps. The details are left as exercises.

- Step 1: Let (c_n) be any sequence. Define the *positive part* of (c_n) to be the sequence (c_n^+) given by

$$c_n^+ = \begin{cases} c_n & \text{if } c_n \geq 0 \\ 0 & \text{if } c_n < 0 \end{cases}$$

and define the *negative part* to be (c_n^-) with

$$c_n^- = \begin{cases} c_n & \text{if } c_n \leq 0 \\ 0 & \text{if } c_n > 0. \end{cases}$$

Exercise: Convince yourself that for each n

$$2c_n^+ = c_n + |c_n| \quad \text{and} \quad 2c_n^- = c_n - |c_n|.$$

Deduce that

If $\sum_{n=1}^{\infty} c_n$ is absolutely convergent then both $\sum_{n=1}^{\infty} c_n^+$ and $\sum_{n=1}^{\infty} c_n^-$ converge.

[245] Or better — see above footnote.

[246] so (b_n) has exactly the same terms as (a_n), just perhaps in a different order.

Next convince yourself that for each n

$$|c_n| = c_n^+ - c_n^-. \quad (\star\star\star)$$

Deduce that

> If both $\sum_{n=1}^{\infty} c_n^+$ and $\sum_{n=1}^{\infty} c_n^-$ converge then $\sum_{n=1}^{\infty} c_n$ is absolutely convergent.

These two facts together say that for any arbitrary sequence (c_n)

> $\sum_{n=1}^{\infty} c_n$ is absolutely convergent if and only if both $\sum_{n=1}^{\infty} c_n^+$ and $\sum_{n=1}^{\infty} c_n^-$ converge.

- Step 2: Now let (a_n) be the conditionally summable sequence hypothesized in the theorem. Let p_n be the positive part of (a_n), and let (q_n) be the negative part. Since (a_n) is not *absolutely* summable, by Step 1 of the proof we know that at least one of $\sum_{n=1}^{\infty} p_n$, $\sum_{n=1}^{\infty} q_n$ diverges (if the former, then to $+\infty$; if the latter, to $-\infty$).

  Exercise: Suppose $\sum_{n=1}^{\infty} p_n$ diverges to $+\infty$, but $\sum_{n=1}^{\infty} q_n$ converges to some fixed number L. Using $a_n = p_n + q_n$ (similar to $(\star\star\star)$ above), argue that $\sum_{n=1}^{\infty} a_n$ diverges to $+\infty$, a contradiction. (We proved something very like this in class.)

  Similarly if $\sum_{n=1}^{\infty} p_n$ converges, but $\sum_{n=1}^{\infty} q_n$ diverges to $-\infty$ then we get the contradiction that $\sum_{n=1}^{\infty} a_n$ diverges to $-\infty$. So the conclusion of Step 2 is

  > If $\sum_{n=1}^{\infty} a_n$ is conditionally convergent then both $\sum_{n=1}^{\infty} p_n$ diverges to $+\infty$ and $\sum_{n=1}^{\infty} q_n$ diverges to $-\infty$.

- Step 3: (**THE MEAT**) Fix $\alpha > 0$. Construct a rearrangement of (a_n) whose sum converges to α, as follows:[247]
  - Start the rearrangement by taking initial terms from (p_n), until the partial sum of the rearrangement thus far constructed either reaches or exceeds α.
    - **Question**: How do we know that such a point can be reached?
    - **Question**: Suppose it is reached exactly when p_{n_1} is added. By how much at most can the partial sum exceed α at this point?
  - Continue the rearrangement by taking initial terms from (q_n) until the partial sum of the rearrangement thus far constructed either reaches or falls below α.
    - **Question**: How do we know that such a point can be reached?
    - **Question**: Before this point is reached, by how much at most can the partial sums exceed α?

[247] Pictures will be *very* helpful here!

* **Question**: Suppose this point is reached exactly when q_{n_1} is "added"[248]. By how much at most can the partial sum fall short of α at this point?

– Continue the rearrangement by going back to where you left off from (p_n), and continuing to take terms from (p_n) until the partial sum of the rearrangement either reaches or exceeds α.

* **Question**: How do we know that such a point can be reached?
* **Question**: Before this point is reached, by how much at most can the partial sums fall short of α?
* **Question**: Suppose this point is reached exactly when p_{n_2} is added. By how much at most can the partial sum exceed α at this point?

– Continue in this manner, swapping back and forth between (p_n) and (q_n) alternately adding from (p_n) until α is reached or exceeded, then adding from (q_n) until α is reached or fallen short of.

* **Question**: How do we know that this process can continue indefinitely?
* **Question**: Suppose that the points where you flip from choosing from one subsequence to the other happen at p_{n_1}, q_{n_1}, p_{n_2}, q_{n_2}, and so on. In terms of these quantities, by how much at most can the partial sums of the rearrangement differ from α?

Exercise: Use what you know[249] about the sequence $(p_{n_1}, q_{n_1}, p_{n_2}, q_{n_2}, \ldots)$ to conclude that the rearrangement just constructed is summable with sum α.

– Step 4: Modify the argument to deal with negative α, $\alpha = 0$, $\alpha = \infty$ and $\alpha = -\infty$.

□

The story for *absolutely* convergent sequences vis-à-vis infinite commutativity is completely different. Here's a theorem that that says that if the sequence (a_n) is absolutely convergent, then the order in which the terms are added does *not* impact the sum.

Theorem 16.3. *If (a_n) is absolutely summable, and (b_n) is any rearrangement of (a_n), then*

- *(b_n) is absolutely summable,*
- *$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} b_n$, and*
- *$\sum_{n=1}^{\infty} |a_n| = \sum_{n=1}^{\infty} |b_n|$.*

Proof: Denote by s_n the partial sums of $(a_k)_{k=1}^{\infty}$ (i.e., $s_n = \sum_{k=1}^{n} a_k$) and by t_n the partial sums of (b_n). Let $\ell = \sum_{n=1}^{\infty} a_n$.

[248] "added" in quotes because q_n is negative.

[249] Something basic, coming from conditional convergence of (a_n).

Fix $\varepsilon > 0$. There is $N \in \mathbb{N}$ such that $|\ell - s_N| < \varepsilon/2$ (by absolute convergence of (a_n)), and any finite sum from $\{|a_{N+1}|, |a_{N+2}|, \cdots\}$ is at most $\varepsilon/2$ (applying the Cauchy criterion to $(|a_n|)$).

Now because (b_n) is a rearrangement of (a_n), there is an $M \in \mathbb{N}$ (perhaps much larger than M) such that all of $a_1, \ldots, a_N$ appear among $b_1, \ldots, b_M$.

For $m > M$, we have

$$\begin{aligned} |\ell - t_m| &= |\ell - s_N - (t_m - s_N)| \\ &\leq |\ell - s_N| + |t_m - s_N| \\ &< \varepsilon \end{aligned}$$

(the last inequality because $|t_m - s_N|$ a finite sum from among $\{|a_{N+1}|, |a_{N+2}|, \cdots\}$). So $\sum_{n=1}^{\infty} b_n = \ell = \sum_{n=1}^{\infty} a_n$.

To deal with $\sum_{n=1}^{\infty} |b_n|$, note that $(|b_n|)$ is a rearrangement of $(|a_n|)$, and since $(|a_n|)$ absolutely summable, by what we just proved, $\sum_{n=1}^{\infty} |b_n|$ converges to $\sum_{n=1}^{\infty} |a_n|$. $\square$

17 Power series

17.1 Introduction to Taylor series

Recall that as an application of the ratio test, we said that

$$\sum_{n\geq 0} \frac{x^{2n}}{(2n)!}$$

converges to a limit for all real x. Let $f(x)$ be limit. We claim that $f(x) = \cosh x$. Indeed, by Taylor's theorem with Lagrange form of the remainder term, we have that for each fixed x there is some t between 0 and x for which

$$\cosh x = 1 + \frac{x^2}{2!} + \cdots + \frac{x^{2n}}{(2n)!} + (\cosh)^{(2n+1)}(t)\frac{x^{2n+1}}{(2n+1)!}.$$

so

$$\left|\cosh x - \left(1 + \frac{x^2}{2!} + \cdots + \frac{x^{2n}}{(2n)!}\right)\right| = \left|\sinh t \frac{x^{2n+1}}{(2n+1)!}\right|.$$

Now between 0 and x, $|\sinh t|$ is never more than $\sinh |x|$, and so

$$\left|\cosh x - \left(1 + \frac{x^2}{2!} + \cdots + \frac{x^{2n}}{(2n)!}\right)\right| \leq |\sinh x| \frac{|x|^{2n+1}}{(2n+1)!}.$$

For each fixed real x, we have that $|\sinh x||x^{2n+1}|/(2n+1)! \to 0$ as $n \to \infty$ and so, by definition of convergence of a sequence to a limit, we get that

$$\left(1 + \frac{x^2}{2!} + \cdots + \frac{x^{2n}}{(2n)!}\right)_{n=0}^{\infty} \to \cosh x$$

for each fixed real x, that is,

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

Note that we used the ratio test to conclude that $(x^{2n}/(2n)!)$ converges for all real x, but in fact that was unnecessary, given Taylor's theorem. Indeed, suppose we have a function $f(x)$, and we know that for all fixed x in some set A, the remainder term $R_{n,a,f}(x)$ goes to 0 as n goes to infinity. Then we have

$$|f(x) - P_{n,a,f}(x)| \to 0$$

as n goes to infinity, so that, directly from the definition of convergence, for each fixed $x \in A$, $(f^{(n)}(a)(x-a)^n/n!)$ is summable, and the sum is $f(x)$. That is, for $x \in A$

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

So, for example, we have

- $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ for all real x,
- $\sin x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!}$ for all real x,
- $\cos x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!}$ for all real x,
- $\tan^{-1} x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1}$ for $-1 \leq x \leq 1$ (and note that when $|x| > 1$, the sum *does not converge*, by the vanishing criterion, although $\tan^{-1}$ is defined for $|x| > 1$).

The expression

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

is called the *Taylor series*[250] *of* f *at* (or *about*) a (and evaluated at x); think of it as an "infinite degree Taylor polynomial". Sometimes, the Taylor series evaluates to the function it is the Taylor series of, at all points in the domain of the function (e.g. when $f(x) = e^x$ and $a = 0$); sometimes, it evaluates to the function at some points of the domain, but not all of the function (e.g. when $f(x) = \tan^{-1}(x)$ and $a = 0$); and sometimes, it evaluates to the function at the single points of the domain a (e.g. when

$$f(x) = \begin{cases} e^{-1/x^2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0, \end{cases}$$

at $a = 0$). We will discuss this in more detail later.

We now consider the degree to which the Taylor series of a function is a useful tool for understanding the function. We start with an example. For clarity, write

- $\sin_n(x)$ for $P_{2n+1,0,\sin}(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!}$

and

- $\cos_n(x)$ for $P_{2n,0,\cos}(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots + (-1)^n \frac{x^{2n}}{(2n)!}$.

As we have earlier observed, we have that for all real x

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

and

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

Notice that

$$\int_0^x \cos_n(t)\, dt = \sin_n(x)$$

[250] When $a = 0$, the Taylor series is often referred to as the *Maclaurin* series.

so that

$$\sin x = \lim_{n\to\infty} \int_0^x \cos_n(t)\ dt. \quad (\star)$$

Can we use $(\star)$ to conclude that $\sin_x = \int_0^x \cos(t)\ dt$ for all real x? In other words, is it true that

$$\lim_{n\to\infty} \int_0^x \cos_n(t)\ dt = \int_0^x \lim_{n\to\infty} \cos_n(t)\ dt?$$

In general, we cannot draw such a conclusion. That is, if $f_n : \mathbb{R} \to \mathbb{R}$ is a sequence of functions, one for each $n \in \mathbb{N}$, with the properties that for all real x, each of

- $\lim_{n\to\infty} f_n(x)$,
- $\int_0^x \lim_{n\to\infty} f_n(t)\ dt$
- $\int_0^x f_n(t)\ dt$ and
- $\lim_{n\to\infty} \int_0^x f_n(t)\ dt$

exist, we *cannot* conclude that

$$\lim_{n\to\infty} \int_0^x f_n(t)\ dt = \int_0^x \lim_{n\to\infty} f_n(t)\ dt \quad (\star\star)$$

for all x.

Here is a counter-example. Define f_n via:

$$f_n(x) = \begin{cases} 0 & \text{if } x \le 0 \text{ or } x \ge 2/n \\ n^2 x & \text{if } 0 \le x \le 1/n \\ 2n - n^2 x & \text{if } 1/n \le x \le 2/n. \end{cases}$$

Observe[251] that for each real x,

- $\lim_{n\to\infty} f_n(x) = 0$ (if $n \le 0$, this is automatic; if $x > 0$ then as long as $n > 2/x$ we have $f_n(x) = 0$),
- $\int_0^x \lim_{n\to\infty} f_n(t)\ dt = 0$ (this is automatic from the last observation),
- $\int_0^x f_n(t)\ dt$ exists for all n (f_n is continuous and bounded) and
- $$\lim_{n\to\infty} \int_0^x f_n(t)\ dt = \begin{cases} 0 & \text{if } x \le 0 \qquad\qquad\qquad\qquad\qquad\qquad \text{(immediate)} \\ 1 & \text{if } x > 0 \quad (\text{for all } n > 2/x,\ \int_0^x f_n(t)\ dt = 1). \end{cases}$$

So

$$\lim_{n\to\infty} \int_0^x f_n(t)\ dt \neq \int_0^x \lim_{n\to\infty} f_n(t)\ dt$$

for all x; only for $x \le 0$.

In the next section, we isolate the condition that is needed to make $(\star\star)$ hold, and explore its consequences.

[251] Draw a graph to see what's going on!

17.2 Pathologies of pointwise convergence

At the end of the last section we saw that if a sequence of integrable functions (f_n) converges *pointwise* to an integrable limit f on some set A, meaning that for each $x \in A$ we have $f_n(x) \to f(x)$ as $n \to \infty$, then it is not necessarily that case that the limit of the integrals of the f_n's is the integral of the limit.

It is fairly clear what is going on in the counter-example: although the functions f_n are converging, at each real x, to the value 0, there is some sense in which the functions f_n as a whole are not converging to 0: we cannot put an ε-width window around the x-axis, and then find an N large enough so that for all $n > N$ the graph of f_n is lying completely inside that window. What we can do is, for each specific x, find an N so that for all $n > N$ we have $f_n(x)$ within ε of $f(x)$; but as the x's get closer to 0, the required N gets larger and larger.

Here's another, perhaps more worrying example. Let $f_n : \mathbb{R} \to \mathbb{R}$ be defined by

$$f_n(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x^n & \text{if } 0 \leq 1 \leq 1 \\ 1 & \text{if } x \geq 1. \end{cases}$$

For each x we have that $f_n(x)$ converges to a limit as n grows. Specifically,

$$\lim_{n\to\infty} f_n(x) = \begin{cases} 0 & \text{if } x < 1 \\ 1 & \text{if } x \geq 1. \end{cases}$$

Notice that

> Each f_n is continuous, but the limit f is not.

So, the limit of continuous functions need not be continuous. This example can be modified slightly, by rounding the corners of f_n, to also yield a sequence of differentiable functions that tends to a limit, but not to a differentiable limit.

It's not as clear what the problem might be in this second (pair of) example(s); but a little thought shows that it suffers from the same issue as the first example. Given $\varepsilon > 0$, for each x there is an N large enough so that $f_n(x)$ is within ε of $\lim_{n\to\infty} f_n(x)$ for all $n > N$. But if we put an ε-width window around the limit function (a strip of width ε around the x-axis, for $x < 1$, and a strip of width ε around the line $y = 1$, for $x \geq 1$), then there is no single N with the property that for all $n > N$ the graph of f_n lies inside this window: for each $n \in \mathbb{N}$, the point $(\sqrt[n]{(1/2)}, 1/2)$ is on the graph of f_n, but it is not in the window as long as $\varepsilon < 1/2$.

So: we can talk about a sequence (f_n) of functions (all with the same domain A) converging to a limit f (also with domain A), if for each $x \in A$ we have $f_n(x) \to f(x)$ as $n \to \infty$. This kind of convergence is called *pointwise* convergence. But pointwise convergence appears to be quite weak; it preserves neither continuity nor integrability.

Or: we can talk about a sequence (f_n) of functions (all with the same domain A) converging to a limit f (also with domain A) if the graphs of the f_n's converge to the graph of f, in the

sense that for all $\varepsilon > 0$ there is N such that for all $n > N$ the graph of f_n lies completely inside a window of width ε around the graph of f. This notion of converges is easily seen to imply pointwise convergence; and as we are about to see, it behaves much better with respect to continuity, integrability and differentiability. Before exploring this, we made the definition of this kind of convergence, which we call *uniform* convergence, precise.

17.3 Definition and basic properties of uniform convergence

We begin with the definition.

Definition of uniform convergence Let (f_n) be a sequence of functions, all defined on some domain A, and let f be a function, also defined on the domain A. Say that (f_n) *converges uniformly to f on A* if for all $\varepsilon > 0$ there is N such that $n > N$ implies that for all $x \in A$,

$$|f_n(x) - f(x)| < \varepsilon.$$

It's worth symbolically comparing the definition of "$f_n \to f$ pointwise on A":

$$(\forall \varepsilon > 0)(\forall x \in A)(\exists N)((n > N) \Rightarrow (|f_n(x) - f(x)| < \varepsilon))$$

with that of "$f_n \to f$ uniformly on A":

$$(\forall \varepsilon > 0)(\exists N)(\forall x \in A)((n > N) \Rightarrow (|f_n(x) - f(x)| < \varepsilon)).$$

In pointwise convergence, N depends on both ε *and* x; but in uniform convergence, N only depends on ε — for each ε, the *same* N works for every x. This should seem familiar — it is essentially the same distinction as between continuity and uniform continuity.

The value of uniform convergence is conveyed in the following three theorems, which (roughly) say that

- the uniform limit of continuous functions is continuous,
- the integral of the uniform limit is the limit of the integrals, and
- (modulo some extra conditions) the derivative of the uniform limit is the limit of the derivatives.

Theorem 17.1. *Let (f_n) be a sequence of functions that are all defined and continuous on $[a, b]$, and that converge uniformly on $[a, b]$ to some a function f. Then f is continuous on $[a, b]$.*

Proof: Unsurprisingly, we will show the slightly stronger statement that f is *uniformly* continuous on $[a, b]$. Given $\varepsilon > 0$ we want a $\delta > 0$ such that $|x - y| < \delta$ (with $x, y \in [a, b]$) implies $|f(x) - f(y)| < \varepsilon$. We write

$$\begin{aligned} |f(x) - f(y)| &= |(f(x) - f_n(x)) + (f_n(x) - f_n(y)) + (f_n(y) - f(y))| \\ &\leq |f(x) - f_n(x)| + |f_n(x) - f_n(y)| + |f(y) - f_n(y)|. \quad (\star) \end{aligned}$$

By *uniform* convergence of f_n to f, we know that there is N such that $n > N$ implies both

$$|f(x) - f_n(x)| < \frac{\varepsilon}{3}, \quad |f(y) - f_n(y)| < \frac{\varepsilon}{3}. \quad (\star\star)$$

Fix any $n > N$. Since f_n is uniformly continuous (it is a continuous function on a closed interval), there is $\delta > 0$ such that $|x - y| < \delta$ (with $x, y \in [a, b]$) implies $|f_n(x) - f_n(y)| < \varepsilon/3$. But this implies, via $(\star)$ and $(\star\star)$ that whenever $|x - y| < \delta$ (with $x, y \in [a, b]$) we have

$$|f(x) - f(y)| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

□

Theorem 17.2. *Let (f_n) be a sequence of functions that are all defined and integrable on $[a, b]$, and that converge uniformly on $[a, b]$ to some a function f. Then f is integrable*[252] *on $[a, b]$ and*

$$\int_a^b f = \lim_{n\to\infty} \int_a^b f_n.$$

Proof: We first check that f is bounded. By uniform convergence of (f_n) to f there is N such that $n > N$ implies $|f_n(x) - f(x)| < 1$ for all $x \in [a, b]$, which says that $|f(x)| < |f_n(x)| + 1$. Now fix such an $n > N$. Since f_n is integrable, it is bounded on $[a, b]$, say by M; but then that says that f is bounded on $[a, b]$ by $M + 1$.

Next we show that f is integrable on $[a, b]$. Fix $\varepsilon > 0$. By uniform convergence of (f_n) to f there is N such that $n > N$ implies

$$|f_n(x) - f(x)| < \frac{\varepsilon}{4(b - a)}$$

for all $x \in [a, b]$. Fix such an $n > N$. Since f_n is integrable on $[a, b]$, there is a partition $P = (t_0, \ldots, t_m)$ of $[a, b]$ with

$$U(f_n, P) - L(f_n, P) = \sum_{i=1}^{m} (M_i^n - m_i^n)(t_i - t_{i-1}) < \varepsilon/2,$$

where M_i^n is the supremum of f_n on $[t_{i-1}, t_i]$, and m_i^n is the infimum.

Because $|f_n(x) - f(x)| < \frac{\varepsilon}{4(b-a)}$ for all $x \in [a, b]$, we have that

$$M_i \leq M_i^n + \frac{\varepsilon}{4(b - a)}, \quad m_i \geq m_i^n - frac\varepsilon 4(b - a)$$

where M_i is the supremum of f on $[t_{i-1}, t_i]$, and m_i is the infimum. It follows that

$$M_i - m_i \leq M_i^n - m_i^n + \frac{\varepsilon}{2(b - a)}$$

[252] Spivak takes integrability of f as a hypothesis, but it is in fact implied by uniform convergence, and is a good exercise in definition of the integral.

for each i, and so

$$\begin{aligned} U(f_n,P)-L(f_n,P) &= \sum_{i=1}^{m}(M_i-m_i)(t_i-t_{i-1}) \\ &\leq \sum_{i=1}^{m}\left(M_i^n-m_i^n+\frac{\varepsilon}{2(b-a)}\right)(t_i-t_{i-1}) \\ &= \sum_{i=1}^{m}(M_i^n-m_i^n)(t_i-t_{i-1})+\frac{\varepsilon}{2(b-a)}\sum_{i=1}^{m}(t_i-t_{i-1}) \\ &< \frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon, \end{aligned}$$

and so we conclude that f is integrable on $[a,b]$.

Finally, we show that $\int_a^b f = \lim_{n\to\infty}\int_a^b f_n$. Fix $\varepsilon>0$. By uniform convergence of (f_n) to f there is N such that $n>N$ implies

$$|f_n(x)-f(x)|<\frac{\varepsilon}{b-a}$$

for all $x\in[a,b]$. So (using integrability of f_n and f to justify the existence of the integrals below)

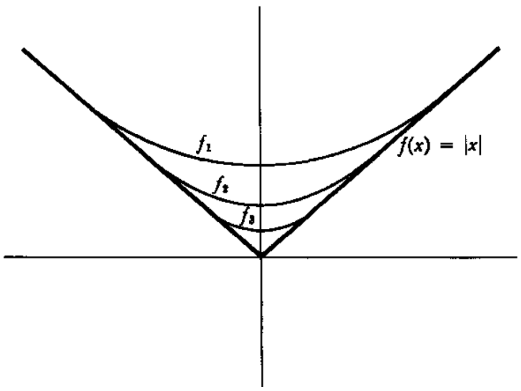
$$\left|\int_a^b f_n-\int_a^b f\right|\leq\int_a^b|f_n-f|<(b-a)\frac{\varepsilon}{b-a}=\varepsilon,$$

which says $\int_a^b f_n\to\int_a^b f$ as $n\to\infty$. □

It would be nice to now have a theorem that says

> Let (f_n) be a sequence of functions that are all defined and differentiable on $[a,b]$, and that converge uniformly on $[a,b]$ to some a function f. Then f is differentiable on $[a,b]$ and $f'(x)=\lim_{n\to\infty}f_n'(x)$ for each $x\in[a,b]$.

Unfortunately, such a theorem is false; the uniform limit of differentiable functions on a closed interval need not be differentiable, as shown in this picture taken from Spivak (Chapter 24, Figure 8):



And Spivak has a further example showing that even if f is differentiable, it need not be the case that $f'(x) = \lim_{n\to\infty} f_n'(x)$ for each $x \in [a,b]$. Some quite careful hypothesis are needed to get an analog of Theorems 17.1 and 17.2 for differentiability. Unlike Theorem 17.2, where the hypothesis were natural but the proof involved, here the proof is very easy, given the correct hypotheses.

Theorem 17.3. *Let (f_n) be a sequence of functions that are all defined and differentiable on $[a,b]$, and that converge pointwise on $[a,b]$ to some a function f. Suppose also that each f_n' is integrable on $[a,b]$, and that the f_n' converge uniformly to* some *continuous function g. Then f is differentiable on $[a,b]$ and $f'(x) = \lim_{n\to\infty} f_n'(x)$ for each $x \in [a,b]$.*

Proof: From Theorem 17.2, then the fundamental theorem of calculus we have that for each $x \in [a,b]$,

$$\begin{aligned}\int_a^x g &= \lim_{n\to\infty} \int_a^x f_n' \\ &= \lim_{n\to\infty} (f_n(x) - f_n(a)) \\ &= f(x) - f(a).\end{aligned}$$

So

$$f(x) = f(a) + \int_a^x g,$$

and (using continuity of g) another application of the fundamental theorem of calculus gives that f is differentiable at x, and that

$$f'(x) = g(x) = \lim_{n\to\infty} f_n'(x).$$

□

17.4 Application to power series

Definition of a power series A *power series* (at, or about, 0) is a series of the form

$$a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n + \cdots$$

where x and $a_0, a_1, \ldots$ are real numbers. More generally, for a real number a a *power series* at, or about, a is a series of the form

$$a_0 + a_1(x-a) + a_2(x-a)^2 + \cdots + a_n(x-a)^n + \cdots .$$

At least initially, we'll only think about power series at 0. An example of a power series is the Taylor series of function f that is infinitely differentiable at 0:

$$f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \cdots + \frac{f^{(n)}(0)}{n!}x^n + \cdots .$$

Consistent with our previous notation, we say that a power series *converges* if the sequence $(a_n x^n)$ is summable, and *converges absolutely* if $(|a_n||x|^n)$ is summable. By definition of summability, saying that a power series converges is equivalent to saying that the sequence $(s_n(x))$ of partial sums converges, where

$$s_n(x) = \sum_{k=0}^{n} a_k x^k,$$

and saying that a power series converges absolutely is equivalent to saying that the sequence $(\bar{s}_n(x))$ of "absolute" partial sums converges, where $\bar{s}_n(x) = \sum_{k=0}^{n} |a_k||x|^k$.

Our aim now is to use the results of the last section to explore situations in which a power series converges, and the degree to which a power series can be manipulated in "natural" ways (specifically, via term-by-term integration and differentiation). Rather than stating a general theorem (which would be rather long) we will develop the theory piece by piece.

So, let $f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n + \cdots$ be a power series, with partial sums $s_n(x) = \sum_{k=0}^{n} a_k x^k$ and "absolute" partial sums $\bar{s}_n(x) = \sum_{k=0}^{n} |a_k||x|^k$. Note that f here is not (yet) naming a function, since the power series may not be summable; it is simply convenient shorthand for the expression $a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n + \cdots$.

Basic working hypothesis Suppose that $x_0 \geq 0$ is a real number satisfying that

$$\lim_{n\to\infty} \frac{|a_{n+1}|}{|a_n|} x_0 \text{ exists and equal some } \ell < 1.$$

We begin by establishing that under this hypothesis,

for all $x \in [-x_0, x_0]$ the power series $f(x)$ converges to a finite limit.

Indeed, for $x = 0$ the claim is trivial, and for $x \neq 0$ apply the ratio test to the series $\sum_{n=0}^{\infty} |a_n x^n|$. We get that

$$\frac{|a_{n+1}x^{n+1}|}{|a_n x^n|} = \left(\frac{|a_{n+1}|}{|a_n|} x_0\right) \frac{|x|}{x_0} \to \ell \frac{|x|}{x_0} < 1,$$

the existence of the limit following from the basic working hypothesis. From the ratio test we conclude that $(a_n x^n)$ is absolutely summable, and so in particular is summable, i.e., the power series is both convergent and absolutely convergent. So now f is a function with domain $[-x_0, x_0]$, and $f(x)$ is more than just shorthand for the power series; it is the value of the function.

Next we establish that

the function f is continuous on $[-x_0, x_0]$.

What we will actually show is that

the sequence of partial sums $s_n(x)$ converges *uniformly* to $f(x)$ on $[-x_0, x_0]$

(we already know that is converges pointwise). From Theorem 17.1, the continuity of f on $[-x_0, x_0]$ follows immediately. To see uniform convergence, note that for $x \in [-x_0, x_0]$ we have

$$\begin{aligned} |f(x) - s_n(x)| &= \left| \sum_{k=n+1}^{\infty} a_k x^k \right| \\ &\leq \sum_{k=n+1}^{\infty} |a_k||x|^k \\ &\leq \sum_{k=n+1}^{\infty} |a_k||x_0|^k. \end{aligned}$$

Because $(a_k x_0^k)$ is absolutely summable (we just proved that above), for every $\varepsilon > 0$, n can be chosen large enough that $\sum_{k=n+1}^{\infty} |a_k||x_0|^k < \varepsilon$,[253] and since the choice of n depends only on x_0 (and ε) it works for all $x \in [-x_0, x_0]$.

Notice that in the first inequality above we have slipped in a "triangle inequality" for series: if (c_n) is absolutely summable (so both $\sum_{n=1}^{\infty} c_n$ and $\sum_{n=1}^{\infty} |c_n|$ exist), then

$$\left| \sum_{n=1}^{\infty} c_n \right| \leq \sum_{n=1}^{\infty} |c_n|.$$

This is an easy exercise.

In summary, so far we have show that if $a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n + \cdots$ is a power series, and if x_0 is a non-negative number for which $\lim_{n\to\infty} |a_{n+1}|x_0/|a_n|$ exists and is less than 1, then

- for each $x \in [-x_0, x_0]$, $a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n + \cdots$ converges to a finite limit $f(x)$, and in fact converges absolutely,
- the convergence is uniform on $[-x_0, x_0]$, and
- f is continuous on $[-x_0, x_0]$.

The situation described above is often referred to as *absolute uniform convergence* of the power series; and the power series is said to converge *absolutely uniformly.*

Next, we establish that

> f is differentiable on $[-x_0, x_0]$, and its derivative is given by
>
> $$f'(x) = a_1 + 2a_2 x + 3a_3 x^2 + \cdots + (n+1)a_{n+1} x^n + \cdots .$$
>
> Moreover, the sequence described above is absolutely uniformly convergent on $[-x_0, x_0]$.

[253] Formally: by the Cauchy condition, n can be chosen large enough so that for all $m > n$ we have $\sum_{k=n+1}^{m} |a_k||x_0|^k < \varepsilon$, so by the boundedness criterion, for suitably large n we have that $\sum_{k=n+1}^{\infty} |a_k||x_0|^k$ exists and is at most ε.

To verify all this, the first thing we do is check that $a_1+2a_2x+3a_3x^2+\cdots+(n+1)a_{n+1}x^n+\cdots$ is absolutely uniformly convergent on $[-x_0, x_0]$. But this follows immediately from what we have just done:

$$\lim_{n\to\infty} \frac{(n+2)|a_{n+2}|}{(n+1)|a_{n+1}|}|x_0| = \lim_{n\to\infty} \frac{|a_{n+1}|}{|a_n|}|x_0| < 1$$

(by the basic working hypothesis for $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n + \cdots$), so the basic working hypothesis is satisfied for $a_1 + 2a_2x + 3a_3x^2 + \cdots + (n+1)a_{n+1}x^n + \cdots$, and we can apply all the results we have proven about $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n + \cdots$ to $a_1 + 2a_2x + 3a_3x^2 + \cdots + (n+1)a_{n+1}x^n + \cdots$.

So now we know that the sequence $(t_n(x))$ of partial sums of $a_1 + 2a_2x + 3a_3x^2 + \cdots + (n+1)a_{n+1}x^n + \cdots$ (where $t_n(x) = \sum_{k=0}^{n}(k+1)a_{k+1}x^k$) converges uniformly to a limit that is continuous on $[-x_0, x_0]$. But $t_n(x) = s'_{n+1}(x)$, where recall $s_n(x) = \sum_{k=0}^{n} a_kx^k$ is the partial sum of $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n + \cdots$. It follows that $(s_n(x))$ converges uniformly to a limit that is continuous on $[-x_0, x_0]$. Now also (s_n) is a sequence of functions that are all defined and differentiable on $[-x_0, x_0]$, and that converge pointwise (in fact, uniformly) on $[-x_0, x_0]$ to f. And finally, clearly each s'_n is integrable on $[-x_0, x_0]$. So all the hypotheses of Theorem 17.3 are satisfied, and we conclude that f is differentiable on $[-x_0, x_0]$ and

$$f'(x) = \lim_{n\to\infty} s'_n(x) = a_1 + 2a_2x + 3a_3x^2 + \cdots + (n+1)a_{n+1}x^n + \cdots$$

for each $x \in [-x_0, x_0]$.

This whole argument can be repeated of f'; we conclude that f is *infinitely* differentiable on $[-x_0, x_0]$, and that the successive derivatives can be found by differentiating the power series term-by-term.

Finally, we turn to integrability.

> f is integrable on $[-x_0, x_0]$, and the function $g : [-x_0, x_0] \to \mathbb{R}$ defined by $g(x) = \int_0^x f(t)\,dt$ is given by
>
> $$g(x) = a_0x + \frac{a_1}{2}x^2 + \frac{a_2}{3}x^3 + \cdots + \frac{a_{n-1}}{n}x^n + \cdots .$$
>
> Moreover, the sequence described above is absolutely uniformly convergent on $[-x_0, x_0]$.

The absolute uniform convergence of the sequence is verified exactly as in the case of differentiation. For the rest, note that since $(s_n(x))$ converges uniformly to f on $[-x_0, x_0]$, we can apply Theorem 17.3 to conclude that f is integrable on $[-x_0, x_0]$[254] and that

$$\begin{aligned}
\int_0^x f(t)\,dt &= \lim_{n\to\infty} \int_0^x s_n(t)\,dt \\
&= \lim_{n\to\infty} \int_0^x \left(a_0x + \frac{a_1}{2}x^2 + \frac{a_2}{3}x^3 + \cdots + \frac{a_n}{n+1}x^{n+1}\right) \\
&= a_0x + \frac{a_1}{2}x^2 + \frac{a_2}{3}x^3 + \cdots + \frac{a_{n-1}}{n}x^n + \cdots .
\end{aligned}$$

[254] We already knew that f was integrable, since it is continuous. But there is no harm in discovering this fact by a second route.

This whole argument can be repeated of g; we conclude that f is *infinitely* integrable on $[-x_0, x_0]$[255], and that the successive integrals can be found by integrating the power series term-by-term.

Time for a second (and final) summary: if $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n + \cdots$ is a power series, and if x_0 is a non-negative number for which $\lim_{n\to\infty} |a_{n+1}|x_0/|a_n|$ exists and is less than 1, then

- for each $x \in [-x_0, x_0]$, $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n + \cdots$ converges absolutely uniformly to a finite limit $f(x)$,
- f is continuous on $[-x_0, x_0]$,
- f is differentiable arbitrarily many times on $[-x_0, x_0]$; each derivative is found by differentiating the power series term-by-term, and the resulting power series is still absolutely uniformly convergent on $[-x_0, x_0]$,
- $g(x) = \int_0^x f(t)\ dt$ is defined on $[-x_0, x_0]$; it is found by integrating the power series term-by-term, and the resulting power series is still absolutely uniformly convergent on $[-x_0, x_0]$ (so the process can be repeated arbitrarily many times).

We refer to all this as the power series being "nice".

Of course, we would like the power series to be nice for as large an x_0 as possible. If $\lim_{n\to\infty} |a_{n+1}|/|a_n|$ doesn't exist, then there is no x_0 (other than $x_0 = 0$) for which $\lim_{n\to\infty} |a_{n+1}|x_0/|a_n|$ exists. So in this case, we will say nothing.

If $\lim_{n\to\infty} |a_{n+1}|/|a_n| = 0$ then for *all* x_0 we have $\lim_{n\to\infty} |a_{n+1}|x_0/|a_n| = 0 < 1$. So in this case, the power series is nice on the whole real line.

If $\lim_{n\to\infty} |a_{n+1}|/|a_n| = 1/R > 0$, then as long as $x_0 < R$, the power series is nice on $[-x_0, x_0]$. On the other hand, it is easy to see (by the ratio test, or just by the fact that the nth term does not go to zero) that if $|x_0| > R$ then the power series does not converge.

If $\lim_{n\to\infty} |a_{n+1}|/|a_n| = +\infty$ then only for $x_0 = 0$ do we have $\lim_{n\to\infty} |a_{n+1}|x_0/|a_n| < 1$, and again it is easy to see (by the fact that the nth term does not go to zero) that if $x_0 \neq 0$ then the power series does not converge. So in this case, the power series is nice only at 0.

So, in summary:

> if $\lim_{n\to\infty} |a_{n+1}|/|a_n| = 1/R > 0$, then the power series is nice on every closed interval contained in $(-R, R)$, but not on any closed interval that includes anything outside $[-R, R]$.[256] R is referred to as the *radius of convergence* of the power series. If $\lim_{n\to\infty} = 0$, then the power series is nice on every closed interval contained in the reals; in this case we say that the radius of convergence is infinite. If $\lim_{n\to\infty} |a_{n+1}|/|a_n| = +\infty$ then the power series is nice only at 0; in this case we say that the radius of convergence is 0. In all three cases, the power series has

[255] No surprise — unlike with differentiation, once a function is integrable once, it is infinitely integrable

[256] What happens at R and $-R$ has to be dealt with on a case-by-case basis.

a radius of convergence, inside of which it is nice, and outside of which it does not converge.

If $\lim_{n\to\infty} |a_{n+1}|/|a_n|$ does not exist, then the power series might still have a radius of convergence; that is a story for another day (year?) involving the notion of lim sup.

For power series at values a other than zero, that is, power series of the form

$$a_0 + a_1(x-a) + a_2(x-a)^2 + \cdots + a_n(x-a)^n + \cdots,$$

the story is essentially the same: with R defined exactly as before, the power series is nice on every closed interval contained in $(a-R, a+R)$; the proof is almost exactly the same, just a little bit more annoying as it involves the extra parameter a.

If a power series $a_0 + a_1x + a_2x^2 + \cdots$ has some radius of convergence R, what can we say about the function $f(x)$ that the power series converges to, inside $(-R, R)$? Well,

- $f(0) = a_0$.
- $f'(x) = a_1 + 2a_2x + 3a_3x^2 + 4a_4x^3 + \cdots$ so $f'(0) = a_1$.
- $f''(x) = 2a_2 + 6a_3x + 12a_4x^2 + \cdots$ so $f''(0) = 2a_2$ or $a_2 = f''(0)/2$.
- $f'''(x) = 6a_3 + 24a_4x + \cdots$ so $f'''(0) = 6a_3$ or $a_3 = f'''(0)/6 = f'''(0)/3!$.
- In general, $a_n = f^{(n)}(0)/n!$

In other words:

> If a power series $a_0 + a_1x + a_2x^2 + \cdots$ has some radius of convergence R, and so converges to a function $f(x)$ inside $(-R, R)$, then $a_0 + a_1x + a_2x^2 + \cdots$ is in fact the Taylor series of f. In particular, that means that is any *other* power series converges to f on $(-R, R)$, that other power series must be identical to $a_0 + a_1x + a_2x^2 + \cdots$.

We finish up this section with some examples.

- $\exp, \sin, \cos$: we already know, from Taylor's theorem, that
  - $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ for all real x,
  - $\sin x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!}$ for all real x,
  - $\cos x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!}$ for all real x,

  and it is easy to apply the theory we have developed to show that each of these power series are absolutely uniformly convergent on every closed interval in the reals (so, radius of convergence is $+\infty$ in each case). This allows us to confirm, by term-by-term differentiation, that all the well-known relations between these functions hold, such as

– the derivative of e^x is e^x,

– an antiderivative of cos is sin, and

– $\sin'' + \sin = 0$, $\cos'' + \cos = 0$.

Recall that we proved that the only continuous functions $f : \mathbb{R} \to \mathbb{R}$ satisfying $f' = f$ are functions of the form $f(x) = ae^x$, so in particular the only $f : \mathbb{R} \to \mathbb{R}$ satisfying $f' = f$ and $f(0) = 1$ is $f(x) = e^x$; and that the only continuous functions $f : \mathbb{R} \to \mathbb{R}$ satisfying $f'' + f = 0$ are functions of the form $f(x) = a \sin x + b \cos x$, so in particular the only $f : \mathbb{R} \to \mathbb{R}$ satisfying $f'' + f =$ and $f(0) = 0$, $f'(0) = 1$ is $f(x) = \sin x$, while the only $f : \mathbb{R} \to \mathbb{R}$ satisfying $f'' + f =$ and $f(0) = 1$, $f'(0) = 0$ is $f(x) = \cos x$. But from the theory we have just developed, we can conclude (without ever knowing anything about the exponential function) that if $f(x)$ is defined by

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

then $f' = f$ and $f(0) = 1$; so we could use this as an alternate *definition* of the exponential function. Similarly, we could formally define sin and cos via power series. Many authors take this approach.

- log: We can easily check that the power series

$$f(x) = 1 + x + x^2 + \cdots + x^n + \cdots$$

has radius of convergence 1. We also already know that $f(x) = 1/(1 - x)$ for all $x \in (-1, 1)$. So, substituting $-x$ for x, we have

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

By the theory we have just developed, we have

$$\int_0^x \frac{dt}{1+t} = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots + (-1)^{n-1} \frac{x^n}{n} + \cdots ,$$

valid for all x between -1 and 1. But $\int_0^x \frac{dt}{1+t} = \log(1 + x)$, and the power series on the right above is the Taylor series at 0 of $\log(1 + x)$; so we have just shown that the Taylor series of $\log(1 + x)$ converges to $\log(1 + x)$, that is

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

for $x \in (-1, 1)$. Recall from a homework exercise that it was only easy to analyze the remainder term of the Taylor polynomial, and obtain the above result, for $-1/2 < x < 1$. The theory of uniform convergence allows us to easily fill in the gap.

- Estimating integrals. Power series can be used to estimate integrals. For example,

$$e^{-x^2} = 1 - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} + \cdots$$

(just evaluate the series for e^x at $-x^2/2$; since the series is nice for all real x, it is nice for $-x^2/2$). The above series for e^{-x^2} is easily seen to have infinite radius of convergence, so we can write

$$\int_0^x e^{-t^2}\, dt = x - \frac{x^3}{3} + \frac{x^5}{5 \cdot 2!} - \frac{x^7}{7.3!} + \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)n!} + \cdots .$$

For example, what is $\int_0^1 e^{-t^2}\, dt$? Setting $x = 1$ above, and using Leibniz test for alternating series (in the strong form that gives an error estimate of the magnitude of the next term after truncation) we get that for each n

$$\int_0^1 e^{-t^2}\, dt \left(\sum_{k=0}^{n} (-1)^k \frac{1}{(2k+1)k!} \right) \pm \frac{1}{(2n+3)(n+1)!}$$

Taking $n = 10$ we get

$$\int_0^1 e^{-t^2}\, dt = 0.746824133 \pm 0.000000001.$$

- More general series. What if $\lim_{n\to\infty} \frac{|a_{n+1}|}{|a_n|}$ doesn't exist? If $(|a_{n+1}|/|a_n|)$ is eventually bounded above by $M > 0$, then whole theory described above goes through, as long as $Mx_0 < 1$[257].

For an example, consider the Fibonacci numbers, defined by $F_0 = 0, F_1 = 1$ and $F_n = F_{n_1} + F_{n-2}$ for $n \geq 2$. We can form the "generating function" of the Fibonacci numbers:

$$F(x) = F_0 + F_1 x + F_2 x^2 + \cdots .$$

Since (F_n) is increasing we have

$$\frac{F_{n+1}}{F_n} \leq \frac{2F_n}{F_n} \leq 2,$$

and so $F(x)$ converges absolutely uniformly on some interval around 0 (for example, on $[-0.49, 0.49]$). We have

$$\begin{aligned} F(x) &= F_0 + F_1 x + F_2 x^2 + \cdots \\ &= x + (F_1 + F_0)x^2 + (F_2 + F_1)x^3 + \cdots \\ &= x + x(F_1 x + F_2 x^2 + \cdots) + x^2(F_0 + F_1 x + F_2 x^2 + \cdots) \\ &= x + xF(x) + x^2 F(x) \end{aligned}$$

[257] Exercise!

(with all manipulations easily seen to be valid inside the interval $[-0.49, 0.49]$) so

$$\begin{aligned} F(x) &= \frac{x}{1-x-x^2} \\ &= \frac{A}{1-\varphi_1 x} + \frac{B}{\varphi_2 x}, \quad \text{(partial fractions)} \end{aligned}$$

where $(1-\varphi_1 x)(1-\varphi_2 x) = 1-x-x^2$, so $\varphi_1 + \varphi_2 = 1$, $\varphi_1\varphi_2 = -1$, so

$$\varphi_1 = \frac{1+\sqrt{5}}{2}, \quad \varphi_2 = \frac{1-\sqrt{5}}{2}$$

and A, B are constants. So

$$F_n = A\varphi_1^n + B\varphi_2^n.$$

(This is obtained by using the obvious power series for $1/(1-\varphi_1 x)$ and $1/(1-\varphi_2 x)$). At $n = 0$ we get $A + B = 0$, and at $n = 1$ we get $A\varphi_1 + B\varphi_2 = 1$, so

$$A = \frac{1}{\sqrt{5}}, \quad B = \frac{-1}{\sqrt{5}},$$

and so finally we get the remarkable *Binet's formula*:

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

All manipulations were valid because all power series involved converge in some non-zero interval around 0, and everything can be done inside an interval of convergence that works for all power series involved.

There is a vast literature on power series; we have only scratched the surface. Hopefully we have done enough to see that power series can be a very useful way of viewing functions.

A A quick introduction to sets

The real numbers, that will be our main concern this semester, is a *set* of objects. Functions of the real numbers have associated with them two *sets* — their domain (the set of possible inputs) and their range (the set of possible outputs). A function itself will be defined as a *set* of ordered pairs. Sets are everywhere in mathematics, and so it will be important to have a good grasp on the standard notations for sets, and on the standard ways by which sets can be manipulated.

Formally defining what a set is is one of the concerns of logic, and goes well beyond the scope of this course. For us, a *set* will simply be a well-defined collection of objects — a collection for which there is an unambiguous test that determines whether something is in the collection of not. So, for example, the collection of good actors will not be a set (it's open to debate who is and isn't good), but the collection of best actor or actress Oscar winners from the last twenty years is a set.

A.1 Notation

We represent sets by putting the elements between braces; thus

$$A = \{1, 2, 3, 4, 5\}$$

is the set of all integers between 1 and 5 inclusive. We can list the elements like this only when the set has finitely many elements, and indeed practical considerations dictate that the set needs to be quite small to admit an explicit listing. (I would not like to have to list all the integers between 1 and $10^{10^{10}}$, for example.) We thus need compact ways to represent sets, and some of these ways are described in the next section.

Two sets A and B will be said to be equal if they have the same elements, that is, if for every x, if x is in A then it is in B, and if it is in B then it is in A. A consequence of this is that if we re-arrange the order in which the elements of a set are presented, we get the same set. So each of

$$\{1, 2, 3, 4, 5\}, \quad \{5, 4, 3, 2, 1\}, \quad \{2, 5, 4, 3, 1\}, \quad \{4, 3, 1, 2, 5\}$$

are the *same* set.

A set cannot contain a repeated element. We do not write $A = \{1, 1, 2, 3, 4, 5\}$. And so, although Hilary Swank has won two Oscars for best actress in the last twenty years, she would only be listed once in the set described in the last section. (There is such a thing as a *multiset* where repeated elements are allowed, but we won't think about this.)

The standard convention for representing sets

In calculus we work mainly with sets of real numbers. The most common notation/way to describe such a set is as follows:

$$S = \{x : p(x)\} \quad (\text{or } \{x | p(x)\})$$

where $p(x)$ is some predicate; the set S consists of all real numbers x such that $p(x)$ is true. The way to read this is

"S is the set of all x such that $p(x)$" (or, "such that $p(x)$ holds").

For example

$$\{x : x \geq 0\}$$

is the set of all non-negative numbers,

$$\{w : \text{the decimal expansion of } w \text{ contains no 3's}\}$$

describes a somewhat complicated set of real numbers, and

$$\{t : 3t^3 - 2t^2 - t > 1\}$$

describes a less complicated, but still hard to pin down, subset of the real numbers.

Sometimes we cheat a little and put an extra condition on the variable before the ":", to make things easier to write. For example, the domain of the function $g(y) = \sqrt{y}/(y-2)$ is all non-negative reals (negative reals don't have square roots), except 2 (we can't divide by 0), so we should write

$$\text{Domain}(S) = \{y : y \geq 0 \text{ and } y \neq 2\},$$

but it makes sense to write, slightly more compactly,

$$\text{Domain}(S) = \{y \geq 0 : y \neq 2\}.$$

The ellipsis notation

We sometimes describe a set by listing the first few elements, enough so that a pattern emerges, and then putting an ellipsis (a ...) to say "and so on". For example

$$\{2, 3, 5, 7, 11, \ldots\}$$

is *probably* the set of prime numbers. I say "probably", because there are plenty of reasonably natural sequences that begin 2, 3, 5, 7, 11, and are *not* the prime numbers. The amazing Online Encyclopedia of Integer Sequence, oeis.org (which is just what its name says) lists 956 such sequences, including the sequence of *palindromic* primes (prime numbers whose decimal expansion is a palindrome; the next is 101).

Because of these possible ambiguities, ellipsis notation should be used only when the context is absolutely clear. In the above example, something like

$$\{n : n \text{ is a palidromic prime}\}$$

should be preferred.

Ellipsis notation is sometimes used for a finite set. In this case, after the ellipsis there should be one or two terms, used to indicate where one should stop with the pattern. For example,

$$\left\{1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{100}\right\}$$

probably indicates the reciprocals (multiplicative inverses) of all the natural numbers between 1 and 100; to make this absolutely clear one should write

$$\{1/n : 1 \leq n \leq 100,\ n \in \mathbb{N}\}.$$

Special sets of reals

There are some special sets of reals that occur so frequently, we give them special names. These are the *intervals* — sets of all numbers between two specified reals, with slightly different notation depending on whether the specified end points belong or don't belong to the set. Here's a list of all the incarnations that occur.

- $[a, b] = \{x : a \leq x \leq b\}$
- $[a, b) = \{x : a \leq x < b\}$
- $(a, b] = \{x : a < x \leq b\}$
- $(a, b) = \{x : a < x < b\}$
- $[a, \infty) = \{x : a \leq x\}$
- $(a, \infty) = \{x : a < x\}$
- $(-\infty, b] = \{x : x \leq b\}$
- $(-\infty, b) = \{x :< b\}$
- $(-\infty, \infty) = \mathbb{R}$.

Notice that a square bracket ("[" or "]") is used to indicate that the relevant endpoint is in the interval, and a round bracket ("(" or ")") is used to indicate that it is not. Notice also that we never put a "[" before $-\infty$ or a "]" after ∞; Neither $-\infty$ nor ∞ are numbers, merely notational symbols.

A.2 Manipulating sets

- It is possible for a set to contain no elements. We use the symbol $\emptyset$ to denote this *null* or *empty* set.

- We use the symbol "$\in$" to indicate membership of a set, so $x \in S$ indicates that x is an element of S. On the other side, $x \notin S$ indicates that x is *not* an element of S.

- When there is a clear universe U of all objects under discussion, we denote by S^c or S' the *complement* of S — the set of all elements in U that are not in S. So, for example, if it is absolutely clear that the universe of objects under discussion is the reals, then

  $$(0, \infty)^c = (-\infty, 0].$$

  If instead it is absolutely clear that the universe of objects under discussion is the set of non-negative reals, then

  $$(0, \infty)^c = \{0\} \quad \text{(the set containing the single element 0)}.$$

- If all the elements of a set A are also elements of a set B, we say that A is a *subset* of B and write $A \subseteq B$. For example, the set of prime numbers is a subset of the set of natural numbers. The two lines at the bottom of the symbol "$\subseteq$" are intended to convey that it is possible that $A = B$, that is, that A and B have exactly the same elements. In order words, any set is a subset of itself.

  If $A \subseteq B$ and $A \neq B$ (so there are some elements of B that are not elements of A) then A is said to be a *proper* subset of B, and this is sometimes written $\subsetneq$, or $\subsetneqq$, or $\subsetneqq$. It is also sometimes written $\subset$, but be warned that for many writers $A \subset B$ and $A \subseteq B$ are identical.

  To prove that $A \subseteq B$ it is necessary to prove the implication $A(x) \implies B(x)$ where $A(x)$ is the predicate $x \in A$ and $B(x)$ is the predicate $x \in B$. To prove that $A = B$, it is necessary to prove the equivalence $A(x) \iff B(x)$, which we know really requires two steps: showing $A(x) \implies B(x)$ ($A \subseteq B$) and $B(x) \implies A(x)$ ($B \subseteq A$).

  The empty set $\emptyset$ is a subset of every set.

  The collection of all subsets of a set S is called the *power set* of a set, written $\mathcal{P}(S)$. We will rarely use this.

A.3 Combining sets

There are a number of ways of combining old sets to form new ones.

- **Union**: The union of sets A and B, written $A \cup B$, is the set of all elements that are in either A or B, or perhaps both:

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

  For example, $[0, 1] \cup [1, 2) = [0, 2)$.

- **Intersection**: The intersection of sets A and B, written $A \cap B$, is the set of all elements that are in both A and B:

$$A \cap B = \{x : (x \in A) \wedge (x \in B)\}.$$

For example, $[0, 1] \cap [1, 2) = \{1\}$.

- It is possible to take the intersection or union of arbitrarily many sets. The notation $\{A_i : i \in I\}$ is used to indicate that we have a family of sets, *indexed* by the set I: for each element i of I, there is a set A_i in our family. Often I is the set of natural numbers, and then we can write the family as

$$\{A_1, A_2, A_3, \ldots\}.$$

The intersection of the sets in a family $\{A_i : i \in I\}$, written $\cap_{i \in I} A_i$, is the set of elements that are in all the A_i, while the union $\cup_{i \in I} A_i$ is the set of elements that are in at least one of the A_i.

For example, if $I = \mathbb{N}$ and $A_i = (-i, i)$, then

$$\cup_{i \in I} A_i = \mathbb{R} \quad \text{and} \quad \cap_{i \in I} A_i = (-1, 1).$$

- The notation $A \setminus B$, sometimes written $A - B$, denotes the set of all elements that are in A but are not in B (the $-$ sign indicating that we have removed those elements). It is not necessary for B to be a subset of A for this to makes sense. So, for example

$$\{1, 2, 3, 4\} \setminus \{3, 4, 5, 6\} = \{1, 2\}.$$

- Notice that we always have the relations $A \setminus B \subseteq A$, $A \cap B \subseteq A \subseteq A \cup B$ and $A \cap B \subseteq B \subseteq A \cup B$.

- The *Cartesian product* of two sets X and Y, denoted by $X \times Y$, is the set of all ordered pairs (x, y) with $x \in X$ and $y \in Y$. Note that when we describe a list of elements with round brackets "(" and ")" on either side, the order in which we present the list matters: (a, b) is not the same as (b, a) (unless $a = b$), whereas $\{a, b\} = \{b, a\}$ since both, as sets, have the same collection of elements. Also note that an ordered pair allows repetitions: $(3, 3)$ is a perfectly reasonable ordered pair.

A.4 The algebra of sets

The relations satisfied by union, intersection and complementation bear a striking resemblance to the relations between the logical operators of OR, AND and negation. [This is essentially for the following reason: given a universe of discourse U, and any predicate $p(x)$, we can associate a subset A of U via $A = \{x \in U : p(x) \text{ is true}\}$. Most of the logical equivalences

that we have discussed between propositions have direct counterparts as equalities between the corresponding sets.]

We list the relations that hold between sets A, B and C that are all living inside a universe U. For comparison, we also list the corresponding relations that hold between propositions p, q and r. You should notice a direct correspondence between $\vee$ and $\cup$, between $\wedge$ and $\cap$, between negation and complementation, between T and U and between F and $\emptyset$. You should be able to come up with proofs of any/all of these identities.

Name of law	Equality/equalities	Equivalence(s)
Identity	$A \cap U = A$ $A \cup \emptyset = A$	$p \wedge T \iff p$ $p \vee F \iff p$
Domination	$A \cup U = U$ $A \cap \emptyset = \emptyset$	$p \vee T \iff T$ $p \wedge F \iff F$
Idempotent	$A \cup A = A$ $A \cap A = A$	$p \vee p \iff p$ $p \wedge p \iff p$
Double negation	$(A^c)^c = A$	$\neg(\neg p) \iff p$
Commutative	$A \cup B = B \cup A$ $A \cap B = B \cap A$	$p \vee q \iff q \vee p$ $p \wedge q \iff q \wedge p$
Associative	$(A \cup B) \cup C = A \cup (B \cup C)$ $(A \cap B) \cap C = (A \cap B) \cap C$	$(p \vee q) \vee r \iff p \vee (q \vee r)$ $(p \wedge q) \wedge r \iff (p \wedge q) \wedge r$
Distributive	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$	$p \vee (q \wedge r) \iff (p \vee q) \wedge (p \vee r)$ $p \wedge (q \vee r) \iff (p \wedge q) \vee (p \wedge r)$
De Morgan's	$(A \cap B)^c = A^c \cup B^c$ $(A \cup B)^c = A^c \cap B^c$	$\neg(p \wedge q) \iff (\neg p) \vee (\neg q)$ $\neg(p \vee q) \iff (\neg p) \wedge (\neg q)$
Absorption	$A \cap (A \cup B) = A$ $A \cup (A \cap B) = A$	$p \wedge (p \vee q) \iff p$ $p \vee (p \wedge q) \iff p$
Tautology	$A \cup A^c = U$	$p \vee (\neg p) \iff T$
Contradiction	$A \cap A^c = \emptyset$	$p \wedge (\neg p) \iff F$
Equivalence	$A = B \iff A \subseteq B$ and $B \subseteq A$	$p \leftrightarrow q \iff (p \to q) \wedge (q \to p)$

Table A1: Set identities.

More generally, the highly useful De Morgan's laws say that for *any* index set I,

$$(\cup_{i \in I} A_i)^c = \cap_{i \in I} A_i^c \quad \text{and} \quad (\cap_{i \in I} A_i)^c = \cup_{i \in I} A_i^c.$$

References

[1] J. Bryant and P. Kirby, Florida State University Course Notes, MAD 2104 (Discrete Mathematics I), retrieved from https://www.math.fsu.edu/~wooland/mad2104/ August 12 2019.

[2] M. Hutchings, Introduction to mathematical arguments, retrieved from https://math.berkeley.edu/~hutching/ August 12 2019.

[3] M. Spivak, *Calculus* (4th edition), Publish or Perish Press, Houston, 2008.