Course notes for Math 10850, fall 2018

following Spivak's *Calculus*

David Galvin, University of Notre Dame

Last updated December 4, 2018

Abstract

This document is a set of notes to accompany lectures for Math 10850 (Honors Calculus I), University of Notre Dame, fall 2018. Together with Math 10860 in the spring, this 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 incoming math majors and potential math majors. 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 edition), Publish or Perish, 2008.

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 discussion. With the exception of one chapter, 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 its first draft. Comments & corrections are welcome! Email `dgalvin1@nd.edu`.

Contents

1 A crash course in logic

Logical thinking — the process of inferring the truth or otherwise of complex statements from knowledge of the truth of simpler ones, using rules of inference — underpins all of mathematics and all of the sciences, and will play a central role in this course.

This section gives a quick, and somewhat informal, "crash-course" in the basics of propositional logic, the suite of tools we will use to establish complex truths from simpler ones.

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.
- "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)[1], 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.

[1] 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

- "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.

- **Negation**: If p is a statement, then the negation of p, which we call "not p" and sometimes write symbolically as $\neg p$, is a statement that has the opposite truth value to p (so $\neg p$ is false whenever p is true, and is true whenever p is false). Here are two clarifying examples:

  - If p is "There is life on Mars" then $\neg p$ 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

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

    (it doesn't matter what "field" might mean), then $\neg p$ 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	$\neg p$ (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. (I put "**AND**" in bold because this is how people in computer science refer to conjunction.) We will typically write "p and q" for the conjunction (the "and" not in bold/all-caps). The symbolic notation is $p \wedge q$.

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.

Here is the truth table for conjunction:

p	q	$p \wedge q$ (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.

- **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 **OR** q are true, and false otherwise. (Again, "**OR**" is a typical computer science notation.) We write "p or q" for this compound statement, and sometimes denote it symbolically by $p \vee q$. Notice that it is an *inclusive* or: $p \vee q$ is true if p is true, or if q is true, *or if both are true*.

  If p is "There is life on Mars" and q is "There is water on Mars", then the disjunction $p \vee q$ 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.

  Here is the truth table for disjunction:

p	q	$p \vee q$ (p or q)
T	T	T
T	F	T
F	T	T
F	F	F

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

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

or

not (p and q) or not (p and r) or not (q and 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 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 truth value of A is the same as the 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 shortly.

1.2 An note on parentheses

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 "$\neg p \vee q$". This could mean "take the disjunction of the following two statements:

- q
- 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 this ambiguity is to decide, once and for all time, on an order of precedence among logical operations. There *is* a fairly standard such order[2]. Two problems with this approach are

- that the order of precedence it is not so universal that it eliminates all ambiguity and

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

- that it is something that one must remember, and there is no obvious motivation behind it to act as a memory aid.

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

$$\text{“}(\neg p) \vee q\text{”}$$

(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

$$\text{“}\neg(p \vee q)\text{”}$$

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

1.3 Implication

This 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 $p \Rightarrow q$. This can be rendered into ordinary English 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).

We have seen an example: "If I go out in the rain without a coat on, I will get wet." This is the implication $p \Rightarrow q$ where p is "I go out in the rain without a coat" and q is "I get wet". Mirroring the list above, the implication can be expressed in English 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
- q is referred to as the *conclusion*.

To make precise that $p \Rightarrow q$ is a *statement*, we assign a truth table to it. Two lines of the truth table 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. Here's the full truth table:

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

To justify this, think of $p \Rightarrow q$ 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: 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

It's easy to check via a truth table that $p \Rightarrow q$ is equivalent to $(\neg p) \vee q$. By De Morgan's law (and the easy fact that we have not yet explicitly stated, that $\neg(\neg p)$ always has the same truth value as p), the negation $\neg(p \Rightarrow q)$ of an implication is equivalent to $p \wedge (\neg q)$.

The contrapositive of an implication

Here is another statement that is equivalent to $p \Rightarrow q$: the statement $(\neg q) \Rightarrow (\neg p)$. Since this is such an important reformulation of implication (as we will see many times as the year progresses) we will go through the equivalence carefully:

- if p, q are both true, so is $p \Rightarrow q$ (from the first line of the truth table of $p \Rightarrow q$); while also $\neg q$ and $\neg p$ are both false, so $(\neg q) \Rightarrow (\neg p)$ is true (from the last line of the truth table of $p \Rightarrow q$).

- if p, q are both false, then $p \Rightarrow q$ is true (from the last line of the truth table of $p \Rightarrow q$); while also $\neg q$ and $\neg p$ are both true, so $(\neg q) \Rightarrow (\neg p)$ is true, too (from the first line of the truth table of $p \Rightarrow q$).

- if p is true and q is false, then $p \Rightarrow q$ is false (from the second line of the truth table of $p \Rightarrow q$); while also $\neg q$ is true and $\neg p$ is false, so $(\neg q) \Rightarrow (\neg p)$ is false (again from the second line of the truth table of $p \Rightarrow q$).

- if p is false and q is true, then $p \Rightarrow q$ is true (from the third line of the truth table of $p \Rightarrow q$); while also $\neg q$ is false and $\neg p$ is true, so $(\neg q) \Rightarrow (\neg p)$ is true (again from the third line of the truth table of $p \Rightarrow q$).

The statement $(\neg q) \Rightarrow (\neg p)$ is called the *contrapositive* of the implication $p \Rightarrow q$, and we will often find that to prove that some statement $p \Rightarrow q$ is true, it will be easier to prove instead that the contrapositive is true, and then use that the implication and its contrapositive are equivalent, so the implication must be true as well.

As an example of the contrapositive, think of the proposition "If I go out in the rain without a coat on, I will get wet". The contrapositive is "If I do not get wet, then I didn't go out in the rain". A little thought should convince you that this is exactly the same as the original statement.

The converse of an implication

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

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

Note that this is *not* the same as the truth table of $p \Rightarrow q$. One strategy we will **NEVER** employ to show $p \Rightarrow q$ is to show $q \Rightarrow p$ and then say that that's enough to deduce $p \Rightarrow q$ — 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 English, 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 you got no dessert 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 English-language wording of the statement formally parsed as an implication in one direction, it's meaning was really an implication in the converse direction.

These kinds of ambiguities occur a lot in ordinary English, 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 $p \Leftrightarrow q$ is shorthand for $(p \Rightarrow q) \wedge (q \Rightarrow p)$ (p implies q, *and* q implies p), and we read it as "p if and only if q" or "p iff q" (an explanation for this terminology appears in the next section). 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	$p \Leftrightarrow q$ (p if and only if q)
T	T	T
T	F	F
F	T	F
F	F	T

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 notation

The field of propositional logic is filled with precisely defined notations, such as $\neg$, $\wedge$, $\vee$, $\Rightarrow$, $\Leftrightarrow$, $\therefore$, $\because$, $\exists$, $\forall$, 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

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

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

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. Much more on this later.

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

The logical implication "p implies q" is often rendered into English 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 English 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 English 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".

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

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

$$\text{“}\neg(p \wedge q)\text{” and “}(\neg p) \vee (\neg q)\text{”}.$$

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

1.7 Predicates

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 — if (x, y) is on the circle of radius 2 centered at the origin then the assertion is true, otherwise it is false. 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 "$p \Rightarrow q$" where p: "$x-2 = 1$" and q: "$x^2-9 = 0$" are both predicates, not statements.

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.

Order of quantifiers

A predicate needs a quantifier for every variable to turn into a statement, and 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”. Notice that we read the quantifiers, as in ordinary reading, from left to right. 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*, and we will discuss it in more detail later.

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 the next section.

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"[3][4][5]:

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.)

(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:

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

[4] Why not just "Geometry"? For over two millennia after Euclid proposed his five axioms, mathematicians struggled with the fifth. 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".

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

> "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[6][7]:

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*[8]:

> **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.

- **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,

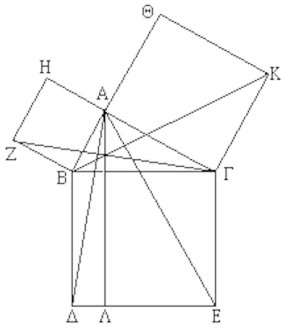
[6]Statements taken from http://www-history.mcs.st-and.ac.uk/HistTopics/Euclid_definitions.html.

[7]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!

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

– 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.

2.2 Basic 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 $p \vee r$	$q \vee s$	$((p \Rightarrow q) \wedge (r \Rightarrow s) \wedge (p \vee r)) \Rightarrow q \vee s$

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.

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

2.4 Various approaches to proving implications

Throughout the semester we will be using the rules of inference to prove more complicated statements from simpler ones. In this section we introduce some of the basic types of proofs. As with the end of the last section, we won't give too many very substantial examples here, as there will be plenty of genuine examples in the near future.[9]

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 $p \Rightarrow q$, and we want to prove that it is valid. Here are the six basic proof techniques we can use:

- **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

[9]For the material in this section and the next I've drawn heavily on Florida State University course notes for course MAD 2104 (Discrete Mathematics I), written by Dr. John Bryant and Dr. Penelope Kirby.

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 $p \Rightarrow q$ is true regardless of the truth value of q.

  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.

- **Direct proof**: Assume p, and then use the rules of inference, axioms, definitions, and logical equivalences to prove q.

  Example: "if m and n are odd then $m+n$ is even".

  Proof: Let us assume that m and n are odd; we will argue directly that $m+n$ is even. Since m and n are odd, there are whole numbers k and ℓ with $m=2k+1$ and $n=2\ell+1$. We have

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

  Since $k+\ell+1$ is a whole number, m is two times a whole number, and so is even.

  A few comments are in order.

  - It's ok to *assume* that m and n satisfy the hypothesis of the theorem, and work throughout only with that assumption; if m and n don't both satisfy the hypothesis, then the premise of the implication is false, so the truth value of the implication is true. In a sense, we are proving the truth of the implication by studying cases (see below), without explicitly mentioning case 0, the simpler of the two cases; case 0 is where the hypothesis is false, and so immediately the

implication is true, and case 1 is where the hypothesis is true, and we need to do some work.

- Notice that the proof is presented, not as a collection of unconnected assertions — along the lines of:

$$\begin{gathered} \text{“}m = 2k + 1 \\ n = 2\ell + 1 \\ m + n = 2k + 1 + 2\ell + 1 \\ m + n = 2(k + \ell + 1) \\ m \text{ is even,”} \end{gathered}$$

  — but as a prose narrative: complete sentences, no symbol introduced without its meaning explained, et cetera. This is how you should be aiming to present proofs.

- We need to prove the theorem for *all* possible odd numbers m and n, so when we write m as $2k + 1$ and n as $2\ell + 1$, we need to introduce two different symbols — k and ℓ. If we used the same symbol, say writing $m = 2k + 1$ and $n = 2k + 1$, then we would have introduced a new truth, one that is not in the hypothesis, namely that $m = n$.

- It's reasonably for a first attempt at most proofs to be an attempt at a direct proof. Write down what it means for the hypothesis to be true, and see where it takes you.

- Are you concerned that although this is supposed to be a course where everything is done precisely, I'm throwing around properties like evenness and oddness of whole numbers, without properly defining them? Don't worry, we'll soon reach a "reset" point, after which *everything* will be done precisely.

Indirect Proof, or Proof by Contradiction: Assume that p is true and q is false (the one situation in which the implication $p \Rightarrow q$ is false) and derive a contradiction, meaning: deduce that some statement r is true, and also that its negation $\neg r$ is true, so that $r \wedge (\neg r)$ is true. This can't be (the truth value of $r \wedge (\neg r)$ is always F), 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 $p \Rightarrow q$ is true.

Example: (Here the universe of discourse for all variables is the set of real numbers). "If $5x + 25y = 2018$, then at least one of x, y is not an integer".

Proof: We argue by contradiction. Let us assume both that $5x + 25y = 2018$ 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 $5x + 25y = 2018$, this says that 2018 is a multiple of 5. But also, by a direct calculation, we see that 2018 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:

- Why didn't we try a direct proof? Because assuming the truth of the hypothesis in this case (that $5x + 25y = 2018$) 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:"$5x+25y = 2018$", q:"at least one of x, y is not an integer", and r:"2018 is a multiple of 5".

- **Proof by Contrapositive**: Give a direct proof of $(\neg q) \Rightarrow (\neg p)$ (the contrapositive of, and equivalent to, $p \Rightarrow q$). That is, assume $\neg q$ and then use the rules of inference, axioms, definitions, and logical equivalences to deduce $\neg p$.

  Example: "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:

  - This could be though of as a special case of proof by contradiction: assume p and $\neg q$ are true, and reach the contradiction $p \wedge (\neg p)$.
  - 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.

- **Proof by Cases**: If the hypothesis p can be written as $p_1 \vee p_2 \vee \cdots \vee p_k$ (think of the p_i's as the "cases" of the hypothesis; they don't have to be mutually exclusive), then prove each of the statements $p_1 \Rightarrow q, p_2 \Rightarrow q, \ldots, p_k \Rightarrow q$ separately (possibly with a different method of proof for each case).

  Example 1: "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:

  - 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 2 ($n = 3$): Here $n^2 - n + 41 = 47$, again a prime number.
– Case 40 ($n = 40$): Here $n^2 - n + 41 = 1601$, which is a prime number[10].

A few comments:

– Proofs by cases are often quite tedious! I've left cases 4 through 39 to the reader.
– There is an equivalence underlying proof by cases. This equivalence has

$$(p_1 \vee p_2 \vee \cdots p_k) \Rightarrow q$$

on one side, and something involving $p_1 \Rightarrow q, p_2 \Rightarrow q, \ldots, p_1 \Rightarrow q$ on the other. It will appear in homework.
– You might ask, "Is it just by dumb luck that n^2-n+41 is prime for $n = 1, 2, \ldots, 40$, or is there some underlying *reason*?" There *is* a reason, and it involves some very substantial number theory. Ask Prof. Jorza if you are intrigued by this.

Example 2: "if $n^2 - n - 2 > 0$ then either $n > 2$ or $n < -1$". (A much more subtle example than the last.)

Proof: We have $n^2 - n - 2 = (n + 1)(n - 2)$. The sign of $(n + 1)(n - 2)$ changes at $n = -1$ and at $n = 2$ (and the product is equal to 0 at these two points), so initially we consider five cases: $n < -1$, $n = -1$, $-1 < n < 2$, $n = 2$ and $n > 2$.

– Case 1 ($n < -1$): In this case $n + 1 < 0$ and $n - 2 < -3 < 0$, so $n^2 - n - 2 = (n+1)(n-2) > 0$. Thus the premise "$n^2 - n - 2 > 0$" is true in this case, and so is the conclusion: in this case $n < -1$, so (by disjunction introduction) it is true that either $n < -1$ or $n > 2$.
– Case 2 ($n = -1$): In this case $n+1 = 0$ and $n-2 = -3$, so $n^2 - n - 2 = 0$. Thus the premise is *false* in this case, and we need not go any further with this case.
– Case 3 ($-1 < n < 2$): Here $n + 1 > 0$ and $n - 2 < 0$ so $n^2 - n - 2 < 0$; again the premise is false in this case.
– Case 4 ($n = 2$): Here $n^2 - n - 2 = 0$, and once again the premise is false.
– Case 5 ($n > 2$): Here $n + 1 > 0$ and $n - 2 > 0$ so $n^2 - n - 2 = 0$; the premise is true in this case. The conclusion is also true ($n > 2$, so at least one of $n > 2$, $n < -1$ holds), so in this case the implication is true.

Since in all cases in which the hypothesis is true, the conclusion is also true, we conclude that the implication is true.

A few comments:

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

- As in a previous example, we have just argued that "when the hypothesis is true, the conclusion is true", and from there jumped to "the implication is true", leaving out the short argument that "when the hypothesis is false, the implication is automatically true". It is a universal convention to leave this out.
- This is another example of proofs by cases being quite tedious: a very simple statement required five cases. We could have been more clever, and reduced to three cases: $n < -1$, $-1 \leq n \leq 2$ and $n > 2$; but this only became clear after the proof was complete.
- In this example, we have p:"$n^2 - n - 2 > 0$", and we discovered that p is equivalent to $p_1 \vee p_2$, where p_1:"$n < -1$" and p_2:"$n > 2$". This wasn't an *obvious* collection of cases, a priori. Once we had discovered the decomposition, the implication was obvious (it became "$(p_1 \vee p_2) \Rightarrow (p_1 \vee p_2)$, which is easily seen to be always true.)
- What's wrong with the following "proof" that $n^2 - n - 2 > 0$ implies either $n > 2$ or $n < -1$?

  > "We have $n^2 - n - 2 = (n-2)(n+1)$. If $n > 2$ then $(n-2)(n+1)$ is positive times positive, so > 0. If $n < -1$ then $(n-2)(n+1)$ is negative times negative, so > 0. So in either case, $n^2 - n - 2 > 0$."

  The fallacy that has been committed here is one that you would never have been tempted to commit in Example 1, but it's a little more tempting (and equally fallacious) here.

- **"If any only if" proofs**: To prove "p if any only if q" we often have to break the bidirectional implication $p \Leftrightarrow q$ into the conjunction of two uni-directional implications: $(p \Rightarrow q) \wedge (q \Rightarrow p)$, and prove these two implications separately.

  Example: "m^2 is even if and only if m is even".

  Proof: First we prove that if m is even then m^2 is even. If m is even then $m = 2k$ for some integer k, so $m^2 = 4k^2 = 2(2k^2)$, so (since $2k^2$ is a whole number), m^2 is even.

  Next we show that if m^2 is even then m is even. It's tempting to do a contrapositive proof — to deduce from "m is odd" that "m^2 is odd". But we don't need to. We have already proven that if mn is even then either m or n is even; the special instance of this when $m = n$ is exactly "if m^2 is even then m is even".

  A comment:

  - As we prove more and more complicated statements, it's important to remember that these statements can then be used as truths in later proofs (just as we used "if mn is even then either m or n is even" as a truth in the above proof). If we don't remember this, then we will find ourselves continually re-inventing the wheel, and we will make very little substantial progress!

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[11].

2.6 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

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

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

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[12], 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. Passing 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 the
   - real numbers, which may be thought of as the *entire* number line:

     ———————————————————————————,

     makes mathematical life much easier.

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 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;

[12]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. We should properly 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.

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[13] ways of parenthesizing $a + b + c + d + e + f$ are the 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.*

[13] 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.

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[14], $a+(b+(c+d)) = a+((b+c)+d)$.

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.[15] □[16]

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 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.

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 $a + x = a$:

[14] 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!

[15] 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.

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

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 in general 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[17], namely

[17] 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.

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

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 commutativity, 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.

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 "$\cdot\cdots$" 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,

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

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 $(x-y)\cdot(x+y) = (x-y)\cdot x + (x-y)\cdot y$. Now applying P9[18] to both terms on the right-hand side of this last expression, we get that it equals $(x \cdot x - y \cdot x) + (x \cdot y + (-y) \cdot y)$. A little rearranging, using P2, P3, P4 and P8, leads to $x \cdot x - y \cdot y$ or $x^2 - y^2$.

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 $b = 0$.*

[18] Actually we are using a version of P9 that follows immediately from P9 using P8: for all a, b, c, $(b+c)\cdot a = (b \cdot a) + (c \cdot a)$.

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

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

□

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.

As a last example of the power of P9, we present a proof that suggests the rule "negative times negative equals positive".

Claim 3.6. *For all numbers a, b, $(-a) \cdot (-b) = ab$*[19].

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[20],

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

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

But now we have

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

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 $(-a) \cdot (-b) = a \cdot b$. □

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

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

[20] 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 only prove the first point, leaving the (very similar) second as an exercise.

Suppose a is positive. Then by trichotomy, a is *not* 0 and $-a$ is *not* positive. Since $-a$ is *not* positive, it must be either 0 or negative. If $-a = 0$ then, since $-(-a) = a$ (a previous exercise), we get $a = -0 = 0$ (the first equality using uniqueness of additive inverses, the second using $x + 0 = x$ applied to $x = 0$ to conclude $-0 = 0$). This is a contradiction (we've already concluded $a \neq 0$), so we conclude that $-a \neq 0$ and so $-a$ must be 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 $(b - a) + (c - b) = c - a$[21] $\in \mathbb{P}$, 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 $c > c$ 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 $a^2 > 0$.*

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.[22] □

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

- $1 > 0$.

[21] 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.

[22] 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.

Indeed, we have $1^2 = 1 \cdot 1 = 1$ by P6, so since $1 \neq 0$[23] 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 any complete our earlier "proof" that $a - b = b - a$ only if[24] $a = b$.

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 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![25]

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.*

[23] 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]em 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.

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

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

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

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

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[26], and by trichotomy they cover all reals, so there are no gaps in the definition.

[26]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:

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

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 $a \geq$ [27]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.

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

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

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

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: $x < 1$, $x = 1$, $1 < x < 2$[28], $x = 2$ 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.[29]

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

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

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[30] 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

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

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[31] — 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

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

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

[31] We will formalize this later.

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

- the set B of

  $$\text{non-positive numbers (number that are 0 or less than 0)}$$

  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

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

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

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

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 and 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

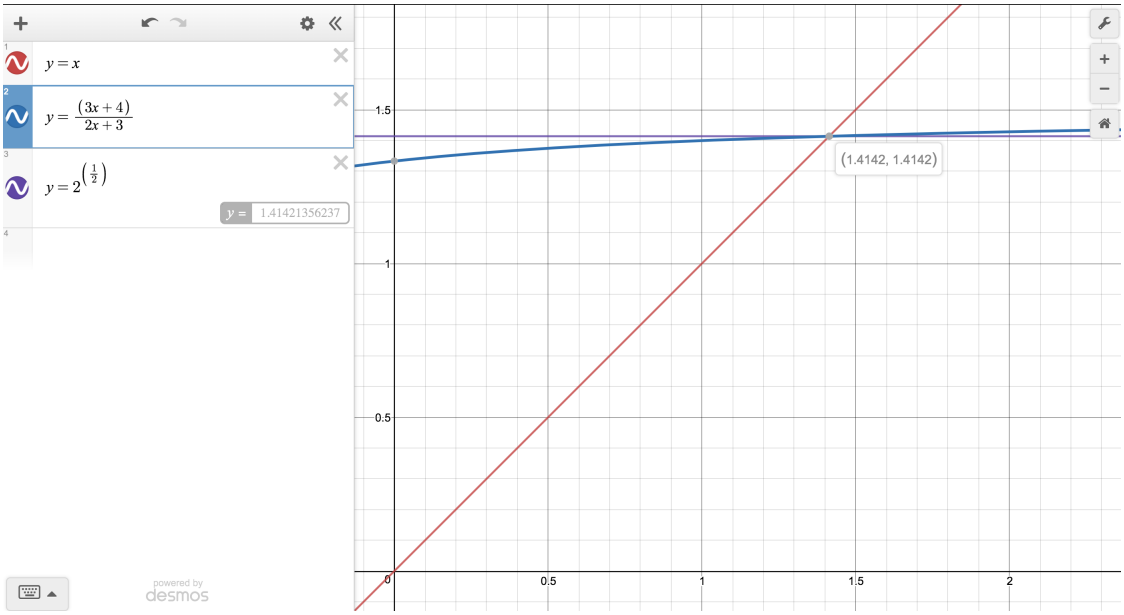
$$\left(\frac{3x+4}{2x+3}\right)^2 < 2,$$

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 $n\varepsilon > r$.[32]

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.

[32] 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).[33] 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).

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 $m \geq y$. Let m_1 be the least such (this is an application of the well-ordering principle). Consider $(m_1 - 1)/n$. We have $(m_1 - 1)/n < y$. 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! □

[33] S stands for *Starbucks* — between any two points in New York City, there is a Starbucks.

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,[34]
- 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 $ab \in \mathbb{P}$.

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.

[34] See the footnote just after the proof of Corollary 3.11.

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 use 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, and 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 out friends 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.[35]

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 manipulating 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

$$\sum_{k=1}^{7} k^2 = 1^2+2^2+3^2+4^2+5^2+6^2+7^2,$$

$$\sum_{k=1}^{2} f(k) = f(1)+f(2)$$

[35] 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!

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[36]

- $f_0 = 0$, $f_1 = 1$ 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[37];
- $f_(n+1)$ 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.[38]

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

$$\begin{aligned} s_{n+1} &= s_n + f_{n+1} \\ &= (f_{n+2} - 1) + f_{n+1} \quad \text{(inductive hypothesis)} \\ &= (f_{n+2} + f_{n+1}) - 1 \\ &= f_{n+3} - 1 \quad \text{(recursive definition of Fibonacci numbers)} \\ &= f_{(n+1)+2} - 1, \end{aligned}$$

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:

[36] Notice that here I'm starting indexing at 0, rather than 1.

[37] See https://en.wikipedia.org/wiki/Hopscotch.

[38] 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.

- (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[39].

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.

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

$$\begin{aligned} (1+x)^{n+1} &= (1+x)(1+x)^n \quad \text{by definition of powers} \\ &\geq (1+x)(1+nx) \quad \text{(by induction hypothesis)} \\ &= 1 + (n+1)x + nx^2 \quad \text{(by some algebra)} \\ &\geq 1 + (n+1)x \quad \text{(since } nx^2 \geq 0\text{)}. \end{aligned}$$

[39] 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.

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:

$$\begin{array}{ccccccccccccccc}
 & & & & & & 1 & & & & & & & & \\
 & & & & & 1 & & 1 & & & & & & & \\
 & & & & 1 & & 2 & & 1 & & & & & & \\
 & & & 1 & & 3 & & 3 & & 1 & & & & & \\
 & & 1 & & 4 & & 6 & & 4 & & 1 & & & & \\
 & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & & & \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1. & &
\end{array}$$

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 $\binom{n}{k} = 1$.[40]

[40] This is a strong justification for declaring $0! = 1$.

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 $n \cdot (n-1) \cdot \cdots \cdot (n-k+1) = n!/(n-k)!$; 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:

- 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*[41]

[41] 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.

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

$$(x+y)^{n+1} = (x+y)\left(x^n + \binom{n}{1}x^{n-1}y + \binom{n}{2}x^{n-2}y^2 + \cdots + \binom{n}{n-2}x^2y^{n-2} + \binom{n}{n-1}xy^{n-1} + y^n\right).$$

Now the right-hand side above is

$$\begin{array}{c} x^{n+1}+ \\ \binom{n}{1}x^ny + \binom{n}{2}x^{n-1}y^2 + \cdots + \binom{n}{n-2}x^3y^{n-2} + \binom{n}{n-1}x^2y^{n-1} + xy^n + \\ x^ny + \binom{n}{1}x^{n-1}y^2 + \cdots + \binom{n}{n-3}x^3y^{n-2} + \binom{n}{n-2}x^2y^{n-1} + \binom{n}{n-1}xy^n + \\ y^{n+1}. \end{array}$$

or

$$\begin{array}{c} x^{n+1}+ \\ \binom{n}{1}x^ny + \binom{n}{2}x^{n-1}y^2 + \cdots + \binom{n}{n-2}x^3y^{n-2} + \binom{n}{n-1}x^2y^{n-1} + \binom{n}{n}xy^n + \\ \binom{n}{0}x^ny + \binom{n}{1}x^{n-1}y^2 + \cdots + \binom{n}{n-3}x^3y^{n-2} + \binom{n}{n-2}x^2y^{n-1} + \binom{n}{n-1}xy^n + \\ y^{n+1}. \end{array}$$

Applying Claim 4.4 to each pair of terms in matching columns in the second and third rows, this becomes

$$\begin{matrix} & & & & & & & & x^{n+1}+ \\ \binom{n+1}{1}x^n y + \binom{n+1}{2}x^{n-1}y^2 + \cdots + & \binom{n}{n-2}x^3 y^{n-2} & + & \binom{n}{n-1}x^2 y^{n-1} & + & \binom{n+1}{n}xy^n & + & \\ & & & & & & & & y^{n+1} \end{matrix}$$

(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.[42]

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 $2 \leq m \leq n-1$[43]. 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

  $$b = q_1 q_2 \cdots q_\ell$$

[42] 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.

[43] 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

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 $n \geq 3$.

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*:

$$R(a_1, \ldots, a_n) = (\cdots(((a_1 a_2)a_3)a_4)\cdots)a_n).$$

We will prove, by strong induction on n, that for all $n \geq 3$, 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 $n = 3$ is axiom P5.

For the inductive step, let $n \geq 4$ be arbitrary, and suppose that the predicate we are trying to prove is true for all values of the variable between 3 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 $1 \geq k \leq n-2$. 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 $P = R(a_1, \ldots, a_n)$.

In either case, $P = R(a_1, \ldots, a_n)$, and so the claim is proven by (strong) induction. □

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.[44]

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.

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 $3/2$.

[44] This proof was discovered by Stanley Tennenbaum; see https://divisbyzero.com/2009/10/06/tennenbaums-proof-of-the-irrationality-of-the-square-root-of-2/ for a lovely visual illustration of it.

- 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 noe 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 tue 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 $n-1 \in \mathbb{N}$. □

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. □

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. □

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 $x_0 \leq S$. 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 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.

5.1 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.

5.2 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.

5.3 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.

5.4 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	$(A \cup B) \cup C = A \cup (B \cup C)$ $(A \cap B) \cap C = (A \cap B) \cap C$	$(p \cup q) \vee r \iff p \vee (q \vee r)$ $(p \wedge q) \wedge r \iff (p \wedge q) \wedge r$
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 $B \subseteq A$	$p \leftrightarrow q \iff (p \to q) \wedge (q \to p)$

More generally, the highly useful De Morgan's laws say that for *any* index set I,

$$(\cup_{i \in I} A_i)^c = \cap_{i \in I} A_i^c \quad \text{and} \quad (\cap_{i \in I} A_i)^c = \cup_{i \in I} A_i^c.$$

6 Functions

6.1 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[45].

As an[46] 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.

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

[45] Or is it?

[46] Possible "another"; see footnote above!

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.

6.2 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

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

6.3 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 6.1. *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[47] 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.

6.4 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: 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

$$a(b(x)) = (x + 1)^2 = x^2 + 2x + 1 \neq x^2 + 1 = b(a(x)),$$

[47] A quite difficult one, using something called Galois theory.

then inequality in the middle being witnessed by any x other than $x = 0$.

Because composition is no 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 g,

and these are also exactly the conditions under which x is in the domain of $(f \circ g) \circ h$.

6.5 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

  $$\sqrt{(x - s)^2 + (y - t)^2}.$$

  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 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[48] 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 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*:

$$f(x) = \begin{cases} 1 & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational.} \end{cases}$$

[48] 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.

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?

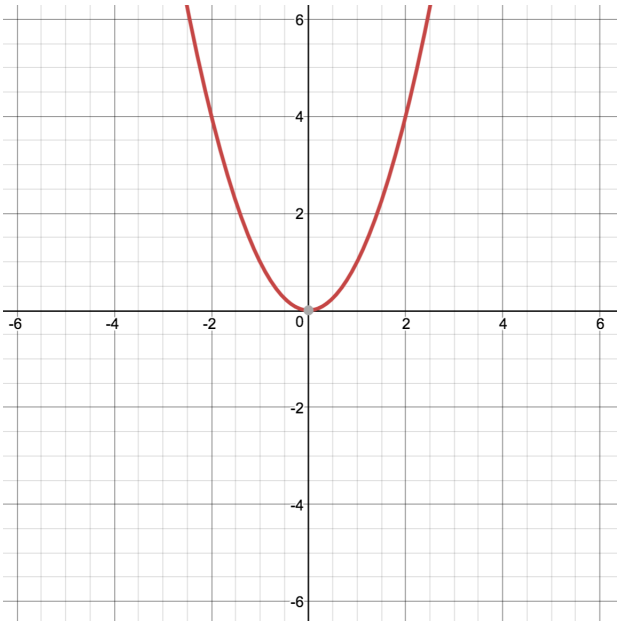
- Is the x part going to infinity faster than the $\sin(1/x)$ part is going to 0, leading to the product g going to infinity?
- Or is the x part going to infinity slower than the $\sin(1/x)$ part is going to 0, leading to the product g going to zero?
- Or are they both going to their respective limits 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.

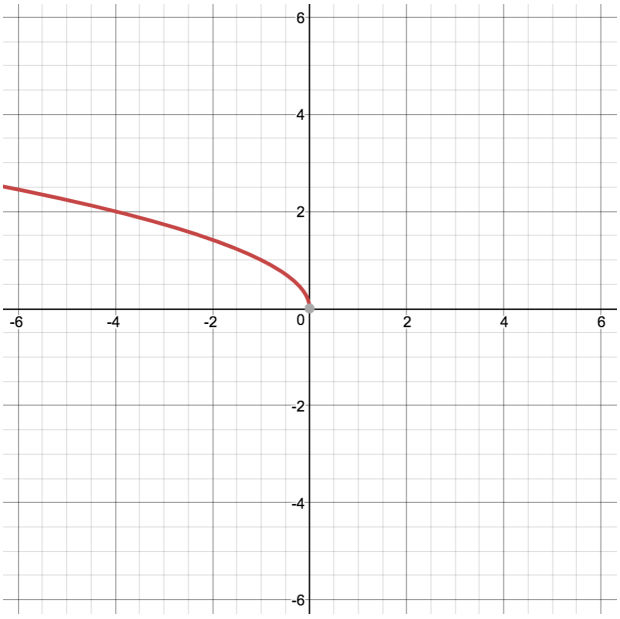
7 Limits

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.

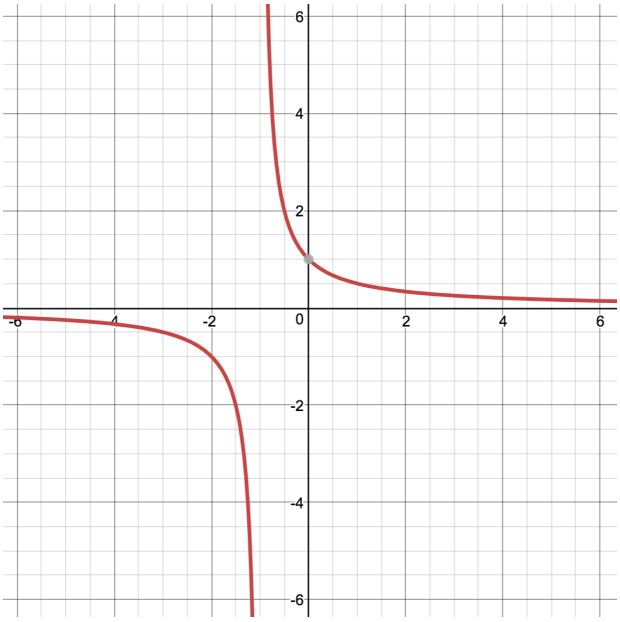
1. Function exhibits no problems at a $f_1(x) = x^2$ at $a = 1$.



2. Function defined nowhere near a $f_2(x) = \sqrt{-x}$ at $a = 1$.

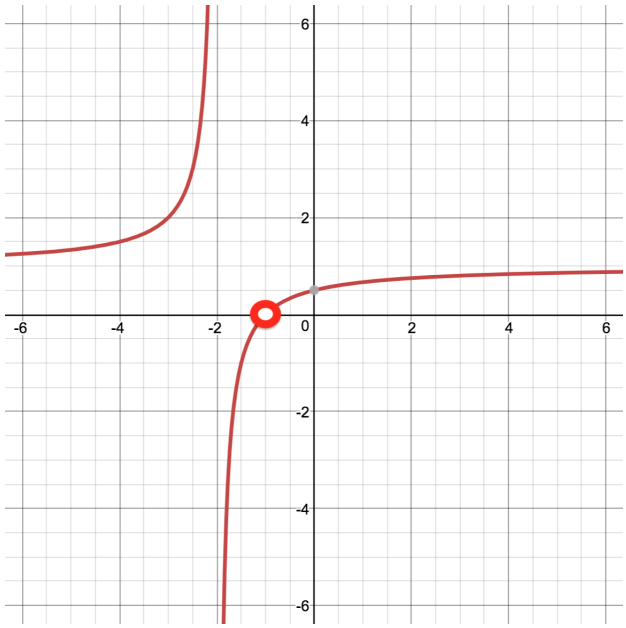


3. Function blows up to infinity approaching a $f_3(x) = 1/(1+x)$ at $a = -1$.

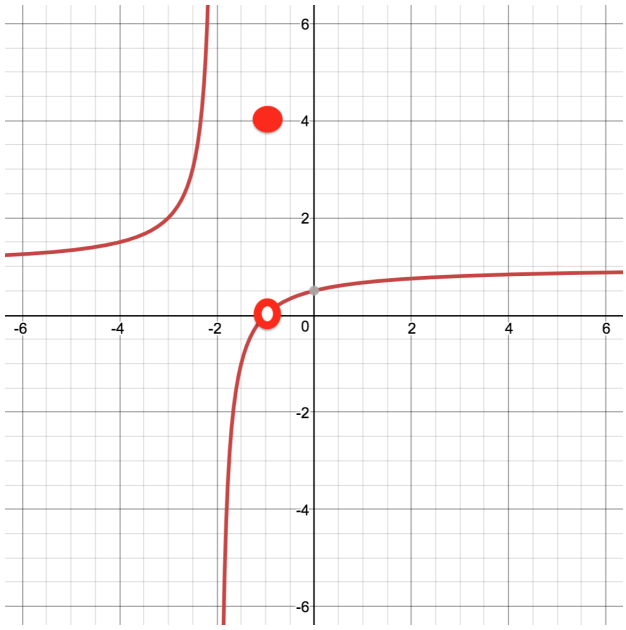


4. Function not defined at a, but otherwise unremarkable $f_4(x) = 1/(1 + (1/(1 + x)))$, $a = -1$. This situation, a function with a "hole", might seem odd, but it can arise naturally. Notice here that $f_4 = f_3 \circ f_3$, and that the expression on the right-hand side of the definition of f_4 can 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.

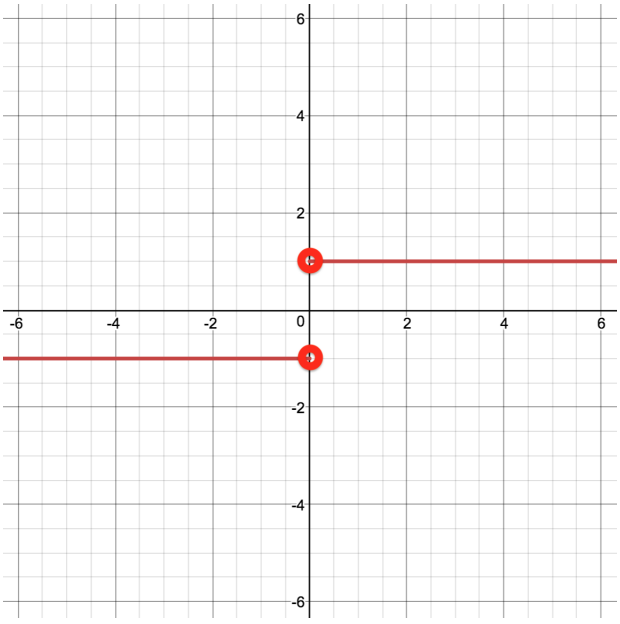
Notice also the graphical notation that we use to indicate the "hole" at -1: literally, a hole.



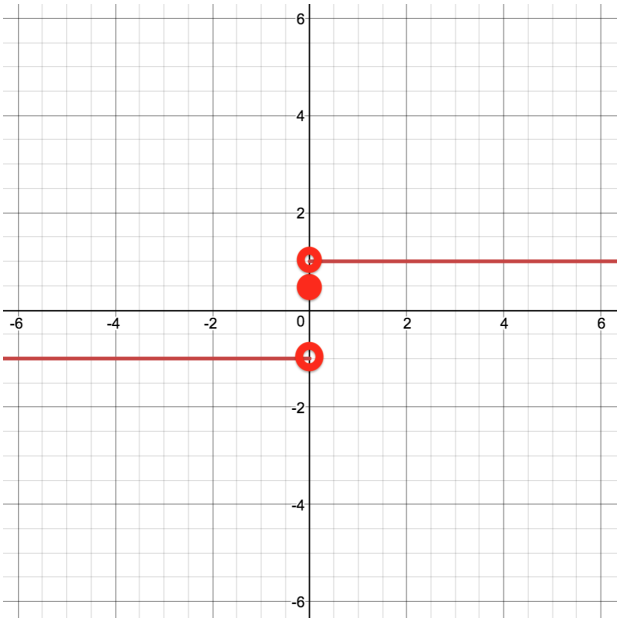
5. Function with "wrong value" at a $f_5(x) = \begin{cases} f_4(x) & \text{if } x \neq -1 \\ 4 & \text{if } x = -1 \end{cases}$ at $a = -1$.



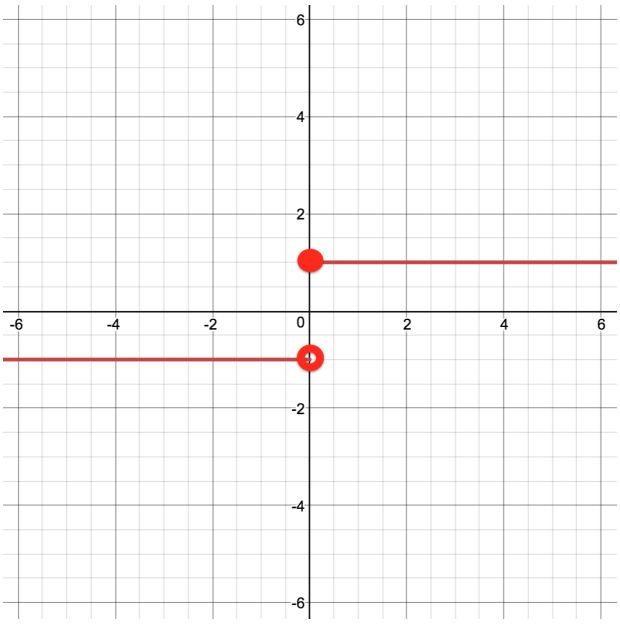
6. Function with a "jump" at a (1) $f_6(x) = \frac{x}{|x|}$ at $a = 0$. The natural domain here is $\mathbb{R} - \{0\}$, and for positive x, $x/|x| = 1$ while for negative x, $x/|x| = -1$. Notice that we graphically indicate the failure of the function to be defined at 0 by *two* holes, one at the end of each of the intervals of the graph that end at 0.



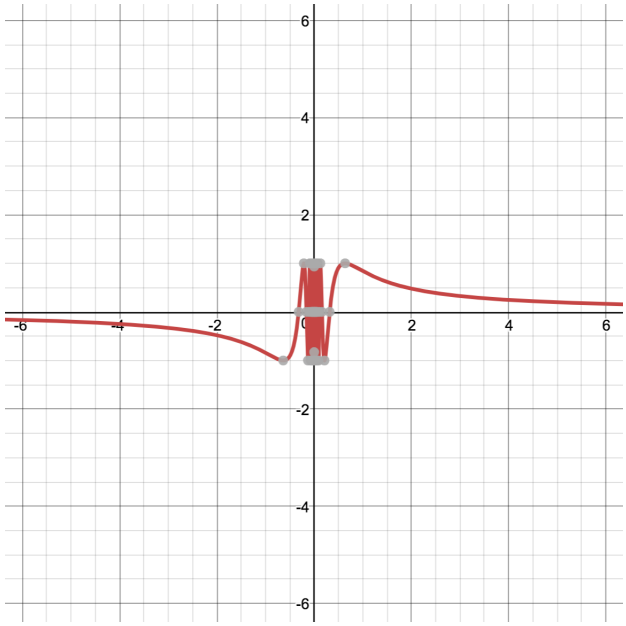
7. Function with a "jump" at a (2) $f_7(x) = \begin{cases} f_6(x) & \text{if } x \neq 0 \\ 1/2 & \text{if } x = 0 \end{cases}$ at $a = 0$. Notice that we graphically indicate the value of the function at 0 with a *solid* holes at the appropriate height.



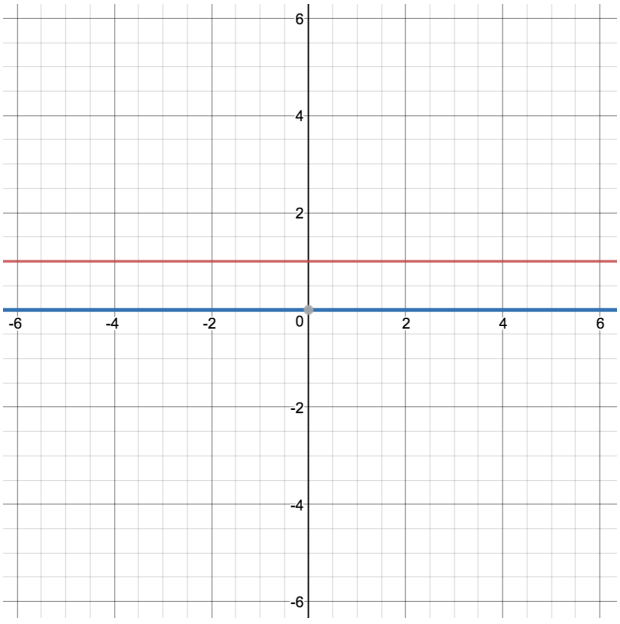
8. Function with a "jump" at a (3) $f_8(x) = \begin{cases} f_6(x) & \text{if } x \neq 0 \\ 1 & \text{if } x = 0 \end{cases}$ at $a = 0$. Notice that here we graphically indicate the behavior of the function around its jump with an appropriate combination of holes and solid holes.



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[49]. Because the rationals are dense in the reals, and so are the irrationals (given any real, there are rationals arbitrarily close to it, and 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.

[49] After the German mathematician Peter Dirichlet, https://en.wikipedia.org/wiki/Peter_Gustav_Lejeune_Dirichlet.

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

7.2 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 $\delta < 1$, 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 smaller 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 $< \varepsilon/(2|a|+1)$" (note that $\varepsilon/(2|a|+1) > 0$, so there *is* such a positive δ). Our choice here was predicated on our promise that $\delta < 1$. So what we really want to do, is choose δ to be any positive number smaller than *both* $\varepsilon/(2|a|+1)$ *and* 1. We can do this, for example, by taking δ to be half the minimum of $\varepsilon/(2|a|+1)$ and 1, or, symbolically,

$$\delta = \frac{1}{2}\min\left\{\frac{\varepsilon}{2|a|+1}, 1\right\}.$$

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 $\delta < 1$ then $0 < |x-2| < \delta$ implies $x \in (1,3)$, so $|x| > 1$ and $3/(2|x|) < 3/2$. So if both $\delta < 1$ *and* $\delta < 2\varepsilon/3$, 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 smaller than $\min\{1, 2\varepsilon/3\}$ verifies $\lim_{x\to 2} 3/x = 3/2$.

Notice that we initially choose $\delta < 1$ 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 did, we would have a lower bound of 0 on x, so *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 comtrol 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.

7.3 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 7.1. *(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[50]

$$\begin{aligned} |(fg)(x) - LM| &= |f(x)g(x) - LM)| \\ &= |f(x)g(x) - Lg(x) + Lg(x) - LM)| \\ &= |g(x)(f(x) - L) + L(g(x) - M)| \\ &\leq |g(x)||f(x) - L| + |L||g(x) - M| \quad \text{by triangle inequality.} \end{aligned}$$

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 $|L||g(x) - M| < |L|(\varepsilon/(2(|L| + 1))) < \varepsilon/2$.[51]

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 $\lim_{x\to a}(fg)(x) = LM$.

[50] 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.

[51] 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.

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.

Claim 7.2. *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 7.3. *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.[52] □

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 out general theorems, and avoiding a nasty derivation from first principles.

What about $\lim_{x\to 1}(2x^2 - 4x)/(x^3 - 8)$? 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

$$\frac{2x^2 - 4x}{x^3 - 8} = \frac{2x(x-2)}{(x-2)(x^2 + 2x + 4)} = \frac{2x}{x^2 + 2x + 4},$$

[52] 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.

valid on the entire domain of $(2x^2 - 4x)/(x^3 - 8)$. So

$$\lim_{x\to 2} \frac{2x^2 - 4x}{x^3 - 8} = \lim_{x\to 2} \frac{2x}{x^2 + 2x + 4} = \frac{4}{14} = \frac{2}{7}.$$

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 7.4. *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[53], 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 $L = M$. □

7.4 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

$$(\forall\varepsilon)(\exists\delta)(\forall x)((0 < |x - a| < \delta) \Rightarrow (|f(x) - L| < \varepsilon))$$[54]

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)

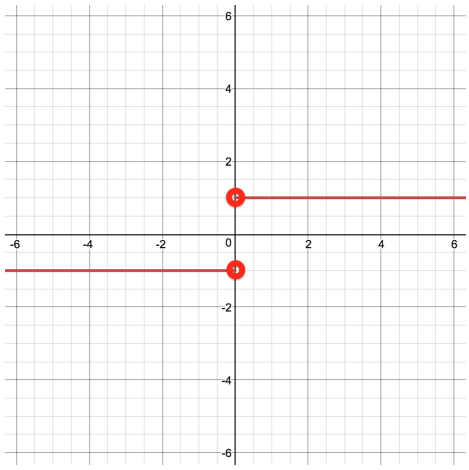
[53] A handy phrase, but one to be used only when you are really saying that no generality is lost.

[54] 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.

or

- 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, $|f_6(x) - 3/4| = |(\delta/2)/(|\delta/2|) - 3/4| = |1 - 3/4| = |1/4| = 1/4 \geq 1/10$[55]

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.[56]

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.

[55] We could have equally well picked $x = -\delta/2$; then $|f_6(x) - 3/4| = 7/4 \geq 1/10$.

[56] 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 $(-\delta, \delta)$[57] (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 $|x|\sin(1/x)|| \leq |x| < \delta = \varepsilon$, 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.

[57] Is there???

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 $g_e ll$ 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 7.5. *(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 $(\Delta - a, \Delta + a)$ 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.

7.5 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)[58] 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.

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:

[58] Note well: as you'll see from the definition, it is a that is being approached from above, not L

- 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 7.6. *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. □

7.6 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

$$\lim_{x\to a} f(x) = \infty^{59}$$

(or sometimes $\lim_{x\to a} f(x) = +\infty$) if f is defined near a, and if

[59] 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!

- for all real numbers M
- there is $\delta > 0$
- such that for all real x,

$$0 < |x - a| < \delta \quad \text{implies} \quad f(x) > M.$$[60]

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).

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

[60] 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.

- for all real numbers $M > M_0$[61]
- 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 $M < 0$[62]), then $x \in (1 - \delta, 1)$ indeed implies $1/(x - 1) < M$.

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,

$$x > M \text{ implies } |f(x) - L| < \varepsilon.$$

[61] As observed after the definition of an infinite limit, this M_0 can be completely arbitrary.

[62] Without this (valid) assumption, the limit calculation would be rather more awkward.

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 $\lim_{x\to\infty} \frac{x}{x+1} = 1$. This entails showing that for all $\varepsilon > 0$ there is M such that $x > M$ implies $x/(x+1) \in (1-\varepsilon, 1+\varepsilon)$. We might as well promise that $M \geq -1$, so that for $x > M$ we have $x > 0$ and $x + 1 > 0$ so $x/(x+1) < 1$; so now we need just ensure $x/(x+1) > 1-\varepsilon$. But now (again remembering $x > 0$, $1 + x > 0$, 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.

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).

Theorem 7.7.

- *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* $q(x) = x^m + b_{m-1}x^{m-1} + \ldots + b_1x + b_0$[63] *(*$n, m \geq 0$*). Then*

$$\lim_{x\to\infty} \frac{p(x)}{q(x)} = \begin{cases} 1 & \text{if } n = m \\ \infty & \text{if } n > m \\ 0 & \text{if } n < m. \end{cases}$$

[63]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.

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[64] 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 $|g(x) - M| < \varepsilon/(|L|+1)$[65]. It follows that if $x > \max\{X_1, X_2, X_3\}$ then

$$\begin{aligned} |(fg)(x) - LM| &\leq |f(x) - L||g(x)| + |L||g(x) - M| \\ &\leq |f(x) - L|(|M|+1) + (|L|+1)|g(x) - M| \\ &< \varepsilon/2 + \varepsilon/2 \\ &= \varepsilon, \end{aligned}$$

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.[66]

[64] We have to change notation slightly from the definition, since M is now being used for something else.

[65] We bound by $\varepsilon/(|L|+1)$ here, rather than $\varepsilon/|L|$, to avoid the possibility of dividing by 0

[66] For plenty of exercises on the kinds of limits introduced in this section, see Spivak, Chapter 5, questions 32-41.

8 Continuity

Looking back at the ten functions that we used at the beginning of Section 7 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.

8.1 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 7.2, 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.[67]

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 8.1. *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.

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 $g(a)$).

[67] 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.

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 $|f(g(x)) - f(g(a))| < \varepsilon$.[68] □

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);
- $f(x) = [x]$[69] 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.

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:

$$f_{10}(x) = \begin{cases} 1 & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational.} \end{cases}$$

[68] 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.

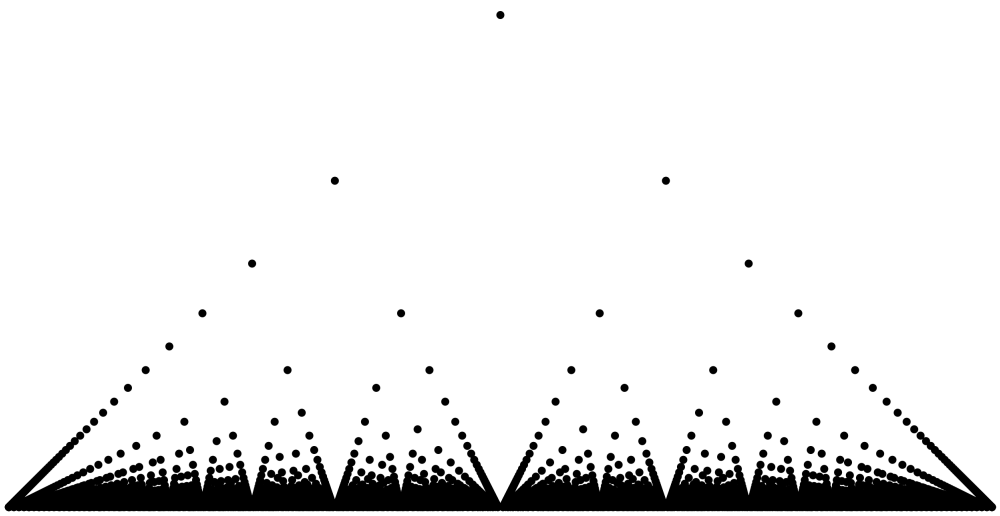
[69] "$[x]$" is the *floor*, or *integer part*, of x — the largest integer that is less than or equal to x. So for example $[2.1] = [2.9] = [2] = 2$ and $[-0.5] = [-.001] = [-1] = -1$.

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[70] 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[71] 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.[72] 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 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

$$1/2, 1/3, 2/3, 1/4, 3/4, \ldots, 1/n, \ldots, (n-1)/n$$

[70] Musn't there be?

[71] Again, musn't there be?

[72] So named by John Conway, for it's unusual graph; it is also called *Thomae's* function, or the *popcorn* function.

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.

8.2 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 8.2. *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 7.2, 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 7.6:

Claim 8.3. *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[73]

$$h(x) = \begin{cases} f(x) & \text{if } x \le c \\ g(x) & \text{if } x \ge c \end{cases}$$

Corollary 8.4. *(of Claim 8.3) 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 $\mathbb{R}$.

[73] Notice that there's no problem with the overlap of clauses here, since $f(c) = g(c)$.

8.3 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.

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 8.5. *(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.[74]

- *Is IVT still true if f is not continuous on all of $[a, b]$?* No. Consider

$$f(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0. \end{cases}$$

[74] 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.

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 $f(x) = 0$.

A second remark is that the IVT quickly gives us the existence of a unique square root of any positive number:

Claim 8.6. *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 8.7. *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 n*th 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 $(a')^n \neq 0$.

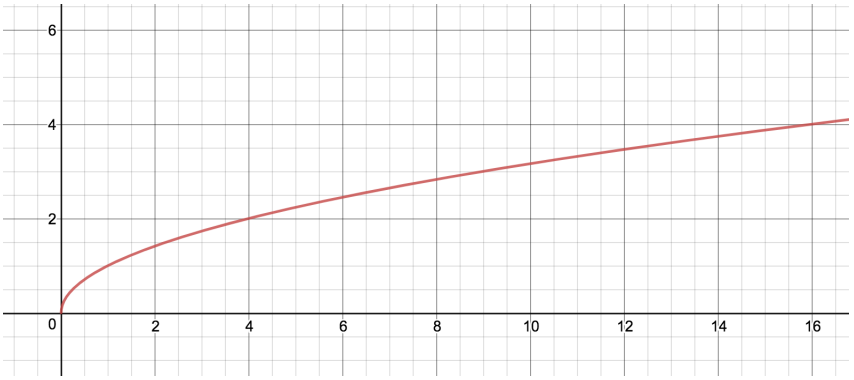
Suppose $a > 0$. Consider the function $f_a : [0, a+1] \to \mathbb{R}$ given by $f_a(x) = x^n - a$. This is a continuous function. Also $f_a(0) = -a < 0$ and (using the binomial theorem)

$$f_a(a+1) = (a+1)^n - a = a^n + \binom{n}{n-1} a^{n-1} + \cdots + \binom{n}{n-k} a^{n-k} + \cdots + \left(\binom{n}{1} - 1 \right) a + 1 > 0.$$

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 8.8. *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 8.9. *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 8.10. *Let $P(x) = x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n$ 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 $a_1x^{n-1} + \cdots + a_{n-1}x + a_n$ is added to it; while if x is sufficiently positive, then x^n is very positive, so much so that it remains positive after $a_1x^{n-1} + \cdots + a_{n-1}x + a_n$ (which may itself be negative) is added to it.

To formalize this, we use the triangle inequality to bound $|a_1x^{n-1} + \cdots + a_{n-1}x + a_n|$. Setting $M = |a_1| + |a_2| + \cdots + |a_n| + 1$ (the $+1$ at the end to make sure that $M > 0$), and

considering only those x for which $|x| > 1$ (so that $1 < |x| < |x|^2 < |x|^3 < \cdots$), we have

$$\begin{aligned} |a_1x^{n-1} + \cdots + a_{n-1}x + a_n| &\leq |a_1x^{n-1}| + \cdots + |a_{n-1}x| + |a_n| \\ &= |a_1||x|^{n-1} + \cdots + |a_{n-1}||x| + |a_n| \\ &\leq |a_1||x^{n-1}| + \cdots + |a_{n-1}||x|^{n-1} + |a_n||x|^{n-1} \\ &< M|x|^{n-1}. \end{aligned}$$

It follows that for any x,

$$x^n - M|x|^{n-1} < P(x) < x^n + M|x|^{n-1}$$

Now take $x_2 = 2M$. 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 $x_1 = -2M$ 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.[75] (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 $[a, b]$.)

8.4 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 subset of the domain. For example, we may wish to consider the function $x \mapsto 1/x$ not at

[75] This is often taken as the statement of the Intermediate Value Theorem.

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 8.11. *(Extreme Value Theorem, or EVT for short) Suppose* $f : [a, b] \to \mathbb{R}$ *is continuous. Then*

- *f is bounded on $[a, b]$*[76]*, 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 8.12. *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 8.10, 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 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.

[76] In other words, a continuous function on a closed interval cannot "blow up" to infinity (or negative infinity).

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.[77] 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 $[a, b]$.

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 $f(y) = \alpha$.

[77] 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

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 $\alpha = \sup A$.

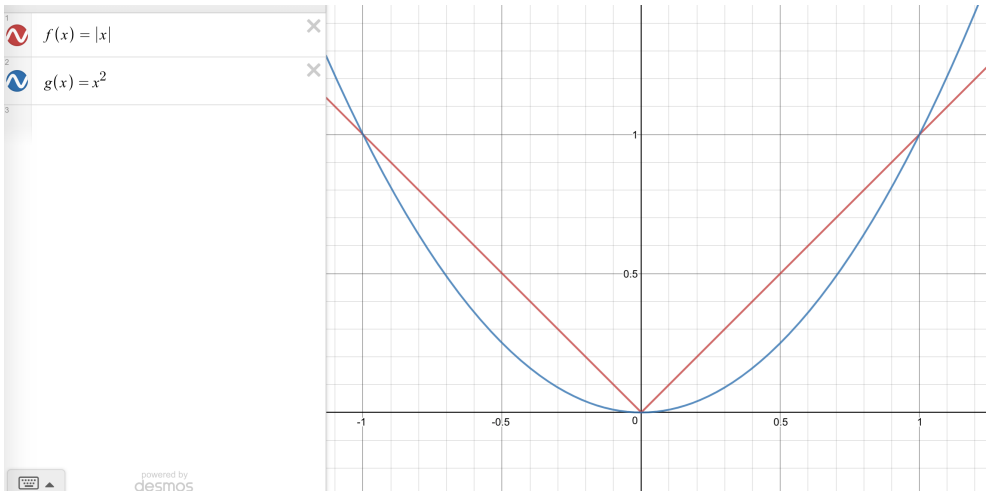
We conclude[78] that there must be a y with $f(y) = \alpha$, completing the proof of the theorem. □

[78] Somewhat miraculously — the function g was quite a rabbit-out-of-a-hat in this proof.

9 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*.

9.1 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, $t(b) - t(a)$, divided by the total time, $b - a$.[79]

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 at time t, ending at time $t + \Delta t$), with $\Delta t > 0$, the average velocity is

$$\frac{\text{displacement}}{\text{time}} = \frac{s(t + \Delta t) - s(t)}{\Delta t}.$$

[79] 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.

Similarly over a small time interval $[t+\Delta t, t]$, with $\Delta t < 0$, the average velocity

$$\frac{s(t) - s(t+\Delta t)}{-\Delta t} = \frac{s(t+\Delta t) - s(t)}{\Delta t}.$$

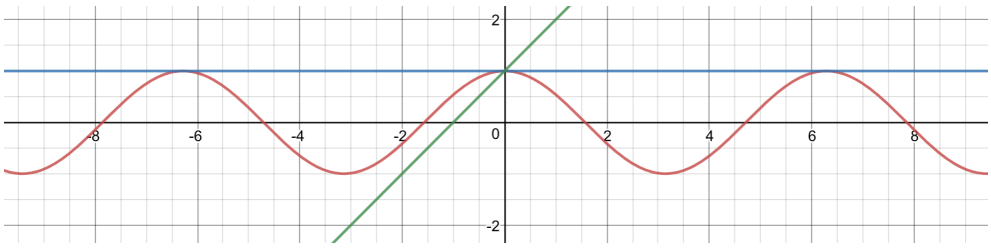
If this common quantity, $(s(t+\Delta t) - s(t))/\Delta t$, is approaching a limit as Δt approaches 0, then it makes sense to *define* instantaneous velocity at time t to be that limit, that is, to be

$$\lim_{\Delta t \to 0} \frac{s(t+\Delta t) - s(t)}{\Delta t}.$$

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.

9.2 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[80].

[80] There are many alternate notations for the derivative:

- $f'(a)$
- $\dot{f}(a)$
- $\frac{d}{dx} f(x) \,|_{x=a}$
- $\frac{df(x)}{dx} \,|_{x=a}$
- $\frac{df}{dx} \,|_{x=a}$
- $\frac{dy}{dx} \,|_{x=a}$ (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.

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'[81], whose domain is $\{a \in D : f \text{ differentiable at } a\}$[82], 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.

9.3 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,

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

[81] or $\frac{df}{dx}$, or $\dot{f}$.

[82] 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.

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.[83]

[83] We actually already did this above, when we wrote "and so

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

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[84]. 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.

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.

[84] 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 $2a$.

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[85] 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

$$\lim_{h\to 0}\frac{(a+h)^n - a^n}{h} = na^{n-1};$$

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: $f(x) = 1/x^2$.

[85] or at least a

As long as $a \neq 0$, we have

$$\begin{aligned} f'(a) &= \lim_{h\to 0} \frac{f(a+h)-f(a)}{h} \\ &= \lim_{h\to 0} \frac{\frac{1}{(a+h)^2} - \frac{1}{a^2}}{h} \\ &= \lim_{h\to 0} \frac{a^2-(a+h)^2}{(a+h)^2a^2h} \\ &= \lim_{h\to 0} \frac{-2ah-h^2}{(a+h)^2a^2h} \\ &= \lim_{h\to 0} \frac{-2a-h}{(a+h)^2a^2} \\ &= \frac{-2a}{a^2a^2} \\ &= \frac{-2}{a^3}. \end{aligned}$$

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

$$\lim_{h\to 0^+} \frac{f(1+h)-f(1)}{h} = \lim_{h\to 0^+} \frac{a+ah+b-(a+b)}{h} = \lim_{h\to 0^+} \frac{ah}{h} = \lim_{h\to 0^+} a = a,$$

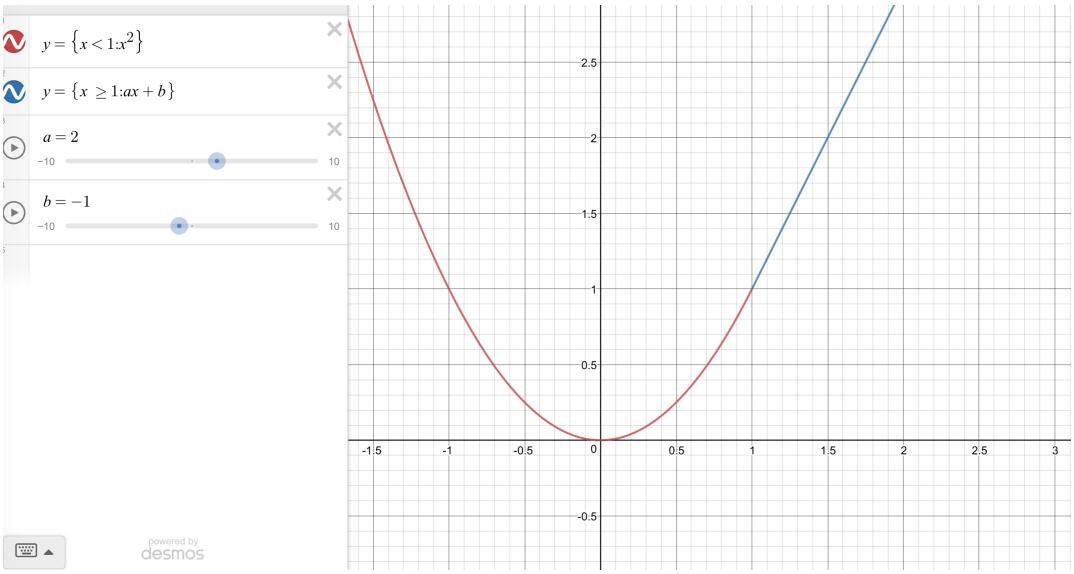
and (recalling that $a+b=1$, since we require f to be continuous at 1)

$$\lim_{h\to 0^-} \frac{f(1+h)-f(1)}{h} = \lim_{h\to 0^-} \frac{(1+h)^2-(a+b)}{h} = \lim_{h\to 0^-} \frac{2h+h^2}{h} = \lim_{h\to 0^-} (2+h) = 2.$$

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

$$\lim_{h\to 0^+} \frac{\sqrt{0+h} - \sqrt{0}}{h} = \lim_{h\to 0^+} \frac{1}{\sqrt{h}}.$$

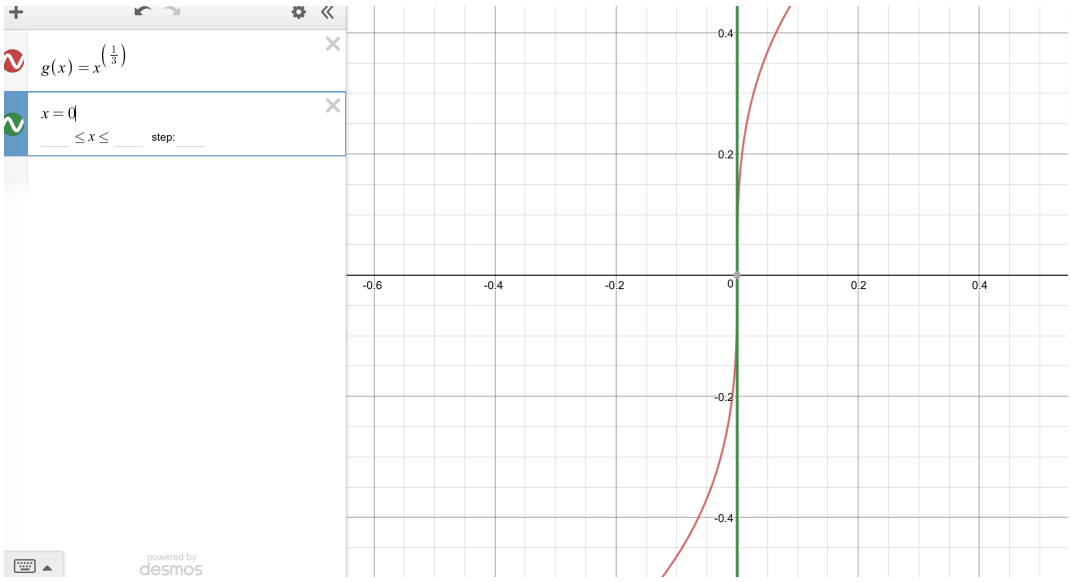
This limit does not exist, so f is not left differentiable at 0.

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.[86]

$\sin(1/x)$ and variants Consider the three functions

$$f_1(x) = \sin(1/x), \ x \neq 0$$
$$f_2(x) = x\sin(1/x), \ x \neq 0$$
$$f_3(x) = x^2\sin(1/x), \ x \neq 0,$$

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.[87]

What about at 0? For f_1, using $f_1(0) = 0$ we have (if the limit exists)

$$f'(0) = \lim_{h\to 0} \frac{\sin(1/h)}{h}.$$

[86] Shortly we will talk about "infinite limits" and rectify this deficiency.

[87] We will verify this informally soon, using or informal/geometric definition of the trigonometric functions.

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

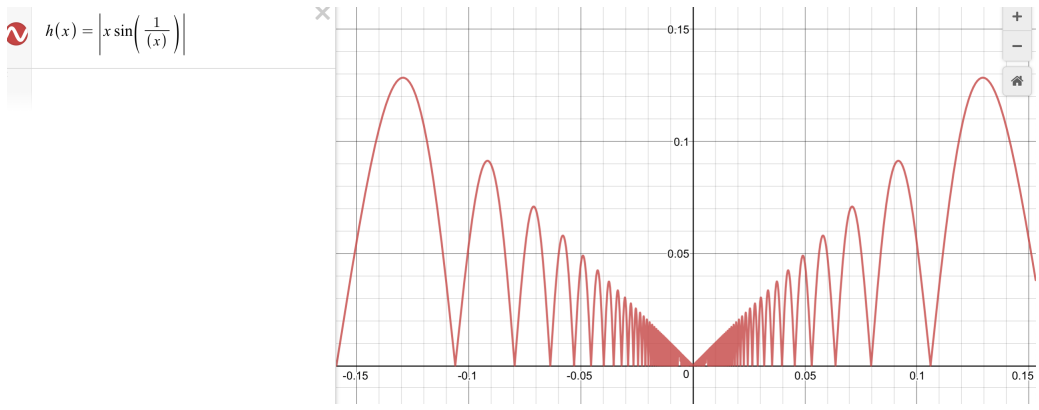
$$f_3'(0)=\lim_{h\to 0}\frac{h^2\sin(1/h)}{h}=\lim_{h\to 0}h\sin(1/h)=0,$$

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.

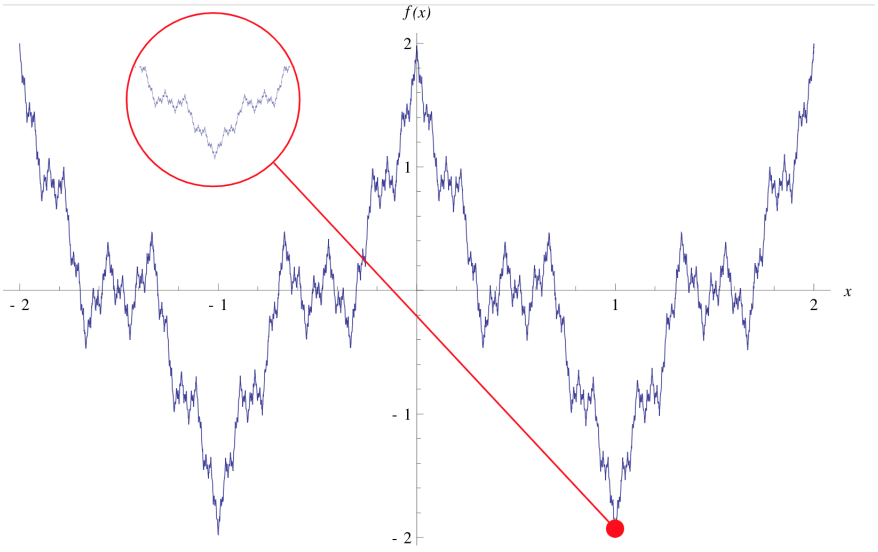
Weierstrass function It's easy to come up with an example of a function that is defined *everywhere* differentiable *nowhere* — the Dirichlet function works well. But this seems a cheat. The Dirichlet function is nowhere continuous, and if we imagine differentiability to be a more stringent notion of smoothness than continuity, then we might expect that non-continuous functions are for some fairly trivial reason non-differentiable.[88]

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 $\text{Domain}(h')=\mathbb{R}\setminus\{0,\pm 1/\pi,\pm 2/\pi,\pm 3/\pi,\ldots\}$.



[88] 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[89], 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 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

[89] In fact, in a quite precise sense *most* continuous function are nowhere differentiable.

- $f'(t)$ 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

$$L_{f,a}(x) = f(a) + f'(a)(x - a).$$

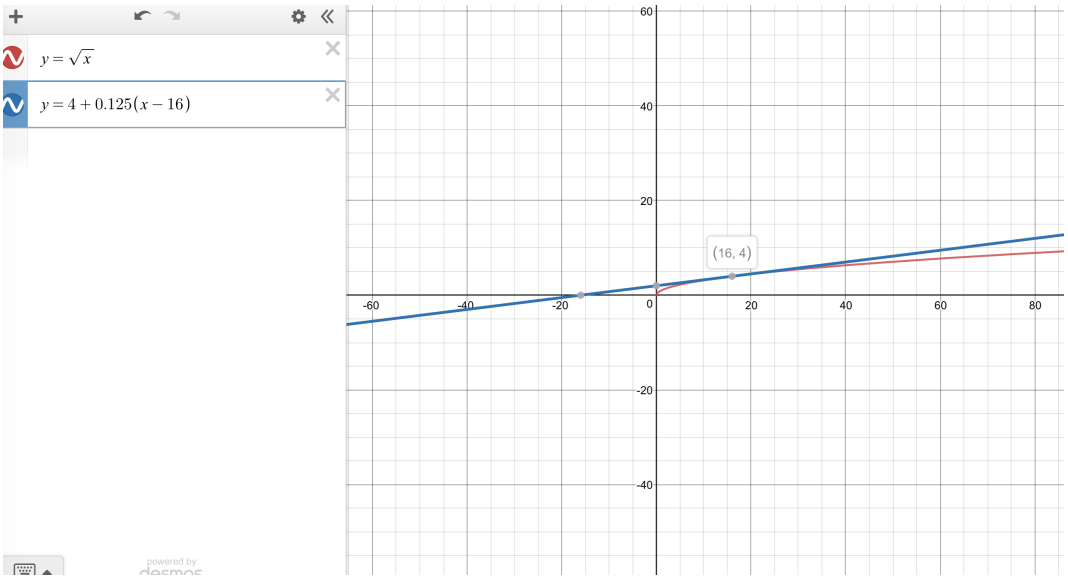
Notice that the linearization of f at a agrees with f at a: $L_{f,a}(x) = f(a)$. The intuition of the definition is that *near* a, $L_{f,a}(x)$ is a good approximation for $f(x)$, and is (often) much easier to calculate than $f(x)$.

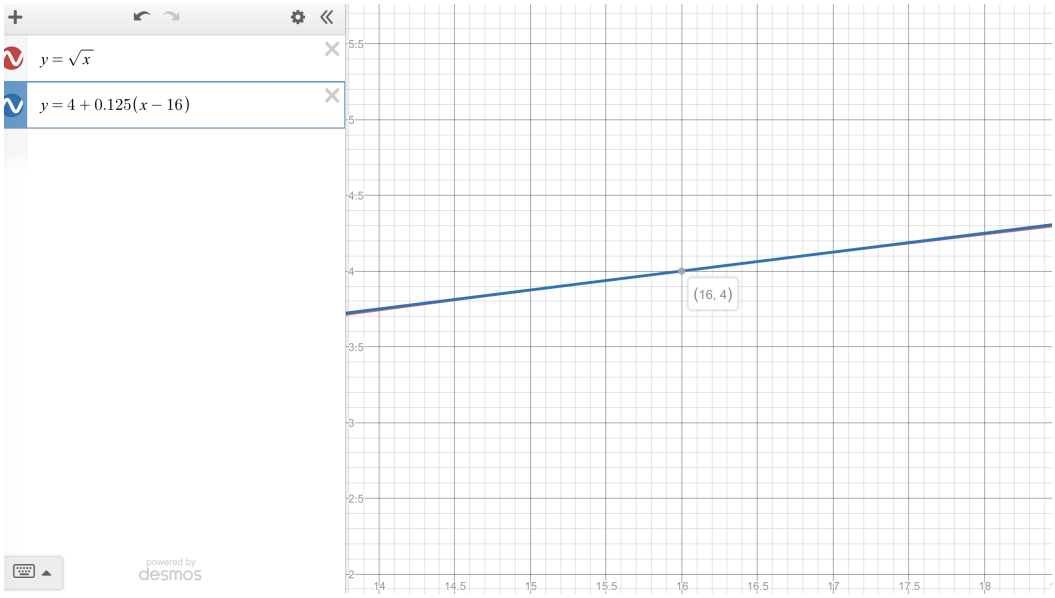
The linearization is particularly useful if the point a is such that both $f(a)$ and $f'(a)$ are particularly "nice". Here's an example: consider $f(x) = \sqrt{x}$. It's not in general easy to

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[90]; 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 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

[90] Not too surprising, since by most measures 100 is *not* close to 16.

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).[91]

9.4 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:

> 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.

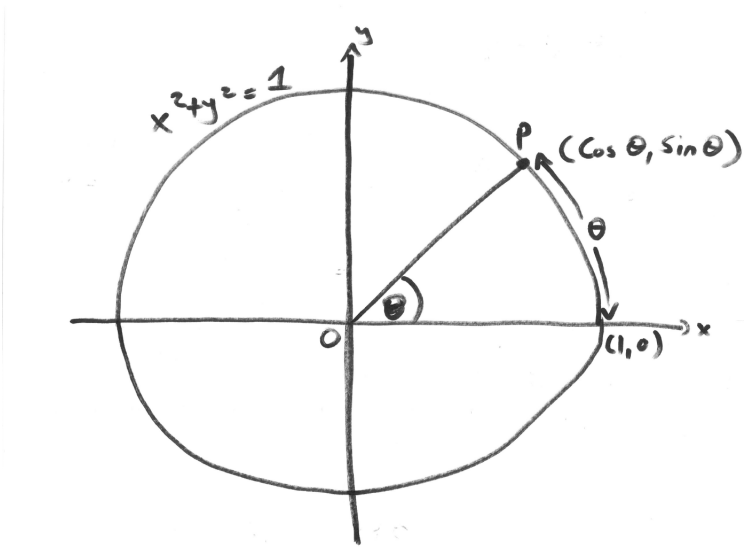


FIGURE 1 → DEF. OF SIN, COS.

[91] 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.

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):

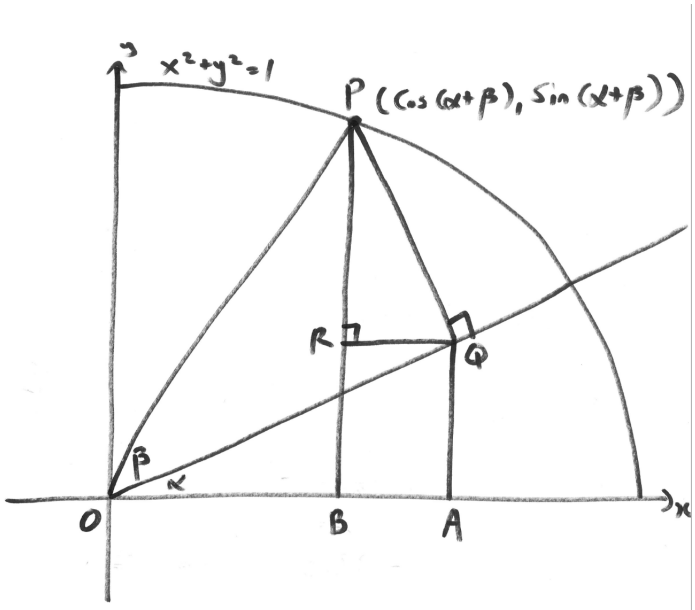


FIGURE 2 → SIN SUM FORMULA

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.

Now using the sin sum formula, we have (throughout assuming that all the various limits 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:

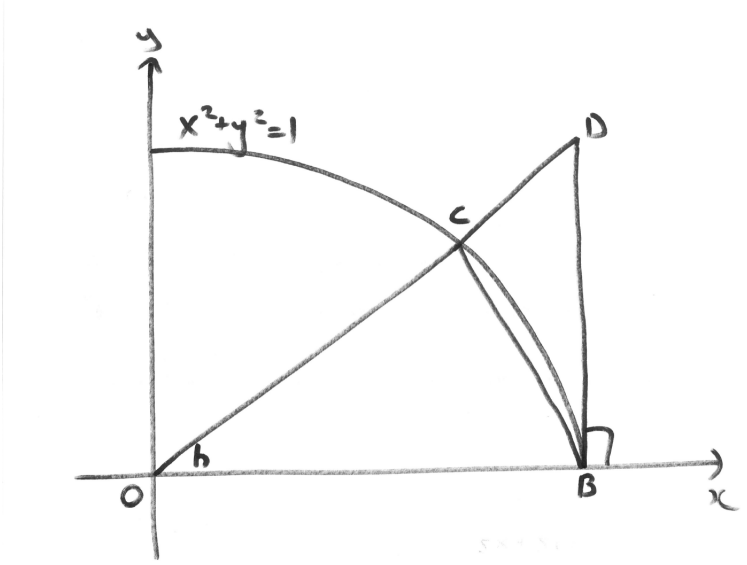


FIGURE 3 → LIMIT OF $\frac{\sin h}{h}$

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.

9.5 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 9.1. *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.[92] □

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 9.2. *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

$$\begin{aligned} f'(a)+g'(a) &= \lim_{h\to 0} \frac{f(a+h)-f(a)}{h} + \lim_{h\to 0} \frac{g(a+h)-g(a)}{h} \\ &= \lim_{h\to 0} \left(\frac{f(a+h)-f(a)}{h} + \frac{g(a+h)-g(a)}{h} \right) \\ &= \lim_{h\to 0} \frac{f(a+h)+g(a+h)-f(a)-g(a)}{h} \\ &= \lim_{h\to 0} \frac{(f+g)(a+h)-(f+g)(a)}{h}, \end{aligned}$$

which exactly says that $f+g$ is differentiable at a, with derivative $f'(a)+g'(a)$.

Similarly,

$$\begin{aligned} cf'(a) &= c\lim_{h\to 0} \frac{f(a+h)-f(a)}{h} \\ &= \lim_{h\to 0} \frac{c(f(a+h)-f(a))}{h} \\ &= \lim_{h\to 0} \frac{(cf)(a+h)-(cf)(a)}{h} \end{aligned}$$

[92] 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 cf is differentiable at a, with derivative $cf'(a)$. $\square$

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 per second squared. 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 9.3. *(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 9.4. *(Product rule for product of n functions)*[93] *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

$$\begin{aligned} (f_1 \cdots f_{n-1})'(a) f_n(a) &= \left(\sum_{k=1}^{n-1} f_1(a) \cdots f'_k(a) \cdots f_{n-1}(a) \right) f_n(a) \\ &= \sum_{k=1}^{n-1} f_1(a) \cdots f'_k(a) \cdots f_{n-1}(a) f_n(a), \end{aligned}$$

[93] 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.

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

$$h_n'(x) = \frac{1}{n} x^{\frac{1}{n}-1}.$$

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 9.5. *(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 9.6. *(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.

9.6 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 f and g 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 9.7. *(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 sin 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 $\cos x$[94],

$$\cos' x = -\sin x$$

(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

$$\begin{aligned}(f_1 \circ f_2 \circ f_3)'(a) &= (f_1 \circ (f_2 \circ f_3))'(a) \\ &= f_1'((f_2 \circ f_3)(a))((f_2 \circ f_3)'(a)) \\ &= f_1'((f_2 \circ f_3)(a))f_2'(f_3(a))f_3'(a) \\ &= f_1'(f_2(f_3(a)))f_2'(f_3(a))f_3'(a).\end{aligned}$$

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. This is a somewhat different justification to the one presented by Spivak.

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

$$g(a+h) - g(a) = g'(a)h + \varepsilon(h)h \tag{1}$$

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.

[94] 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.

Similarly

$$f(g(a)+k)-f(g(a))=f'(g(a))k+\eta(k)k \tag{2}$$

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 (2) remains true at $k=0$.

We now study the expression $f(g(a+h))-f(g(a))$. Applying (1) we get

$$f(g(a+h))-f(g(a))=f(g(a)+g'(a)h+\varepsilon(h)h)-f(g(a)). \tag{3}$$

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 (2) to (3) 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.

10 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.

10.1 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*[95] *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.

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

[95] "the": the maximum value is easily checked to be unique.

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 10.1. *(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[96], but at neither point in the derivative zero.[97]
- 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

$$f'_-(x) = \lim_{h \to 0^-} \frac{f(x+h) - f(x)}{h} \geq \lim_{h \to 0^-} 0 = 0.$$

Now consider the derivative of f from above at x. We have, for $h > 0$,

$$\frac{f(x+h) - f(x)}{h} \leq 0$$

[96] As differentiable as it can be ... differentiable from above at 0 and from below at 1.

[97] 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.

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*[98] *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[99]). 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 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 10.2. *(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*

[98] "a": a local maximum value is clearly not necessarily unique; see examples below.

[99] Physically, a local maximum is the summit of a mountain, and of course different mountains in general have different heights.

- *f differentiable at x*

then $f'(x) = 0$.

We do not present the proof here; it is in fact just a corollary of Claim 10.1. 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 10.1 shows that $f'(x) = 0$.[100]

As with Claim 10.1, Claim 10.2 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.[101]

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 $f'(x) = 0$.[102] 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 10.3. *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).*

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:

[100] I wrote "We do not present the proof here"; but then it seems I went and gave the proof.

[101] 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.

[102] Many authors also say that x is a critical point for f if f is not differentiable at x.

- $f : [-1, 1] \to \mathbb{R}$ 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

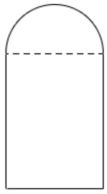
$$f(x) = \frac{1}{x^2 - 4x + 3}.$$

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

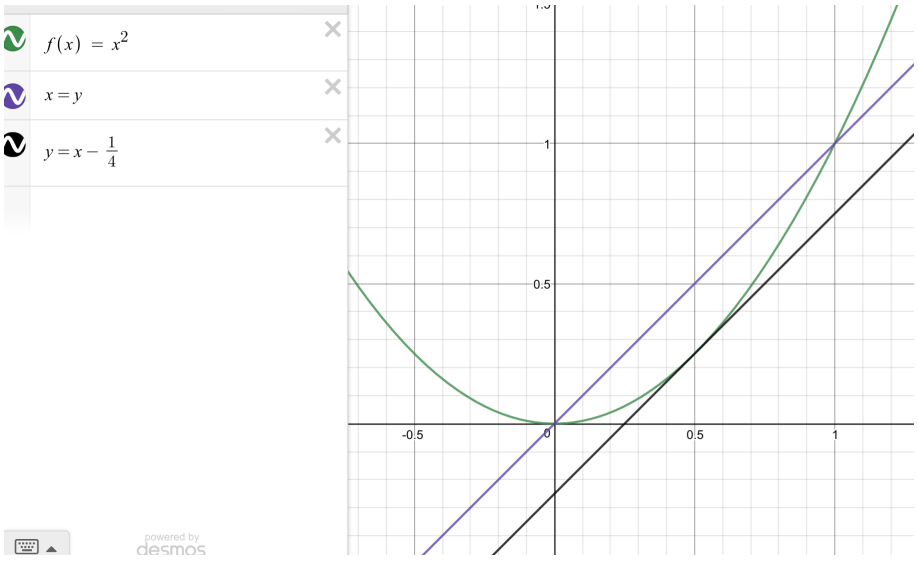
- $A(0) = 0$,
- $A(L/(2 + \pi/2)) = \frac{L^2}{2(2+\pi/2)^2} + \frac{\pi L^2}{8(2+\pi/2)^2}$ and
- $A(L/(1 + \pi/2)) = \frac{\pi L^2}{8(1+\pi/2)^2}$.

The second of these is the largest, so is the largest achievable area.

10.2 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.[103] 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 10.4. *(Mean value theorem) Suppose* $f : [a, b] \to \mathbb{R}$ *is*

- *continuous on* $[a, b]$*, and*
- *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 $f'(c) = 0$.[104]

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)

[103] Find one! (It will be on the homework $\cdots$).

[104] 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.

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.

But of course, it is not[105] 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

$$f(x) = \begin{cases} 0 & \text{if } x^2 < 2 \\ 1 & \text{if } x^2 > 2 \end{cases} cc$$

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 10.5. *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). □

[105] at least, should not

Corollary 10.6. *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 10.5 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*[106], if whenever $a < b$ in I, $f(a) < f(b)$. Say that f is *weakly increasing* on I if whenever $a < b$ in I, $f(a) \leq f(b)$.

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 10.7. *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.

[106] 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.

Claim 10.8. *If f is weakly increasing on an interval, and differentiable on the whole interval, then $f'(x) \geq 0$ on the interval.*[107]

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 10.9. *(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*[108]*. If $f'(y) < 0$ 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 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 10.7, 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 10.10. *(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

$$0 < f''(x) = f''_-(x) = \lim_{h \to 0^-} \frac{f'(a+h) - f'(a)}{h} = \lim_{h \to 0^-} \frac{f'(a+h)}{h}.$$

[107] 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.

[108] Recall that "near x" means: in some interval $(x-\delta, x+\delta)$, $\delta > 0$.

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 10.9, 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. □

10.3 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[109], 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.

[109] Ask Desmos to graph the function $f(x) = [x \cdot (1/x)]$.

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 $(-\infty, \infty)$.

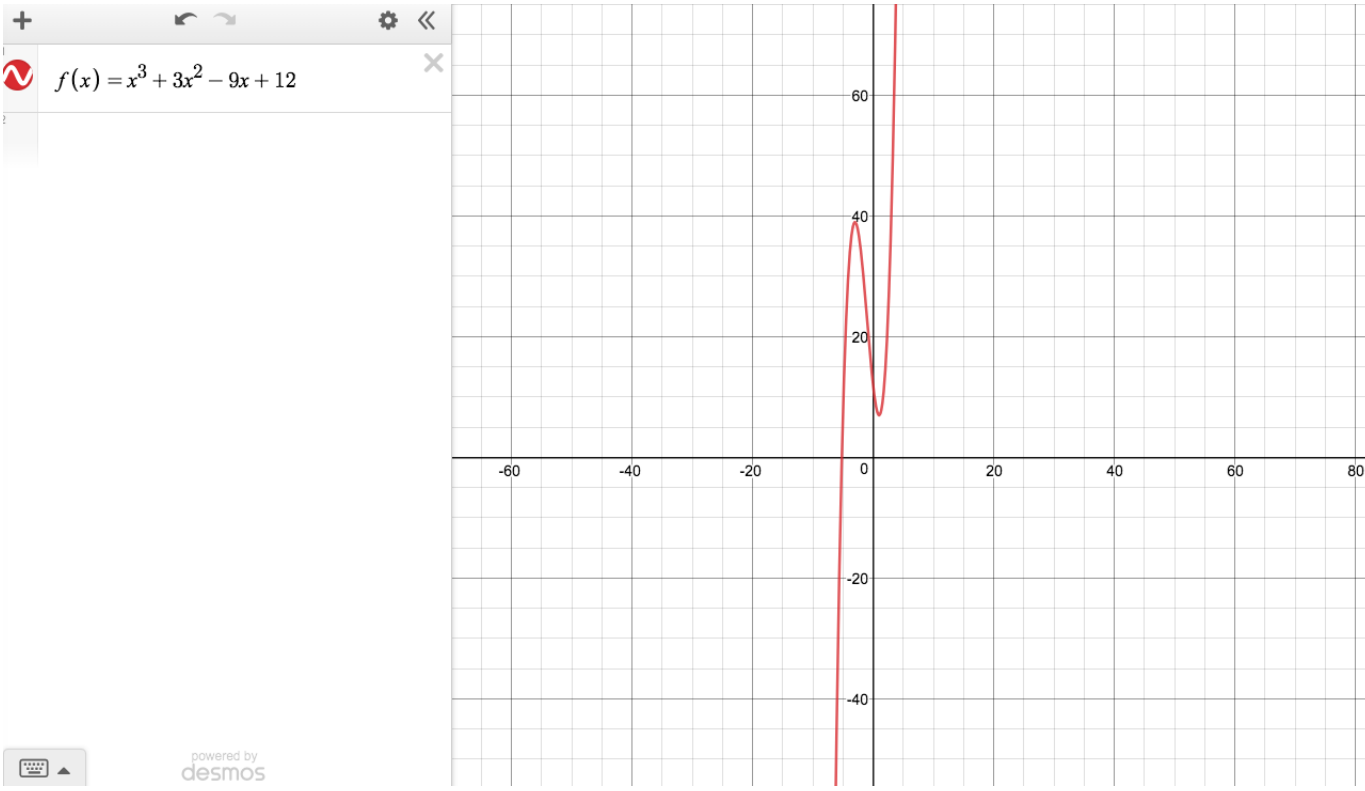
Step 2 $\lim_{x\to\infty} x^3 + 3x^2 - 9x + 12 = \infty$ and $\lim_{x\to-\infty} x^3 + 3x^2 - 9x + 12 = -\infty$.

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

$$\lim_{x \to 0^-} \frac{x^2}{1 - x^2} = +\infty,$$

$$\lim_{x \to 0^+} \frac{x^2}{1 - x^2} = -\infty$$

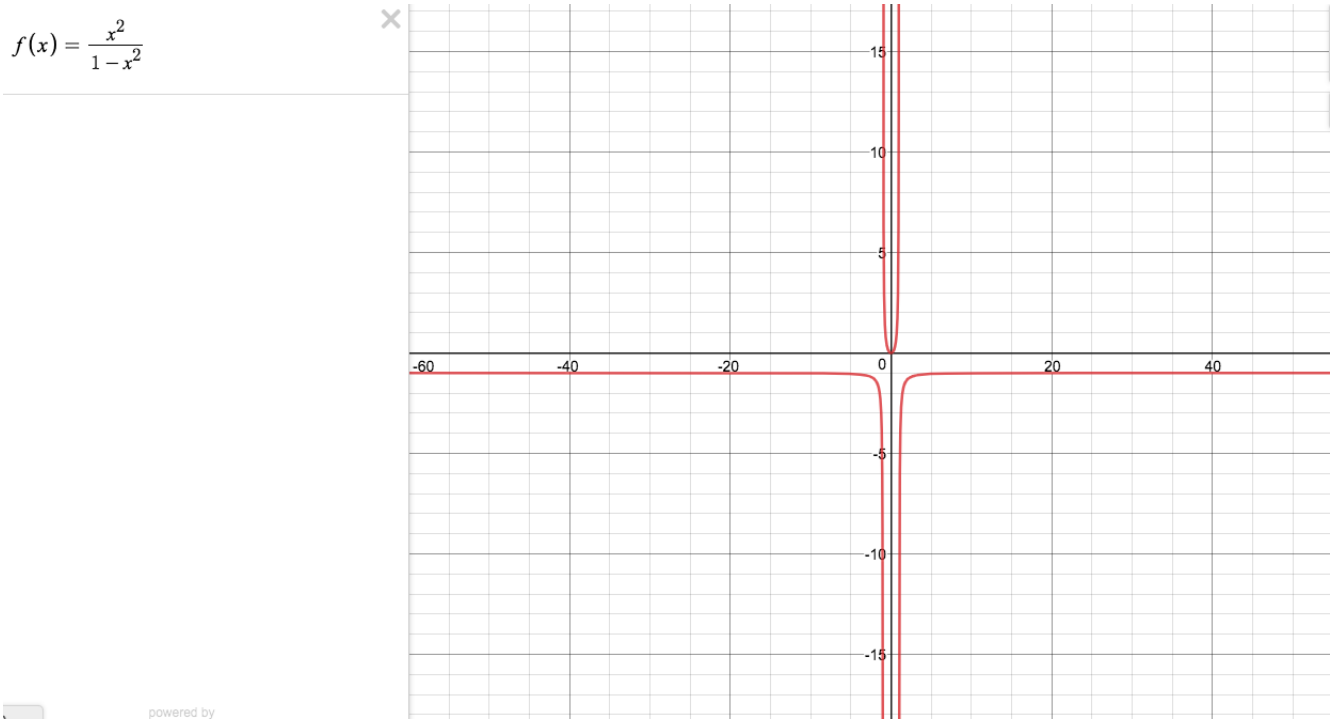
and

$$\lim_{x \to \infty} \frac{x^2}{1 - x^2} = -1.$$

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 $1 : 1$):



10.4 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[110]. 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 $L_f(x) = f(a) + f'(a)(x-a) = f'(a)(x-a)$, and the linearization of g near a is $L_g(x) = g(a) + g'(a)(x-a) = g'(a)(x-a)$.[111] 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,

$$\frac{f(x)}{g(x)} \approx \frac{L_f(a)}{L_g(a)} = \frac{f'(a)(x-a)}{g'(a)(x-a)} = \frac{f'(a)}{g'(a)}$$[112]

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,

$$\begin{aligned}
\lim_{x\to a} \frac{f(x)}{g(x)} &= \lim_{x\to a} \frac{f(x)-f(a)}{g(x)-g(a)} \quad (f(a) = g(a) = 0) \\
&= \lim_{x\to a} \frac{\frac{f(x)-f(a)}{x-a}}{\frac{g(x)-g(a)}{x-a}} \\
&= \frac{\lim_{x\to a} \frac{f(x)-f(a)}{x-a}}{\lim_{x\to a} \frac{g(x)-g(a)}{x-a}} \quad \text{(adding assumption here: bottom limit is non-zero)} \\
&= \frac{f'(a)}{g'(a)}.
\end{aligned}$$

[110] 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.

[111] We're making the assumption here that f, g are both differentiable at a.

[112] We're making another assumption here — that $g'(a) \neq 0$.

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 10.11. *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.

$\lim_{x\to 0} \frac{\sin x}{x}$ 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.[113]

$\lim_{x\to 0} \frac{x}{\tan x}$ 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

$$\lim_{x\to 1} \frac{x^3 - x^2 - x + 1}{x^3 - 3x + 2} = \lim_{x\to 1} \frac{x+1}{x+2} = \frac{2}{3}.$$

[113] 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!

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.[114]

Theorem 10.12. *(L'Hôpital's rule) Let f and g be functions defined and differentiable near a[115]. 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.*[116]

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:

$$\begin{aligned} \text{“}\lim_{x\to 1} \frac{x^3 - x^2 - x + 1}{x^3 - 3x + 2} &= \lim_{x\to 1} \frac{3x^2 - 2x - 1}{3x^2 - 3} \quad \text{(by L'Hôpital's rule)} \\ &= \lim_{x\to 1} \frac{6x - 2}{6x} \quad \text{(by L'Hôpital's rule)} \\ &= \frac{2}{3}, \end{aligned}$$

[114] The proof is quite messy, and will only appear in these notes, not in class.

[115] But not necessarily even defined *at* a.

[116] Note that we don't require $g'(a) \neq 0$: $g'(a)$ might not even exist!

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.

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 $\{(t, f(t)) : t \in [a, b]\}$[117] 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 10.13. *(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).

[117] 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.

- 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.

- 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)[118]: 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:

[118]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

$$\lim_{x \to a^+} \frac{f(x)}{g(x)} = \lim_{x \to a^+} \frac{f'(x)}{g'(x)}, \tag{4}$$

and by similar reasoning on the interval $(a - \delta, a)$ we should have

$$\lim_{x \to a^-} \frac{f(x)}{g(x)} = \lim_{x \to a^-} \frac{f'(x)}{g'(x)}. \tag{5}$$

L'Hôpital's rule follows from a combination of (4) and (5).

Thus to complete the proof of L'Hôpital's rule we need to verify (4). 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 (4). □

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 10.14. *(General form of L'Hôpital's rule)*[119] *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.

- At all sufficiently large number, f is continuous;
- the same for g;
- for all sufficiently large x, $g'(x) \neq 0$; and

[119] 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.

- 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 $f(x) = (\log x)/x$[120] 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 0, it follows that

$$\lim_{x\to\infty} \frac{\log x}{x} = 0.$$[121]

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

[120] 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 $\log'(x) = 1/x$.

[121] 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.

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

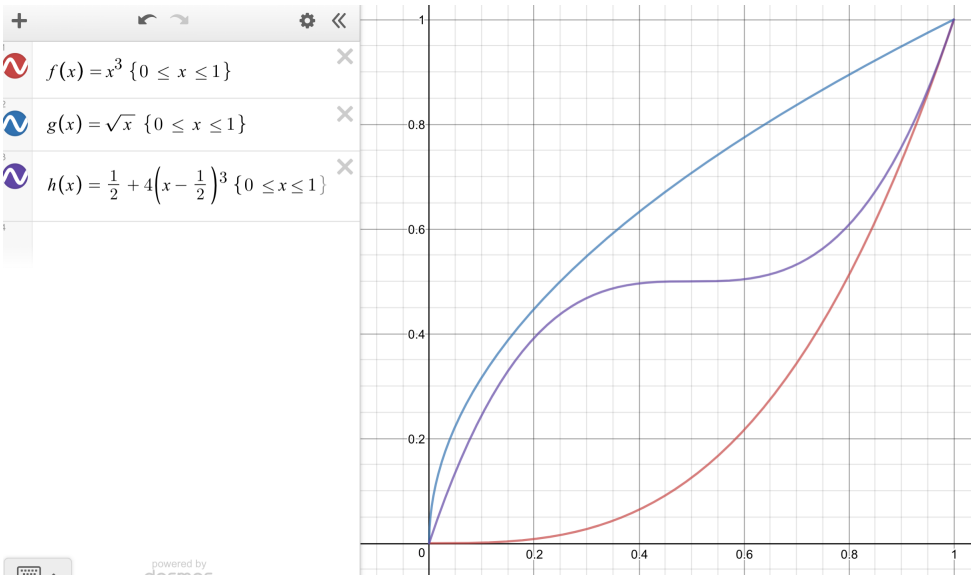
$$\begin{aligned}
\lim_{x\to\infty} \frac{(\log x)^2}{\sqrt{x}} &= \lim_{x\to\infty} \frac{2(\log x)(1/x)}{1/(2\sqrt{x})} \\
&= \lim_{x\to\infty} \frac{4\log x}{\sqrt{x}} \\
&= \lim_{x\to\infty} \frac{4/x}{1/(2\sqrt{x})} \\
&= \lim_{x\to\infty} \frac{8}{\sqrt{x}} \\
&= 0.
\end{aligned}$$

10.5 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.



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*[122], 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*[123], 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

$$((1-t)a + tb, f((1-t)a + tb))$$

[122]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 I) 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.

[123]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.

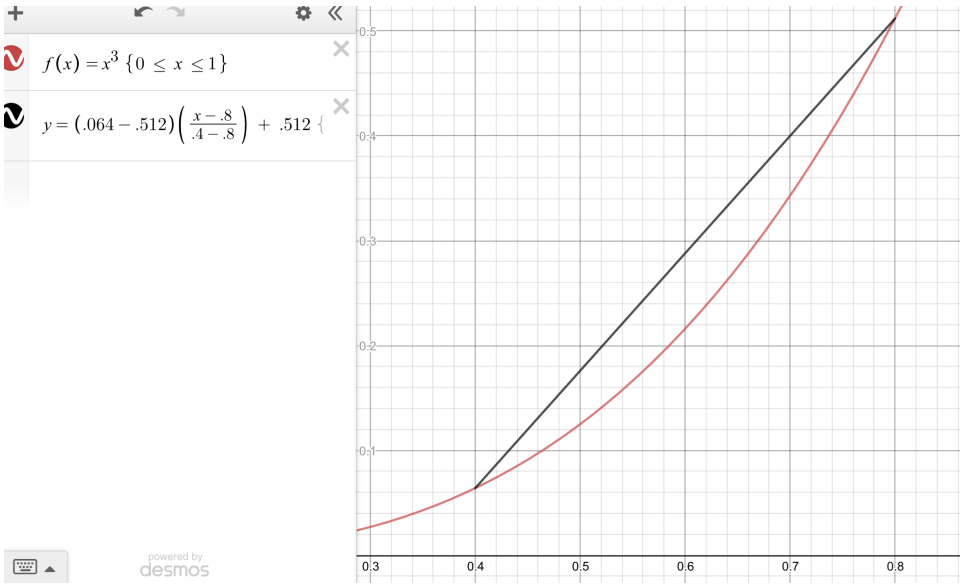
is a point on the graph of the function f, while the point

$$((1-t)a+tb, (1-t)f(a)+tf(b))$$

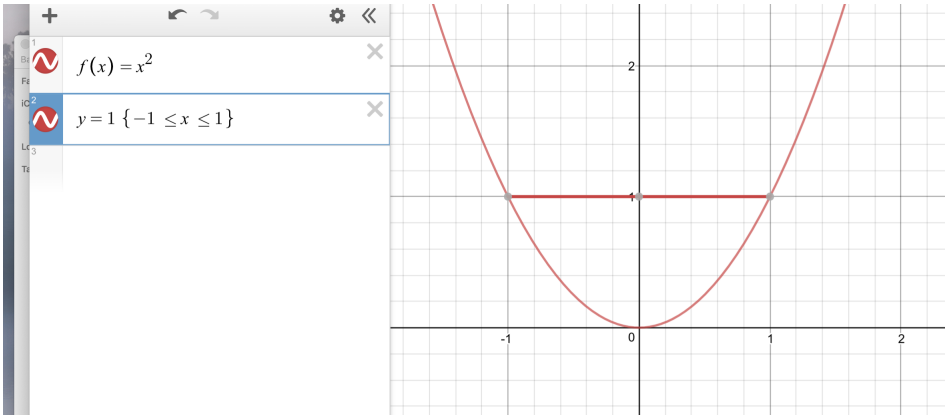
is a point on the secant line to the graph of the function f between the points $(a, f(a))$ and $(b, f(b))$. So the graphical sense of convexity is that

> f is convex on I if the graph of f lies below the graphs of all its secant lines on I.

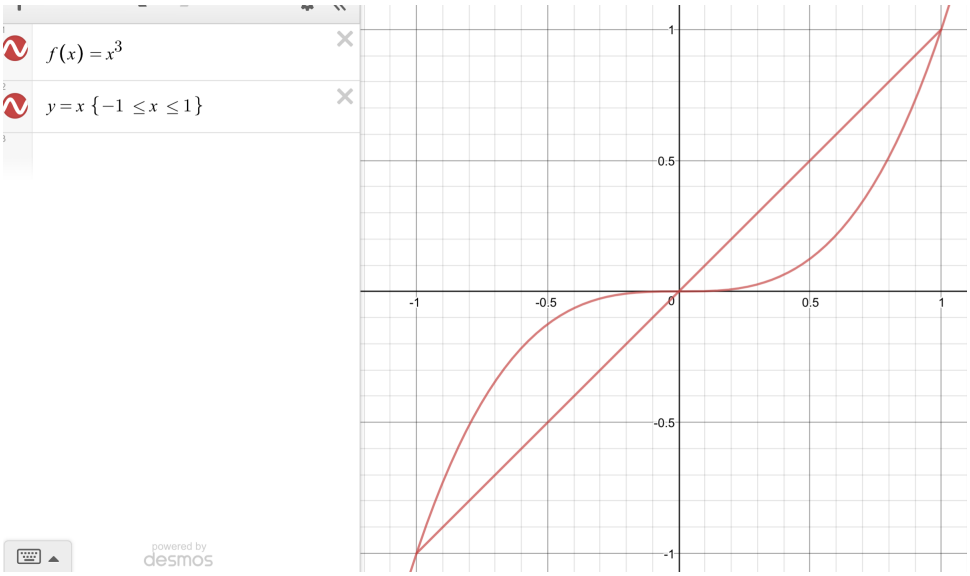
Illustrated below is the graph of $f(x) = x^3$, which lies below all its secant lines between 0 and 1, and so fairly clearly is convex on that interval. The picture below shows one secant line, from $(0.4, 0.64)$ to $(0.8, 0.512)$.



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).



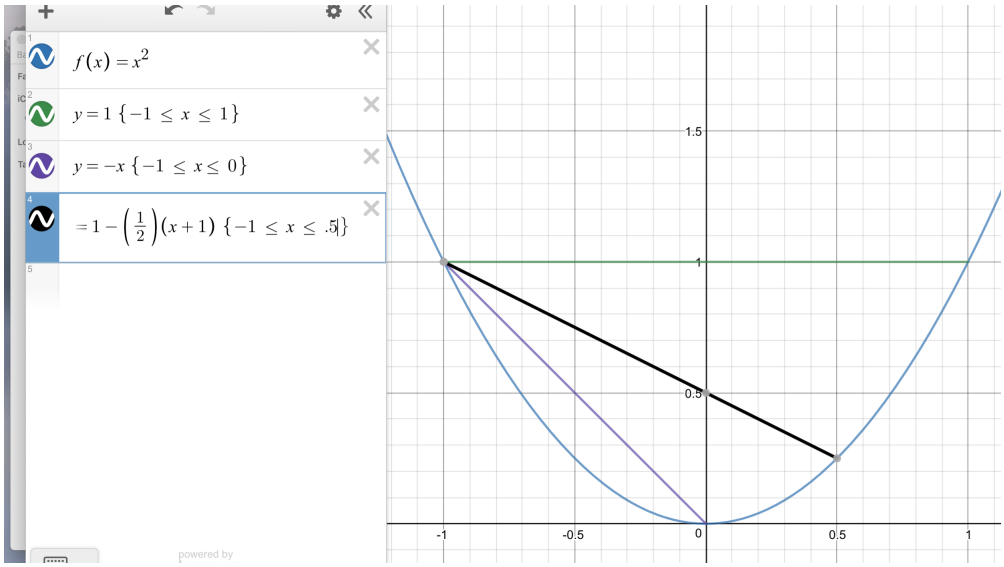
As was implicitly assumed in the last example discussed, just as convexity graphically means that secant lines lie above the graph, we have a graphical interpretation of concavity:

> 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 10.15. *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 10.16.

- *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[124]*, 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 $a < x < b \in I$,

$$f(x) < \left(1 - \frac{x - a}{b - a}\right) f(a) + \left(\frac{x - a}{b - a}\right) f(b).$$

[124]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 10.17. *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[125] 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 10.18. *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 10.19. *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 10.20. *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.*

[125] 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 $(1, \infty)$).

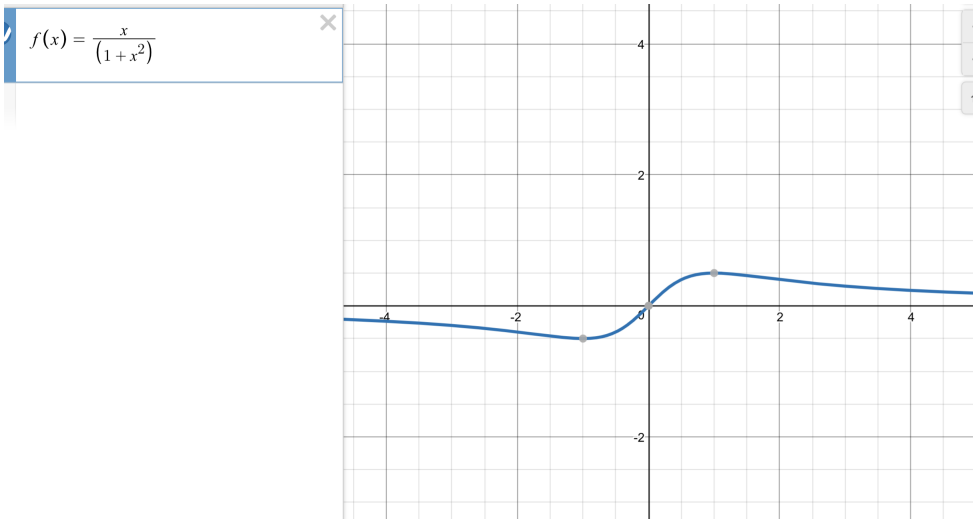
It follows that there is a local minimum at $(-1, -1/2)$ and a local maximum at $(1, 1/2)$. We also have

$$f''(x) = \frac{2x(x^2-3)}{(1+x^2)^3},$$

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 10.19, 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 10.21. *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 $x_0 \in (a, b)$. Since f is differentiable everywhere, by Fermat principle $f'(x_0) = 0$. By the Mean Value Theorem applied to $[a, x_0]$, there is $x_1 \in (a, x_0)$ with

$$f'(x_1) = \frac{f(x_0) - f(a)}{(x_0 - a}. \quad (\star)$$

Now f' is increasing on $[a, b]$ (by hypothesis), so $f'(x_1) < f'(x_0) = 0$. But $f(x_0) \geq f(a)$ (since x_0 is a maximum point for f on $[a, b]$), so $(f(x_0) - f(a))/(x_0 - a) \geq 0$. This contradicts the equality in $(\star)$ above. □

Proof (of Theorem 10.19): 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.[126] □

[126] The second inequality, $g(x) < g(b)$, similarly reduces to the other characterization of convexity, but that isn't needed here.