Math 1231 Course Notes

Summer 2024

Contents

1 Limits

1.1 Functions

We begin by recalling what functions are: rules for assigning outputs to inputs. Anything that takes in a piece of input and returns a specific output according to some rule is a function. The rule might be expressed algebraically, as in $f(x) = 3x+1$, which says that the output $f(x)$ assigned to the input x is $3x+1$. This is an example of a linear equation—if we plot all of the pairs (x, y) satisfying the rule $y = f(x) = 3x+1$, we get a line in the plane. We are using the letter f here to refer to the function itself—f is the rule, and we apply it to an input x to get an output. For instance, we could have an input of 4, and applying the function f we get $f(4) = 13$. The point $(4, 13)$ is then a point on the line that is the graph of this function.

The rule for a function doesn't always have to look like this though. Sometimes we might combine multiple rules into one overall rule, where the kind of input determines which rule we use. This happens when we define *piecewise functions*, e.g.

$$f(x) = \begin{cases} x-2 & x \leq 4 \\ 10-2x & x > 4 \end{cases} \qquad \text{or} \qquad g(x) = \begin{cases} 3x+1 & x \text{ is odd} \\ x/2 & x \text{ is even.} \end{cases}$$

Note that this second piecewise funtion (which is quite famous!) has a rule that we can really only use with integers, since while it makes sense to ask whether a number like 10 is odd or even (so we can say $g(10) = 5$), it doesn't make sense to ask whether a number like 10.1 is odd or even. We would say that the **domain** of this function is the set of all integers (usually denoted $\mathbb{Z}$). The first function is defined for any real number, so it has domain we could write either as $(-\infty, \infty)$, which tells us any real number strictly between $-\infty$ and ∞ is in the domain (so, all of the real numbers), or we could also use the symbol $\mathbb{R}$. The rule

$$h(x) = \frac{(x-2)(x+1)}{x-3}$$

also defines a function where it makes sense to try plugging in real numbers, but here we have to be a little bit careful. Most real numbers work just fine, but if we try to ask what output the rule assigns to the input $x = 3$, we run into a problem. When x is 3, the denominator of the expression we've given for y becomes 0, and division by zero is undefined, meaning the rule we have doesn't yield any output for $x = 3$. We would say the domain of this function is therefore all real numbers except for 3, which we can write as $(-\infty, 3) \cup (3, \infty)$. The first interval here tells us anything strictly between $-\infty$ and 3 is included in the domain, and the second interval says anything strictly between 3 and ∞ is in the domain. The symbol $\cup$ means *union*, so what this is saying is that the domain of the function is both of these two intervals taken together.

This isn't the only way to write down the domain of a function, of course. A domain is just a set of numbers, and we have lots of ways of writing down sets. A common one is to use notation like $\{x : \text{-some condition-}\}$, which means the set of all numbers x such that the specified condition is true. For example, if we wanted a different way to write the domain of the function above, we could say $\{x : x \neq 3\}$, i.e., all numbers x such that x is not 3. The interval (a, b) could also be written as $\{x : a < x < b\}$ (called an **open** interval), and the interval $[a, b]$ would be $\{x : a \leq x \leq b\}$ (called a **closed** interval—square brackets mean a and b are included). An interval like $[6, 7)$ is the set $\{x : 6 \leq x < 7\}$, and this is **neither open nor closed**.

Functions and their domains can also be a lot more abstract. Inputs and outputs don't necessarily have to be things we would usually think of as being mathematical objects. For instance, GW assigns students a GWID—this is a function whose inputs are people and outputs are numbers. The domain of this function is the set of GW students, $\{x : x \text{ is a GW student}\}$.

In this course, we'll be focusing on functions whose inputs and outputs are mathematical objects of some kind, usually real numbers (although when we start talking about derivatives, we'll be talking about a function whose inputs and outputs are themselves functions!). Here are some familiar kinds of functions we'll be spending a lot of time with this summer:

Polynomial functions are functions like $f(x) = 5x^6 + x^2 + 4x - 2$, sums of powers of a variable x with some coefficients, written very generally like

$$y = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0$$

where the $a_0, \ldots, a_n$ are some numbers and n is called the **degree** of the polynomial. When $n = 1$ we have a linear equation, $n = 2$ is a quadratic equation, and $n = 3$ is a cubic equation. You should remember the quadratic formula, which tells you how to find the inputs x that get assigned outputs $f(x) = 0$. It says that if $ax^2 + bx + c = 0$ then

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

It is also useful to recall that

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

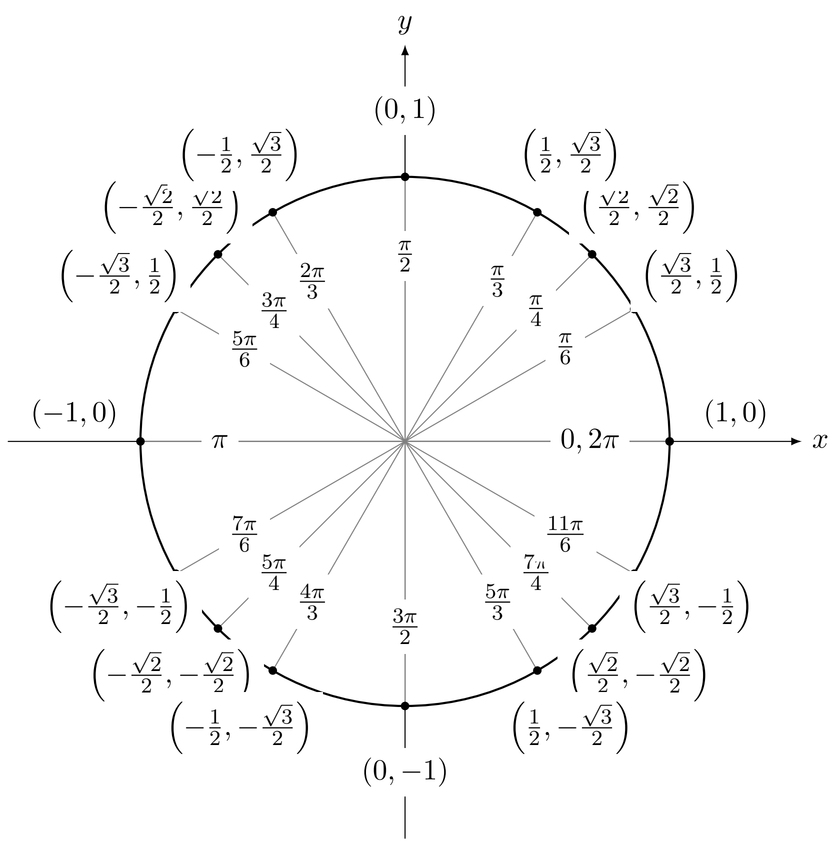
Rational functions are the ratio of two polynomials, e.g.

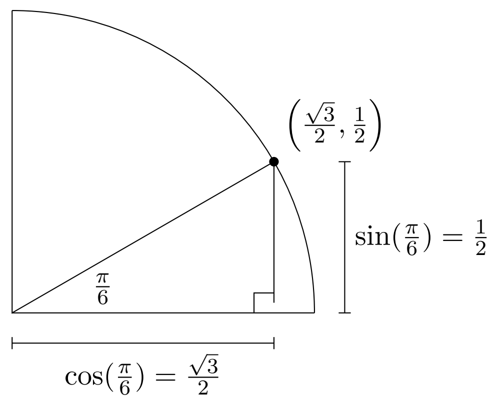
$$f(x) = \frac{x^2 + 2x}{x^3 - 6}.$$

Trigonometric functions: In this course we will *always* use radians, because they are unitless and thus easier to track (especially when using the chain rule). Useful facts include:

- The most important trigonometric identity, and really the only one you probably need to remember, is $\cos^2(x) + \sin^2(x) = 1$.
- From this you can derive the fact that $1 + \tan^2(x) = \sec^2(x)$.
- $\sin(-x) = -\sin(x)$. We call functions like this "odd".
- $\cos(-x) = \cos(x)$. We call functions like this "even."
- $\sin(x + \pi/2) = \sin(\pi/2 - x) = \cos(x)$
- A fact that we will probably use exactly twice is the sum of angles formula for sine: $\sin(x+y) = \sin(x)\cos(y) + \cos(x)\sin(y)$.
- Similarly, $\cos(x + y) = \cos(x)\cos(y) + \sin(x)\sin(y)$

Throughout the course you will be asked to recall various values of the trig functions, that is, you will have to be able to say that, e.g., $\cos(\pi/6) = \sqrt{3}/2$. For this reason, you should always have the unit circle written down or committed to memory:





And we recall that the sine of the angle is the y-coordinate of the point and the cosine of the angle is the x-coordinate of the point.

1.2 Limits, intuitively

Intuitively, we know what it means for the value of a function to approach or get closer to some particular value, and we'll be relying on this intuitive understanding to develop the tools around limits.

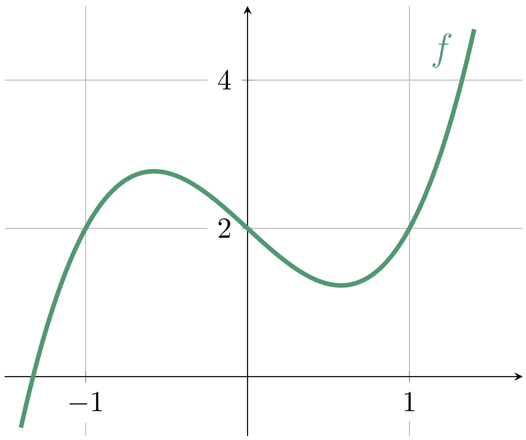
Given a function f that is defined at all the points *near* some point a, we can define the limit of f at a, should it exist. Here's the definition given in OpenStax Section 2.2:

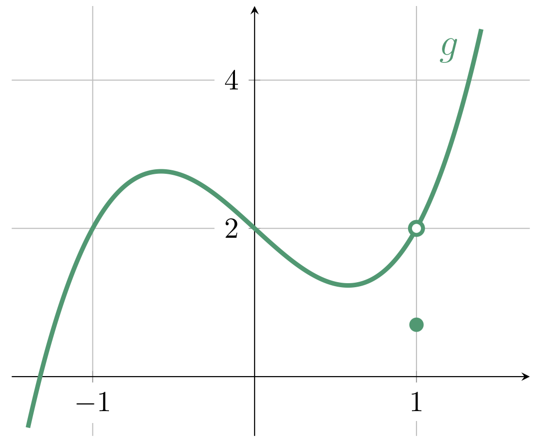
Definition. Let $f(x)$ be a function defined at all values in an open interval containing a, with the possible exception of a itself, and let L be a real number. If all values of the function $f(x)$ approach the real number L as the values of x approach the number a, then we say that the limit of $f(x)$ as x approaches a is L. (More succinct, as x gets closer to a, $f(x)$ gets closer and stays close to L.) Symbolically, we express this idea as

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

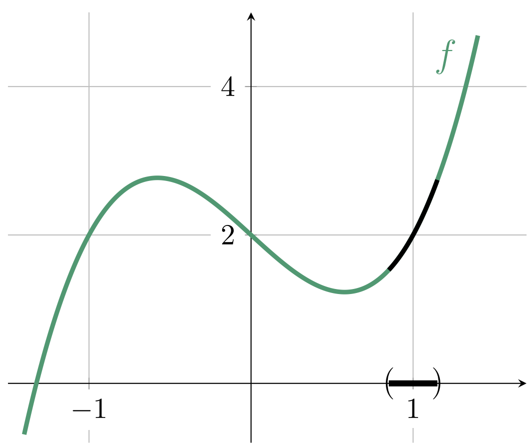
Note that this really does rely on our intuition—we're not formally defining what it means to "get closer" and "stay close," even though these are key to understanding limits. This is okay though, and in fact, the formal way we might define limits (in say, section 2.5 of OpenStax) didn't show up until almost 200 years after Newton and Leibniz began their work on the main ideas of calculus.

The kind of intuition we should have about things getting close and staying close is a visual intuition, so for all of our first examples, we'll consider graphs of some functions without actually saying what equations define the functions. We should be able to use our intuitive definition right away on the two functions pictured below:





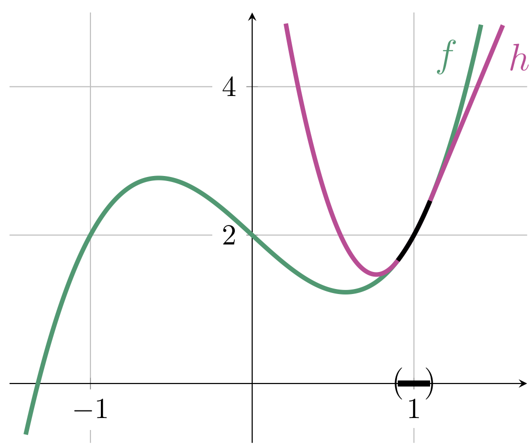
Let's see what we can say about $\lim_{x\to 1} f(x)$ and $\lim_{x\to 1} g(x)$. Per our intuitive definition, the question we need to be asking is the following: when x "gets close" to 1, is there some value L such that $f(x)$ "gets close" and "stays close" to L? The picture suggests that when x is close to 1, $f(x)$ looks pretty close to 2, so we should have in mind that $L = 2$ is a good candidate for the limit. Let's check both parts of the definition. If we consider the points close to $x = 1$ (say, points in a small open interval around 1) on the x axis and then look at their corresponding outputs, we get the following picture:



The outputs $f(x)$ *get close* to $L = 2$ because no matter how close to 2 we get, say, 2.1, there's some input x' in this little interval around $x = 1$ with an output even closer to 2 than 2.1 is. The outputs $f(x)$ also *stay close* to $L = 2$ since the outputs we get from this small interval around $x = 1$ are all contained in their own nice little interval around $L = 2$—they don't suddenly wander off away from $L = 2$ with some outputs close to 2 and some farther away. If this is true for all the intervals we could draw around $L = 2$ and $x = 1$, then we know that $f(x)$ really does get close and stay close to $L = 2$ when x gets close to 1, and looking at the picture it should be easy to convince ourselves that this really is the case: there's nothing special about the little interval drawn here—it could get much smaller and the same "closeness" considerations would apply. Therefore we can say with confidence that $\lim_{x\to 1} f(x) = 2$.

The discussion here with these small intervals around the input $x = 1$ and the limit $L = 2$ is already pushing us in the direction of a more formal definition of limits, so we won't pursue this any further. Looking at the first drawing of the function f, we can already say just based on the graph that the limit as x approaches 1 should clearly be 2, and this business with small intervals can help give us a little bit more to back up our intuition about what "closeness" should mean, but we don't need it. It'll be more helpful in a moment when we see some limits that fail to exist.

Before we do that, let's look for a second at the function g. It looks nearly identical to f, except we have these little dots telling us that even though the function g looks like it should have $g(1) = 2$, the function is actually defined so that $g(1) \neq 2$. Instead, $g(1)$ looks like something maybe a bit less than 1. And in fact, when we say that g "looks like it should have $g(1) = 2$" this is appealing to exactly the fact that as x gets close to 1, $g(x)$ gets close to 2 and stays close to 2, except at the point $x = 1$ itself. And by our intuitive definition, this is all we need to say that we also have $\lim_{x\to 1} g(x) = 2$. (Recall that in our definition, the point $x = a$ is explicitly excluded from consideration when we're deciding if outputs are getting close and staying close to L.) Because we really only care about the points that are close to $x = a$ in general, this same argument means that if there's some interval around $x = a$ on which two functions agree, except maybe at a itself, then the two functions have to have the same limit at a, if the limit exists. For instance, we've already seen that f and g agree everywhere except $x = 1$, so they have the same limit at $x = 1$. Here's another function h, graphed now on top of f:



We see that they agree on an open interval around $x = 1$, and since we only care about what's happening close to $x = 1$, we can ignore the fact that they are different functions—as far as $\lim_{x\to 1}$ is concerned, f and h are exactly the same:

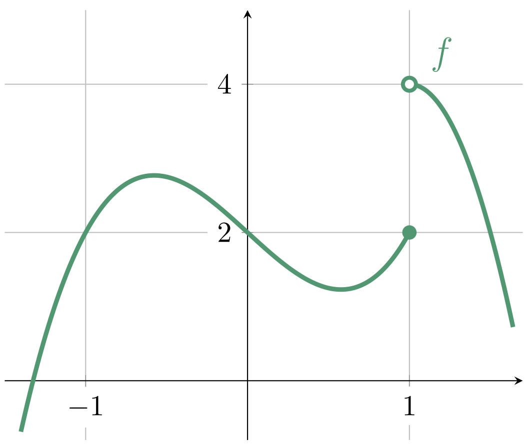
$$\lim_{x\to 1} f(x) = \lim_{x\to 1} g(x) = \lim_{x\to 1} h(x) = 2.$$

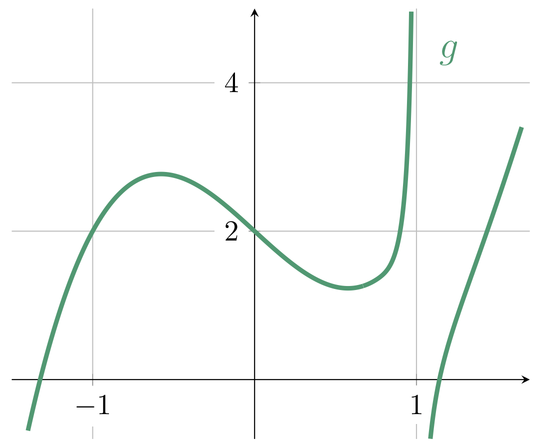
Stating this a little more formally, we get the following definition and fact about limits:

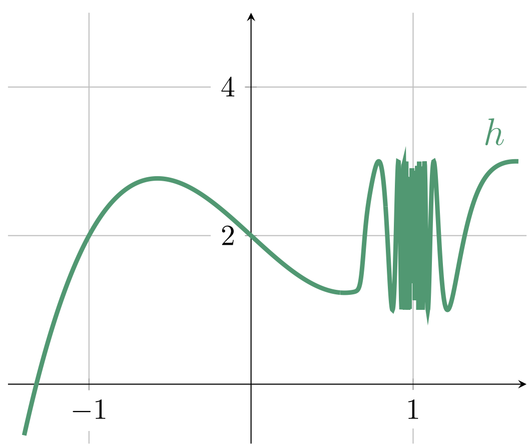
Definition. Two functions f and h are said to be **almost identical near** a if there exists some open interval around $x = a$ such that $f(x) = h(x)$ for all x in the interval, except possibly at $x = a$ itself.

Almost Identical Functions Property. If f and h are almost identical near a, then whenever the limit exists, we have that $\lim_{x \to a} f(x) = \lim_{x \to a} h(x)$.

This property will be key to several algebraic tricks we'll discuss later on, but before we do this, let's look at some of the ways limits can fail to exist. The pictures below show three different ways this can happen:







The function f is very close to having a limit at $x = 1$. If we approach 1 from the left, as x gets closer to 1, $f(x)$ looks to be getting closer and staying close to $L = 2$. However, if we approach from the right, x getting closer to 1 means $f(x)$ getting closer and staying close to $L = 4$. Our definition requires that there is a single number L that *all* the values $f(x)$ approach, whereas we have two numbers that are approached depending on the direction of approach. We must therefore conclude that the limit $\lim_{x\to 1} f(x)$ does not exist. However, we can still say that the **one-sided limits** exist, and we have the notation

$$\lim_{x\to 1^-} f(x) = 2 \qquad \text{and} \qquad \lim_{x\to 1^+} f(x) = 4,$$

where the minus sign superscript next to the 1 means approaching from the left and the plus sign superscript means approaching from the right. The (two-sided) limit does not exist because even though the one-sided limits exist, they are not the same.

The function g exhibits similar behavior in that the one-sided limits do not agree, but the situation here is even worse, because the one-sided limits aren't even real numbers. Approaching from the left, $g(x)$ looks to be getting arbitrarily large and positive, and so we would say that $\lim_{x\to 1^-} g(x) = \infty$, and similarly from the right we have $\lim_{x\to 1^+} g(x) = -\infty$ since the values $g(x)$ are getting arbitrarily large in the negative direction. We'll discuss this sort of example more later, but one thing to note here is that because ∞ and $-\infty$ aren't real numbers, it wouldn't be incorrect to say that these one-sided limits don't exist. This is true, but it doesn't tell the full story—we can be more specific about how the limits fail to exist, and so any time we can say what sort of infinite limit we're dealing with, we should do that. Since g is approaching both infinities from different directions, the most specific answer for what the two-sided limit is should be $\lim_{x\to 1} g(x) = \pm\infty$, rather than simply saying it doesn't exist.

For h, the best possible answer really is that the limit just doesn't exist. This is a case where we see why it matters that part of our intuitive definition requires the function to *stay* close to a particular value. As x approaches 1, the outputs $h(x)$ begin to oscillate wildly, repeatedly getting very close to all sorts of different values between about 1 and 3, but it doesn't stay close to any of them. For instance, it's true that as x goes to 1, $h(x)$ gets close to $L = 2$ infinitely often, but because it never stays close, we cannot say that the limit is 2. Instead, $\lim_{x\to 1} h(x)$ does not exist.

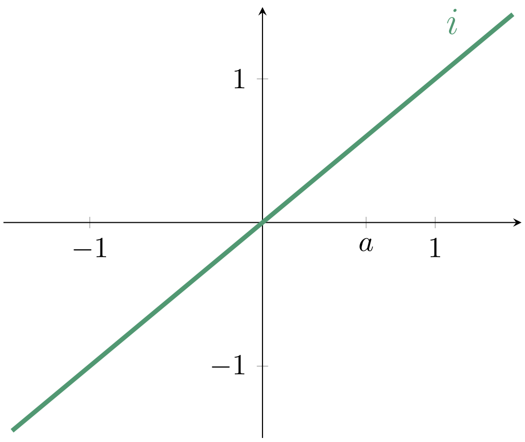
1.3 Limits laws and continuity

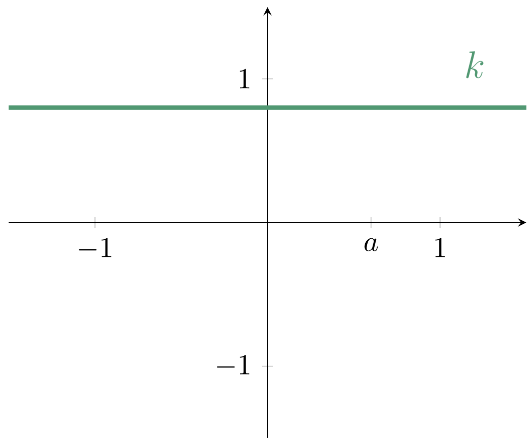
So far the only tool we have for computing limits beside looking at the graph of a function is the Almost Identical Functions (AIF) property. If we have a graph to look at, this is usually fine—our definition of a limit is essentially based on visual intuition about "closeness"—but in this course, the more common situation will be that we are given a rule or equation defining a function and asked to find some limit directly from the equation. We therefore introduce the following *limit laws*, rules for breaking a function down into smaller pieces whose limits we know. For instance, given a complicated limit like

$$\lim_{x\to\pi} \frac{x^2(x-\pi)+7x}{\sqrt{7x/\pi}},$$

we'd like some way of splitting this into more manageable pieces.

We begin by considering some of the smallest, most manageable pieces: the function $i(x) = x$ and the function $k(x) = c$ where c is some real number. The function i is called the identity function, and the function k is called a constant function. These functions are shown below, and have easy limits to understand:





Looking at $\lim_{x\to a} i(x)$, we see that as x gets close to a, the outputs $i(x)$ are by definition just the values x, so these are also getting close and staying close to a. Therefore the limit is $L = a$. Looking at $\lim_{x\to a} k(x)$, no matter where x is close to, the outputs $k(x)$ are always exactly the number c, whatever it is, so in particular as x gets close to a, $k(x)$ gets close and stays close to $L = c$. This gives us the first two limit laws:

$$\lim_{x\to a} x = a \qquad \text{and} \qquad \lim_{x\to a} c = c.$$

For the rest of the limit laws, let f and g be functions defined near the point $x = a$ (we can think there's some open interval around a on which the functions are defined) such that $\lim_{x\to a} f(x) = L$ and $\lim_{x\to a} g(x) = M$. We'll just state the rest of the laws without proof or much analysis—they should mostly seem quite natural. That said, you should think a little bit about *why* these are things we expect to be true (e.g. for the first one here, the claim is that if we add $f(x)+g(x)$, we're adding something close to L to something close to M, so we should get something close to $L+M$). Here's the full list:

Proposition. *For functions f and g defined near a with $\lim_{x\to a} f(x) = L$ and $\lim_{x\to a} g(x) = M$ for real numbers L and M, we have the following:*

- ***Sum law:*** $\lim_{x\to a} f(x) + g(x) = L + M$
- ***Difference law:*** $\lim_{x\to a} f(x) - g(x) = L - M$
- ***Constant multiple law:*** *for a real number c,* $\lim_{x\to a} cf(x) = cL$

- ***Product law:*** $\lim_{x\to a} f(x)g(x) = LM$
- ***Quotient law:*** *if* $M \neq 0$*, then* $\lim_{x\to a} f(x)/g(x) = L/M$
- ***Power law:*** *for all positive integers* n*,* $\lim_{x\to a}(f(x))^n = L^n$
- ***Root law:*** *for odd positive integers* n*,* $\lim_{x\to a} \sqrt[n]{f(x)} = \sqrt[n]{L}$*, and also for even* n *when* $L \geq 0$
- ***Comparison law:*** *if* $f(x) \leq g(x)$ *for all* x *near* a *(except possibly at* a *itself), then* $L \leq M$*.*

Let's now use these laws to compute the more complicated limit from earlier,

$$L = \lim_{x\to\pi} \frac{x^2(x-\pi)+7x}{\sqrt{7x/\pi}}.$$

First, the quotient law tells us this is the same as

$$L = \frac{\lim_{x\to\pi}(x^2(x-\pi)+7x)}{\lim_{x\to\pi}\sqrt{7x/\pi}}.$$

To use this, we're assuming that the denominator has a nonzero limit, and this is justified by the comparison law: we know that for x near π, $7x/\pi$ is near 7, which is greater than 1, so the comparison law says the limit must be greater than or equal to 1. Then $\lim_{x\to\pi} 7x/\pi \geq 1 > 0$. Since this limit is positive, the root law says that the limit of the root is the root of the limit, and so the limit of the denominator is therefore also positive, and so cannot be equal to zero, justifying our use of the quotient law. Writing out what we've just said, we now have that

$$L = \frac{\lim_{x\to\pi}(x^2(x-\pi)+7x)}{\sqrt{\lim_{x\to\pi} 7x/\pi}}.$$

Continuing to work on the denominator, the limit inside the square root is the limit of the function $7x/\pi = (7/\pi)x$, a constant multiple of the identity function. Using the constant multiple law and our rule about the limit of the identity function, we get that

$$L = \frac{\lim_{x\to\pi}(x^2(x-\pi)+7x)}{\sqrt{\lim_{x\to\pi}(7/\pi)x}} = \frac{\lim_{x\to\pi}(x^2(x-\pi)+7x)}{\sqrt{(7/\pi)\lim_{x\to\pi} x}} = \frac{\lim_{x\to\pi}(x^2(x-\pi)+7x)}{\sqrt{(7/\pi)\cdot\pi}}.$$

Inside the square root, we are both dividing and multiplying by π, so this cancels out and we get

$$L = \frac{\lim_{x\to\pi}(x^2(x-\pi)+7x)}{\sqrt{7}}.$$

Let's look now at the numerator. Since the numerator is the limit of a sum of two functions, the sum law tells us this is the same as the sum of the limits. Then we can use the product rule on the first term and the scalar multiple law on the second term to get

$$L = \frac{\lim_{x\to\pi} x^2(x-\pi) + \lim_{x\to\pi} 7x}{\sqrt{7}} = \frac{\lim_{x\to\pi} x^2 \cdot \lim_{x\to\pi}(x-\pi) + 7\lim_{x\to\pi} x}{\sqrt{7}}.$$

Using the power law on the first term and the difference law on the second gives us

$$\frac{\left(\lim_{x\to\pi} x\right)^2 \left(\lim_{x\to\pi} x - \lim_{x\to\pi} \pi\right) + 7\lim_{x\to\pi} x}{\sqrt{7}}.$$

At this point, the only limits left to calculate are the two easy ones that became our very first limit laws: the identity law and the constant law. Using these, we get

$$L = \frac{(\pi)^2(\pi-\pi)+7\pi}{\sqrt{7}} = \frac{\pi^2\cdot 0 + 7\pi}{\sqrt{7}} = \frac{7\pi}{\sqrt{7}}.$$

Recalling some exponent rules, this is

$$L = \frac{7}{\sqrt{7}}\cdot\pi = \frac{7^1}{7^{\frac{1}{2}}}\cdot\pi = 7^{1-\frac{1}{2}}\cdot\pi = 7^{\frac{1}{2}}\cdot\pi = \sqrt{7}\pi.$$

As is readily apparent, actually writing out and explaining all the various limit laws we use is sort of a pain. **You will not have to do this unless explicitly prompted.** In fact, for this particular limit, there's an easier property to appeal to. You may have noticed that $\sqrt{7}\pi$ is exactly what we get if we plug in the number we approach (π) into our original function:

$$\frac{x^2(x-\pi)+7x}{\sqrt{7x/\pi}} \text{ at } x=\pi: \qquad \frac{(\pi)^2((\pi)-\pi)+7(\pi)}{\sqrt{7(\pi)/\pi}} = \frac{7\pi}{\sqrt{7}} = \sqrt{7}\pi$$

This reflects the fact that this function is *continuous* at $x=\pi$. Informally, we can think of a function f being continuous if we can draw its graph without picking up our pencil. What this means is that every value of the function is equal to the value you'd expect based on the values of the points nearby—you don't need to pick up the pencil at the value $f(a)$ for some a if $f(a)$ is already really close to all the values $f(x)$ you've just drawn. To make this formal, we just need to complete this translation into the language of limits:

Definition. A function f is said to be **continuous at the point** a if all of the following hold:

(i) f is defined at a, that is, a is in the domain of f;
(ii) $\lim_{x\to a} f(x)$ exists;
(iii) $\lim_{x\to a} f(x) = f(a)$.

Otherwise, we say f is **discontinuous at** a. A function f is said to be **continuous** if it is continuous at every point in its domain.

The calculation we just did above showed that the function

$$\frac{x^2(x-\pi)+7x}{\sqrt{7x/\pi}}$$

is defined at π, has a limit as x approaches π, and that the limit is equal to the value the function takes at π. That is, we have just checked all the conditions for the function to be continuous at π. This was a somewhat laborious process, and as we've already seen, it would have been a lot easier to calculate the limit by just plugging in π to the function. Had we known in advance that this function was continuous at π, we could have been completely justified in doing exactly this. In general, if we have some function f and we happen to know in advance that f is continuous at some point a, then by definition we know that the limit as x approaches a exists and that it's equal to $f(a)$.

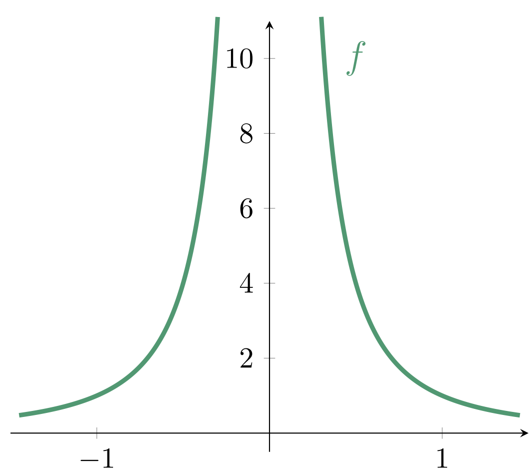
In order to make the most use out of this, we'd like to know which kinds of functions are continuous. When we have a function we know is continuous, we can compute its limits very easily. Fortunately, a lot of the functions we're used to working with are, in fact, continuous!

Fact. Any function built out of algebra and trig is continuous (i.e. continuous at any point where it's defined).

Here are some examples of such functions:

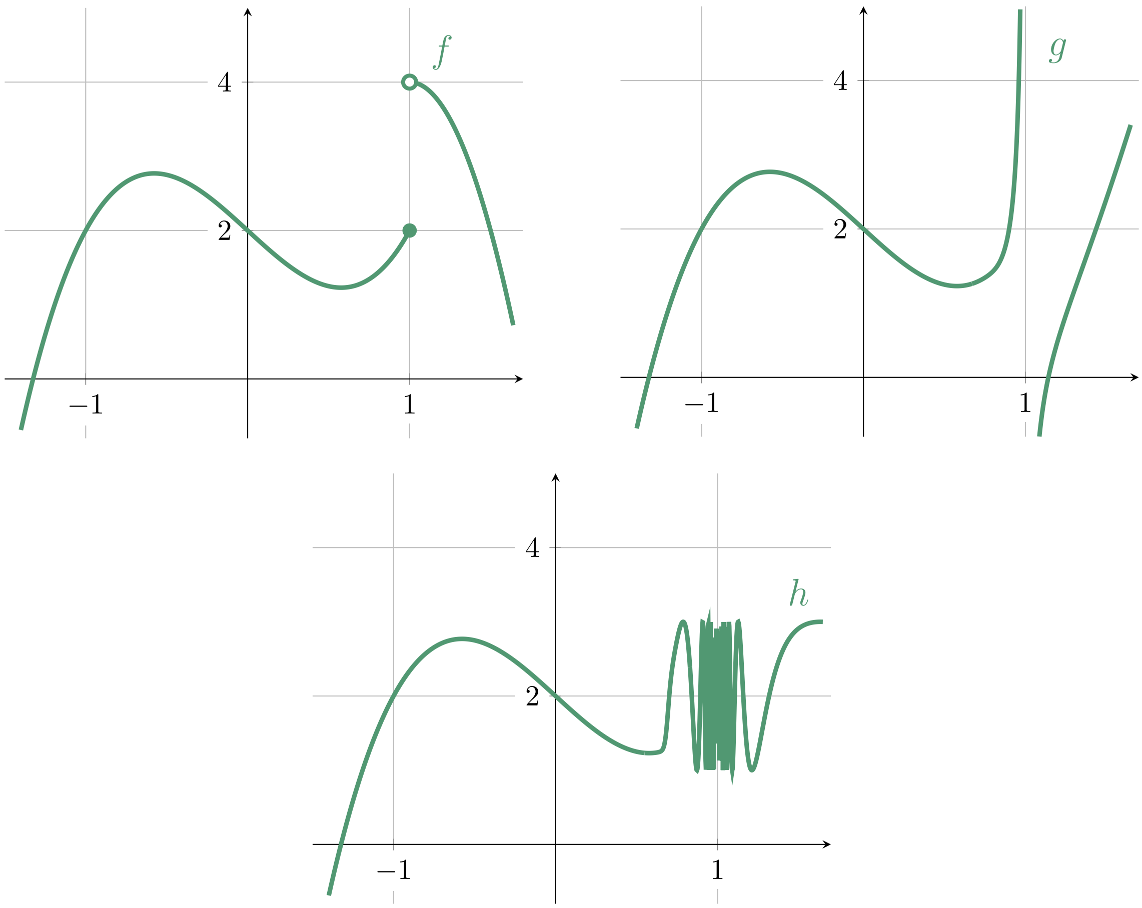
- All trigonometric functions
- All polynomial and rational functions
- Root and power functions
- Any function that is the sum, difference, product, quotient, or composition of these kinds of functions—this is what is meant by "built out of algebra and trig"

More concretely, the functions $\sqrt{\sin(x)}$, $\frac{x^2-1}{x^2+1}$, and $\frac{1}{x^2}$ are all continuous where they are defined. Looking at a graph of $f(x) = 1/x^2$, we have the following:

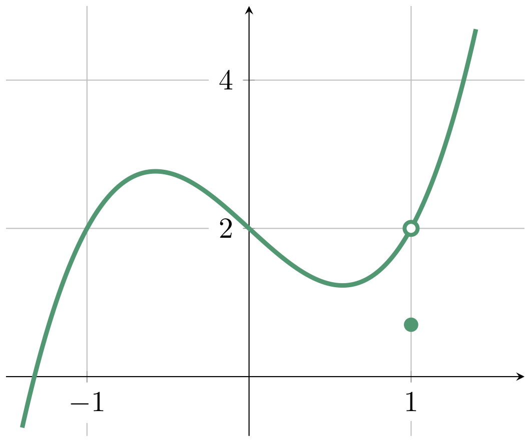


We can clearly see there's a discontinuity here, even though we just said the function is continuous. This is where it's important to recall that being continuous only means being continuous at every point in the domain. The problem with this function then is that the point $x = 0$ is not in the domain (why?), and this allows the function to have a discontinuity there even though it still qualifies as being "built out of algebra and trig." Hence it's accurate to say both that this function is continuous (on its domain) and that it has a discontinuity at $x = 0$.

Since the problem here is that the function suddenly shoots off to infinity near $x = 0$, this kind of discontinuity is called an **infinite discontinuity**. The three functions



we looked at earlier also exhibit discontinuities, this time at $x = 1$. The function f has what is called a **jump discontinuity** (which happens when both one-sided limits exist but aren't equal), and g again has an infinite discontinuity. The function h fails to have one-sided limits existing since again it's not getting close and staying close to any one real number, and we call this behavior an **essential discontinuity**. (Since ∞ and $-\infty$ also aren't real numbers, all infinite discontinuities are also technically essential discontinuities—the one-sided limits aren't real numbers.) One more kind of discontinuity we've seen is called a **removable discontinuity**, which is when the limit does exist but isn't equal to the value of the function. Here's that example again:



We said the limit of this function as $x \to 1$ was 2, but we see that this limit isn't the same as the value the function actually takes at $x = 1$, so it's not continuous there.

Functions with non-infinite essential discontinuities can look especially crazy. Here are a couple examples you might want to think about to convince yourself that not all functions look so nice:

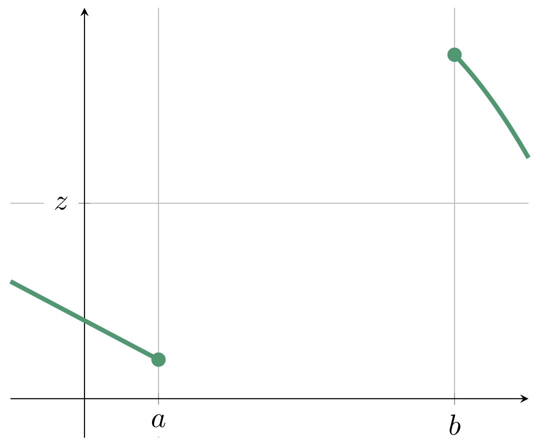
$$d(x) = \begin{cases} 1 & x \text{ is rational} \\ 0 & x \text{ is irrational} \end{cases} \qquad t(x) = \begin{cases} 1/q & x = p/q \text{ is rational in lowest terms} \\ 0 & x \text{ is irrational} \end{cases}$$

Again, these really are both functions! They have a clearly defined rule for how to turn input into a single output. For instance, $d(1/2) = 1$ because $1/2$ is a rational number, and $t(4/6) = t(2/3) = 1/3$, while $d(\pi) = t(\pi) = 0$ because π is irrational. It turns out that d is discontinuous *at every single point*, and t is discontinuous at every rational number, but continuous at every irrational number. We won't worry too much about functions like this for this course, but they're still out there, and the truth is that our intuitive definition of limits doesn't help us much when it comes to determining where these functions are continuous and discontinuous—there is a good reason to want a very formal version of our intuitive definition, and if any of this has piqued your interest, you can find a definition in OpenStax 2.5 that actually can help us see why these functions are discontinuous where they are (and then by all means come talk to me about it in office hours)! But to reiterate, there's no requirement that you look at the formal definition or spend any time with these kinds of functions.

Now that we've introduced continuity, we have a limit law that generalizes the power and root laws. The power and root laws both involve the composition of two functions, and tell us that the limit of a power or root is the power or root of the limit. What this is using really is that powers and roots are continuous functions, and so it turns out that the same holds whenever the outside function of a composition is continuous.

Composition Limit Law: If f is defined near a with limit $\lim_{x\to a} f(x) = L$ and g is continuous at L, then $\lim_{x \to a} g(f(x)) = g(L)$.

Another nice property of continuous functions is that the Intermediate Value Theorem (IVT) holds. This says that if we have a continuous function f and some values $f(a) \neq f(b)$, f has to take on every value between $f(a)$ and $f(b)$ at some point. If we have the following setup



with $f(a) < z < f(b)$, there's no way for us to fill in the missing part of the graph without crossing this line $y = z$ if I require the function I draw to be continuous. Try it! What you'll find is that you have to cross the line, and this means there's at least one point, call it c, between a and b such that $f(c)$ is exactly z. Formally, we have the following:

Intermediate Value Theorem: Let f be continuous on the closed interval $[a, b]$. If z is any real number between the numbers $f(a)$ and $f(b)$, then there exists some point c with $a \leq c \leq b$ such that $f(c) = z$.

Previously we've probably learned a lot of ways to solve certain polynomial equations algebraically, for instance with the quadratic formula. But some polynomials don't have roots that can be written down in a nice algebraic way (e.g. it's a theorem that there's no "quintic formula" for finding the zeros of a degree five polynomial using just the usual algebraic operations and roots). For example, $f(x) = x^5 + x^3 - 1$ is unsolvable (in this algebraic sense), but we can use the IVT to show it's got at least one real root. We have

$$f(0) = (0)^5 + (0)^3 - 1 = -1$$
$$f(1) = (1)^5 + (1)^3 - 1 = 1,$$

so $f(0) < 0 < f(1)$ implies that f has at least one real root on the interval $[0, 1]$ since f is a polynomial and is therefore continuous.

1.4 Beyond the limit laws—algebraic tricks and trig limits

Now that we have the continuity of algebraic and trigonometric functions established, we're ready to address the main sorts of limits we'll be encountering in the course. We can look at a graph of a function and determine a limit that way, but if the function is presented as an equation, we've only discussed how to compute a limit at a point where the function is continuous. We'd like to have a way of computing more interesting limits, and this is where the Almost Identical Function property helps us tremendously. Consider the following: what is

$$\lim_{x\to 1} \frac{x^2-1}{x-1}?$$

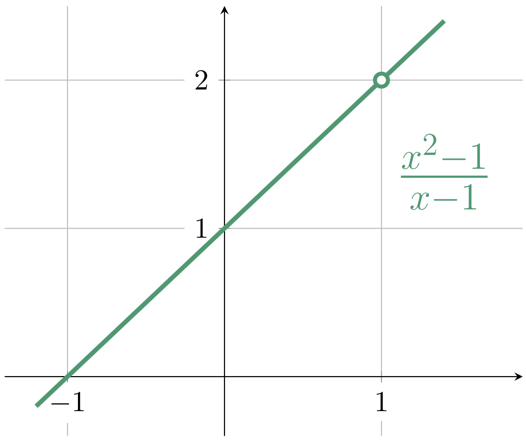
We're taking the limit as x approaches 1, so the first thing we should do is ask if this function is continuous at 1. If it is, we can just plug 1 in and the limit is whatever this comes out to be. However, this is a function that fits into the category "built out of algebra and trig," so we can determine continuity at $x = 1$ by checking if 1 is in the domain of the function. We see that it unfortunately isn't, since the expression is undefined due to division by zero. Hence this isn't continuous at $x = 1$, and so we have to work harder to compute the limit.

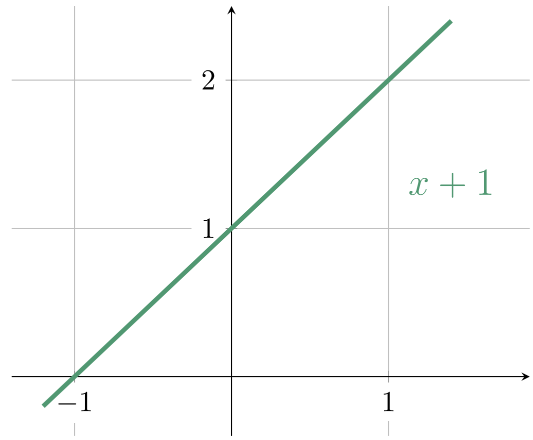
What to try? One thing to note is that in plugging in $x = 1$ to check if the function is defined, not only is the denominator 0, so is the numerator. We have a $0/0$ situation, which is an example of an **indeterminate form**. That is to say, we can't really figure anything out about the limit just knowing what the function looks like right at $x = 1$—a $0/0$ could be anything at all. This is in contrast to something like a function we looked at earlier, $1/x^2$ as $x \to 0$. Here we say we had an infinite limit, and this is because when we try and plug in $x = 0$, only the denominator is zero, we get a $1/0$, which is not an indeterminate form—we always get some kind of infinite limit when we have something nonzero divided by zero (more on this in the next section). When we have a genuine indeterminate form like $0/0$, the trick is usually to do some kind of algebraic manipulation to find a function that is almost identical near where we're taking the limit. In this case, we're looking for a function that agrees with $(x^2-1)/(x-1)$ near 1, with the exception of 1 itself. Because continuous functions are the only functions we can really take limits of directly, our goal will be to find an almost identical function that is continuous.

Something that we should always give a try whenever we have a rational function like this is to factor the numerator and denominator and look for anything that will cancel out. If we do this, we get that

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

If $x \neq 1$, then $x - 1$ is nonzero and we can divide both numerator and denominator to cancel the $(x - 1)$ factor, leaving us with the new function $x + 1$. This really is a different function! The function we started with wasn't defined at $x = 1$, but this new function certainly is. Compare their graphs below:





Even though they are different functions, the Almost Identical Functions property guarantees they have the same limit at $x = 1$, so the full solution to the limit problem would look like the following:

$$\lim_{x\to 1} \frac{x^2-1}{x-1} = \lim_{x\to 1} \frac{(x-1)(x+1)}{x-1} \overset{\text{AIF}}{=} \lim_{x\to 1} x+1 = 1+1 = 2,$$

where we put AIF above the equals sign to remind us that this equality holds because we're using the AIF property. Once we've done this, we're left with the continuous function $x+1$ and so we can evaluate the limit just by evaluating the function at 1.

Here are a few more examples of this type:

$$\lim_{x\to 3} \frac{x^2-4x+3}{x-3} = \lim_{x\to 3} \frac{(x-1)(x-3)}{x-3} \overset{\text{AIF}}{=} \lim_{x\to 3} x-1 = 3-1 = 2.$$
$$\lim_{x\to -2} \frac{x^2+5x+6}{x^2-4x-12} = \lim_{x\to -2} \frac{(x+2)(x+3)}{(x+2)(x-6)} \overset{\text{AIF}}{=} \lim_{x\to -2} \frac{x+3}{x-6} = \frac{-2+3}{-2-6} = -\frac{1}{8}.$$
$$\lim_{x\to 2} \frac{x^3-64}{x-4} = \lim_{x\to 2} \frac{(x-4)(x^2+4x+16)}{x-4} \overset{\text{AIF}}{=} \lim_{x\to 4} x^2+4x+16 = (4)^2+4(4)+16 = 48.$$

Sometimes when we encounter a $0/0$ situation in computing a limit, there is no way to factor and cancel to find an almost identical function, and we have to employ some clever tricks. For instance, if we were to try and plug in a zero to the numerator and denominator of the limit below, we'll get a $0/0$ with no way to factor:

$$\lim_{x\to 0} \frac{\sqrt{x+9}-3}{x}.$$

The tool we need here is to take advantage of the way a difference of squares factors: $a^2 - b^2 = (a-b)(a+b)$. Our numerator has the form $a - b$, so we multiply by what's called the *conjugate*,

$a+b$, on both numerator and denominator:

$$\lim_{x\to 0}\frac{\sqrt{x+9}-3}{x}=\lim_{x\to 0}\frac{\sqrt{x+9}-3}{x}\cdot\frac{\sqrt{x+9}+3}{\sqrt{x+9}+3}.$$

The point is that the numerator now has the form $(a-b)(a+b)$, which we know is the same as a^2-b^2 (double check this by FOILing it out). Our limit is then

$$\lim_{x\to 0}\frac{\sqrt{x+9}-3}{x}=\lim_{x\to 0}\frac{\sqrt{x+9}-3}{x}\cdot\frac{\sqrt{x+9}+3}{\sqrt{x+9}+3}=\lim_{x\to 0}\frac{(\sqrt{x+9})^2-(3)^2}{x(\sqrt{x+9}+3)}.$$

Simplifying the numerator, we have $(x+9)-(9)=x$. Hence

$$\lim_{x\to 0}\frac{\sqrt{x+9}-3}{x}=\lim_{x\to 0}\frac{x}{x(\sqrt{x+9}+3)}\overset{\text{AIF}}{=}\lim_{x\to 0}\frac{1}{\sqrt{x+9}+3}$$

where we notice for $x\neq 0$ that we can cancel the x in the numerator and denominator, producing a function that is almost identical to the original. This new function $1/(\sqrt{x+9}+3)$ is built out of algebra and trig, so it's continuous where it's defined. Since it's defined at $x=0$ (having cancelled the x, we're no longer dividing by zero), it's continuous there, and so we can now finish off the computation by plugging zero into the new function. Putting all of this work together, our work looks like this:

$$\begin{aligned}\lim_{x\to 0}\frac{\sqrt{x+9}-3}{x}&=\lim_{x\to 0}\frac{\sqrt{x+9}-3}{x}\cdot\frac{\sqrt{x+9}+3}{\sqrt{x+9}+3}=\lim_{x\to 0}\frac{(\sqrt{x+9})^2-(3)^2}{x(\sqrt{x+9}+3)}\\&=\lim_{x\to 0}\frac{x}{x(\sqrt{x+9}+3)}\overset{\text{AIF}}{=}\lim_{x\to 0}\frac{1}{\sqrt{x+9}+3}=\frac{1}{\sqrt{0+9}+3}=\frac{1}{3+3}=\frac{1}{6}.\end{aligned}$$

Even though we're multiplying *and dividing* by the conjugate to avoid completely changing the function, this trick is usually just referred to as "multiplying by the conjugate." It comes up frequently, so here are a couple more examples of the trick in action:

$$\begin{aligned}\lim_{x\to 6}\frac{2-\sqrt{x-2}}{x-6}&=\lim_{x\to 6}\frac{2-\sqrt{x-2}}{x-6}\cdot\frac{2+\sqrt{x-2}}{2+\sqrt{x-2}}=\lim_{x\to 6}\frac{2^2-(x-2)}{(x-6)(2+\sqrt{x-2})}\\&=\lim_{x\to 6}\frac{6-x}{(x-6)(2+\sqrt{x-2}}\overset{\text{AIF}}{=}\lim_{x\to 6}\frac{-1}{2+\sqrt{x-2}}=\frac{-1}{\sqrt{6-2}+2}=-\frac{1}{4}.\\\lim_{x\to 14}\frac{x-14}{\sqrt{x-5}-3}&=\lim_{x\to 14}\frac{x-14}{\sqrt{x-5}-3}\cdot\frac{\sqrt{x-5}+3}{\sqrt{x-5}+3}=\lim_{x\to 14}\frac{(x-14)(\sqrt{x-5}+3)}{(x-5)-3^2}\\&=\lim_{x\to 14}\frac{(x-14)(\sqrt{x-5}+3)}{x-14}\overset{\text{AIF}}{=}\lim_{x\to 14}\sqrt{x-5}+3=\sqrt{14-5}+3=6.\end{aligned}$$

Note that in the first example here, we've used that $6-x=-(x-6)$ and then cancelled to get an almost identical function.

Here's another limit that results in a 0/0 indeterminate form:

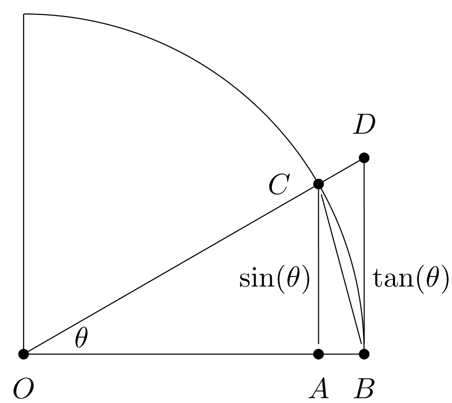
$$\lim_{x\to 0} \frac{\sin(x)}{x}.$$

This is a tough limit to evaluate since there's no factoring we could do and no conjugates we could multiply by. Instead, we'll need to use some geometry. The following proposition gives what we'll refer to as the **small angle approximation**. You won't need to recreate the proof, but you will need to know how to use the result.

Proposition. *When θ is very small,* $\sin(\theta)$ *is approximately θ. Formally,*

$$\lim_{\theta\to 0} \frac{\sin(\theta)}{\theta} = 1.$$

Proof. Consider the following portion of the unit circle:



Recall that since we're working with radians, θ is the measure of the angle, but also the length of the curved arc between points B and C. The full circle has length (i.e. circumference) 2π, so the wedge BOC must have area $\theta/2\pi$ of the area of the full circle. The unit circle has radius 1, hence area π, so the wedge BOC has area

$$\frac{\theta}{2\pi} \cdot \pi = \frac{\theta}{2}.$$

This wedge completely contains the triangle BOC, so the area of the wedge must be greater than the area of the triangle. BOC has the line segment BO as its base, and the line segment AC as its height. BO has length 1 because it is a radius of the unit circle, and AC is the side opposite to the angle θ and therefore has length $\sin(\theta)$. Hence

$$\frac{\theta}{2} = \text{ area of wedge } \geq \text{ area of triangle } = \frac{1}{2} \cdot \text{ base } \cdot \text{height} = \frac{\sin(\theta)}{2}.$$

Multiplying through by 2, we get that $\theta \geq \sin(\theta)$, meaning $\frac{\sin(\theta)}{\theta} \leq 1$. By the limit comparison law, we have that

$$\lim_{\theta \to 0} \frac{\sin(\theta)}{\theta} \le \lim_{\theta \to 0} 1 = 1.$$

Of course, our goal is to show that the limit is *equal* to 1, not just less than or equal to 1, so our next step is to show that the limit is also greater than or equal to 1. This will rule out the possibility that the limit is strictly less than 1, allowing us to conclude the limit is exactly 1. To do this, we'll need to use the fact that the segment BD has length $\tan(\theta)$. This is because the triangle BOD is similar to triangle BOC (both triangles share the same angle θ at the tip and then have right angles next to it), so the ratio of their side lengths are the same. In particular, this means

$$DB = \frac{DB}{1} = \frac{DB}{OB} = \frac{CA}{AC} = \frac{\text{opp.}}{\text{adj.}} = \tan(\theta).$$

Now to get our inequality, we just need to notice that the wedge BOC, which we said has area $\theta/2$, is completely contained within triangle BOD, meaning

$$\frac{\theta}{2} \le \text{area of } BOD = \frac{1}{2} \cdot \text{base} \cdot \text{height} = \frac{\tan(\theta)}{2}.$$

Again multiplying by 2 we get that $\theta \le \tan(\theta)$. But $\tan(\theta)$ is the same as $\sin(\theta)/\cos(\theta)$, so we have the inequality

$$\theta \le \frac{\sin(\theta)}{\cos(\theta)}.$$

If we then divide both sides by θ and multiply both sides by $\cos(\theta)$, we get that

$$\cos(\theta) \le \frac{\sin(\theta)}{\theta}.$$

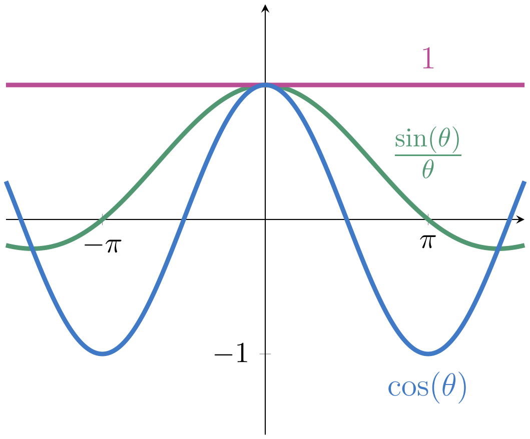
Therefore, again using the comparison law for limits,

$$\lim_{\theta \to 0} \frac{\sin(\theta)}{\theta} \ge \lim_{\theta \to 0} \cos(\theta) = \cos(0) = 1$$

where we use the fact that cosine is continuous, and so in particular continuous at zero. Since we have shown that the limit is less than or equal to 1 and *also* greater than or equal to 1, the only possibility is that the limit is exactly 1. $\square$

In the proof here, we're really only showing that $\lim_{\theta \to 0^+} \sin(\theta)/\theta = 1$ since we've only considered angles θ in the first quadrant ($\theta > 0$), but flipping our picture upside down and giving the same argument works just as well and gives the other one-sided limit.

The proof isn't something you need to memorize, but there's one aspect of it that's worth knowing about. We wanted to show a limit was equal to a certain number (1), and so to do this, we found a function ($\cos(\theta)$) always less than the function we were interested. and a function (1) always greater than the function we were interested in. Because the smaller function and the bigger function both had the same limit, we could conclude by the comparison law that the middle function $(\sin(\theta)/\theta)$ also had the same limit. Here's the picture of all these different functions:



That this approach to proving a limit works is called the **Squeeze Theorem** for the way that we find two functions (in pink and blue) whose limit we know and that squeeze the function we're actually interested in (green). Here's the full statement:

Theorem. *Let f, g, and h be functions defined over some open interval containing $x = a$, except possibly at a itself. If $f(x) \leq g(x) \leq h(x)$ for all $x \neq a$ in this interval and there is some real number L such that*

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

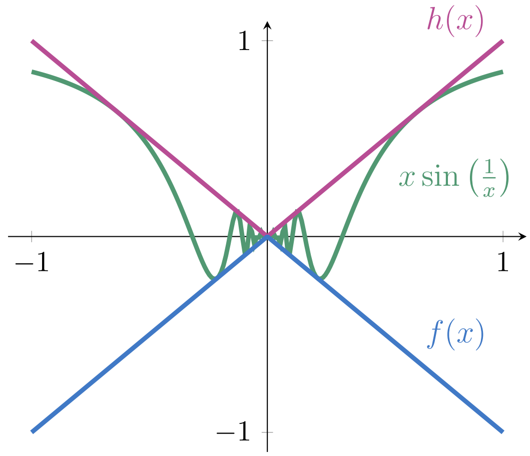
then $\lim_{x \to a} g(x) = L$ also.

As another example of the usefullness of this technique, we said earlier that the function $\sin(1/x)$ has an essential discontinuity at $x = 0$. As x gets close to 0, $\sin(1/x)$ oscillates wildly and gets close to every value between -1 and 1. If we multiply by x to get $g(x) = x \sin(1/x)$, the function still oscillates wildly, but now we can use the Squeeze Theorem to show that it does actually have a limit that exists as x approaches 0.

No matter what the input is to the sine function, the output is always between -1 and 1. Therefore if we multiply the inequality $-1 \leq \sin(1/x) \leq 1$ by x, we get that $-x \leq x \sin(1/x) \leq x$ when $x \geq 0$, and since multiplying by a negative reverses inequalities, also that $x \leq x \sin(1/x) \leq -x$ when $x \leq 0$. The functions we need to bound $g(x) = x \sin(1/x)$ are then the piecewise functions

$$f(x) = \begin{cases} -x & x \geq 0 \\ x & x \leq 0 \end{cases} \qquad \leq \qquad g(x) \qquad \leq \qquad h(x) = \begin{cases} x & x \geq 0 \\ -x & x \leq 0. \end{cases}$$

Here's our picture of the situation:



Hopefully from the picture we recognize $h(x)$ as the absolute value function and $f(x)$ as its negative. They may not be smooth with the sharp point at $x = 0$, but they're still continuous there—we can find the one-sided limits from their piecewise description and see that they are all equal to 0. This means the limit at 0 of f and h is 0, and so the Squeeze Theorem then says that

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

So even though both functions oscillate infinitely often as x approaches 0, only $\sin(1/x)$ has an essential discontinuity. For $x\sin(1/x)$, the limit does exist, the only problem is that the function isn't defined there—we say $x\sin(1/x)$ has a removable discontinuity at $x = 0$. The discontinuity is removable in the sense that we can define a new function that is almost identical but doesn't have a discontinuity:

$$\begin{cases} x\sin\left(\frac{1}{x}\right) & x \neq 0 \\ 0 & x = 0 \end{cases}$$

is a continuous function almost identical to the one we started with.

Not only is the Squeeze Theorem a powerful tool, so is the limit we initially used it to compute, the Small Angle Approximation (SAA):

$$\lim_{\theta\to 0} \frac{\sin(\theta)}{\theta} = 1.$$

In this limit, θ is just a number approaching zero, but we can take θ also to be some function $f(x)$ that approaches zero by appealing to our limit composition law. The function $\sin(\theta)/\theta$ has a removable discontinuity at 0 since the limit does exist, so if we patch that hole, we get the function

$$g(\theta) = \begin{cases} \frac{\sin(\theta)}{\theta} & \theta \neq 0 \\ 1 & \theta = 0, \end{cases}$$

which is continuous. Since this is continuous at $\theta = 0$, we can use the limit composition law to say that if f is any function with $\lim_{x\to a} f(x) = 0$, then

$$\lim_{x\to a} \frac{\sin(f(x))}{f(x)} = 1.$$

For instance, we showed above that

$$\lim_{x\to 0} x\sin\left(\frac{1}{x}\right) = 0, \quad \text{meaning by SAA} \quad \lim_{x\to 0} \frac{\sin(x\sin(1/x))}{x\sin(1/x)} = 1.$$

The point is that the Small Angle Approximation allows us to calculate more complicated limits involving trig functions. Frequently the way we'll do this is by first doing some algebraic manipulation to introduce something that looks like $\sin(f(x))/f(x)$ for some function f that goes to 0 as x approaches whatever point we're interested in. For example, if we want to compute a limit like

$$\lim_{x\to 0} \frac{\sin(2x)}{\sin(3x)},$$

we should expect the answer to be $2/3$ since $\sin(2x)$ is roughly $2x$ and $\sin(3x)$ is roughly $3x$ when x is very small. To actually show this is the case and use the SAA formally, we'll need to do a little algebraic manipulation:

$$\lim_{x\to 0} \frac{\sin(2x)}{\sin(3x)} = \lim_{x\to 0} \frac{\frac{\sin(2x)}{2x}\cdot 2x}{\frac{\sin(3x)}{3x}\cdot 3x} = \frac{\lim_{x\to 0}\frac{\sin(2x)}{2x}}{\lim_{x\to 0}\frac{\sin(3x)}{3x}} \cdot \lim_{x\to 0} \frac{2x}{3x} = \frac{1}{1}\cdot \lim_{x\to 0} \frac{2x}{3x} \overset{\text{AIF}}{=} \lim_{x\to 0} \frac{2}{3} = \frac{2}{3}.$$

Another important trig limit we can derive from the Small Angle Approximation concerns the cosine function:

$$\begin{aligned}
\lim_{\theta\to 0} \frac{1-\cos(\theta)}{\theta} &= \lim_{\theta\to 0} \frac{1-\cos(\theta)}{\theta}\cdot\frac{1+\cos(\theta)}{1+\cos(\theta)} = \lim_{\theta\to 0} \frac{1-\cos^2(\theta)}{\theta(1+\cos(\theta))} \\
&= \lim_{\theta\to 0} \frac{\sin^2(\theta)}{\theta(1+\cos(\theta))} = \lim_{\theta\to 0} \frac{\sin(\theta)}{\theta}\cdot\lim_{\theta\to 0}\frac{\sin(\theta)}{1+\cos(\theta)} \\
&= \lim_{\theta\to 0} \frac{\sin(\theta)}{1+\cos(\theta)} = \frac{\sin(0)}{1+\cos(0)} = \frac{0}{1+1} = 0.
\end{aligned}$$

This says that $\cos(\theta)$ and 1 are extremely close to one another near zero—even closer than θ itself is to zero when we let θ approach zero.

1.5 Infinite limits and limits at infinity

We've already seen that some functions have infinite discontinuities. For example, $f(x) = 1/x$ has an infinite discontinuity at $x = 0$. This means that if we take a limit as x approaches 0, we get some kind of infinity. Previously, we were looking at a graph to see this, but we can just as well find the limits from the equation $1/x$ itself. This is the function that takes the reciprocal of a number, and the key things we need to remember are the following: a) the reciprocal preserves the sign of a number, and b) the reciprocal of something large (i.e. something greater than 1) is small (something less than 1). This helps us think about what the one-sided limits should be. If we let x approach zero from the right, all of our x values are positive. This means all of their reciprocals $1/x$ are also positive. Moreover, the values x are getting arbitrarily small since we're letting x approach zero, the smallest number (in absolute value). This means the reciprocals are getting arbitrarily large, and we just said they all still have to be positive, so what we've argued here is that

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

A similar argument gives us the other one-sided limit: $x \to 0^-$ means our inputs are getting arbitrarily small, but are now negative, meaning our outputs are getting arbitrarily large and negative:

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

Putting these together, it would be correct to say that the limit does not exist, but the best answer here is the following:

$$\lim_{x \to 0} \frac{1}{x} = \pm\infty,$$

which says that as x approaches 0, the function $1/x$ approaches both kinds of infinities, positive and negative, depending on the direction of approach.

If we look instead at the function $1/x^2$, we have a slightly different situation. This time, whether we approach from right or left doesn't matter. If x is approaching 0, it's some number getting increasingly small, so x^2 is also getting increasingly small, and $1/x^2$ is therefore getting increasingly large. As for the sign, if x is a little bit less than 0 or a little bit greater than 0, when we square it, we always have $x^2 > 0$, so $1/x^2$ is also *always positive*. This argument tells us that

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

If we were to plug in $x = 0$ to these functions directly, in both cases we would get the expression $1/0$, which we know doesn't make sense. We saw earlier that when we had an expression $0/0$, the answer to the limit could conceivably be any number, so we called this an *indeterminate form*. Limits of the form $1/0$ are **not indeterminate forms**. These are the prototypical examples of infinite limits, because whenever we try to plug in the number we're approaching to a function and get something like $1/0$ or $c/0$ for some real number $c \neq 0$, the answer will always be some kind of infinity. More formally, we have the following proposition:

Proposition. *Let f and g be functions such that $\lim_{x\to a} f(x) = c \neq 0$ and $\lim_{x\to a} g(x) = 0$. Then*

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \pm\infty.$$

If $c > 0$ and then the limit is ∞ if and only if $g(x) \geq 0$ near a and is $-\infty$ if and only if $g(x) \leq 0$ near a. If $c < 0$, then the opposite is true.

Here are a few examples of how to use this proposition. The notation we'll be using to show our work needs some explanation. Always with limits, we want to start by plugging in the number we're approaching (so in the first example below, we want to try plugging in 3). Either the function is defined and continuous at this point (so we can stop), or we'll get something undefined, often due to division by zero. In this latter case, the proposition above says that when we are dealing with division by zero and the numerator is approaching something *nonzero*, we must have some kind of infinity. Our task then is to determine whether the answer is $\pm\infty$ (approaching both kinds of infinity), ∞ (approaching positive infinity from both directions), or $-\infty$ (approaching negative infinity from both directions). To show our work on a problem like this, we use arrows in the numerator and denominator showing what each part of the fraction approaches as x approaches whatever the limit suggests it should approach (3 in the case of the first example below), and when the denominator is approaching zero, we want to put a superscript to denote how it's approaching zero ($\pm$ for both sides, $+$ for only from the positive direction, or $-$ for only from the negative direction). In the first example below, we put a $\pm$ superscript next to the zero, because if x is a bit larger than 3, the denominator is positive, but if x is a bit less than 3, the denominator is negative. In the second example, we put a $+$ superscript because the denominator gets squared and so is always positive. For the first example, our arrows indicate we have a positive number (6) divided by a number approaching zero from both directions, so our answer is $\pm\infty$. On the left, you'll see what the problem statement would look like, and on the right, you'll see how the notation works for completely showing work solving the problem:

$$\lim_{x\to 3} \frac{2x}{x-3}: \qquad \lim_{x\to 3} \frac{2x^{\nearrow 6}}{x-3_{\searrow 0^{\pm}}} = \pm\infty$$

$$\lim_{x\to 3} \frac{2x}{(x-3)^2}: \qquad \lim_{x\to 3} \frac{2x^{\nearrow 6}}{(x-3)^2_{\searrow 0^{+}}} = \infty$$

$$\lim_{x\to -3} \frac{2x}{x+3}: \qquad \lim_{x\to 3} \frac{2x^{\nearrow -6}}{x+3_{\searrow 0^{\pm}}} = \pm\infty$$

$$\lim_{x\to -3} \frac{2x}{(x+3)^2}: \qquad \lim_{x\to 3} \frac{2x^{\nearrow -6}}{(x+3)^2_{\searrow 0^{+}}} = -\infty$$

Here's another example we'll need a little bit more detail for:

$$\lim_{x\to 2} \frac{x-1}{x^3-12x^2+36x-32}.$$

Trying to plug in $x = 2$ to get a first idea, we see the numerator is approaching $2 - 1 = 1$, and the denominator is approaching

$$2^3 - 12(2^2) + 36(2) - 32 = 8 - 48 + 72 - 32 = 80 - 80 = 0.$$

Since the denominator approaches zero and the numerator approaches something nonzero, we know it's going to have an infinite limit of some kind. However, the way our function is written at the moment makes it difficult to reason about which direction the denominator approaches zero from. What we'd like to do is be able to easily determine what sign the denominator has near $x = 2$, and for this, our best tool is to factor. Usually factoring a cubic like this would be difficult, but notice we already know one of the roots—we just calculated that the cubic is zero when $x = 2$. This means $(x-2)$ is a factor, and we know when we divide by it, we should be left with some quadratic equation. To find this quadratic, we could use any polynomial division method. Here's synthetic division:

$$\begin{array}{c|cccc} 2 & 1 & -12 & 36 & -32 \\ & & 2 & -20 & 32 \\ \hline & 1 & -10 & 16 & 0 \end{array}$$

This tells us $x^3-12x^2+36x-32 = (x-2)(x^2-10x+16)$, and to completely factor this thing, we can now use whatever approach we'd like on the quadratic term to see that $x^2-10x+16 = (x-2)(x-8)$. We're therefore trying to compute the limit

$$\lim_{x\to 2} \frac{x-1}{x^3-12x^2+36x-32} = \lim_{x\to 2} \frac{x-1}{(x-2)^2(x-8)}.$$

If x is approaching 2, we know that the numerator is approaching 1, and now we can see how the denominator works more clearly: the $(x-2)^2$ term is the part that makes the whole thing approach zero, and because we're squaring, this term is always positive. The $(x-8)$ term, however, is approaching -6 when x approaches 2, so when x is near 2, the denominator as a whole is negative. Now we can fill in our arrows to complete this solution:

$$\lim_{x\to 2} \frac{x-1}{x^3-12x^2+36x-32} = \lim_{x\to 2} \frac{x-1 \nearrow^{1}}{(x-2)^2(x-8)\searrow_{0^-}} = -\infty.$$

One last thing to note about this proposition is how it compares to our previous limit laws. Earlier, we had a limit law for quotients, but using this required the denominator to have a nonzero limit. Now we have a tool that covers this case, provided the numerator has nonzero limit. When both numerator and denominator have limit equal to zero, we have to look for some almost identical function and hope its limit is easier to handle, as in the previous section.

Recall also that all the limit laws we gave previously required the functions f and g we were working with to have *real number* limits L and M. When f and g have infinite limits, the limit laws can break down. For a simple example, consider the function $1/x - 1/x$. Everywhere this is defined (i.e. for $x \neq 0$, we are subtracting $1/x$ from itself, and so we get 0. Hence

$$\lim_{x\to 0} 1/x - 1/x = 0.$$

The limit law we had for the difference of two functions, if applied here, would suggest the limit should be $\pm\infty \pm \infty$, which is another example of an indeterminate form—any sum or difference of infinities is indeterminate, and so we often need to rewrite the functions involved to avoid this situation. For instance, another difference of infinities is

$$\lim_{x\to 0} \frac{1}{x} - \frac{1}{x^2}.$$

We said $\lim_{x\to 0} 1/x = \pm\infty$ and that $\lim_{x\to 0} 1/x^2 = \infty$, but the answer to this question is not $\pm\infty - \infty$. Instead, we need to rewrite the equation by finding a common denominator to see what's really going on:

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

Here are a couple more examples:

$$\lim_{x\to 0} \frac{1^{\nearrow 1}}{x_{\searrow 0\pm}} = \pm\infty \quad \text{and} \quad \lim_{x\to 0} \frac{2^{\nearrow 2}}{x(x-2)_{\searrow 0\pm}} = \pm\infty,$$

$$\text{but} \quad \lim_{x\to 0} \frac{1}{x} + \frac{3}{x(x-3)} = \lim_{x\to 0} \frac{x-3}{x(x-3)} + \frac{3}{x(x-3)} = \lim_{x\to 0} \frac{x}{x(x-3)} \overset{\text{AIF}}{=} \lim_{x\to 0} \frac{1}{x-3} = -\frac{1}{3}$$

$$\lim_{x\to 0} \frac{1^{\nearrow 1}}{x_{\searrow 0\pm}} = \pm\infty \quad \text{and} \quad \lim_{x\to 0} \frac{x^2 - 3x + 2^{\nearrow 2}}{x^2 - 2x_{\searrow 0\pm}} = \pm\infty,$$

$$\text{but} \quad \lim_{x\to 0} \frac{1}{x} + \frac{x^2 - 3x + 2}{x^2 - 2x} = \lim_{x\to 0} \frac{x-2}{x^2 - 2x} + \frac{x^2 - 3x + 2}{x^2 - 2x} = \lim_{x\to 0} \frac{x^2 - 2x}{x^2 - 2x} \overset{\text{AIF}}{=} \lim_{x\to 0} 1 = 1.$$

We've seen that a limit of the form $\pm\infty \pm \infty$ can equal 0, $-\infty$, $-1/3$, or 1, and in fact, it can equal any real number or any kind of infinity. It all depends on what the function we're taking the limit of is, which is why this is an example of an indeterminate form. The best approach usually is trying to find a common denominator and simplifying what we can.

Not only can we let x approach some number a and get outputs that get increasingly large, we can also ask about what happens when x itself gets increasingly large. You may recall from a high school math class the idea of "end behavior" of polynomial functions. There you might have learned that if a polynomial f has a positive leading coefficient, then $f(x)$ goes to infinity as x goes to infinity, and goes to ∞ or $-\infty$ as x goes to negative infinity, depending on the degree of the polynomial. This is the idea of taking a limit *at infinity*, and with the limit notation, we would denote this as $\lim_{x\to\infty} f(x) = \infty$.

It can also happen that $f(x)$ approaches some real number when x approaches infinity, an example of this being the function $1/x$. We have the following two limits, both of which we can see by looking at a graph of the function:

$$\lim_{x\to\infty} \frac{1}{x} = \lim_{x\to-\infty} \frac{1}{x} = 0.$$

This says that the function $1/x$ gets arbitrarily small as x gets arbitrarily large in either direction. Another way you might have talked about this fact in a high school math class is that this function has a *horizontal asymptote* that it approaches.

The problems we'll look at in this class concerning limits at infinity will all boil down to this limit, or something that looks a lot like it:

Fact. For any positive integer n, we have

$$\lim_{x\to\pm\infty} \frac{1}{x^n} = 0.$$

With this fact, we can use our original limit laws to take limits like the following, where the tool we have is to divide everything by the highest power of x appearing and employ the fact above:

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

$$\lim_{x\to\infty} \frac{2x^3-2x+1}{3x^3+x^2} \overset{\text{AIF}}{=} \lim_{x\to\infty} \frac{2-2/x^2+1/x^3}{3+1/x} = \frac{2-2\lim_{x\to\infty} 1/x^2+\lim_{x\to\infty} 1/x^3}{3+\lim_{x\to\infty} 1/x} = \frac{2-2\cdot 0+0}{3+0} = \frac{2}{3}$$

$$\lim_{x\to\infty} \frac{\sqrt{x^2-2}}{x-2} \overset{\text{AIF}}{=} \lim_{x\to\infty} \frac{\sqrt{x^2-2}/x}{1-2/x} \overset{\text{AIF}}{=} \lim_{x\to\infty} \frac{\sqrt{x^2-2}/\sqrt{x^2}}{1-2/x} = \lim_{x\to\infty} \frac{\sqrt{1-2/x^2}}{1-2/x}$$
$$= \frac{\sqrt{1-2\lim_{x\to\infty} 1/x^2}}{1-2\lim_{x\to\infty} 1/x} = \frac{\sqrt{1-2\cdot 0}}{1-2\cdot 0} = \frac{1}{1} = 1$$

$$\lim_{x\to-\infty} \frac{\sqrt{x^2-2}}{x-2} \overset{\text{AIF}}{=} \lim_{x\to-\infty} \frac{\sqrt{x^2-2}/x}{1-2/x} \overset{\text{AIF}}{=} \lim_{x\to-\infty} \frac{\sqrt{x^2-2}/-\sqrt{x^2}}{1-2/x} = \lim_{x\to-\infty} \frac{-\sqrt{1-2/x^2}}{1-2/x}$$
$$= \frac{-\sqrt{1-2\lim_{x\to-\infty} 1/x^2}}{1-2\lim_{x\to-\infty} 1/x} = \frac{-\sqrt{1-2\cdot 0}}{1-2\cdot 0} = \frac{-1}{1} = -1$$

The first $=$ of each line has an 'AIF' over it to remind us that we're changing the function by dividing by whatever power of x is necessary. The new function is almost identical, but now undefined at $x=0$. In the last two examples, there's a much more subtle use of the Almost Identical Functions property being used. Here, we have a square root of something whose highest power of x is 2, so the "highest" power in the numerator is therefore a half of this (because of the square root). Similarly, the highest power in the denominator is 1, so we know we need to divide top and bottom by x^1. To move the x inside the square root, though, we need to figure out some equivalent expression that involves a square root. This is because we have the rule

$$\frac{\sqrt{a}}{\sqrt{b}} = \sqrt{\frac{a}{b}}, \quad \text{and it is not true that} \quad \frac{\sqrt{a}}{b} = \sqrt{\frac{a}{b}}.$$

Therefore we replace x by $\sqrt{x^2}$, since the squaring cancels the square root. However, while it's true that $\sqrt{3^2} = \sqrt{9} = 3$, we can't do this with negative numbers: $\sqrt{(-3)^2} = \sqrt{9} \neq -3$. So it turns out that replacing x by $\sqrt{x^2}$ is only valid for $x \geq 0$. Fortunately, we are taking the limit as $x \to \infty$, so we only care about positive x anyway. Nonetheless, we are changing the function and using the AIF property to say that the limit stays the same, so we denote this we another use of $\stackrel{\text{AIF}}{=}$. For the final example, we cannot replace x by $\sqrt{x^2}$ since we are taking the limit as $x \to -\infty$, meaning we instead use $x = -\sqrt{x^2}$ for $x \leq 0$. The limit is then otherwise the same, but because of this we end up with a -1 at the end instead of a 1.

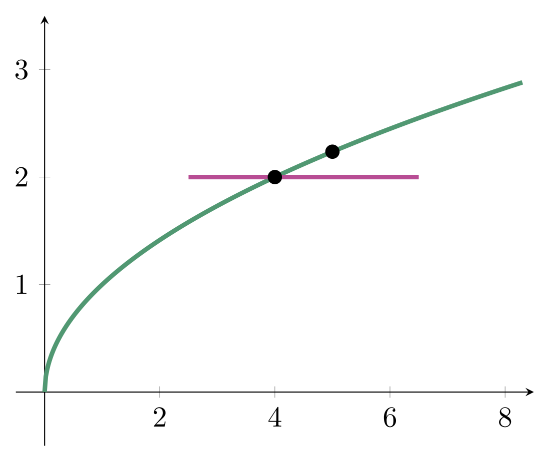
2 Derivatives

2.1 Linear approximation, tangent lines, and the definition of the derivative

In the previous section on limits, we talked about continuous functions. Recall that if some function f is continuous at a point a, this means the limit of f as we approach a is actually equal to $f(a)$. In symbols, $\lim_{x\to a} f(x) = f(a)$. By our intuitive definition of limits, this means that "as x gets close to a, $f(x)$ gets close and stays close to $f(a)$." Another way of phrasing this is that $f(x)$ is *approximately* $f(a)$ when x is close to a.

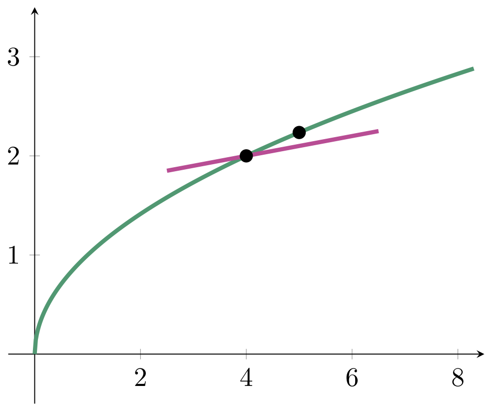
Indeed, when f is continuous at a point a, we call the constant function $g(x) = f(a)$ the *zeroth-order approximation to f at a*. The goal of this section will be to introduce another, even nicer property a function can have, called *differentiability*, and show how to use this to produce a *first-order approximation* to a differentiable function f at a point a. To see what this first-order approximation should look like, consider function $f(x) = \sqrt{x}$.

There are some inputs x (e.g. 4, 9, 16) whose corresponding outputs $\sqrt{x}$ we can easily figure out (2,3,4, respectively), but other inputs (non square numbers) whose outputs aren't nearly as easy to calculate. If we wanted to know $\sqrt{5}$, out goal would be to approximate this function at 5 in a reliable way. One approach we could take is to pick the nearest value whose output we know (4 in this case), and use the zeroth-order approximation. Here's a graph of $\sqrt{x}$ along with the zeroth-order approximation at $x = 4$:

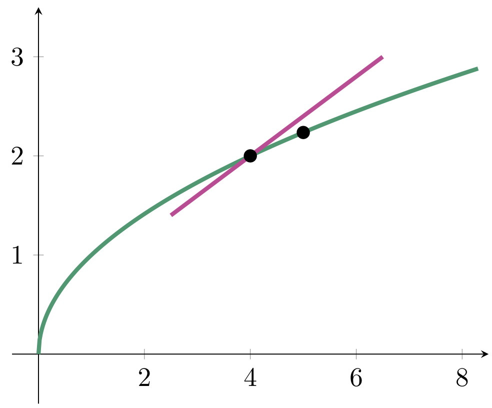


A we can see, the zeroth-order approximation isn't really that good. The actual value of $\sqrt{5}$ is around 2.236, but the approximation we're using here suggests $\sqrt{5} \approx 2$. Intuitively, even though we don't know by how much, we know that 5 being a bit more than 4 means that $\sqrt{5}$ should in fact be a bit more than $2 = \sqrt{4}$. What we're using to say that is that when we think about the square root function or look at its graph, we see it's *increasing*. The larger the input, the larger the output. Right now with our zeroth-order approximation, we're trying to match a completely flat line up with our curvy, increasing square root function. If we want to improve our approximation, we want to find a way to make it sensitive to changes in the input, since right now the input is irrelevant—no matter x what we plug in, the approximation will always be that $\sqrt{x} \approx 2$.

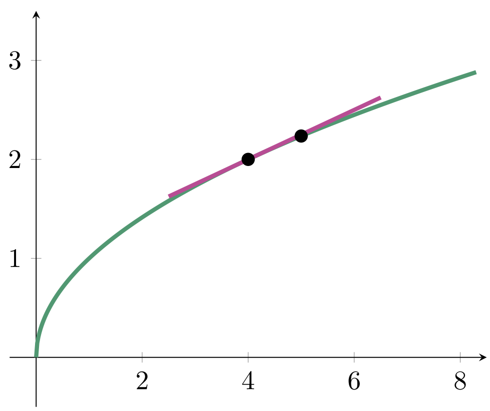
What we want then is a simple function that's increasing. Increasing because the function we're trying to approximate is increasing, and simple because we'd like the be able to do calculations really easily. From our zeroth-order approximation, which is a constant function, the next simplest thing we can try is a linear function. Here are a few linear functions, all of which are better fits than the zeroth-order approximation:



Better, but not steep enough.



Better, but a bit too steep.



Perfect! Just the right slope.

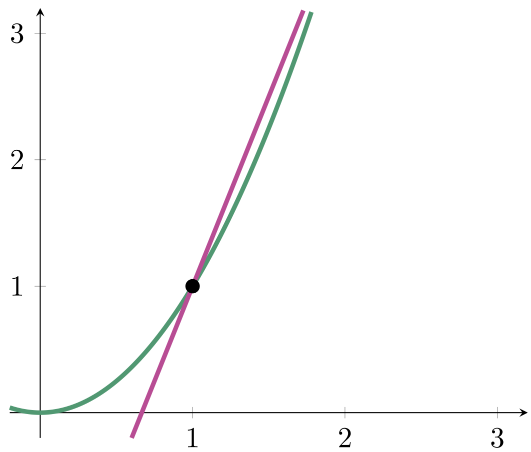
In the first graph, the function is increasing faster than the line, and in the second, the function is increasing slower than the line. What we want is a line that increases at exactly the same rate as the function—the quality of the approximation depends on how well the slope of this line matches the function we're approximating.

Let's take a moment to recall what we know about lines. We can write a line in a few different ways:

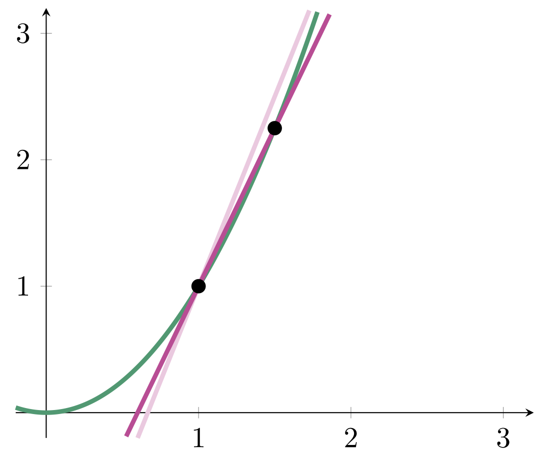
$y = mx + b$	slope-intercept form
$y - y_0 = m(x - x_0)$	point-slope form
$y - y_0 = \frac{y_1 - y_0}{x_1 - x_0}(x - x_0)$	two-point form

Note that for our approximations, we generally won't know either the y-intercept of a line or any second point on the line. Really all we have to work with is the equation of the function and then the point at which we want to approximate it. This means we want to be working with point-slope form. We've got a point already, so all we need now is a way to figure out which slope best matches the function we want to approximate. When we do this, the line that we'll end up with is the line through the point of interest that is *tangent* to the graph of the function.

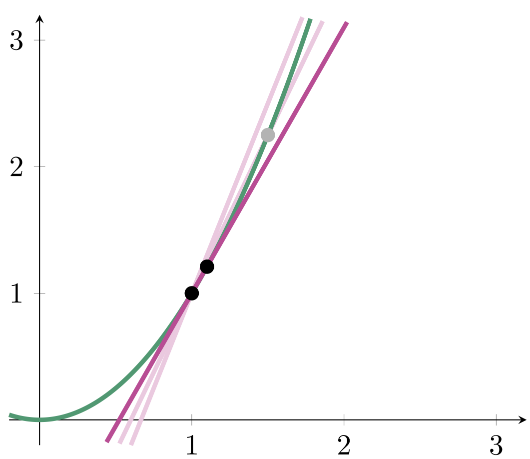
Here's another function we can already compute pretty easily that will allow us to demonstrate the principal involved here: take $f(x) = x^2$, and lets say we want to find the line tangent to its graph at the point $(1, 1)$. We know from the point-slope form that this line should have the form $y - 1 = m(x - 1)$, and now we just need to figure out what the missing slope ought to be. Here's the technique: we want to find a nearby input/output pair and see what slope we get from the two-point form of the line between them. Then we want to find an even closer point on the curve, and again use the two-point form and see what the slope is:



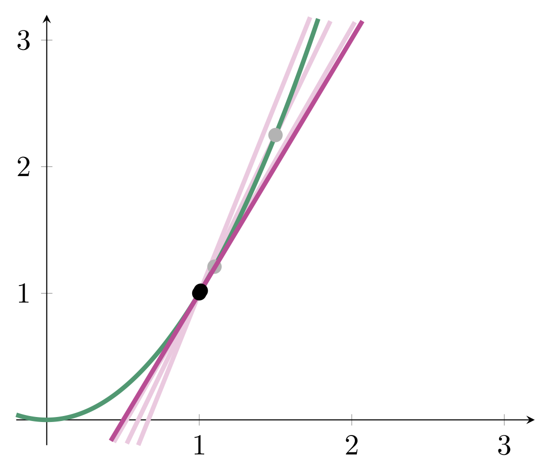
Using $(2, 4)$—slope: $\frac{4-1}{2-1} = 3$



Using $(1.5, 2.25)$—slope: $\frac{2.25-1}{1.5-1} = 2.5$



Using $(1.1, 1.21)$—slope: $\frac{1.21-1}{1.1-1} = 2.1$



Using $(1.01, 1.0201)$; slope—$\frac{1.0201-1}{1.01-1} = 2.01$

When we have a line through two distinct points on a graph like this, we call it a *secant* line, and as we can see, the tangent line we're looking for is the limit of these secant lines. That is, the slopes of all the secant lines seem to be approaching 2, which is exactly the slope we're looking for.

To make this a little more precise, let's figure out how to actually write down this limit like we did in section 1 and then compute it. We have the point $(1,1)$ and a line going through it as well as some other point. Since this other point is on the graph, it has the form (t,t^2) for some t, and so to get the slope of the line connecting these two points, we just need to do rise over run:

$$m(t) = \frac{t^2-1}{t-1},$$

where we use the name $m(t)$ for this function since it is the slope of the secant line we get when we use the point on the graph with input x. Now we want to take the limit of these slopes, which is

$$\lim_{x\to 1} m(t) = \lim_{t\to 1} \frac{t^2-1}{t-1} = \lim_{t\to 1} \frac{(t-1)(t+1)}{t-1} \overset{\text{AIF}}{=} \lim_{t\to 1} t+1 = 2.$$

So indeed, the slopes of the secant lines really do approach 2. Our tangent line is therefore the line that passes through the point $(1,1)$ and has slope 2, which we can write in point-slope form as $y-1 = 2(x-1)$.

We can take this example and generalize what we did here to define the derivative of a function f at a certain point $x=a$. This number is the slope of the line tangent to the function's graph at the point $(a, f(a))$. Here's the precise definition.

Definition. Let f be a function defined in some open interval around a point a. Then we define the **derivative of f at a** to be

$$f'(a) = \lim_{t\to a} \frac{f(t)-f(a)}{t-a} = \lim_{h\to 0} \frac{f(a+h)-f(a)}{h},$$

if it exists. When these limits do exist, we say that f is **differentiable** at a.

The two limits in the definition here always give the same number since the second one is just the first one with the substitution $h = t-a$, which approaches 0 as t approaches a. Let's use the second limit to redo the earlier example with $f(x) = x^2$ where we found that the derivative of f at 1 was 2:

$$\lim_{h\to 0} \frac{f(1+h)-f(1)}{h} = \lim_{h\to 0} \frac{(1+h)^2-1}{h} = \lim_{h\to 0} \frac{1+2h+h^2-1}{h} = \lim_{h\to 0} \frac{2h+h^2}{h} \overset{\text{AIF}}{=} \lim_{h\to 0} 2+h = 2.$$

The notation $f'(a)$ for the derivative of f at a is due to Newton, sometimes we'll also use Leibniz' notation $\frac{df}{dx}(a)$, which means the same thing.

Before we turned our attention to the function $f(x) = x^2$, we were looking at $f(x) = \sqrt{x}$ and wanting to approximate $\sqrt{5}$. Our goal was to be able to approximate this better than the zeroth-order approximation where we said that $\sqrt{5} \approx 2$. To get our first-order (linear) approximation, we know we want a line that passes through the point $(4, 2)$, and to get the slope, we need the derivative of the function $f(x) = \sqrt{x}$ at the point 4:

$$f'(4) = \lim_{t \to 4} \frac{\sqrt{t} - \sqrt{4}}{t - 4} = \lim_{t \to 4} \frac{\sqrt{t} - 2}{t - 4} \cdot \frac{\sqrt{t} + 2}{\sqrt{t} + 2} = \lim_{t \to 4} \frac{t - 4}{(t-4)(\sqrt{t} + 2)} \stackrel{\text{AIF}}{=} \lim_{t \to 4} \frac{1}{\sqrt{t} + 2} = \frac{1}{\sqrt{4} + 2} = \frac{1}{4}.$$

This means the line tangent to $\sqrt{x}$ at $x = 4$ is $y - 2 = \frac{1}{4}(x - 4)$, so the best linear approximation to $\sqrt{x}$ for x close to 4 is the function

$$L(x) = 2 + \frac{1}{4}(x - 4).$$

Then 5 is relatively close to 4, so we get that $\sqrt{5} \approx L(5) = 2 + (1/4) \cdot 1 = 11/5$ or 2.25. The exact answer is around 2.236, so we're only off by about .014, whereas the zeroth-order approximation was much worse. Note though that if we try approximating outputs that aren't so close to 5, the approximation doesn't do very well: $\sqrt{1} = 1$ but the linear approximation suggests $\sqrt{1} \approx 5/4 = 1.25$, and $\sqrt{100} = 10$ but the linear approximation suggests $\sqrt{100} \approx 26$, which is way off.

When we reviewed functions briefly at the beginning of the notes, it was claimed that when we started talking about derivatives, we'd really be talking about a function whose inputs and outputs were functions. Right now, having just defined derivatives, this claim doesn't yet make sense. The derivative we have at the moment is indeed a function, but it takes a function f and a value a as input and outputs the number $f'(a)$. However, if we forget about specifying a particular point we'd like to take the derivative at, we can instead ask about f' as being a function itself, and this was what was alluded to previously. We think of the derivative as a function whose inputs and outputs are functions: if we have the input function f, which is a function taking real numbers to real numbers, we get an output function f', which is also a function taking real numbers to real numbers. For instance, consider again $f(x) = x^2$. Previously, we calculated $f'(1) = 2$, but to view f' as a function taking inputs to outputs, we'd like to know what $f'(x)$ is for any x, not just for $x = 1$. Here's the calculation:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \to 0} \frac{(x+h)^2 - x^2}{h} = \lim_{h \to 0} \frac{2xh + h^2}{h} \stackrel{\text{AIF}}{=} \lim_{h \to 0} 2x + h = 2x + 0 = 2x.$$

We could just as well have done the computation with the other limit:

$$f'(x) = \lim_{t \to x} \frac{f(t) - f(x)}{t - x} = \lim_{t \to x} \frac{t^2 - x^2}{t - x} = \lim_{t \to x} \frac{(t-x)(t+x)}{t - x} \stackrel{\text{AIF}}{=} \lim_{t \to x} t + x = x + x = 2x.$$

The limit we use doesn't change the outcome, but to do what we did above in the second limit, we needed to remember how to factor a difference of squares. In the first version, all we needed to know was to expand the $(x+h)^2$ term and then look for something to cancel. Thinking of the process of taking a derivative as a function, this function takes the input $f(x) = x^2$ and returns the output $f'(x) = 2x$.

Even though when we write the derivative of a function, we'll more often use Newton's $f'(x)$ notation, when we consider the function that takes a function to its derivative, it is more common to use Leibniz' notation. We think of $\frac{d}{dx}$ as being the name of this function of functions, in the same way that f is the name of a function of real numbers. To introduce even more notation, we usually write $f : \mathbb{R} \to \mathbb{R}$ to say that f is a function that takes a real number input to a real number output. We might then write

$$\frac{d}{dx} : \{\text{functions } \mathbb{R} \to \mathbb{R}\} \to \{\text{functions } \mathbb{R} \to \mathbb{R}\}$$

to be clear about what the inputs and outputs of the $\frac{d}{dx}$ function are—$\frac{d}{dx}$ takes a function on the real numbers as input to some other function on the real numbers as output. For example, we could write $\frac{d}{dx}(x^2) = 2x$ based on what we calculated above. (But we could also use Newton's notation and write something like $(x^2)' = 2x$, and we'll use this notation at times as well.) This notation isn't so important for us to remember, but it does help clarify exactly how we're thinking about the process of taking derivatives as a function.

Since the derivative of a function is another function, we can also take this function's derivative. Taking the derivative of a derivative gives a *second derivative*, and in general if we keep doing this, we're computing what are collectively called *higher-order derivatives*, higher-order here meaning anything greater than 1. The normal derivative of a function is that function's *first derivative*. The notation here is $f''(x)$ for the second derivative using Newton's notation, and $\frac{d^2 f}{dx^2}$ using Leibniz' notation. Since Newton's notation involves adding a bunch of apostrophes, for higher-order derivatives beyond the second derivative, we'll usually use Leibniz' notation, which is $\frac{d^n f}{dx^n}$ for the nth derivative of a function. For example, if we take the derivative of $2x$, we compute that

$$\lim_{h \to 0} \frac{2(x+h) - 2x}{h} = \lim_{h \to 0} \frac{2x + 2h - 2x}{h} = \lim_{h \to 0} \frac{2h}{h} \overset{\text{AIF}}{=} \lim_{h \to 0} 2 = 2.$$

We can therefore say that for $f(x) = x^2$, the second derivative is $f''(x) = 2$ or $\frac{d^2 f}{dx^2} = 2$.

Let's see a few more examples. For each, we'll do the calculation twice, once using each of the limits in the definition of the derivative.

$$\begin{aligned} \frac{d}{dx}\sqrt{x} &= \lim_{h \to 0} \frac{\sqrt{x+h} - \sqrt{x}}{h} = \lim_{h \to 0} \frac{\sqrt{x+h} - \sqrt{x}}{h} \cdot \frac{\sqrt{x+h} + \sqrt{x}}{\sqrt{x+h} + \sqrt{x}} = \lim_{h \to 0} \frac{(x+h) - x}{h(\sqrt{x+h} + \sqrt{x})} \\ &= \lim_{h \to 0} \frac{h}{h(\sqrt{x+h} + \sqrt{x})} \overset{\text{AIF}}{=} \lim_{h \to 0} \frac{1}{\sqrt{x+h} + \sqrt{x}} = \frac{1}{\sqrt{x+0} + \sqrt{x}} = \boxed{\frac{1}{2\sqrt{x}}} \end{aligned}$$

$$
\begin{aligned}
\frac{d}{dx}\sqrt{x} &= \lim_{t\to x}\frac{\sqrt{t}-\sqrt{x}}{t-x} = \lim_{t\to x}\frac{\sqrt{t}-\sqrt{x}}{t-x}\cdot\frac{\sqrt{t}+\sqrt{x}}{\sqrt{t}+\sqrt{x}} = \lim_{t\to x}\frac{t-x}{(t-x)(\sqrt{t}+\sqrt{x})} \\
&\overset{\text{AIF}}{=} \lim_{t\to x}\frac{1}{\sqrt{t}+\sqrt{x}} = \frac{1}{\sqrt{x}+\sqrt{x}} = \boxed{\frac{1}{2\sqrt{x}}} \\
\frac{d}{dx}(x^3) &= \lim_{h\to 0}\frac{(x+h)^3-x^3}{h} = \lim_{h\to 0}\frac{(x^3+3x^2h+3xh^2+h^3)-x^3}{h} = \lim_{h\to 0}\frac{3x^2h+3xh^2+h^3}{h} \\
&\overset{\text{AIF}}{=} \lim_{h\to 0} 3x^2+3xh+h^2 = 3x^2+3x\cdot 0+0^2 = \boxed{3x^2} \\
\frac{d}{dx}(x^3) &= \lim_{t\to x}\frac{t^3-x^3}{t-x} = \lim_{t\to x}\frac{(t-x)(t^2+tx+x^2)}{t-x} \overset{\text{AIF}}{=} \lim_{t\to x} t^2+tx+x^2 = x^2+x^2+x^2 = \boxed{3x^2} \\
\frac{d}{dx}\left(\frac{x}{x+1}\right) &= \lim_{h\to 0}\frac{1}{h}\cdot\left(\frac{x+h}{x+h+1}-\frac{x}{x+1}\right) = \lim_{h\to 0}\frac{(x+h)(x+1)-x(x+h+1)}{h(x+h+1)(x+1)} \\
&= \lim_{h\to 0}\frac{(x^2+xh+x+h)-(x^2+xh+x)}{h(x+h+1)(x+1)} = \lim_{h\to 0}\frac{h}{h(x+h+1)(x+1)} \\
&\overset{\text{AIF}}{=} \lim_{h\to 0}\frac{1}{(x+h+1)(x+1)} = \frac{1}{(x+0+1)(x+1)} = \boxed{\frac{1}{(x+1)^2}}
\end{aligned}
$$

$$
\begin{aligned}
\frac{d}{dx}\left(\frac{x}{x+1}\right) &= \lim_{t\to x}\frac{1}{t-x}\left(\frac{t}{t+1}-\frac{x}{x+1}\right) = \lim_{t\to x}\frac{t(x+1)-x(t+1)}{(t-x)(t+1)(x+1)} \\
&= \lim_{t\to x}\frac{(tx+t)-(tx+x)}{(t-x)(t+1)(x+1)} = \lim_{t\to x}\frac{t-x}{(t-x)(t+1)(x+1)} \\
&\overset{\text{AIF}}{=} \frac{1}{(t+1)(x+1)} = \frac{1}{(x+1)(x+1)} = \boxed{\frac{1}{(x+1)^2}}
\end{aligned}
$$

As we can see, starting with a function like $x/(x+1)$, which is defined everywhere except at $x=-1$, means we get a function $1/(x+1)^2$ which is also not defined at $x=-1$. Similarly, taking the derivative of x^3, a function which is defined everywhere, we get the function $3x^2$, which is also defined everywhere. However, just because a function is defined at a particular point does not mean its derivative must also be defined there. Per the definition, we say that a function f is *differentiable* at a if $f'(a)$ exists. The functions x^3 and $x/(x+1)$ have the same domain as their derivatives, but our first example, $\sqrt{x}$, does not. The function $\sqrt{x}$ itself is defined for all $x\geq 0$, so its domain is $[0,\infty)$. However, looking at its derivative $1/2\sqrt{x}$, we see that we still can't plug in anything because of the square root, but now we also can't plug in 0 since this would make the denominator 0. Thus, $\sqrt{x}$ is defined and continuous at 0, but is not differentiable there! What's happening is that as we approach 0, the function is getting steeper and steeper—infinitely steep in fact. To get a line tangent to the function at the point $(0,0)$, the line would need to be completely vertical, somehow having infinite slope.

Getting infinitely steep like this is just one thing that can cause a function to fail to be differentiable at a certain point, but there are others as well. Anywhere the function fails to be defined is a point at which the function is not differentiable, and similarly any point at which the function fails to be continuous is not differentiable. Indeed, being differentiable is a strictly stronger condition than being continuous:

Proposition. *Let f be a function differentiable at a. Then f is continuous at a.*

Proof. We want to show that $\lim_{x\to a} f(x) = f(a)$ to show that f is continuous at a. By the limit laws, it is equivalent to show that $\lim_{x\to a} f(x) - f(a) = 0$. Indeed,

$$\lim_{x\to a} f(x) - f(a) \stackrel{\text{AIF}}{=} \lim_{x\to a} \frac{f(x)-f(a)}{x-a} \cdot (x-a) = \left(\lim_{x\to a} \frac{f(x)-f(a)}{x-a}\right)\left(\lim_{x\to a} x-a\right) = f'(a)\cdot 0 = 0,$$

where we are using the differentiability of f at a to say that

$$\lim_{x\to a} \frac{f(x)-f(a)}{x-a}$$

equals the real number $f'(a)$. This allows us to use the limit law for products to split the limit into two pieces. □

Notice also that the first step of the limit algebra here involves multiplying and dividing by the same thing (here it's $x-a$). This is just like multiplying by the conjugate in that it doesn't change the function in a way that affects the limit. Finding a nice way to multiply and divide by the same thing is one of mathematicians' favorite tricks, along with adding and subtracting the same thing.

Therefore if we have a function like

$$f(x) = \begin{cases} 3x+2 & x \leq 1 \\ 4-x & x > 1 \end{cases}$$

with a jump discontinuity, the point of discontinuity (here $x = 1$) is also a point at which f is not differentiable. There is no linear function we could pick here to approximate $f(x)$ well. If the approximation is good for values to the left of 1, it must be bad for values to the right of 1 and vice versa.

Another thing that can go wrong and cause a function to fail to be differentiable at a point is if the graph has a corner at the point. Consider the absolute value function

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

This is continuous everywhere, and if $x \neq 0$ it is also differentiable. We have that

$$f'(x) = \lim_{t\to x} \frac{|t|-|x|}{t-x}$$

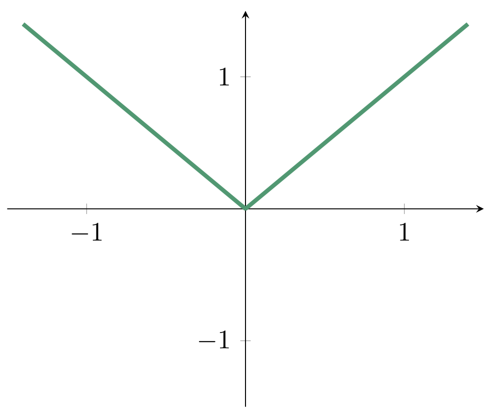
If $x > 0$ and t is close enough to x (so that $t > 0$ also), then the above is

$$\frac{d}{dx}|x| = \lim_{t\to x} \frac{|t| - |x|}{t - x} = \lim_{t\to x} \frac{t-x}{t-x} \overset{\text{AIF}}{=} \lim_{t\to x} 1 = 1.$$

If instead $x < 0$ and t is close enough to x, then

$$f'(x) = \lim_{t\to x} \frac{|t| - |x|}{-t - (-x)} = \lim_{t\to x} \frac{-(t-x)}{t-x} \overset{\text{AIF}}{=} \lim_{t\to x} -1 = -1.$$

Hence for $x > 0$ the derivative is 1 and for $x < 0$ the derivative is -1. This should make sense since the absolute value function is just two lines smushed together, one with a $+1$ slope and one with a -1 slope:



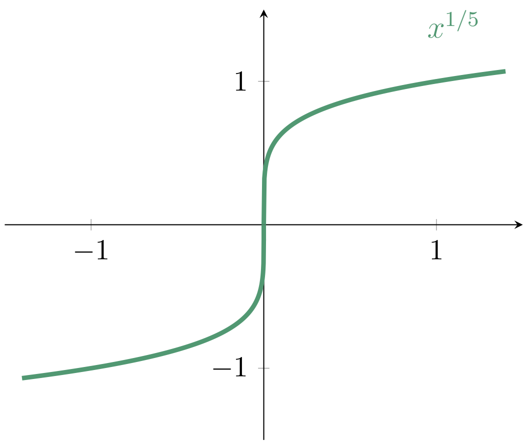
Right at $x = 0$ where the function has its corner, the rate of change suddenly jumps, and we see this reflected in the fact that there's no way to assign a "slope" to the function. The derivative at 0 would be the limit

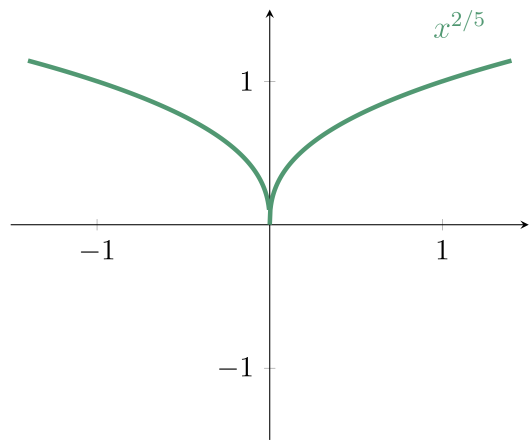
$$f'(0) = \lim_{t\to 0} \frac{|t| - |0|}{t - 0} = \lim_{t\to 0} \frac{|t|}{t},$$

but this does not exist. This is because the one-sided limits exist but are different—we have that

$$\lim_{t\to 0^+} \frac{|t|}{t} = \lim_{t\to 0^+} \frac{t}{t} \overset{\text{AIF}}{=} \lim_{t\to 0^+} 1 = 1 \quad \text{but} \quad \lim_{t\to 0^-} \frac{|t|}{t} = \lim_{t\to 0^-} \frac{-t}{t} = \lim_{t\to 0^-} -1 = -1.$$

In addition to corners, we can also have a sort of infinitely sharp corner, called a *cusp*, where the slope would need to simultaneously be both $+\infty$ on one side and $-\infty$ on the other. Compare, for example, the graphs of $x^{\frac{1}{5}}$ and $x^{\frac{2}{5}}$ below:





As x approaches 0, the slope of the graphs both become infinite. However, in the case of $x^{\frac{1}{5}}$, the secant lines we draw between $(0, 0)$ and any point of the form $(t, t^{\frac{1}{5}})$ will have positive slope, so the slope right at 0 should somehow be $+\infty$. For $x^{\frac{2}{5}}$, the sign of the slope depends on the direction of approach. If we have a secant line drawn between $(0, 0)$ and $(t, t^{2/5})$ for $t > 0$, then again the slope is positive, whereas if $t < 0$, the slope is negative. Once we have some more tools for computing derivatives (our next section), we'll see that the functions we get when taking the derivatives of $x^{\frac{1}{5}}$ and $x^{\frac{2}{5}}$ agree with the analysis in that they will be undefined at $x = 0$ with limits $+\infty$ and $\pm\infty$, respectively.

2.2 Computing derivatives

Now that we've defined the derivative of a function and seen a few calculations directly from this definition, we'd like a way of computing the derivative more easily, without having to appeal directly to the limit definition. The limit definition can be a pain to work with, and in practice when taking derivatives, it's really not what we use unless our function is sufficiently strange that the rules we develop in this section can't help us. Here, we'll state and (sometimes) prove several differentiation rules that will allow us to mostly forget about the limit definition.

We start with the simplest rule possible. If $f(x) = c$ is some constant function, then the line $y = c$ is already the best possible linear approximation since it's always equal to $f(x)$ for any input x. This line has slope 0, so we expect $f'(x) = 0$ for all x. Indeed, we have the following:

Constant rule: *If $f(x) = c$ is a constant function, then $f'(x) = 0$.*

Proof. We use the limit definition:

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{c - c}{h} = \lim_{h\to 0} \frac{0}{h} \stackrel{\text{AIF}}{=} \lim_{h\to 0} 0 = 0,$$

where we have used that the function $0/h$ is almost identical to the function that is 0 everywhere.

They are the same except right at $h = 0$ itself. $\square$

Example. If $f(x) = 99^{99}$, then $f'(x) = 0$.

The derivative also behaves very nicely with respect to sums and constant multiples. The derivative of a sum is the sum of the derivatives, and the derivative of some constant multiple of a function is that same constant multiple of the function's derivative:

Sum rule: *If f and g are differentiable, then $(f(x) + g(x))' = f'(x) + g'(x)$.*

Proof. We have that

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

where we have used the limit law for sums since the differentiability of f and g guarantees that the two limits each exist individually. $\square$

Example. We said earlier that $(x^2)' = 2x$, so using this together with the constant rule tells us that $(x^2 + 6)' = 2x$.

Constant multiple rule: *If f is differentiable and c is some real number, then $(cf(x))' = cf'(x)$.*

Proof. We have that

$$(cf(x))' = \lim_{h\to 0} \frac{cf(x+h) - cf(x)}{h} = \lim_{h\to 0} \frac{c(f(x+h) - f(x))}{h} = c \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = cf'(x)$$

this time using the constant multiple limit law. $\square$

Example. We can now say that $(4x^2)' = 4(x^2)' = 4 \cdot 2x = 8x$.

We also have the difference rule $(f(x) - g(x))' = f'(x) - g'(x)$ for differentiable f and g, and we could give a proof very similar to the proof for the sum rule, or we could note that $f(x) - g(x) = f(x) + (-1) \cdot g(x)$ so that the difference rule follows from the sum and constant multiple rules. A function that takes functions to functions is often called an *operator*, and when it respects sums and constant multiples, we call it *linear*. To be fancy, then, we could say that having the sum and constant multiple rules makes the derivative a *linear operator*.

As with limits, we also have rules to handle products and quotients of functions, but these rules are more complicated. **It is not true that the derivative of a product is the product of**

the derivatives, nor that the derivative of a quotient is the quotient of the derivatives. Instead, we have the following product and quotient rules:

Product rule: *If f and g are differentiable, then $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$.*

Proof. We have that

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

using the trick of adding and subtracting the same thing in the numerator as our first step. We have also used the fact that differentiability implies continuity to say that $\lim_{h\to 0} f(x+h) = f(x)$. □

Example. If we want to find the derivative of x^4, we will soon have a rule to compute this directly, but we can also do it now using the product rule and our earlier calculation that $(x^2)' = 2x$. We have that

$$(x^4)' = (x^2 \cdot x^2)' = (x^2)' \cdot x^2 + x^2 \cdot (x^2)' = 2x \cdot x^2 + x^2 \cdot 2x = 2x^3 + 2x^3 = 4x^3.$$

Quotient rule: If f and g are differentiable and $g(x) \neq 0$, then

$$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}.$$

Proof. We have that

$$\begin{aligned}
\left(\frac{f(x)}{g(x)}\right)' &= \lim_{h\to 0} \frac{\frac{f(x+h)}{g(x+h)} - \frac{f(x)}{g(x)}}{h} = \lim_{h\to 0} \frac{\frac{f(x+h)g(x) - f(x)g(x+h)}{g(x+h)g(x)}}{h} = \lim_{h\to 0} \frac{\frac{f(x+h)g(x) - f(x)g(x+h)}{h}}{g(x+h)g(x)} \\
&= \lim_{h\to 0} \frac{\frac{f(x+h)g(x) - f(x)g(x) + f(x)g(x) - f(x)g(x+h)}{h}}{g(x+h)g(x)} = \lim_{h\to 0} \frac{\frac{f(x+h) - f(x)}{h} \cdot g(x) - f(x) \cdot \frac{g(x+h) - g(x)}{h}}{g(x+h)g(x)} \\
&= \frac{\left(\lim_{h\to 0} \frac{f(x+h) - f(x)}{h}\right) g(x) - f(x) \left(\lim_{h\to 0} \frac{g(x+h) - g(x)}{h}\right)}{g(x) \lim_{h\to 0} g(x+h)} = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}
\end{aligned}$$

again adding and subtracting the same thing and using the continuity of g. □

Example. We can now calculate

$$\left(\frac{x^2-1}{x^2+1}\right)' = \frac{2x(x^2+1)-(x^2-1)\cdot 2x}{(x^2+1)^2} = \frac{2x^3+2x-2x^3+2x}{(x^2+1)^2} = \frac{4x}{(x^2+1)^2}.$$

Our last rule tells us how to take the derivative of a power of x. So for instance, we should be able to use the following rule to recover our previous results that $(x^2)' = 2x$ and $(x^4)' = 4x^3$. For the proof, we will need the formula for factoring a difference of nth powers:

$$a^n - b^n = (a-b)(a^{n-1} + a^{n-2}b + a^{n-3}b^2 + \cdots + ab^{n-2} + b^{n-1}).$$

Power rule: For an integer n, $(x^n)' = nx^{n-1}$.

Proof. We first prove this for the case $n > 0$. We have

$$\begin{aligned}(x^n)' &= \lim_{t\to x}\frac{t^n - x^n}{t-x} = \lim_{t-x}\frac{(t-x)(t^{n-1}+t^{n-2}x+\cdots+tx^{n-2}+x^{n-1})}{t-x}\\ &\overset{\text{AIF}}{=} \lim_{t\to x}(t^{n-1}+t^{n-2}x+\cdots+tx^{n-2}+x^{n-1}) = \underbrace{x^{n-1}+\cdots+x^{n-1}}_{n\text{ times}} = nx^{n-1}.\end{aligned}$$

For $n = 0$, we have $x^n = x^0 = 1$, and so the formula says $(1)' = (x^n)' = nx^{n-1} = 0\cdot x^{-1} = 0$, which is true by the constant rule.[1] Finally, for $n < 0$, we need the quotient rule. We also use the power rule for positive integers we just proved, since $-n > 0$ when $n < 0$. We have

$$\begin{aligned}(x^n)' &= \left(\frac{1}{x^{-n}}\right)' = \frac{(1)'\cdot x^{-n} - 1\cdot(x^{-n})'}{(x^{-n})^2} = \frac{0\cdot x^{-n} - (-n)x^{-n-1}}{x^{-2n}}\\ &= \frac{nx^{-n-1}}{x^{-2n}} = nx^{-n-1-(-2n)} = nx^{-n-1+2n} = nx^{n-1},\end{aligned}$$

so the formula holds for all integers n. $\square$

Example. We have that $(x^{100})' = 100x^{99}$, that $(x^{-100})' = -100x^{-101}$, and that $(x^4)' = 4x^3$ as we calculated earlier.

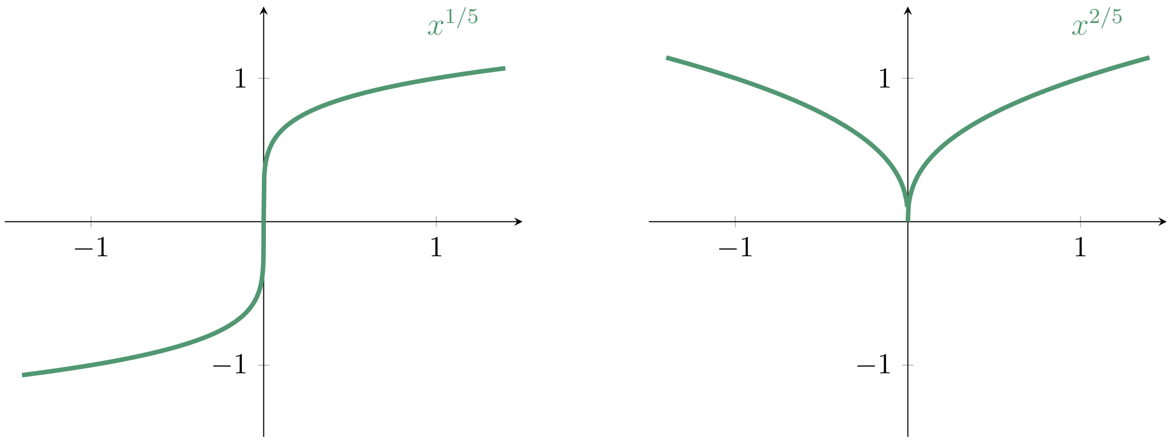
In fact, this formula holds not only for all integers n, but also for all real numbers r. It's considerably more involved to prove this, and we really won't have the tools we need (the exponential function) until Calculus II, but we're going to use the formula freely in this class anyway: for any real number r, $(x^r)' = rx^{r-1}$. So for example, we could say that $(x^\pi)' = \pi x^{\pi-1}$. Since the formula holds for all real numbers, it also holds for rational numbers, and we see that the derivative of $\sqrt{x}$

[1] Things are weird if we let $x = 0$ though, in which case our formulas are telling us that $0^0 = 1$ and that $0/0 = 0$. We'll just go with this since the formulas we have at least have limits that exist and are equal to the values we want as x approaches 0.

we calculated earlier was correct:

$$(\sqrt{x})' = (x^{\frac{1}{2}}) = \frac{1}{2}x^{-\frac{1}{2}} = \frac{1}{2x^{\frac{1}{2}}} = \frac{1}{2\sqrt{x}}.$$

We also claimed earlier in section 2.1 that the functions $x^{\frac{1}{5}}$ and $x^{\frac{2}{5}}$ exhibited related but different behavior at $x = 0$. We said $x^{\frac{1}{5}}$ was to have 'slope' $+\infty$ at $x = 0$, but that $x^{\frac{2}{5}}$ should have 'slope' $\pm\infty$:



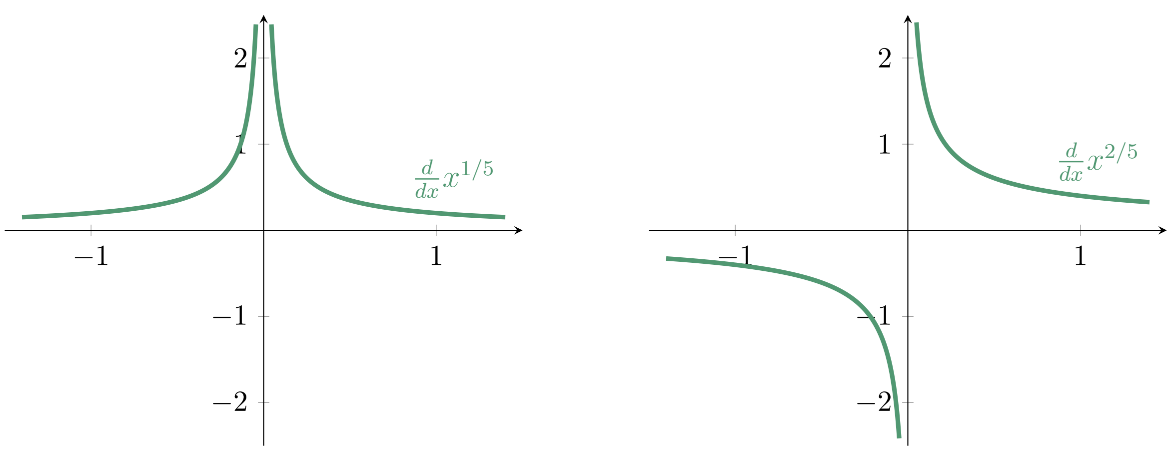
We can now use the power rule to verify this. We get that

$$(x^{\frac{1}{5}})' = \frac{1}{5}x^{-\frac{4}{5}} = \frac{1}{5x^{\frac{4}{5}}} = \frac{1}{5\sqrt[5]{x^4}} \quad \text{and that} \quad (x^{\frac{2}{5}})' = \frac{2}{5}x^{-\frac{3}{5}} = \frac{2}{5x^{\frac{3}{5}}} = \frac{2}{5\sqrt[5]{x^3}}.$$

Therefore the slopes of $x^{\frac{1}{5}}$ and $x^{\frac{2}{5}}$ as x approaches 0 are

$$\lim_{x\to 0} \frac{1^{\nearrow 1}}{5\sqrt[5]{x^4}_{\searrow 0+}} = \infty \quad \text{and} \quad \lim_{x\to 0} \frac{2^{\nearrow 2}}{5\sqrt[5]{x^3}_{\searrow 0\pm}} = \pm\infty,$$

as claimed, and this can also clearly be seen in the graphs of the derivatives:



Useful as they are now, these rules will begin to show their usefulness even more when we start combining functions built out of algebra with trig functions. Before we can do this however, we need to figure out how to take derivatives of the basic trig functions, $\sin(x)$ and $\cos(x)$. The tools we need to do this are formulas you might have seen in high school—the angle addition formulas for sine and cosine:

$$\begin{aligned}\sin(\alpha+\beta) &= \sin(\alpha)\cos(\beta)+\sin(\beta)\cos(\alpha)\\ \cos(\alpha+\beta) &= \cos(\alpha)\cos(\beta)-\sin(\alpha)\sin(\beta).\end{aligned}$$

We will use these in proving the following proposition, along with two limits we previously computed:

$$\lim_{\theta\to 0}\frac{\sin(\theta)}{\theta}=1 \quad \text{and} \quad \lim_{\theta\to 0}\frac{\cos(\theta)-1}{\theta}=0.$$

Proposition. *We have that* $\frac{d}{dx}\sin(x)=\cos(x)$ *and* $\frac{d}{dx}\cos(x)=-\sin(x)$.

Proof. We have that

$$\begin{aligned}\frac{d}{dx}\sin(x) &= \lim_{h\to 0}\frac{\sin(x+h)-\sin(x)}{h} = \lim_{h\to 0}\frac{\sin(x)\cos(h)+\sin(h)\cos(x)-\sin(x)}{h}\\ &= \lim_{h\to 0}\left(\sin(x)\frac{\cos(h)-1}{h}+\frac{\sin(h)}{h}\cos(x)\right) = \sin(x)\cdot 0+1\cdot\cos(x)=\cos(x)\end{aligned}$$

and that

$$\begin{aligned}\frac{d}{dx}\cos(x) &= \lim_{h\to 0}\frac{\cos(x+h)-\cos(x)}{h} = \lim_{h\to 0}\frac{\cos(x)\cos(h)-\sin(x)\sin(h)-\cos(x)}{h}\\ &= \lim_{h\to 0}\left(\cos(x)\frac{\cos(h)-1}{h}-\frac{\sin(h)}{h}\sin(x)\right) = \cos(x)\cdot 0-1\cdot\sin(x)=-\sin(x)\end{aligned}$$

as claimed. □

We can combine this proposition with the quotient rule to compute the derivatives of the other trig functions (and this means that as long as you remember the derivatives of sine and cosine, you're only one step away from computing the others):

$$\begin{aligned}\frac{d}{dx}\tan(x) &= \frac{d}{dx}\left(\frac{\sin(x)}{\cos(x)}\right) = \frac{\cos^2(x)+\sin^2(x)}{\cos^2(x)} = \frac{1}{\cos^2(x)} = \sec^2(x)\\ \frac{d}{dx}\cot(x) &= \frac{d}{dx}\left(\frac{\cos(x)}{\sin(x)}\right) = \frac{-\sin^2(x)-\cos^2(x)}{\sin^2(x)} = \frac{-1}{\sin^2(x)} = -\csc^2(x)\\ \frac{d}{dx}\csc(x) &= \frac{d}{dx}\left(\frac{1}{\sin(x)}\right) = \frac{0-\cos(x)}{\sin^2(x)} = -\frac{\cos(x)}{\sin(x)}\cdot\frac{1}{\sin(x)} = -\cot(x)\csc(x)\\ \frac{d}{dx}\sec(x) &= \frac{d}{dx}\left(\frac{1}{\cos(x)}\right) = \frac{0-(-\sin(x))}{\cos^2(x)} = \frac{\sin(x)}{\cos(x)}\cdot\frac{1}{\cos(x)} = \tan(x)\sec(x).\end{aligned}$$

We can also compute the higher-order trig derivatives:

$$\begin{aligned}
\frac{d^2}{dx^2}\sin(x) &= \frac{d}{dx}\cos(x) = -\sin(x) \\
\frac{d^3}{dx^3}\sin(x) &= \frac{d}{dx} -\sin(x) = -\cos(x) \\
\frac{d^4}{dx^4}\sin(x) &= \frac{d}{dx} -\cos(x) = \sin(x).
\end{aligned}$$

That $\sin(x)$ has fourth derivative equal to itself is very nice, because we can now compute all higher-order derivatives with ease. For example, the 97th derivative of $\sin(x)$ is the derivative of the 96th derivative of $\sin(x)$, and because 96 is divisible by 4, we know that $\frac{d^{96}}{dx^{96}}\sin(x) = \sin(x)$, meaning the 97th derivative is $\cos(x)$.

Now that we have all the trig derivatives, let's do a few more examples:

$$\begin{gathered}
\frac{d}{dx}(4\sin(x) + \csc(x)) = 4\cos(x) - \csc^2(x) \\
\frac{d}{dx}\left(x\sin(x) + \frac{\cos(x)}{x}\right) = \sin(x) + x\cos(x) + \frac{-x\sin(x) - \cos(x)}{x^2} \\
\frac{d}{dx}\left(\frac{\overbrace{x^5\cos(x)}^{f(x)}}{\underbrace{1+\tan(x)}_{g(x)}}\right) = \frac{\overbrace{(5x^4\cos(x) - x^5\sin(x))}^{f'(x)}\overbrace{(1+\tan(x))}^{g(x)} - \overbrace{x^5\cos(x)}^{f(x)}\overbrace{\sec^2(x)}^{g'(x)}}{(1+\tan(x))^2}
\end{gathered}$$

Notice that in this last problem, there's some simplification that could in theory be done: we have a $\cos(x)$ term times a $\sec^2(x)$ term, and so the $\cos(x)$ will cancel with one of the $\sec(x)$ to give us

$$\frac{d}{dx}\left(\frac{x^5\cos(x)}{1+\tan(x)}\right) = \frac{(5x^4\cos(x) - x^5\sin(x))(1+\tan(x)) - x^5\sec(x)}{(1+\tan(x))^2}.$$

That said, you'll never have to actually do this when taking the derivatives. The problems I assign will look a lot like this one, and the point is that you can use these rules to take derivatives, however unsimplified the answer ends up being. (Leaving it unsimplified also makes my life easier when I go to grade your work.)

Notice that we now have a derivative rule for almost all of the situations we had limit laws for—products, sums, quotients, etc.—except for compositions of functions. Recall that a composition of functions f with g is denoted $f \circ g$ and given by $(f \circ g)(x) = f(g(x))$. That is, to calculate $f \circ g$, we take our input x and apply the function g first, then the function f to the output $g(x)$. We think of g as being our "inside function" and f as being our "outside function." There are lots of processes that can be nicely modeled by function composition. For example, if you're driving in the mountains gaining 2km of altitude every hour, then if the temperature decreases by 6.5° C every km of altitude gained, we can think of this is an instance of function composition. Temperature here is a function of altitude, and altitude is in turn a function of time, giving temperature as a composite function of time. If we want to figure out the rate of change of temperature as a function of time, we sort of know intuitively to multiply the rates of change we've just been given:

$$\frac{\text{change in temp } (^\circ \text{ C})}{1 \text{ hr}} = \frac{\text{change in temp } (^\circ \text{ C})}{1\text{km}} \cdot \frac{\text{change in altitude (km)}}{1 \text{ hr}} = 6.5^\circ \text{ C/km} \cdot 2\text{km/hr}.$$

So we expect that in one hour of driving, the temperature around us will decrease by 13° C. That we can get the rate of change of a composition of functions by multiplying the rates of change of the individual contributing functions is captured in the **chain rule** for derivatives:

Chain rule: If f and g are differentiable, then $(f \circ g)'(x) = f'(g(x))g'(x)$.

Proof. We have that

$$\begin{aligned}(f \circ g)'(x) &= \lim_{h\to 0} \frac{f(g(x+h)) - f(g(x))}{h} = \lim_{h\to 0} \frac{f(g(x+h)) - f(g(x))}{g(x+h) - g(x)} \cdot \frac{g(x+h) - g(x)}{h} \\ &= \lim_{h\to 0} \frac{f(g(x+h)) - f(g(x))}{g(x+h) - g(x)} \cdot \lim_{h\to 0} \frac{g(x+h) - g(x)}{h}.\end{aligned}$$

The second limit is $g'(x)$, which we want, and to deal with the first limit, we can make the substitution $k = g(x+h) - g(x)$. Because g is differentiable, it is continuous, so $\lim_{h\to 0} k = g(x+0) - g(x) = 0$. Hence letting h approach zero is the same as letting k approach zero, so we can rewrite the first limit as

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

The first calculation therefore shows that $(f \circ g)'(x) = f'(g(x))g'(x)$, as claimed.[2] □

Here's a bunch of examples of our new rule:

$$\begin{aligned}\frac{d}{dx}(2x+1)^{100} &= 100(x+1)^{99} \cdot 2 = 200(x+1)^{99} \\ \frac{d}{dx}\sqrt{x^2 - 2} &= \frac{1}{2\sqrt{x^2-2}} \cdot 2x = \frac{x}{\sqrt{x^2-2}} \\ \frac{d}{dx}\sin^2(x) &= 2\sin(x) \cdot \cos(x) \\ \frac{d}{dx}\sin(x^2) &= \cos(x^2) \cdot 2x\end{aligned}$$

[2] In fact, there's something subtle that's wrong with this proof. This is good enough for our purposes, so we don't need to worry about it too much, but if it sounds interesting to you, see if you can figure out what goes wrong here.

$$\frac{d}{dx}\sin(\sin(x)) = \cos(\sin(x)) \cdot \cos(x)$$

$$\frac{d}{dx}\sin(\sin(\sin(x))) = \cos(\sin(\sin(x))) \cdot \cos(\sin(x)) \cdot \cos(x)$$

$$\frac{d}{dx}\sin^2(\sin(x)) = 2\sin(\sin(x)) \cdot \cos(\sin(x)) \cdot \cos(x)$$

$$\frac{d}{dx}\sin(x\sin(x)) = \cos(x\sin(x)) \cdot (\sin(x) + x\cos(x))$$

$$\frac{d}{dx}\sqrt{\frac{\sqrt{\sin(x)}}{x^2+\tan(x)}} = \frac{1}{2\sqrt{\frac{\sqrt{\sin(x)}}{x^2+\tan(x)}}} \cdot \frac{\frac{1}{2\sqrt{\sin(x)}} \cdot \cos(x) \cdot (x^2+\tan(x)) - \sqrt{\sin(x)} \cdot (2x + \sec^2(x))}{(x^2+\tan(x))^2}$$

$$\frac{d}{dx}\sec(\tan(\sin(x^2))) = \sec(\tan(\sin(x^2)))\tan(\tan(\sin(x^2))) \cdot \sec^2(\sin(x^2)) \cdot \cos(x^2) \cdot 2x.$$

2.3 Derivatives as rates of change

An important concept in physics is *speed*, which is defined to be distance covered divided by time spent. That is, $v = \frac{\Delta x}{\Delta t}$. In particular, if your position at time t is given by the function $p(t)$, then your average speed between time t_0 and time t_1 is

$$v = \frac{p(t_1) - p(t_0)}{t_1 - t_0}.$$

This formula should look familiar. It is the slope of a line through the points $(t_0, p(t_0))$ and $(t_1, p(t_1))$. It is *not* the derivative of p, because we didn't take a limit. It is instead a "difference quotient", which is really a fancy way of saying the slope of a line.

Example. For example, on Earth dropped objects fall about $p(t) = 5t^2$ meters after t seconds. The average speed between time $t = 1$ and time $t = 2$ is

$$v = \frac{p(2) - p(1)}{2-1} = \frac{20-5}{1} = 15\text{m/s}$$

and the average speed between time $t = 3$ and time $t = 1$ is

$$v = \frac{p(3) - f(1)}{3-1} = \frac{45-5}{3-1} = 20\text{m/s}.$$

It's useful here to look at the units. We know that the result is a speed, so comes out in m/s. But how do we know we get those units? We have to think a bit about what the function p is actually doing.

The function p gives us position as a function of time. Thus the *inputs* to p are given in seconds, and the *outputs* are given in meters. So it's not really fully correct to say that $p(t) = 5t^2$; that would suggest that $p(1\text{s}) = 5(1\text{s})^2 = 5\text{s}^2$. But your position isn't described in square seconds!

Instead, we would write something like $p(t\text{s}) = 5t^2\text{m}$. The function takes in seconds as inputs, and gives meters as outputs. Thus our last calculation properly should have been

$$v = \frac{p(3\text{s}) - f(1\text{s})}{3\text{s} - 1\text{s}} = \frac{45\text{m} - 5\text{m}}{3\text{s} - 1\text{s}} = 20\text{m/s}.$$

We see that the numerator—which is made up of the outputs of p—has units of meters, while the denominator, which is made up of the inputs of p, has units of seconds. So the entire fraction has units of m/s, which is what it should be.

We can give a more general formula. What's the average speed between time $t_0 = 1$ and time $t_1 = t$? We have

$$v = \frac{p(t\text{s}) - p(1\text{s})}{t\text{s} - 1\text{s}} = \frac{5t^2\text{m} - 5\text{m}}{t\text{s} - 1\text{s}} = 5(t+1)\frac{t-1}{t-1}\text{m/s}.$$

As long as $t \neq 1$, this gives us a formula for average speed between time t and time 1: the average speed is $5(t+1)$m/s. But what if we want to know the speed "at" the time $t = 1$?

On some level, this question doesn't make any sense. Speed is defined as the change in distance divided by the change in time; if time doesn't change, and distance doesn't change, then this doesn't really mean anything. Maybe what we really mean is, what's a good estimate of our average speed, as long as our time is close to $t = 1$? Our average speed depends on the exact interval we choose; the speed from $t = 1$ to $t = 2$ isn't the same as the speed from $t = 1$ to $t = 1.1$. But can we find one number that gives a good estimate?

This should make you think of limits. We can find a good estimate of the speed from time 1 to time t by taking a limit as t approaches 1. Thus we define your *instantaneous speed* or *speed at time* t_0 to be

$$\lim_{t_1 \to t_0} \frac{p(t_1) - p(t_0)}{t_1 - t_0} = \lim_{h \to t_0} \frac{p(t_0 + h) - p(t_0)}{h}.$$

Notice that since the function p has input in seconds and output in meters, the instantaneous speed will be in m/ s, as it should be. But also notice that this formula is just the definition of the derivative of p.

Thus from the previous example, we can see that the instantaneous speed at time $t_0 = 1$ is

$$v(1\text{s}) = p'(1\text{s}) = \lim_{t \to 1} 5(t+1)\frac{t-1}{t-1}\text{m/s} = 10\text{m/s}.$$

Alternatively, we know that $p(t) = 5t^2$, so by our derivative rules we know that $p'(t) = 10t$ and thus $p'(1) = 10$. Once we add units, we have $p'(t\text{s}) = 10t\text{m/s}$ and thus $p'(1\text{s}) = 10\text{m/s}$.

The derivative of a function has different units from the original function. Since the derivative is given by a formula with output in the numerator and input in the denominator, the derivative will have the units of the output per units of input.

We can take this one step further and look at the derivative of p'. The function p' takes in a time and outputs a speed; its derivative will be

$$p''(t_0\mathrm{s}) = \lim_{t \to t_0} \frac{p'(t\mathrm{s}) - p'(t_0\mathrm{s})}{t\mathrm{s} - t_0\mathrm{s}}.$$

The units of the denominator are still seconds; but the units of the top are m/s, so the second derivative takes in seconds and outputs meters per second *per second*, or m/s^2. This makes sense: the second derivative is the change in the first derivative, so p'' tells us how quickly the speed is changing. So it tells us how many meters per second your speed changes each second. This is otherwise known as "acceleration."

Once we have the speed of a particle in terms of its derivative, we can apply it to do the sort of things we've already been doing. So for instance, we can ask how far a dropped object will have fallen after 2.2 seconds. We could calculate this exactly, but we can also approximate:

$$p(2.2s) \approx p(2s) + p'(2s)(2.2s - 2s) = 20m + 10m/s(.2s) = 22m.$$

How does all this relate to linear approximation? We know that speed is change in distance over time. Another way of saying that is that our final position is our initial position, plus speed times time.

$$p(t) = p(0) + v_{\text{average}}(t - 0).$$

If our speed varies over time, this isn't terribly helpful: we can only compute average speed by knowing our initial and final position. If we only know our speed "at" each moment, this doesn't work—and making it work precisely involves *integrals*, which we will develop in later sections.

But if the length of time is small, we can make a pretty good guess by assuming our speed is constant. Thus we compute our instantaneous speed at time 0, and wee have the approximate formula

$$p(t) \approx p(0) + v_0(t - 0).$$

And this is precisely the linear approximation formula we started with when we first introduced derivatives.

This is basically how we reason about speed in real life. If you're driving fifteen miles and your friend calls you and asks how long you'll take, you might say "Well, traffic isn't too bad; I'm going about 30 miles per hour. So I should be there in about half an hour". This doesn't mean you'll get there in exactly half an hour. Traffic might get better or worse, and you might speed up or slow down. But your best guess of your average speed is your speed right now.

Of course, that's not always your best guess. If you're driving into the city you might know that you're about to hit bad traffic. Or if you can see the end of your traffic jam, you might know you're about to speed up. In either case, this is like having information about the second derivative, and you can refine your guess.

The worst-case version of this thought process is the old Windows download boxes, which would give an estimate of how long a file transfer would take. But this estimate was a simple linear approximation of remaining file size divided by your current download speed—and download speeds would vary wildly from second to second. So you'd see an estimate jump from thirty minutes to two hours to five minutes and back up to forty minutes, all within the space of thirty seconds.

So far, we used this to think about physical speed as we move from one location to another. But the same logic applies to basically any time we have a physical process with change over time. If you know how quickly the output is changing "right now", you can use that to build a linear model of what the output will look like over time. And that means that any rate of change is, fundamentally, a derivative.

Another way of thinking about the derivative is the difference between "stocks" and "flows". If your function measures the *level* or something, then the derivative measures the rate at which the level is changing. If the function measures the amount of something you have in stock, then the derivative measures the rate at which new stock is flowing in or out of your warehouse.

Example (Debt and Deficit)**.** A lot of discussions of economics and public policy address the deficit and the debt. The "deficit" and the "debt" are easy to confuse but importantly different, in a way that maps cleanly to the idea of a derivative.

A "debt" is the amount of money that is currently owed; it is measured in dollars (or euro or yen or some other currency). The current US national debt is approximately \$34 trillion.

A "deficit" is the rate at which the debt is increasing. So the national deficit is currently a little over \$1 trillion. This means we expect the debt next year to be about \$1 trillion bigger than the debt this year.

Mathematically we can define a function $D(t)$ which takes in the year and outputs the number of dollars owed. Then the annual deficit is

$$\frac{D((t+1)y) - D(ty)}{1y}.$$

This isn't a derivative, since there's no limit; this is a *difference quotient* that measures a discrete change in debt over a discrete time. It's analogous to average speed, not instantaneous speed.

But we could imagine asking how the deficit is changing from month to month, or from week to week, or from hour to hour. We can take a limit as the time between $t+h$ and t goes to zero, and then the deficit would be the derivative of debt. The function $D'(t)$ will take in years, and output dollars per year.

What about the second derivative? The function D'' will take in years, and output the yearly change in the deficit, measured in dollars per year per year. When people talk about whether the deficit is going up or down, they are looking at the second derivative of the debt.

Example (Inflation)**.** We can make a similar point about inflation, and make fun of Richard Nixon at the same time.

Roughly speaking, inflation is the change in the *price level*, which measures how the value of money changes over time. Thus inflation is a rate of change, and thus a derivative. If we oversimplify and measure the price level as the number of liters of gas you can buy with a dollar, then inflation is measured in liters per dollar per year.

In the seventies, inflation was a major political topic, because inflation was both high and rising. What does it mean to say inflation is rising? That's a *second derivative*. Inflation is the rate at which the price level is changing, but that rate is itself increasing.

In Nixon's reelection campaign, he couldn't say inflation was low, because it wasn't. And he couldn't even say it was falling, because it wasn't. So instead he said that "the rate at which the rate of inflation is increasing is decreasing". That's terrible sentence, even before we unpack it into "the rate at which the rate at which the price level is increasing is increasing is decreasing".

I've heard that this is the only known use of the third derivative in political messaging.

Both of these examples have one very important trait in common. The position function $p(t)$ and the debt function $D(t)$ output different types of things with different units, but they both take *time* as an input. But it's easy for a function to take inputs other than time, and these functions are often physically important and meaningful.

One common place they show up is in economics. Economics cares a lot about so called "marginal" effects.

Example (Marginal Revenue)**.** If you're deciding how many machines to buy, what really matters isn't the total cost of the machines and the total revenue they'll make you. Instead, you need to ask how much more you'll have to spend to get *one* more machine, and how much more revenue that one machine will get you. (This is called "marginal thinking", because we care about the effect of getting one more machine on the margin.)

Any of these marginal effects are implicitly asking for a derivative. So suppose we have some revenue curve where $R(m) = 100m - m^2$: your total revenue is \$100 for every machine, minus upkeep costs of the square of the number of machines you have. So with one machine, you make \$99; with two machines, you make \$196; with ten machines you make \$900. The units of the input is "machines" and the units of the output are "dollars".

We compute $R'(m) = 100 - 2m$; each new machine adds roughly \$100 of revenue, minus 2 times the number of machines you already have. Thus the marginal revenue of the first machine is about \$98, and the marginal revenue of the tenth machine is about \$80. We can see that the fiftieth machine has a marginal revenue of \$0; this is our break-even point, where adding another machine neither helps nor hurts. The sixtieth machine has a marginal revenue of about $-$\$20, and we actually lose money by adding it! The units of this derivative are "dollars per machine"; how many more dollars will you get by adding a machine?

But of course the actual revenue of 50 machines is $R(50) = 5000 - 2500 = 2500$ dollars. The actual revenue of 60 machines is $R(60) = 6000 - 3600 = 2400$ dollars, which is less than $R(50)$ but still positive.

Example (Marginal Cost)**.** We also often talk about marginal cost. Suppose the cost of buying m machines is $C(m) = 5000 + 10m + .05m^2$. There's some start-up cost to having any machines at all; then each machine costs a bit more than the previous one. The units of the input are "machines" and the units of output are "dollars".

We can see that $C(1) = 5010.05$, and $C(10) = 5105$. Even $C(100) = 6500$ is not that much bigger than $C(1)$.

The marginal cost would be $C'(m) = 10 + .1m$. We have to pay a huge sum to have any machines at all, but each new machine we add costs only 10 plus a tenth of the number of machines we have. So the cost of adding the hundredth machine is about $C'(100) = 20$, which checks out with the numbers we computed earlier. The units of the derivative are, again, dollars per machine.

This shows a really big separation between marginal and average cost. The total cost of all our machines is really high; if this cost is paired with the revenue from the previous example, we'll continually lose money no matter what we do. But once we've already eaten our sunk costs, the marginal cost of adding one more machine is pretty low, so we should go ahead and get a lot of them.

Example (Price Elasticity of Demand)**.** Another common economics question is to see how the demand for a product relates to its price. We can define a function $Q(p)$ that takes in a price in dollars, and outputs the quantity of items that will be bought. So if $Q(p) = 10000 - 10p$, this means that if the price is \$100 then people will buy $Q(100) = 10000 - 1000 = 9000$ widgets.

What's the derivative here? The function $Q'(p)$ takes in a price in dollars and outputs a number of widgets per dollar. It tells you how the quantity demanded changes in response to changes in the price. Thus we see that since $Q'(p) = -10$, we expect to sell ten fewer widgets for each dollar we raise the price.

(Economists call this the Price Elasticity of Demand: "elasticity" is how quickly one thing responds to changes in another thing. So any time the term "elasticity" shows up in economics, there's a derivative inolved somewhere.)

What if instead we had the function $Q(p) = 10000 - 5p^2$? Now we see that changing the price doesn't have a huge effect if the price is already small, but it has a dramatic effect if the price is big. We compute that $Q'(p) = -10p$. This means that increasing the price by one dollar will decrease the quantity demanded by ten widgets for every dollar of the price.

Thus if the current price is \$10, we expect raising the price to \$11 to reduce sales by about a hundred widgets. If the current price is \$30 then raising the price will lose us nine hundred widgets in sales.

2.4 More on linear approximation

When we first defined the derivative in section 2.1, our motivation came from finding good linear approximations and lines tangent to graphs. We want to revisit these ideas now that we have some tools for actually computing derivatives.

We know that if we have a function $f(x)$ and know what it looks like at a point a, we can use the derivative to give a linear (first-order) approximation

$$f(x) \approx f(a) + f'(a)(x-a).$$

Example. We can find an estimate of $(2.1)^5$.

This means approximating the function x^5 near the input $x = 2$, and we know how to calculate 2^5. In fact, 2^5 is itself an approximation to $(2.1)^5$, and we had been calling this the *zeroth-order approximation*. We can now use the derivative to refine this estimate. We take $f(x) = x^5$ and $a = 2$. Then $f'(x) = 5x^4$, so we have $f(2) = 32, f'(2) = 80$, and

$$f(2.1) \approx 80(2.1-2) + 32 = 40.$$

The exact answer is 40.841, so this estimate is pretty good!

What if we approximate $(2.5)^5$ using $a = 2$. What if we approximate 3^5? We have

$$\begin{aligned} (2.5)^5 &\approx 80 \cdot (2.5-2) + 32 = 72 \\ 3^5 &\approx 80 \cdot (3-2) + 32 = 112. \end{aligned}$$

The true answers are 97.6563 and 243. These estimates are not especially good. This is because 3 is actually not very close to 2—especially proportionately. Of course, it's not that hard to compute 3^5 directly.

These methods are best when $x - a$ is very small relative to everything else. We often use them in the real world for $x - a < .1$ or so.

Example. Let's approximate $\sqrt[3]{28}$ and $\sqrt[4]{82}$.

We take $a = 27$ and $a = 81$ respectively.

$$\begin{aligned} \sqrt[3]{28} &\approx \frac{1}{3}(27)^{-2/3}(28-27) + 3 = \frac{1}{27} + 3 \approx 3.03704 \\ \sqrt[4]{82} &\approx \frac{1}{4}(81)^{-3/4}(82-81) + 3 = \frac{1}{108} + 3 \approx 3.00926. \end{aligned}$$

The true answers are approximately 3.03659 and 3.00922 respectively.

Now we'll approximate 28^3 and 82^4 using the same base points

We have

$$28^3 \approx 3(27)^2(28-27) + 27^3 = 21870$$
$$82^4 \approx 4(81)^3(82-81) + 81^4 = 45172485$$

In contrast the true answers are 21952 and 45172485.

These approximations aren't *terrible* but they aren't very good either. Since the derivatives themselves are changing quickly here (the second derivatives are $6 \cdot 27$ and $12 \cdot 81^2$ respectively), the approximation won't be very good—a tangent line is going to match the function best when the function itself sort of looks like a line, i.e. when the second derivative is really small.

There are a few specific linear approximation *formulas* that come up really frequently in other applications, enough to get their own names. Let's take a look at each of them.

Example (Binomial Approximation)**.** As a warmup, let's approximate $(1.01)^{10}$. Our function is $f(x) = x^{10}$ and our $a = 1$. So $f(a) = 1$ and $f'(a) = 10a^9 = 10$. Then we have

$$f(1.01) \approx 10(1.01-1) + 1 = 1.1.$$

The true answer is about 1.10462.

Now let's approximate $(1.01)^\alpha$ where $\alpha \neq 0$ is some constant. (The letter α is a Greek lower-case "a". We'll use it here instead of the friendlier n because it's fairly standard for the formula we're developing.)

We have $f(x) = x^\alpha$, so $f'(x) = \alpha x^{\alpha-1}$. We again have $f(1) = 1$ and $f'(1) = \alpha(1)^{\alpha-1} = \alpha$, so

$$f(1.01) \approx \alpha(1.01-1) + 1 = 1 + \alpha/100.$$

Now let's get the fully general useful formula: approximate $(1+x)^\alpha$ where x is some small number and $\alpha \neq 0$ is a constant. (This rule is the called the "binomial approximation" and is often useful in physics and engineering).

We still take $f(x) = x^\alpha$ and $a = 1$. But we compute

$$f(1+x) \approx 1 + \alpha(1+x-1) = 1 + \alpha x.$$

So the binomial approximation is that for small x, $(1+x)^\alpha \approx 1 + \alpha x$.

Example (Small Angle Approximation)**.** Let's find a formula to approximate $\sin(x)$ when x is small. Recall from the section on trig limits that we already had a "Small Angle Approximation." There, this was phrased in terms of limits, and we said that

$$\lim_{x \to 0} \frac{\sin(x)}{x} = 1$$

meant that for x small, $\sin(x) \approx x$.

We see now that using linear approximations gives us the same result: we take $a = 0$, and then since $\sin'(x) = \cos(x)$ and so $\sin'(0) = \cos(0) = 1$, we have

$$\sin(x) \approx 1(x - 0) + 0 = x.$$

Thus for small angles, $\sin(x)$ is approximately just x! For instance, our formula says that $\sin(.05) \approx .05$, where the true answer is about .04998. So this is pretty good. In fact, we compute that $\sin''(0) = -\sin(0) = 0$. Since the second derivative is zero, we expect the linear approximation to work well.

That means that in a lot of calculations, if we have a formula with a lot of sines in it, as long as our angles are small we can replace every $\sin(x)$ with an x without losing too much. And that's much easier to think about.

We can do the same thing for cosine. We compute that $\cos'(x) = -\sin(x)$ so $\cos'(0) = 0$. Then

$$\cos(x) \approx 0(x - 0) + 1 = 1.$$

This is actually a constant! The line that fits $\cos(x)$ best near 0 is just the horizontal line $y = 1$.

We can calculate, e.g., that $\cos(.05) \approx 1$, where the true answer is about .9986. This is also pretty good, but the approximation isn't quite as good as the one for sine. We compute that $\cos''(0) = -\cos(0) = -1$; while the second derivative isn't huge, it isn't trivial either.

Example (Geometric Series)**.** Let's find a formula to linearly approximate $f(x) = \dfrac{1}{1-x}$ near $x = 0$.

We compute that $f'(x) = (1-x)^{-2} = \frac{1}{(1-x)^2}$. Then

$$f(x) \approx 1 + x.$$

This is a special case of what's known as the geometric series formula.

You might ask why we did the slightly funky $\frac{1}{1-x}$ instead of the more normal $\frac{1}{x}$. After thinking about it for a bit, you'll notice that wee can't approximate $\frac{1}{x}$ near zero at all! We see that f is undefined at 0, and equally importantly, $f'(x) = -1/x^2$ is also undefined at zero. So there's no linear approximation.

But if we want to,, we can linearly approximate $f(x) = 1/x$ near 1. We have $f(1) = 1$ and $f'(1) = -(1)^{-2} = -1$ so

$$f(x) \approx 1 - (x - 1) = 2 - x.$$

Finally, a bonus fun fact to notice:

Example. Let's find a formula to approximate $f(x) = x^3 + 3x^2 + 5x + 1$ near $a = 0$. What do you notice? Why does that happen?

We have $f(0) = 1$ and $f'(x) = 3x^2 + 6x + 5$ so $f'(0) = 5$. Thus

$$f(x) \approx 1 + 5x.$$

This is exactly what you get if you take the original polynomial and cut off all the terms of degree higher than 1.

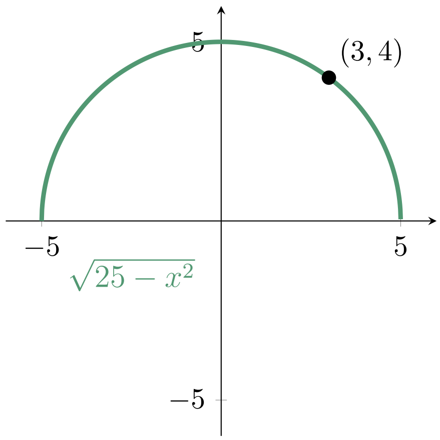
This makes sense, because we're looking for the closest we can get to f without using terms of degree higher than 1.

2.5 Implicit differentiation

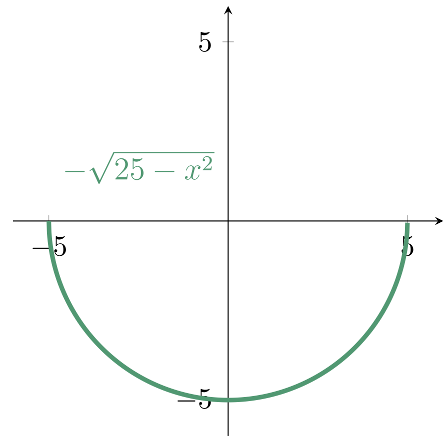
Up to this point, we've been discussing *functions*. When we initially motivated the derivative, it was through linear approximation, which led us to the idea of finding a line tangent to the graph of the function we were trying to approximate. But there are lots of curves in the plane that aren't the graphs of any function, an example you're likely familiar with from high school being a circle. Recall that $x^2 + y^2 = r^2$ gives a circle of radius r centered at $(0, 0)$ in the xy plane. We'd like to be able to use the derivative tools we've just developed

This isn't a function, though, and right now we only know how to find lines tangent to graphs of functions. A function is supposed to be a rule that turns input into a single output, and if you provide me with an input x, there's no rule here I can use to turn it into an output y. Sometimes we have equations that are written in this non-function format, e.g. $x^2 + y = 25$, and we can rearrange them (solving for y) to write them in a format that makes it clear we really can think of y as a function of x. In this example, the equation is equivalent to $y = 25 - x^2$, which *does* give a rule for turning inputs x into outputs y. Written in this way, we say that the equation $x^2 + y = 25$ gives y as a function of x **implicitly**.

If we go back to our circle equation, say $x^2 + y^2 = 25$, this looks pretty similar, and we could try to solve for y again, maybe saying that $y^2 = 25 - x^2$, so $y = \pm\sqrt{25 - x^2}$. This is okay, although we've turned a pretty simple equation into something a bit more complicated. Still, if we want to find a line tangent to the circle at the point $(3, 4)$ (which is on the circle), we can use this to make some progress. The positive square root describes the top half of the circle, and the negative square root describes the bottom half. Our point of interest $(3, 4)$ lies on the top half, so we can sort of forget about the bottom half and just pretend we're working with the equation $f(x) = \sqrt{25 - x^2}$ from the start, trying to find a tangent line at $a = 3$.



Top half: contains point, can find tangent

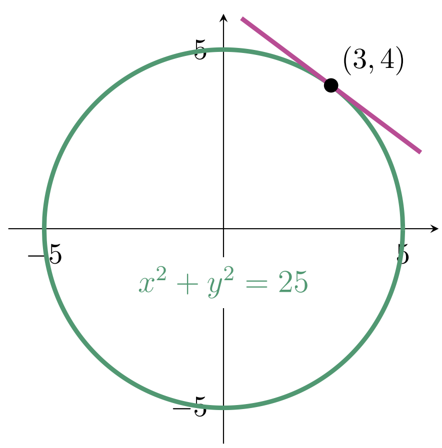


Bottom half: doesn't contain point, can ignore

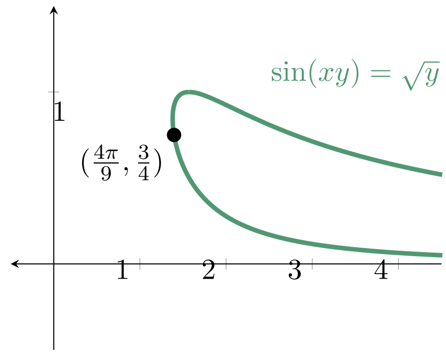
We have our tangent line formula $y = f(a) + f'(a)(x - a)$, so we just need to determine $f'(3)$. We have

$$f'(x) = \frac{1}{2\sqrt{25 - x^2}} \cdot (-2x), \quad \text{so} \quad f'(3) = \frac{-2(3)}{2\sqrt{25 - (3)^2}} = \frac{-3}{\sqrt{16}} = -\frac{3}{4}$$

and the tangent line is therefore $y = 4 - \frac{3}{4}(x - 3)$. Now we can just put the pieces of the circle back together and graph our tangent line to get a picture that looks like the following:



This is feasible for this particular example, but it relied on being able to solve for y as a function of x. If I have an equation like $\sin(xy) = \sqrt{y}$ (pictured below) it suddenly becomes much more difficult to perform the same procedure, even though looking at the picture makes it clear there's a well-defined tangent line at, e.g. the point $(\frac{4\pi}{9}, \frac{3}{4})$, which does lie on the curve (do check this for yourself).



The key idea of what we did in the circle example was to use fact that we *can* write y as a function of x near the point we're interested in. Assuming we can do this, we can write $x^2 + y(x)^2 = 25$. Whatever the function $y(x)$ actually is, we know that the functions $x^2 + y(x)^2$ and 25 are equal. This means their derivatives must also be equal! Hence we can try and solve for $y'(x)$ by taking the derivative of both sides. If we do this, we note that $(y(x))^2$ is function composition with the squaring function being the outside and y being the inside, so we need the chain rule. The derivative of this term is therefore $2y(x)y'(x)$, and so the equation we now have after differentiating both sides is $2x + 2y(x)y'(x) = 0$. Now all we have to do is solve for $y'(x)$:

$$2y(x)y'(x) = -2x \implies y'(x) = -\frac{2x}{2y(x)} = -\frac{x}{y(x)}.$$

Now to find the slope of the line tangent to the curve at $(3, 4)$, we just use that $y'(3) = -\frac{3}{4}$ and get the same $y = 4 - \frac{3}{4}(x - 3)$ equation for the tangent line we did before. And doing all of this only required that we know there's some way to write down y as a function of x, we didn't have to actually calculate that nearby $(3, 4)$, the curve $x^2 + y^2 = 25$ is the same as $y = \sqrt{25 - x^2}$ like we did previously.

Approaching the problem in this way is called **implicit differentiation**, and it allows us to find slopes of all sorts of curves, not just functions. Here's another example, this time we won't try to find just one tangent line, but instead we'd like a general formula for the slope of the curve given by $2xy^2 = 2x^3 - y^4$, so that if we have a point (x_0, y_0) that sits on the curve, we can just plug x_0 and y_0 into this formula to get the slope at that point. Rephrasing slightly, we want a formula for $\frac{dy}{dx}$ (i.e. $y'(x)$) in terms of x and y.

We begin by differentiating both sides, keeping in mind that every time we see a y, this is really some function of x and so to take it's derivative we need an extra chain rule. Doing this, we get that

$$2y^2 + 2x \cdot 2y\frac{dy}{dx} = 6x^2 - 4y^3\frac{dy}{dx}.$$

We can then solve for $\frac{dy}{dx}$, starting by collecting all the $\frac{dy}{dx}$ terms on one side:

$$4xy\frac{dy}{dx} + 4y^3\frac{dy}{dx} = 6x^2 - 2y^2.$$

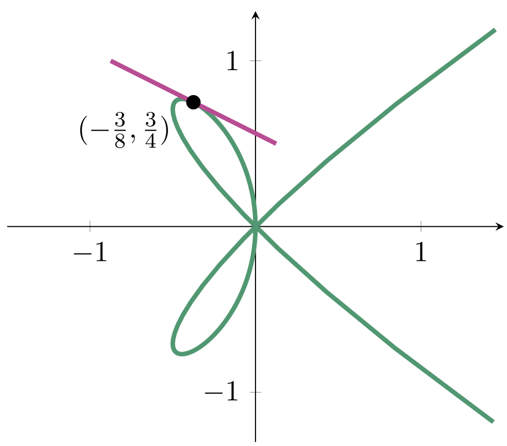
We can then factor out the $\frac{dy}{dx}$ and divide to get our expression for the slope anywhere on the curve:

$$(4xy + 4y^3)\frac{dy}{dx} = 6x^2 - 2y^2, \quad \text{so} \quad \frac{dy}{dx} = \frac{6x^2 - 2y^2}{4xy + 4y^3} = \frac{3x^2 - y^2}{2xy + 2y^3}.$$

If we then want to know the slope of the curve at the point $(-\frac{3}{8}, \frac{3}{4})$, for example, we can then plug these coordinates in:

$$\frac{3\left(-\frac{3}{8}\right)^2 - \left(\frac{3}{4}\right)^2}{2\left(-\frac{3}{8}\right)\left(\frac{3}{4}\right) + 2\left(\frac{3}{4}\right)^3} = \frac{\frac{27}{64} - \frac{9}{16}}{-\frac{18}{32} + \frac{54}{64}} = \frac{\frac{27}{64} - \frac{36}{64}}{\frac{54}{64} - \frac{36}{64}} = \frac{-\frac{9}{64}}{\frac{18}{64}} = -\frac{1}{2}.$$

The tangent line is then $y = (3/4) - (1/2)(x + (3/8))$. The algebra isn't the nicest looking thing, but the picture is just fine:



For another example, we return to the question of $\sin(xy) = \sqrt{y}$. We use the same procedure, and take derivatives on both sides to get that

$$\begin{aligned}
\cos(xy) \cdot \left(y + x\frac{dy}{dx}\right) &= \frac{1}{2\sqrt{y}} \cdot \frac{dy}{dx} \\
y\cos(xy) + x\cos(xy)\frac{dy}{dx} &= \frac{1}{2\sqrt{y}} \cdot \frac{dy}{dx} \\
y\cos(xy) &= \left(\frac{1}{2\sqrt{y}} - x\cos(xy)\right)\frac{dy}{dx} \\
\frac{y\cos(xy)}{\frac{1}{2\sqrt{y}} - x\cos(xy)} &= \frac{dy}{dx}.
\end{aligned}$$

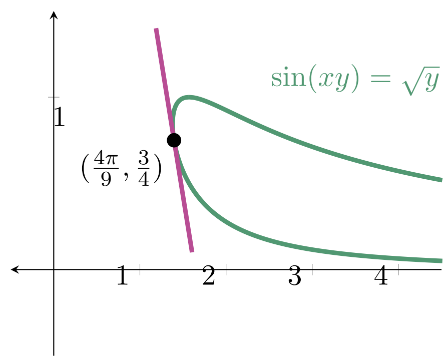
If we'd now like to find the slope of the line tangent to $\sin(xy) = \sqrt{y}$ at the point $(\frac{4\pi}{9}, \frac{3}{4})$, we just need to plug in these coordinates to the above:

$$\left.\frac{dy}{dx}\right|_{(\frac{4\pi}{9},\frac{3}{4})} = \frac{\frac{3}{4}\cos\left(\frac{4\pi}{9}\cdot\frac{3}{4}\right)}{\frac{1}{2\sqrt{\frac{3}{4}}} - \frac{4\pi}{9}\cos\left(\frac{4\pi}{9}\cdot\frac{3}{4}\right)} = \frac{\frac{3}{4}\cos\left(\frac{\pi}{3}\right)}{\frac{1}{\sqrt{3}} - \frac{4\pi}{9}\cos\left(\frac{\pi}{3}\right)} \cdot = \frac{\frac{3}{4}\cdot\frac{1}{2}}{\frac{1}{\sqrt{3}} - \frac{4\pi}{9}\cdot\frac{1}{2}} = \frac{27}{24\sqrt{3} - 16\pi}$$

where in the very last step we've multiplied the numerator and denominator both by 72. This result is about -3.1. Hence the tangent line equation we get is

$$y = \frac{3}{4} + \frac{27}{24\sqrt{3} - 16\pi}\left(x - \frac{4\pi}{9}\right),$$

and if we graph this together with the equation $\sin(xy) = \sqrt{y}$, we get the following:



We can also take second derivatives in this way. The calculation will be a little bit easier if we return to the previous example $2xy^2 = 2x^3 - y^4$, where we found the first derivative was

$$\frac{3x^2 - y^2}{2xy + 2y^3}.$$

To take a another derivative, we use the quotient rule on the expression we just found:

$$\frac{d^2y}{dx^2} = \frac{(6x - 2y\frac{dy}{dx})(2xy + 2y^3) - (3x^2 - y^2)(2y + 2x\frac{dy}{dx} + 6y^2\frac{dy}{dx})}{(2xy + 2y^3)^2}.$$

This gives us a formula for the second derivative in terms of x, y, and $\frac{dy}{dx}$, but we'd really like a formula just in terms of x and y so that knowing the coordinates of a point on the curve is enough to calculate the second derivative completely. Fortunately, we do have a formula for $\frac{dy}{dx}$ in terms of x and y, so we can just plug that in now, and get the following mess:

$$\frac{d^2y}{dx^2} = \frac{\left(6x - 2y\left(\frac{3x^2-y^2}{2xy+2y^3}\right)\right)(2xy + 2y^3) - (3x^2 - y^2)\left(2y + 2x\left(\frac{3x^2-y^2}{2xy+2y^3}\right) + 6y^2\left(\frac{3x^2-y^2}{2xy+2y^3}\right)\right)}{(2xy + 2y^3)^2}.$$

If we plug in the point $(-\frac{3}{8}, \frac{3}{4})$ we looked at earlier, we see that the second derivative comes out to be -5. The fact that this number is negative should make sense from the picture, since it tells us that as x increases, the slope is going to decrease. Indeed, the graph appears to get even steeper in the negative direction as x increases.

2.6 Related rates

Finally, let's apply a version of implicit differentiation to physical problems, or word problems.

It's good to take a moment here to talk about why we do word problems, and how to approach them. On a philosophical level, math does not tell us anything about the physical world. It only tells us that if certain properties hold, other things also have to be true. It's our job to take the aspect of the world we care about and translate it into math. Then we can see what the math implies, and hopefully that will still be true when translated back into the world.

Word problems are training for this process. We take verbal (or pictorial etc.) information, and try to turn it into a mathematical description. Then we see the mathematical consequences, and translate those back into a verbal description of physics.

So how do we approach this? **Checklist of steps for solving word problems:**

1. Draw a picture.
2. Think about what you expect the answer to look like. What is physically plausible?
3. Create notation, choose variable names, and label your picture.
   (a) Write down all the information you were given in the problem.
   (b) Write down the question in your notation.
4. Write down equations that relate the variables you have.
5. Abstractly: "solve the problem." (In the context of related rates: differentiate your equation.)
6. Plug in values and read off the answer.
7. Do a sanity check. Does you answer make sense? Are you running at hundreds of miles an hour, or driving a car twenty gallons per mile to the east?

Example. A spherical balloon is inflating at $12\text{cm}^3/\text{s}$. How quickly is the radius increasing when the radius is 3cm?

In a perfect world, we want to relate the rate at which the volume is changing to the rate at which the radius is changing. But we don't have any formulas lying around that relate those rates. What we *do* have are formulas that relate the levels: we can relate the "current" volume of the sphere to the "current" radius.

Specifically, we know that a sphere has volume $V = \frac{4}{3}\pi r^3$. Then we can differentiate both sides of that equation: using the chain rule, we compute that

$$\frac{dV}{dt} = \frac{d}{dt}\left(\frac{4}{3}\pi r^3\right) = 4\pi r^2 \frac{dr}{dt}.$$

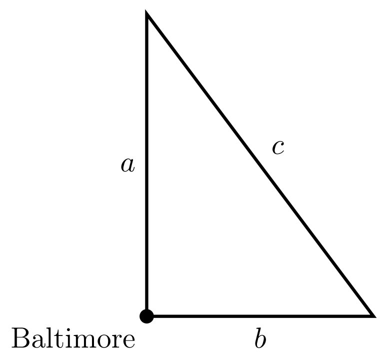
We know that $\frac{dV}{dr} = 12\text{cm}^3/\text{s}$, and $r = 3\text{cm}$. Substituting those facts in gives us

$$12\text{cm}^3/\text{s} = 4\pi(3\text{cm})^2 r'$$
$$r' = \frac{1}{3\pi}\text{cm/s}$$

So the radius is increasing by $1/3\pi$ cm per second.

Example. Suppose one car leaves Baltimore at noon, heading due north at 40 mph, and at 1 PM another car leaves Baltimore heading due west at 60 mph. At 2PM, how quickly is the distance between them increasing?

Write a for the distance the first car has traveled, and b for the distance the second car has traveled. We have that $a = 80\text{mi}, b = 60\text{mi}, a' = 40\text{mi/h}, b' = 60\text{mi/h}$. We want a formula that will relate the distances the cars have traveled to the distance between them; after drawing a picture we see this is the pythagorean theorem $a^2 + b^2 = c^2$, where c is the distance between the two cars:



Differentiating both sides of the Pythagorean Theorem with respect to t (thinking of a, b, and c all as being functions of time t), we get

$$c^2 = a^2 + b^2$$
$$2c\frac{dc}{dt} = 2a\frac{da}{dt} + 2b\frac{db}{dt}.$$

We can use the pythagorean theorem to tell that $c = 100\text{mi}$ when it's 2PM, and thus we get

$$2 \cdot (100\text{mi}) \cdot \frac{dc}{dt} = 2 \cdot (80\text{mi}) \cdot (40\text{mi/h}) + 2 \cdot (60\text{mi}) \cdot (60\text{mi/h})$$
$$200\text{mi}\frac{dc}{dt} = 6400\text{mi}^2/\text{h} + 7200\text{mi}^2/\text{h}$$
$$\frac{dc}{dt} = 68\text{mi/h}$$

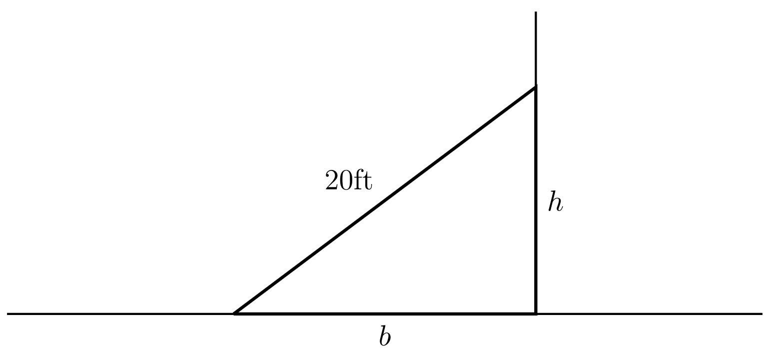
so the distance between the cars is increasing at 68 mph.

The last thing we want to do is ask ourselves if this answer seems basically reasonable. The units are correct; we are expecting the distance to be increasing, so that checks out; and the size of the answer seems basically reasonable, because the cars are traveling at 40 mph and 60 mph and 68 is on the same scale as 40 and 60.

In the picture for the above example, note that we've labeled all the sides of the triangle with the variable names rather than the lengths we know they have at 2PM. This is because we only know those lengths at a fixed moment in time, and we prefer to have a picture that captures the situation well at *any* moment in time. In the next example, we will again use variable names when the lengths are changing, and we'll see an example where it makes sense to label part of the picture with a concrete measurement (20ft) since this represents something that does not change with time.

Example. A twenty foot ladder rests against a wall. The bit on the wall is sliding down at 1 foot per second. How quickly is the bottom end sliding out when the top is 12 feet from the ground?

Let h be the height of the ladder on the wall, and b be the distance of the foot of the ladder from the wall. Here's the picture:



Then at this particular moment in time when the top of the ladder is 12ft above the ground, we see that $h = 12$, $h' = -1$, and $b = \sqrt{400 - 144} = 16$. We have

$$\begin{aligned} h^2 + b^2 &= 400 \\ 2hh' + 2bb' &= 0 \\ 2 \cdot 12 \cdot (-1) + 2 \cdot 16 \cdot b' &= 0 \\ b' &= \frac{24}{32} = \frac{3}{4} \end{aligned}$$

so the foot of the ladder is sliding away from the wall at $\frac{3}{4}$ ft/s. Again, the direction of the sliding is correct (away from the wall), and the number seems plausible.

Example. A rectangle is getting longer by one inch per second and wider by two inches per second. When the rectangle is 5 inches long and 7 inches wide, how quickly is the area increasing?

We have $\ell = 5\text{in}, w = 7\text{in}, \frac{d\ell}{dt} = 1\text{in/s}, \frac{dw}{dt} = 2\text{in/s}$. We can relate all our quantities with the formula for the area of a rectangle: $A = \ell w$ relates the area, which we want to know about, to the length and width, which we do know about.

Taking a derivative gives us

$$\begin{aligned}\frac{dA}{dt} &= \ell\frac{dw}{dt} + w\frac{d\ell}{dt} \\ &= 5\text{in} \cdot 2\text{in/s} + 7\text{in} \cdot 1\text{in/s} \\ &= 17\text{in}^2/\text{s}.\end{aligned}$$

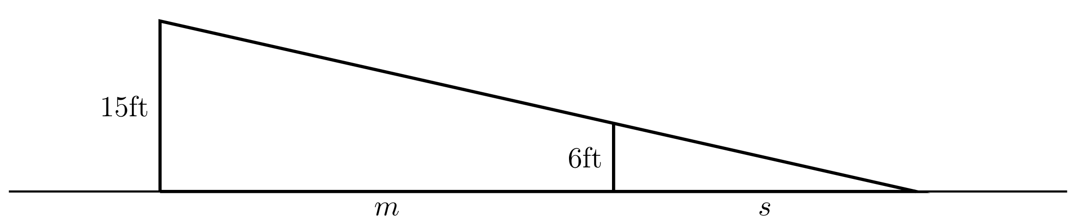
The units are right (the rate at which area is changing per second), and the direction is right (the area should be increasing, and this derivative is positive). It's really hard to see if the size is right using our intuition; people in general have bad intuition for the rate at which area changes in response to lengths.

But we can ask what would happen after a full second. One second later, we'd have $\ell = 6\text{in}$ and $w = 9\text{in}$ for a total area of 54in^2. This is an increase of 19in^2 over our starting area of 35in^2, and 17 is a pretty good approximation of 19.

As one final note, this is a problem we've basically seen before, in a different guise. The derivative of the area formula is just the product rule; we saw basically this same picture during the proof of the product rule in section 2.2.

Example. A street light is mounted at the top of a 15-foot-tall pole. A six-foot-tall man walks straight away from the pole at 5 feet per second. How fast is the distance between the pole and the tip of his shadow changing when he is forty feet from the pole?

There are actually two ways to set this up. The more obvious is to find an equation that will relate the length of the man's shadow to his distance from the pole, because we know how quickly the man is moving and we want to know how the shadow is changing.



We see that we have a similar triangles situation, so if we say that m is the distance between the man and the pole, and s is the length of his shadow, we get the equation

$$\frac{6\text{ft}}{15\text{ft}} = \frac{s}{s+m}.$$

We know from the problem statement that $\frac{dm}{dt} = 5\text{ft/s}$, and we are told we are interested in the specific moment in time when $m = 40\text{ft}$. We could differentiate the equation above using the quotient rule, but it's way easier if we collect terms first:

$$\begin{aligned}
6(s+m) &= 15s \\
6m &= 9s \\
6\frac{dm}{dt} &= 9\frac{ds}{dt} \\
\frac{2}{3} \cdot 5\text{ft/s} &= \frac{ds}{dt}.
\end{aligned}$$

So the shadow is growing at a rate of $\frac{ds}{dt} = \frac{10}{3}\text{ft/s}$.

However, that is *not* the answer to the question we were asked! We don't want to know how fast the shadow is growing; we want to know how fast the tip of the shadow is moving away from the pole. So we need to add $\frac{ds}{dt}$ the rate at which the shadow is growing, to the rate at which the base of the shadow is moving away from the pole, which is $\frac{dm}{dt}$. So our final answer is that the tip of the shadow is moving away from the pole at $(5 + 10/3)\text{ft/s} = \frac{25}{3}\text{ft/s}$.

But once we realize that $\frac{ds}{dt}$ isn't actually the thing we need to know, maybe we can set the question up differently. Let m be the distance between the man and the pole, and let L be the distance from the pole to the tip of the shadow—which is the thing that we actually care about. We can make the same similar triangles equation, but this time we get

$$\begin{aligned}
\frac{6\text{ft}}{15\text{ft}} &= \frac{L-m}{L} \\
6L &= 15(L-m) \\
15m &= 9L \\
15\frac{dm}{dt} &= 9\frac{dL}{dt} \\
15 \cdot 5\text{ft/s} &= 9\frac{dL}{dt} \\
\frac{dL}{dt} &= \frac{25}{3}\text{ft/s}
\end{aligned}$$

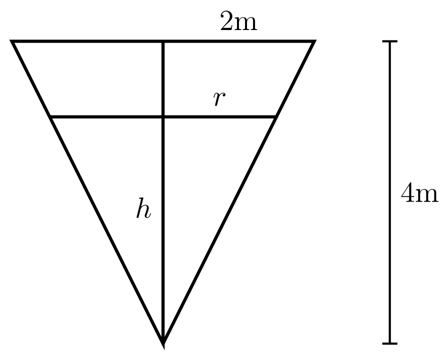
and thus the distance between the pole and the tip of the shadow is increasing at $\frac{25}{3}$ feet per second.

Example. An inverted conical water tank with radius 2m and height 4m is being filled with water at a rate of $2\text{m}^3/\text{min}$. How fast is the water rising when the water level is 3m high in the tank?

We know that we want to relate the height of the water h to the volume of water V. The obvious equation to use here is the formula for the volume of a cone:

$$V = \frac{1}{3}\pi r^2 h.$$

But this has a problem; in addition to V and h, this equation includes an r, which we don't know anything about. (The problem gives us a radius, but it's the radius of the *tank*, not the water filling the tank.) Here's the (flattened) picture of the situation:



The more naive approach is to plunge boldly ahead. We can take a derivative, and we get

$$\begin{aligned}\frac{dV}{dt} &= \frac{\pi}{3}\left(2r\frac{dr}{dt}h + r^2\frac{dh}{dt}\right)\\ 2\mathrm{m}^3/\min &= \frac{\pi}{3}\left(2r\frac{dr}{dt}\cdot 3m + r^2\frac{dh}{dt}\right),\end{aligned}$$

but we still don't have values for r or $\frac{dr}{dh}$.

We need a new equation, that will relate r to something we already know about. But we know that the water is *in the conical tank*, and should have the same shape. In particular, the sides of our cones are similar triangles! The ratio of the radius of the tank to the height of the tank must be the same as the ratio of the radius of the water to the height of the water. So we get

$$\begin{aligned}\frac{2\mathrm{m}}{4\mathrm{m}} &= \frac{r}{h}\\ r &= \frac{h}{2} \qquad \big(\text{which is } \frac{3}{2}\mathrm{m} \text{ at the moment we're interested in}\big)\\ \frac{dr}{dt} &= \frac{1}{2}\frac{dh}{dt}.\end{aligned}$$

We can substitute this back into our original equation, and we get

$$\begin{aligned}2\mathrm{m}^3/\min &= \frac{\pi}{3}\left(2\cdot\frac{3}{2}\mathrm{m}\cdot\frac{1}{2}\frac{dh}{dt}\cdot 3\mathrm{m} + \left(\frac{3}{2}\mathrm{m}\right)^2\frac{dh}{dt}\right)\\ &= \frac{\pi}{3}\left(\frac{9}{2}\frac{dh}{dt}\mathrm{m}^2 + \frac{9}{4}\frac{dh}{dt}\mathrm{m}^2\right)\\ &= \frac{9\pi}{4}\mathrm{m}^2\frac{dh}{dt}\\ \frac{dh}{dt} &= \frac{8}{9\pi}\mathrm{m}/\min.\end{aligned}$$

So we conclude that the water level is rising at $\frac{8}{9\pi}$ meters per minute.

However, while that worked, it was a huge mess algebraically. If we're smart we can do this much more easily. We start with the volume equation

$$V = \frac{1}{3}\pi r^2 h.$$

At this point, we can *notice* that the r will be a problem, so we get rid of it now. We make the similar triangles and see that

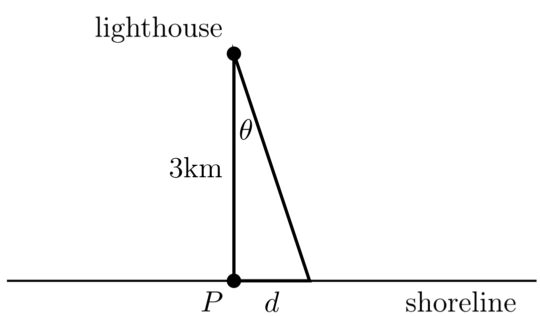
$$\begin{aligned} \frac{2\text{m}}{4\text{m}} &= \frac{r}{h} \\ r &= \frac{h}{2} \end{aligned}$$

and substituting that back into the original equation gives

$$\begin{aligned} V &= \frac{1}{3}\pi\left(\frac{h}{2}\right)^2 h = \frac{\pi}{12}h^3 \\ \frac{dV}{dt} &= \frac{\pi}{12}3h^2\frac{dh}{dt} \\ 2\text{m}^3/\min &= \frac{\pi}{4}\cdot(3\text{m})^2\frac{dh}{dt} \\ \frac{8}{9\pi}\text{m}/\min &= \frac{dh}{dt}. \end{aligned}$$

So we conclude that the water level is rising at $\frac{8}{9\pi}$ meters per minute.

Example. A lighthouse is located three kilometers away from the nearest point P on shore, and its light makes four revolutions per minute. How fast is the beam of light moving along the shoreline 1 kilometer from P?

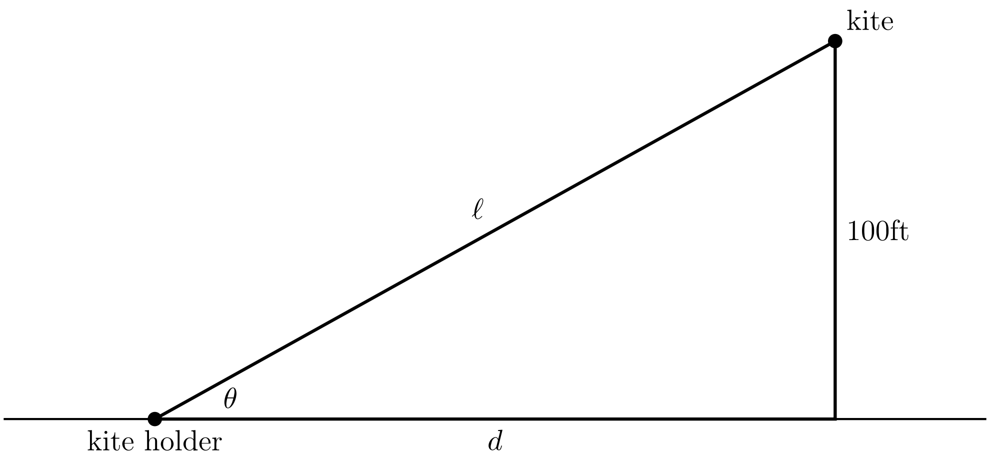


Let's say the angle of the light away from P is θ, and the distance from P is d. Then we're told that the beam makes 4 full rotations a minute, meaning $\theta' = 8\pi/\text{min}$ (this is radians per minute), and that at the moment in time we're interested in, $d = 1$.

We also have the relationship that $\tan\theta = \frac{d}{3}$ (opposite over adjacent), and taking the derivative gives us $\sec^2(\theta) \cdot \theta' = d'/3$. We need to work out $\sec^2(\theta)$, but looking at our triangle we see that the adjacent side is length 3 and the hypotenuse is length $\sqrt{10}$ (by the Pythagorean theorem), so we have $\sec^2(\theta) = (\sqrt{10}/3)^2 = 10/9$.

Thus we have $d' = 3\sec^2(\theta) \cdot 8\pi = \frac{80\pi}{3}$ kilometers per minute.

Example. A kite is flying 100 feet over the ground, moving horizontally at 8 ft/s. At what rate is the angle between the string and the ground decreasing when 200ft of string is let out?



Call the ground distance between the kite-holder and the kite d and the angle between the string and the ground θ. When the length of string is $\ell = 200$ then $d = \sqrt{200^2 - 100^2} = 100\sqrt{3}$. We have that $d' = 8$ (since the angle is decreasing, the kite must be getting farther away). And finally we have the relationship $\tan\theta = \frac{100}{d}$ by the definition of tan in terms of right triangles. Then we have

$$\begin{aligned}
\tan\theta &= 100d^{-1} \\
\sec^2(\theta)\theta' &= -100d^{-2}d' \\
\theta' &= \frac{-100 \cdot 8\cos^2(\theta)}{d^2} \\
&= \frac{-800\cos^2(\theta)}{(100\sqrt{3})^2} \\
&= \frac{-8\cos^2(\theta)}{300}
\end{aligned}$$

We see that $\cos(\theta) = \frac{100\sqrt{3}}{200} = \frac{\sqrt{3}}{2}$, so we have

$$\theta' = \frac{-8 \cdot \frac{3}{4}}{300} = -\frac{8 \cdot 3}{300 \cdot 4} = -\frac{1}{50},$$

so the angle between the string and the ground is decreasing at a rate of $1/50$ per second (radians being unitless).

3 Optimization

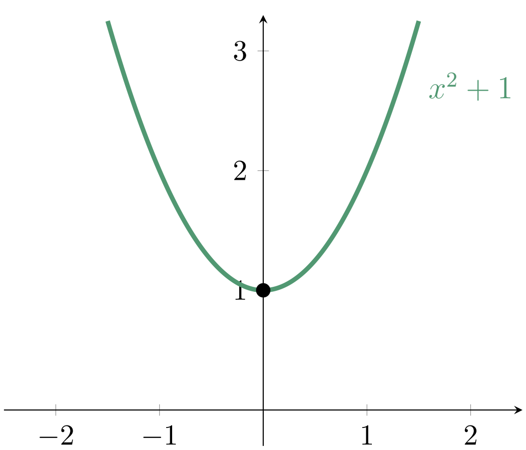
In this section, we'll discuss how we can use the derivative as a tool to understand where a function has its maxima and minima. We'll be working in the more abstract world of mathematical functions until the last part of the section, but this is a very useful tool in the real world, too. If you're running a factory, you may want to ask how you can make as much money as possible. Or you may want to keep your costs as low as possible. Or, if you're feeling pro-social, you may want to minimize the level of pollution you create. If you're a biologist studying an ecosystem, you may want to know what the maximum population of wolves you can expect to see is. If you're doing medical research, you may want to know what drug dose will be most effective. If you're a physicist studying the motion of an object, you may want to see where the highest point of its trajectory is, or where it reaches its fastest speed, or what the shortest path it can take is.

All of these questions are problems of *optimization*: we have some function or relationship, and we want to find the maximum (or minimum) value it can take on. And so for the next few sections we'll talk about maximizing or minimizing a function, but we always want to keep in mind that there are myriad applications in the background.

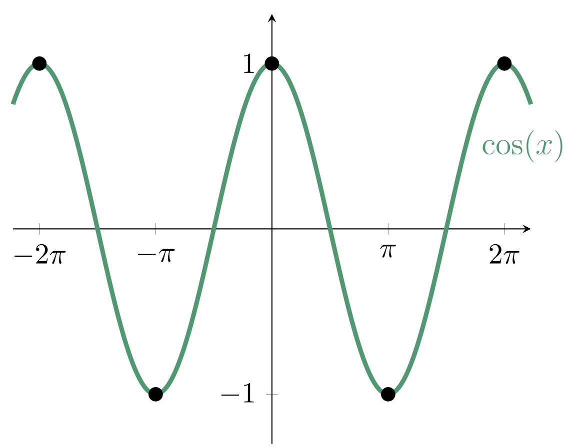
3.1 Maxima and minima

We begin by defining maxima and minima precisely. These are the largest and smallest values a function ever takes on. To say that L is a maximum of some function f means two things: 1) there is some input c for which L is an output, and 2) $f(x) \leq L$ for all inputs x.

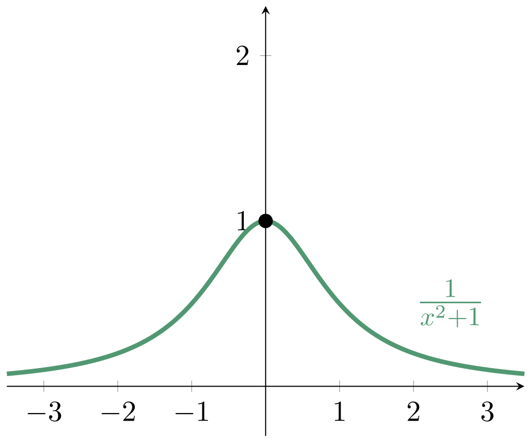
Definition. If $f(c) \geq f(x)$ (respectively, if $f(c) \leq f(x)$) for all inputs x in the domain of f, then we say that $f(c)$ is an **absolute maximum** (absolute minimum) or a **global maximum** (global minimum) for f. We say that f has a absolute maximum (absolute minimum) at c. We refer to absolute minima and maxima together as **absolute extrema**.



Absolute max: none
Absolute min: 1 at 0

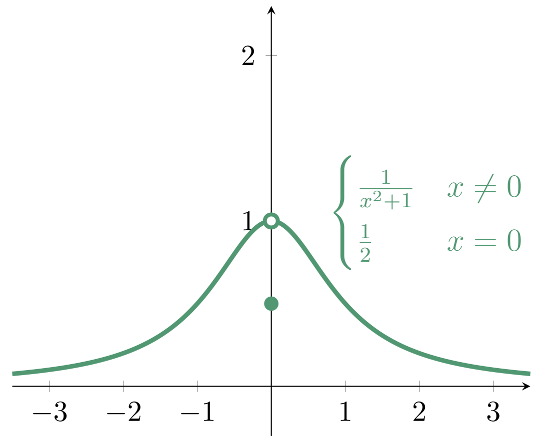


Absolute max: 1 at $2k\pi$
Absolute min: -1 at $(2k+1)\pi$



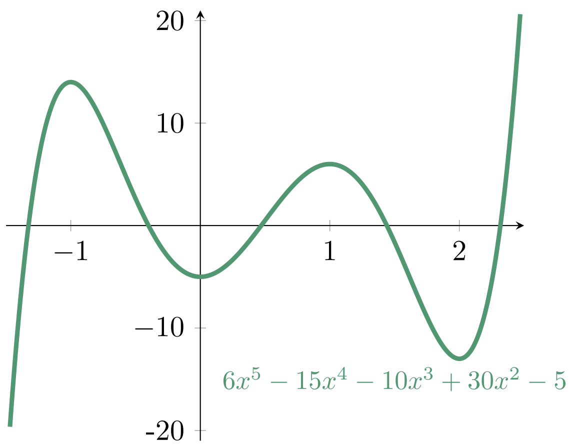
Absolute max: 1 at 0

Absolute min: none



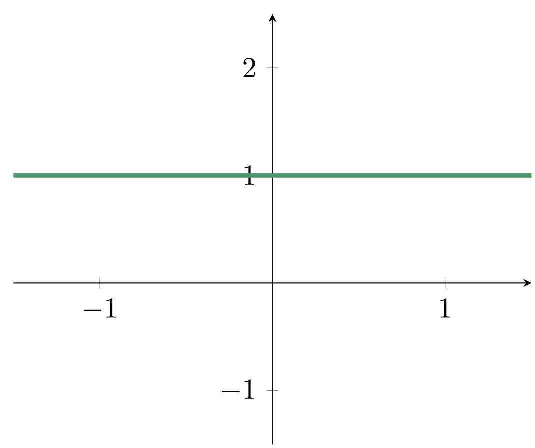
Absolute max: none

Absolute min: none



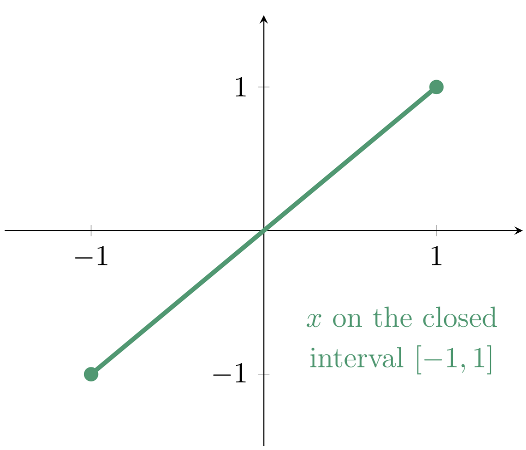
Absolute max: none

Absolute min: none



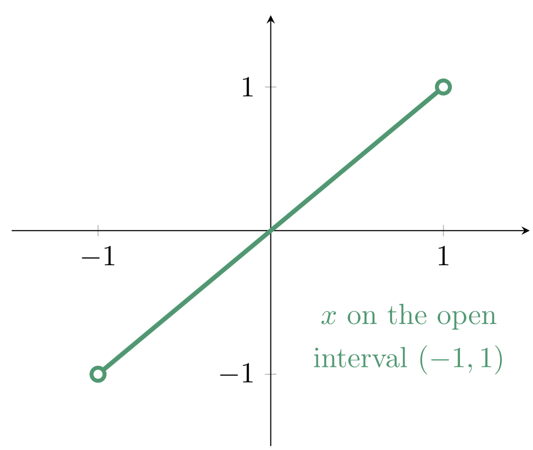
Absolute max: 1 at every point

Absolute min: 1 at every point



Absolute max: 1 at 1

Absolute min: -1 at -1



Absolute max: none

Absolute min: none

We see that a function *might* fail to have minima or maxima if it has a discontinuity (example 4) or if its domain is something other than a closed interval (examples 1, 3, 4, 5). As it turns out, these are the only things that can cause this:

Theorem (Extreme Value Theorem)**.** *Let f be a continuous function defined on a closed interval $[a, b]$. Then there exists a point c in $[a, b]$ such that $f(c)$ is an absolute maximum for f, and a point d in $[a, b]$ such that $f(d)$ is an absolute minimum for f.*

The Extreme Value Theorem (EVT), like the Intermediate Value Theorem from section 1.3, is a very powerful tool. However, it only guarantees the existence of extrema; it doesn't actually tell us how to find any. To find absolute extrema, we'll first try to address a different but related question. You may have noticed that functions like the one in example 5 have lots of peaks and valleys, even though they don't have any absolute extrema. These peaks and valleys are examples of *relative* or *local* extrema:

Definition. If $f(c) \geq f(x)$ (respectively, if $f(c) \leq f(x)$) for all x *near* c, we say that $f(c)$ is a **relative maximum** (relative minimum) for f and that f has a relative maximum (relative minimum) at c.

Again when we lump maxima and minima together, we call them *extrema*, so what we have just defined are **relative extrema**. At first glance, it doesn't seem like it should be all that much easier to find these sorts of points. However, we've just developed a really powerful tool, the derivative, that allows us to get a good sense of what even very complicated functions look like nearby a certain point. Looking at the examples above, we see that all the relative maxima and minima have flat tangent lines, and this is (mostly) not a coincidence! This suggests that points a for which a function f has a relative extremum should satisfy $f'(a) = 0$.

More algebraic intuition for this idea is that we know $f(x) \approx f(a) + f'(a)(x - a)$—this is our best linear approximation. If $f'(a)$ is a positive number, then when we plug some $x > a$ into our approximation, we get something greater than $f(a)$, and when we plug some $x < a$ into the approximation, we get something less than $f(a)$. Since $f(x)$ should have the same behavior, we expect that near a, inputs to the right of a make f bigger than $f(a)$, and inputs to the left of a make f smaller than $f(a)$. Together, this means $f(a)$ isn't a maximum or a minimum, and the story is similar if $f'(a)$ is a negative number. Whether this intuition or the geometric intuition is more appealing, we have the following theorem that tells us the intuition is good:

Theorem (Fermat's Theorem)**.** *If f has a relative extremum at c, then either $f'(c) = 0$ or $f'(c)$ does not exist.*

Proof. To prove this, we assume that f is differentiable at c so that $f'(c)$ does exist, and then we need to prove it's equal to zero. (We can make this assumption because if it's true, we're in the $f'(c) = 0$ case, and if it's not true, we're in the $f'(c)$ doesn't exist case, and the theorem just says we end up in one of these cases.) Therefore suppose $f(c)$ is a local extremum, say a maximum for simplicity, and that $f'(c)$ exists. Then this means from the limit definition that

$$\lim_{t \to c} \frac{f(t) - f(c)}{t - c}$$

exists, meaning also both one-sided limits exist and are equal to $f'(c)$. Now because $f(c)$ is a local maximum, for any t near x, we get that $f(t) \leq f(x)$. Therefore for $t \geq c$, we get that

$$\frac{f(t) - f(c)}{t - c} \leq 0, \quad \text{meaning} \quad f'(c) = \lim_{t \to c^+} \frac{f(t) - f(c)}{t - c} \leq 0$$

by the limit comparison law. Similarly, if $t \leq c$, then

$$\frac{f(t) - f(c)}{t - c} \geq 0, \quad \text{meaning} \quad f'(c) = \lim_{t \to c^-} \frac{f(t) - f(c)}{t - c} \geq 0,$$

again by the limit comparison law. We've just shown that $f'(c) \leq 0$ *and* that $f'(c) \geq 0$, and we conclude that $f'(c) = 0$ exactly. $\square$

Fermat's Theorem is excellent news for us. Instead of checking all (likely infinitely many) points in the domain of a function to see if their outputs are relative extrema, we can narrow our search to to the set of points c in the domain where the derivative is zero or fails to exist, and this collection of points is often small enough that we can check through them by hand to see which are maxima, minima, or in some cases neither. (Note that the theorem doesn't say that $f'(c) = 0$ implies that $f(c)$ is a relative extremum—even if we find a point c where $f'(c)$ is zero or doesn't exist, $f(c)$ could still be **neither** a relative minimum nor a relative maximum.) Because these points are so special, we give them a name:

Definition. Let f be a function. A point c in the domain of f is called a **critical point** if $f'(c) = 0$ or if $f'(c)$ does not exist. When c is a critical point, we call the number $f(c)$ a **critical value**.

With this definition, we can rephrase Fermat's Theorem as the following: **if $f(c)$ is a relative extremum, then c is a critical point.** (And again, this doesn't work the other way around—you can have critical points c for which $f(c)$ is not a relative extremum.) Here are a few examples of functions and their critical points:

- Let $f(x) = x^3 - x$. Then $f'(x) = 3x^2 - 1$; this is defined everywhere, and $f'(x) = 0$ when $x = \pm\frac{\sqrt{3}}{3}$. So the critical points are $\pm\frac{\sqrt{3}}{3}$.
- If $f(x) = x^2$, then $g'(x) = 2x$ and is 0 when $x = 0$. So the only critical point is 0.
- If $h(x) = \sin(x)$ then $h'(x) = \cos(x)$, which is 0 when $x = (n + 1/2)\pi$ for any integer n. Thus the critical points are $\pi/2, 3\pi/2, 5\pi/2, \ldots$.

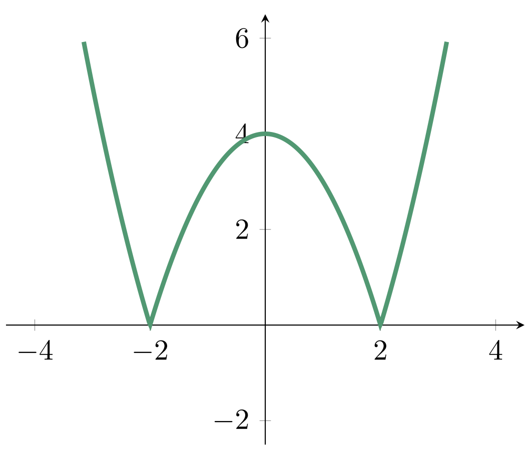
- If $f(x) = x^3$ then $f'(x) = 3x^2$ which is 0 when x is 0. Thus the only critical point is at 0.
- If $g(x) = |x|$ then

$$g'(x) = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \\ DNE & x = 0 \end{cases}$$

  and thus has a critical point at $x = 0$ since the derivative does not exist there.
- If $f(x) = |x^2 - 4|$, then taking the derivative seems a little intimidating, but this is really just the chain rule: if we let $g(x) = |x|$ as in the example above, then $f'(x) = g'(x^2 - 4) \cdot 2x$. We know from the above that g' is undefined at 0, so f' is undefined whenever $x^2 - 4 = 0$, i.e. when $x = \pm 2$. For the other critical point, we see that the derivative has a factor of $2x$, which makes $f'(x) = 0$ for $x = 0$. Therefore the critical points are -2, 0, and 2.

In each case, we manage to reduce the number of potential points at which there could be extrema to just a few (or in the case of the $\sin(x)$ function, an infinite family of points that has a nice description). The $f'(c) = 0$ case was motivated by our geometric reasoning since this indicates that the line tangent to the curve at the point c is horizontal. The case where $f'(c)$ does not exist didn't come up in our geometric intuition, but recall that f can fail to be differentiable at points with cusps and corners, and these can certainly be extrema. The absolute value function is the basic example of this, with its minimum of 0 at $x = 0$ detected only by checking for points at which the derivative does not exist.

The very last example here also shows how we need to consider both cases to find all the candidates for a max or min. Here's a graph of $f(x) = |x^2 - 4|$:



As we can see, there is a relative maximum of 4 at $x = 0$, an absolute minimum of 0 at $x = \pm 2$. We would not have found the relative max without checking where the derivative was zero, and we would not have found the absolute min without checking where the derivative was undefined (note the sharp corners).

Now that we know how to find a list of points where the might be relative extrema, we'd like some way to find out which of these points the function actually *does* have extrema at, and we'd like a way to categorize them as maxima and minima. This turns out to be somewhat more complicated, and we'll develop the theory we need to do this in the following sections (3.2 and 3.3).

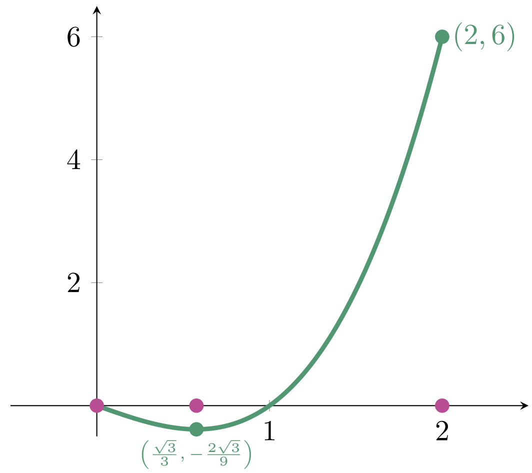
What we *can* do now though is finish off the question of absolute extrema. If $f(c)$ is an absolute maximum, by definition $f(c) \geq f(x)$ for all x in the domain of f. In particular then $f(c) \geq f(x)$ for all x *near* c, so $f(c)$ is automatically also a relative maximum. That is, an absolute max is also a relative max, and similarly, an absolute min is also a relative min. This means that Fermat's Theorem also applies to the question of finding absolute extrema, and so if we know in advance that absolute extrema exist (e.g. by the Extreme Value Theorem), we can find the critical points to get a list of values c for which $f(c)$ may be an extremum, and then just check the values by hand to see which value is the largest and which is the smallest.

Let's see this process in action. Suppose we'd like to find the absolute extrema of the function $f(x) = x^3 - x$ from our first example on the interval $[0, 2]$. The first thing we need to check is whether we should expect this function to have absolute extrema on this interval to begin with. The function is continuous on the closed interval $[0, 2]$, so by the EVT, it must have an absolute max and an absolute min somewhere in the interval.

Next, we saw that the critical points were $\pm\sqrt{3}/3$. However, since we're restricting the domain of the function to the interval $[0, 2]$, we no longer know about any of the points just to the right of 2 or just to the left of 0, which means derivatives can no longer be taken at these points (recall from the definition that f needs to be defined on an open interval around a point to be differentiable at that point). Thus we need to add our endpoints to the list of points we need to check for absolute extrema. We also need to remove the point $-\sqrt{3}/3$ from our list, because we only care about what's happening with the function on the interval $[0, 2]$. Hence we check $x = 0$, $x = \sqrt{3}/3$, and $x = 2$:

- $f(0) = 0^3 - 0 = 0$
- $f(\sqrt{3}/3) = 3\sqrt{3}/27 - \sqrt{3}/3 = \sqrt{3}/9 - 3\sqrt{3}/9 = -2\sqrt{3}/9$
- $f(2) = 2^3 - 2 = 6$

The function f when restricted to $[0, 2]$ therefore has an absolute minimum of $-2\sqrt{3}/9$ at $x = \sqrt{3}/3$ and an absolute maximum of 6 at $x = 2$. Here's a picture, with the critical points highlighted in pink and the maximum and minimum shown in green:



Looking at the procedure we've just performed, we have the following series of steps for finding absolute extrema of a function:

1. Verify that the function *has* absolute extrema. (Your answer for this step should explicitly invoke the Extreme Value Theorem!)

2. Find all the critical points of the function, including the endpoints of whatever interval the function is defined on.

3. Evaluate the function at each critical point (again including endpoints).

4. Determine which value is the largest and which is the smallest, since this gives the absolute max and min.

Lets do another example with these steps in mind. Consider the function $f(x) = |x^3 - 3x^2|$ on the interval $[1, 4]$:

1. First off, the function $x^3 - 3x^2$ is continuous, as is $|x|$, so $f(x)$ is continuous since it is the composition of two continuous functions. It's also defined on $[1, 4]$, which is a closed interval. Since f is continuous on a closed interval, by the Extreme Value Theorem, there must exist an absolute minimum and an absolute maximum.

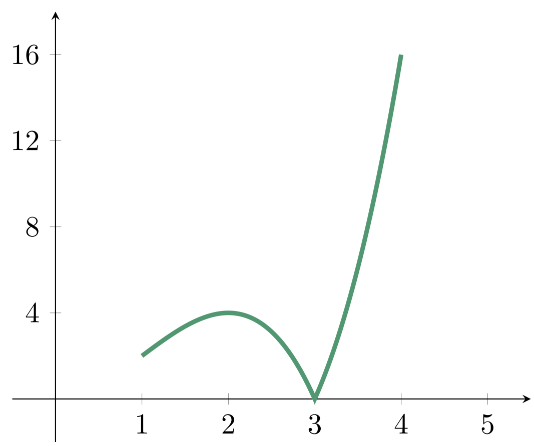
2. To find the critical points, we need to take a derivative. Recall that for $g(x) = |x|$, the derivative is the function
$$g'(x) = \frac{x}{|x|} = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \end{cases}$$
which is undefined at zero. By the chain rule,
$$f'(x) = \frac{x^3 - 3x^2}{|x^3 - 3x^2|} \cdot (3x^2 - 6x),$$

and so f' is undefined anywhere $x^3 - 3x^2 = 0$, which means $x = 0$ and $x = 3$ are critical points. The other factor $3x^2 - 6x = 3x(x - 2)$ gives us $x = 2$ as another critical point. Since we're on the interval $[1, 4]$, the complete list of points to check is therefore $x = 1$, $x = 2$, $x = 3$, and $x = 4$, where 1 and 4 are included because they are endpoints, and $x = 0$ is excluded because it doesn't lie in the interval $[1, 4]$.

3. We evaluate $f(x)$ at each one of the points we've identified:

   - $f(1) = |1^3 - 3(1)^2| = |1 - 3| = 2$
   - $f(2) = |2^3 - 3(2)^2| = |8 - 12| = 4$
   - $f(3) = |3^3 - 3(3)^2| = |27 - 27| = 0$
   - $f(4) = |4^3 - 3(4)^2| = |64 - 48| = 16$

4. From the above, $f(x) = |x^3 - 3x^2|$ on $[1, 4]$ has an absolute maximum of 16 at $x = 4$ and an absolute minimum of 0 at $x = 3$. Indeed, we have the following graph:



Our plan for the rest of the optimization section is as follows. In 3.2 we're going to fill in a lot of theory about the derivative and finally prove that it does what we're saying it does; in 3.3 we'll use this theory to see how we can classify *all* the critical points we find; in 3.4 we'll put all of this together to show how we can use the derivative to sketch graphs of functions ourselves with a good degree of accuracy; and finally in 3.5 we'll use the optimization tools developed in this section and in 3.3 on functions that model some real world scenarios.

3.2 The Mean Value Theorem

Our initial motivation for the derivative was through linear approximation and tangent lines. We showed how we could think of the derivative as the limit of the slopes of secants lines through a point of interest and a nearby point, and presented lots of pictures to convince ourselves that the derivative really was measuring slope. From this, we've been freely using the idea that a positive derivative at a point a means a function is sloping up near a and that a negative derivative means a function is sloping down near a. These are things we should expect to be true based on how we defined the derivative, and indeed, they are true, but we now need to take some time and actually prove our definition works they way we want it to. This is a pretty common pattern in math: we define something interesting based on some kind of heuristic, and often the first theorems we have about whatever we've just defined tell us that the definition we wrote down really does capture what we want it to capture.

We begin with a technical (but relatively easy) result:

Theorem (Rolle's Theorem)**.** *If f is continuous on $[a, b]$ and differentiable on (a, b) such that $f(a) = f(b)$, then there is some point c in (a, b) with $f'(c) = 0$.*

Proof. There are a couple cases to consider here. If $f(x)$ is a constant function on $[a, b]$, then $f'(c) = 0$ for all c in (a, b) and we are done. If instead $f(x)$ is nonconstant, then there exists an input t in (a, b) such that $f(t) \neq f(a)$.

Recall that f being continuous on $[a, b]$ means by the EVT that there exists an absolute max and an absolute min. If $f(t) > f(a)$, then $f(a) = f(b)$ is not the absolute max, so the point c at which the absolute max occurs is within the interval (a, b). Now by Fermat's Theorem, $f(c)$ being an absolute max implies that either $f'(c) = 0$ or $f'(c)$ does not exist. However, one of our hypotheses is that f is differentiable on (a, b), and therefore at c, so we conclude that $f'(c) = 0$.

If instead $f(t) < f(a)$, then $f(a) = f(b)$ is not the absolute min, and by a similar argument with Fermat's Theorem, we conclude that the point c in (a, b) at which the absolute min occurs satisfies $f'(c) = 0$. $\square$

If we suppose $p(t)$ is a function giving the path of an object thrown up into the air and then caught, denote the time of release by t_i and the time of the catch by t_f. Then assuming $p(t)$ is differentiable (not an unreasonable assumption for modeling a trajectory), all Rolle's theorem is saying is that there's some point in time where $p'(t)$, the instantaneous velocity of the object, is 0. Physically, we already know this will happen at the peak of the trajectory, so Rolle's Theorem can be viewed as a piece of evidence that the derivative we've defined does a good job modeling our physical reality, as we've been claiming it does.

Sometimes we can also use Rolle's Theorem to restrict the number of roots a polynomial can have. Recall that in section 1.3 we used the Intermediate Value Theorem to prove that—even though it would be impossible to write down—the function $f(x) = x^5 + x^3 - 1$ has a real root between 0 and 1, call it a. We can note that for $x < 0$, both x^5 and x^3 are negative, so $f(x) < -1$ for $x < 0$, meaning there aren't any roots. Similarly, for $x > 1$, x^5 and x^3 are both greater than 1, meaning $f(x) > 1 + 1 - 1 = 1$ and again there can't be any roots. This means that if the function has more than just the one root a we found with the IVT, the other roots would also have to be in the interval

$[0,1]$. Now f is continuous on $[0,1]$ and differentiable on $(0,1)$ with derivative $f'(x) = 5x^4 + 3x^2$. If there were a second root here, call it b, we would have $f(a) = 0 = f(b)$ and would be in a position to apply Rolle's Theorem. Rolle's Theorem here would say that $f'(c) = 0$ for some c in the interval $(0,1)$, but we just said that $f'(c) = 5c^4 + 3c^2 > 0$. This would contradict Rolle's Theorem, so there cannot be a second root. Hence $f(x) = x^5 + x^3 - 1$ has only one real root.

The big use we have for Rolle's Theorem, though, is that we can use it to prove a more powerful tool, the Mean Value Theorem. Here's the kind of intuitive result that the Mean Value Theorem will allow us to prove: suppose again we have $p(t)$ modeling position as a function of time. Let's also suppose that the object has a max speed of 60mph. Intuitively, it's clear that if we let the object move around for an hour, there's no way for it to cover more than 60 miles. If we interpret the condition that the maximum speed is 60mph as saying that $p'(t) \leq 60\text{mi/h}$, then $p(1\text{h}) - p(0\text{h})$ gives the distance traveled in one hour, and we'd like to be able to say that this means $p(1\text{h}) - p(0\text{h}) \leq 60\text{mi}$. We'll need to Mean Value Theorem to be sure.

The Mean Value Theorem generalizes Rolle's Theorem, allowing us to do without the assumption that $f(a) = f(b)$. Of course, this means the result is slightly different. Instead of finding a point c in (a,b) with $f'(c) = 0$, we find a c where $f'(c)$ is exactly the slope of the secant line between $(a, f(a))$ and $(b, f(b))$.

Theorem (Mean Value Theorem)**.** *If f is continuous on $[a,b]$ and differentiable on (a,b), then there is some point c in (a,b) with*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Proof. Consider the secant line between $(a, f(a))$ and $(b, f(b))$. The line has slope

$$\frac{f(b) - f(a)}{b - a}$$

and passes through $(a, f(a))$, so the equation for the line is

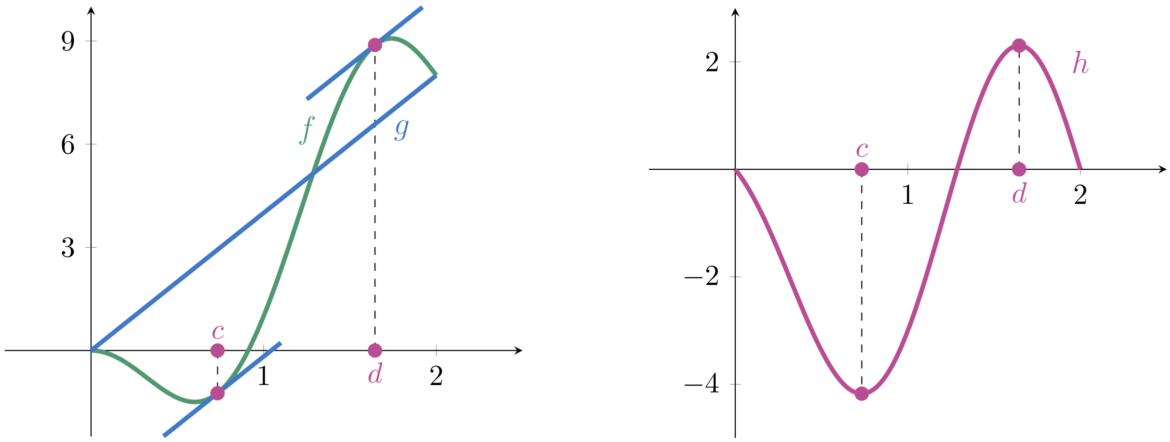
$$g(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a).$$

Now consider the difference $h(x) = f(x) - g(x)$ between the function and the secant line. Both f and g pass through the points $(a, f(a))$ and $(b, f(b))$, so we have that $h(a) = h(b) = 0$. Because f and g are both continuous on $[a,b]$ and differentiable on (a,b), so is h, meaning h satisfies the conditions of Rolle's Theorem on $[a,b]$. Therefore there exists a point c in (a,b) such that $h'(c) = 0$. That is, $0 = h'(c) = f'(c) - g'(c)$, so $f'(c) = g'(c)$. But $g(x)$ is just a line, so its derivative is its slope, which we said was

$$f'(c) = g'(c) = \frac{f(b) - f(a)}{b - a},$$

and so we are done. $\square$

Here's a picture of what's happening in the Mean Value Theorem:



We see that h satisfies the conditions of Rolle's Theorem, and that there are actually two points that fit the conclusion of the theorem (only one is guaranteed), and we've labeled these c and d. We then see that the tangent lines to f at the points c and d are parallel to g, that is, $f'(c) = f'(d)$ is the slope of the secant line g.

Let's go back to our physical example of an object with a max speed of 60mph and see how the MVT helps us prove the object can't travel more than 60 miles if we let it go for an hour (again, note how obviously true this sounds). We said that the total distance traveled was $p(1\text{h}) - p(0\text{h})$, and since p is assumed continuous on $[0, 1]$ and differentiable on $(0, 1)$, the MVT implies that there exists some point c in $(0, 1)$ such that

$$p'(c) = \frac{p(1) - p(0)}{1 - 0} = \frac{\text{distance traveled}}{1\text{h}}.$$

Now the restriction that the object's top speed is 60mph means that

$$\frac{\text{distance traveled}}{1\text{h}} = p'(c) \leq 60\text{mi/h},$$

so indeed, the distance traveled is less than or equal to 60 miles, the result we expected to get.

Here's another example of how the MVT lets us translate our knowledge of the derivative of a function into knowing something about the values of the function itself:

Example. Let's say we've got some differentiable function f and we know both that $f(0) = 7$ and that $|f'(x)| \leq 2$ for all x. What can we figure out about $f(5)$?

Recall that we can rewrite $|f'(x)| \leq 2$ as $-2 \leq f'(x) \leq 2$. Then the MVT says that there's a c in $(0, 5)$ such that

$$\frac{f(5) - f(0)}{5 - 0} = f'(c), \quad \text{meaning} \quad -2 \leq \frac{f(5) - 7}{5} \leq 2.$$

We can solve for $f(5)$ so say that $-10 \leq f(5) - 7 \leq 10$, so $-3 \leq f(5) \leq 17$. We can summarize all of this with our physical intuition by saying that if we're traveling along a number line starting at 7 and can't move more than 2 units per hour in either direction, then after 5 hours the farthest left we could end up is -3 and the farthest right we could end up is 17.

The Mean Value Theorem has several important corollaries we'll now state and prove. (A corollary is a result that follows relatively easily from one just proven.)

Corollary 1 (Derivative Zero Theorem)**.** *Let f be differentiable on some interval I. If $f'(x) = 0$ for all x in I, then f is constant on I.*

Proof. Suppose f is not constant on I. Then there are points a and b in I such that $a < b$ and $f(a) \neq f(b)$. Since f is differentiable on I, it is continuous, hence it is continuous on $[a, b]$ and differentiable on (a, b), satisfying the conditions of the MVT. Then there exists a point c in (a, b) such that

$$f'(c) = \frac{f(b) - f(a)}{b - a} \neq 0,$$

so $f'(x)$ is not zero at every point of I. This means that if we have a function f where it *is* true that $f'(x) = 0$ on an interval I, then f must be constant on I as otherwise the above argument would contradict $f'(x) = 0$.[3] $\square$

Corollary 2 (Constant Difference Theorem)**.** *If f and g are differentiable on an interval I such that $f'(x) = g'(x)$ for all x in I, then $f(x) = g(x) + C$ for some constant C.*

Proof. We use Corollary 1. If $h(x) = f(x) - g(x)$, then $h'(x) = f'(x) - g'(x) = 0$, so Corollary 1 implies $h(x) = C$ for some constant C on the interval I. Therefore $f(x) - g(x) = C$, so $f(x) = g(x) + C$ as claimed. $\square$

Before we finish the section with the next corollary, we stop and carefully define what it means for a function to be increasing or decreasing.

Definition. We say a function is **increasing** on an interval I if for all $a < b$ in I, $f(a) < f(b)$. It is instead **decreasing** on I if $a < b$ for a and b in I implies $f(a) > f(b)$. (If we just say something like "f is increasing," the interval is understood to be the domain of the function.)

Notice that these definitions make sense if you assume we're moving to the right; an increasing function is one where $f(x)$ increases as x increases.

[3] Before moving on to the next corollary, it's worth noting what we did in the proof there. We had a statement of the form P implies Q (in our case, P was the proposition that $f'(x) = 0$ for all x in I and Q was the proposition that f is constant on I), and what we did to prove it was instead to prove the statement (not Q) implies (not P) that if f is not constant on I, then there must be some point in I where f' is nonzero. This statement is the *contrapositive* of the original one (P implies Q), and is logically equivalent, so we can always prove whichever one seems easier to prove. You'll learn more about this in a course on proofs.

Corollary 3 (Increasing and Decreasing Functions)**.** *Let f be differentiable on an interval I.*

(i) If $f'(x) > 0$ for all x in I then f is increasing on I.
(ii) If $f'(x) < 0$ for all x in I then f is decreasing on I.

Proof. We'll just prove (i), the proof of (ii) is very similar. Suppose $f'(x) > 0$ for all x in I, and let a and b be points in I with $a < b$. Then by the MVT, there's a point c in (a, b) with

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Therefore $f(b) - f(a) = f'(c)(b - a) > 0$ since both $f'(c)$ and $b - a$ are positive. □

Corollary 1 is useful in part because it gets us to Corollary 2, which we'll need when we talk about integration and antiderivatives. The main focus of the optimization section though will be Corollary 3, since this provides us an excellent tool for determining the qualitative behavior of a function and allows us to categorize critical points when we find them. We'll see how to do this in the following section 3.3.

3.3 Classifying critical points

In 3.1, we had figured out how to find the absolute extrema of a continuous function on a closed interval. To do this, we found the function's critical points, and then checked them along with the endpoints of the interval to see which values were the biggest and smallest. We left open the question of *relative extrema*, which are the points where the function has peaks and valleys that aren't necessarily the largest ones.

Now that we have Corollary 3 to the Mean Value Theorem, we can find all the intervals on which a function is increasing and decreasing, and this is one of our tools for determining when a critical point corresponds to a relative max or a relative min:

First derivative test: Suppose f has a critical point c.

(i) If $f' > 0$ on some interval left of c and $f' < 0$ on some interval right of c, then $f(c)$ is a relative maximum (f changes from increasing to decreasing).
(ii) If $f' < 0$ on some interval left of c and $f' > 0$ on some interval right of c, then $f(c)$ is a relative minimum (f changes from decreasing to increasing).
(iii) If f' has the same sign on an interval left of c as on an interval right of c, then $f(c)$ is neither a relative min nor a relative max (either f is increasing on both sides or decreasing on both sides).

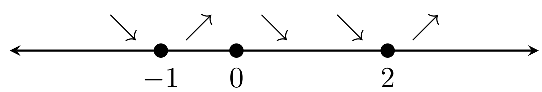
Example. Find and classify the critical points of the function $f(x) = 3x^4 - 4x^3 - 12x^2 + 5$, using the first derivative test to identify intervals of increase and decrease.

We begin by taking a derivative: $f'(x) = 12x^3 - 12x^2 - 24x$. We want to find critical points, and we note that this is defined everywhere, so we just need to see when this is equal to zero. Factoring gives $f'(x) = 12x(x+1)(x-2)$, so the points -1, 0, and 2 are critical points.

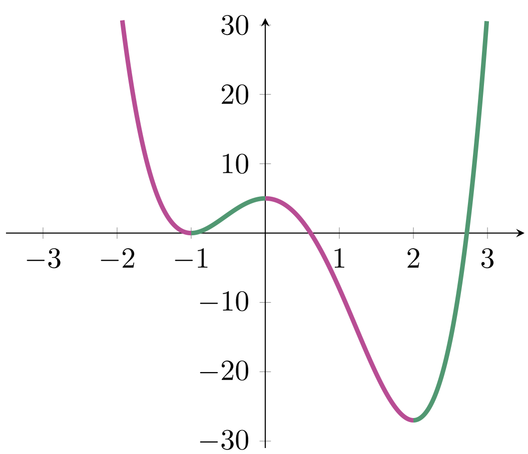
Next, we want to figure out when f is increasing and decreasing, which by the Corollary to the MVT means figuring out when f' is positive and negative. From the factored form of f', there are three terms contributing to the sign of f', and f' is positive if and only if all three of these terms are positive, or if just one is positive and the other two are negative. We can make a little table to determine where f' is positive and negative, with a column for each of these factors. The rows of our table are the intervals f' might have different signs on:

	$12x$	$(x+1)$	$(x-2)$	$f'(x)$
$(-\infty, -1)$	$-$	$-$	$-$	$-$
$(-1, 0)$	$-$	$+$	$-$	$+$
$(0, 2)$	$+$	$+$	$-$	$-$
$(2, \infty)$	$+$	$+$	$+$	$+$

If we like, we can turn our results into a little number line like so:



where we think of this as being the x-axis of a plot of f, with little arrows telling us whether f is increasing or decreasing on the intervals between the critical points. At the critical point $x = -1$, we see that f is decreasing towards $f(-1)$ for $x < -1$ and then increasing away from $f(-1)$ for $x > -1$, meaning $f(-1)$ is a relative minimum. Similarly, f increases towards $f(0)$ and decreases away from it, so $f(0)$ is a relative maximum. Finally, f decreases towards $f(2)$ and increases away from it, so $f(2)$ is a relative minimum. We can also say that f is increasing on $(-1, 0) \cup (2, \infty)$ and decreasing on $(-\infty, -1) \cup (0, 2)$, where the $\cup$ symbol means union (taking the two intervals together).



Indeed, if we graph the function to check our work, we get the above, where regions of increase have the function in green, and regions of decrease have the function in pink. We see that our analysis was spot on, although the work we did didn't say anything about relative sizes of any of the peaks and valleys, only that they existed and could be correctly categorizes as peaks and valleys.

The first derivative test can also handle critical points that come from the derivative being undefined. Consider again the absolute value function $g(x) = |x|$, and recall that

$$g'(x) = \begin{cases} 1 & x > 0 \\ -1 & x < 0. \end{cases}$$

Since g' is undefined at $x = 0$, this is a critical point of the function. We also see that g' goes from negative to positive at 0, so g itself goes from decreasing to increasing and therefore has a relative minimum at $x = 0$.

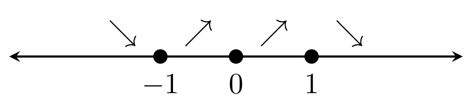
Example. Find and classify the critical points of the function $f(x) = 5x^{\frac{1}{3}} - x^{\frac{5}{3}}$.

The derivative is

$$f'(x) = \frac{5}{3}x^{-\frac{2}{3}} - \frac{5}{3}x^{\frac{2}{3}} = \frac{5(1 - x^{\frac{4}{3}})}{3x^{\frac{2}{3}}}.$$

Then since f' is undefined when $3x^2 = 0$ and equal to zero when $x^{\frac{4}{3}} = 1$, we have the three critical points -1, 0, and 1.

Because the denominator is $3x^{\frac{2}{3}} = 3(x^{\frac{1}{3}})^2$ is always nonnegative, the sign of f' is determined by the sign of the numerator. The numerator is negative when $x^{\frac{4}{3}}$ is greater than 1 and positive when it is less than 1, i.e. is positive for $|x| < 1$ and negative for $|x| > 1$. This tells us that f' is negative on $(-\infty, -1) \cup (1, \infty)$ (so f is decreasing there) and positive on $(-1, 0) \cup (0, 1)$ (so f is increasing there), where we have broken the interval $(-1, 1)$ into two pieces because f' is not defined at $x = 0$. The number line with the critical points of f therefore looks like



which tells us that $f(-1)$ is a minimum, $f(1)$ is a maximum, and $f(0)$ is neither a minimum nor a maximum. (And note that a complete answer to a question of this type should always include mentioning explicitly when a critical point is neither a minimum nor a maximum.)

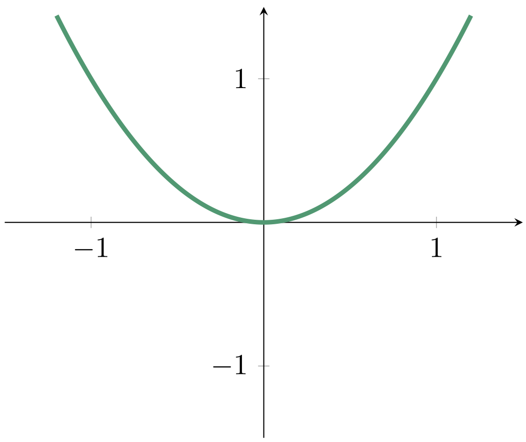
The first derivative test is highly reliable, but it's a fair amount of work. Having to figure out intervals of increase and decrease can be a pain, but it turns out that if we take one more derivative, we get another test that depends only on the critical point itself, and not the points around it. The reason for this is that the second derivative also captures certain information about the shape of a graph: *concavity.*

Definition. We say a function f is **concave upward** on an interval I if f' is increasing on I.

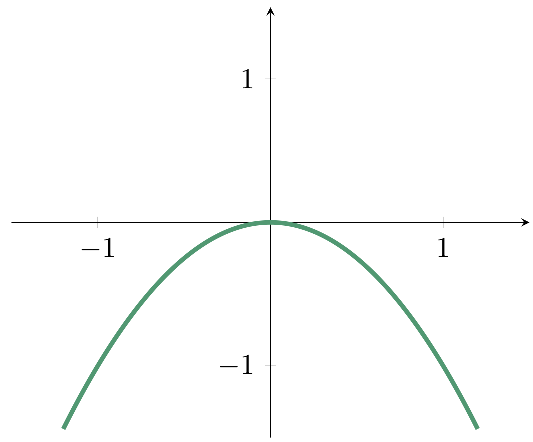
We say a function f is **concave downard** on I if f' is decreasing on I.

We say a point c is an **inflection point** for a function f if the graph of f changes from concave up to concave down, or concave down to concave up, at c.

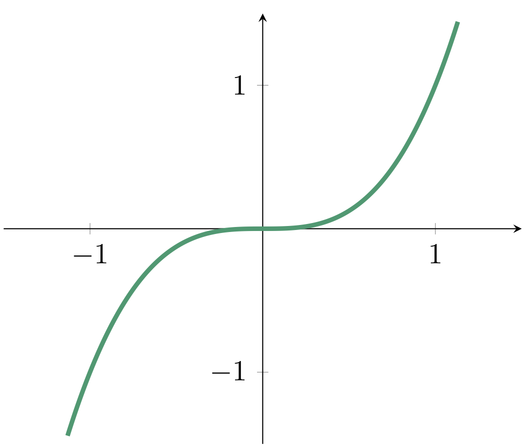
Note that a function with an increasing derivative is one whose instantaneous slopes are increasing, and so must be curving "upwards" like a bowl (hence the name concave *upwards*), and similarly, to have a decreasing derivative means it must be curving downwards like an umbrella. Here are a few examples:



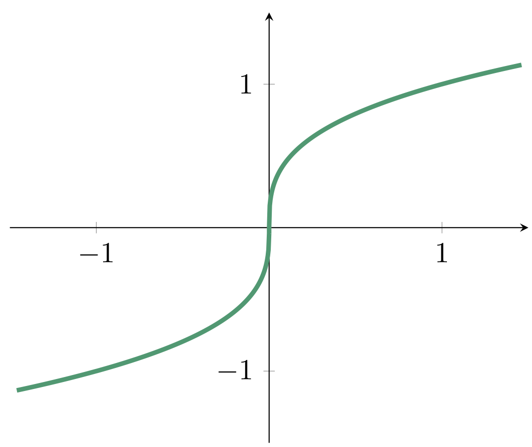
x^2: always concave up



$-x^2$: always concave down



x^3: concave up for $x > 0$, down for $x < 0$



$x^{\frac{1}{3}}$: concave down for $x > 0$, up for $x < 0$

Note how in each of the concave up examples, if we picture tangent lines at several points and move left to right, the slopes are increasing, and if we do the same in the concave down examples, the slopes are decreasing.

Because the second derivative of a function f is just the derivative of f', we can apply the third Corollary to the MVT yet again to say that when f'' is positive on some interval, f' is increasing there, and similarly if f'' is negative. This gives us the following proposition:

Proposition. *If $f''(x) > 0$ on an interval I, then f is concave up on I. If $f''(x) < 0$ on I, then f is concave down on I.*

Note that the proposition has nothing to say about what happens at a point c for which $f''(c) = 0$. As we will see, this could mean just about anything. Returning to our examples of what being concave up or down looks like, we see that $\frac{d^2}{dx}x^2 = 2$ which is positive everywhere and $\frac{d^2}{dx} - x^2 = -2$ which is negative everywhere. This matches what we said about our first two examples.

For x^3, we have that $\frac{d^2}{dx}x^3 = 6x$ is positive for $x > 0$ and negative for $x < 0$, so the function changes from concave down to concave up at the point $x = 0$, and so this is an inflection point.

For $x^{\frac{1}{3}}$, we get

$$\frac{d^2}{dx}x^{\frac{1}{3}} = \frac{d}{dx}\left(\frac{1}{3}x^{-\frac{2}{3}}\right) = -\frac{2}{9x^{\frac{5}{3}}},$$

which is positive when $x < 0$ and negative when $x > 0$, meaning 0 is again an inflection point. Note that this happens even though the second derivative is undefined at $x = 0$. In the same way we defined critical points to be places where the first derivative was either zero or undefined, any point where the second derivative is zero or undefined is a *potential* point of inflection.

A second derivative of zero really can mean anything though. We already saw with x^3 that a second derivative of zero can happen at an inflection point, but if we consider also the functions x^4 and $-x^4$, we get second derivatives $12x^2$ and $-12x^2$. Both second derivatives are zero at $x = 0$, but since $12x^2$ is positive for $x \neq 0$, x^4 is concave up everywhere and $x = 0$ is not a point of inflection. Similarly, 0 is not an inflection point for $-x^4$, since this is concave down everywhere.

If we have a critical point, the second derivative can often be an excellent tool for categorizing it as a min or a max. If we know the second derivative is negative, we know the function is concave down. A concave down function is umbrella-shaped, the shape of a maximum! Similarly, if the second derivative is positive, we know the function is curving up in a bowl shape near the critical point, like a minimum. This intuition isn't quite a proof, but it leads us to stating the correct proposition:

Proposition (Second Derivative Test)**.** *Suppose $f'(c) = 0$ so that c is a critical point of f.*

(i) If $f''(c) > 0$, then $f(c)$ is a relative minimum of f.
(ii) If $f''(c) < 0$, then $f(c)$ is a relative maximum of f.

Proof. By definition, the second derivative is just the derivative of the first derivative, so

$$f''(c) = \lim_{t \to c} \frac{f'(t) - f'(c)}{t - c} = \lim_{t \to c} \frac{f'(t)}{t - c}$$

since we are assuming $f'(c) = 0$. Now if $f''(c) > 0$, this means that $f'(t)/(t - c)$ gets close and stays close to the positive number $f''(c)$ as t gets close to c. This means that close enough to c, $f'(t)/(t - c)$ is itself positive. When t is a little to the left of c, $t - c$ is negative, and so in order for the whole fraction $f'(t)/(t-c)$ to be positive, $f'(t)$ must also be negative. Similarly, when t is a little to the right of c, $f'(t)$ is positive. Therefore $f'(t)$ changes from negative to positive at c, and so $f(c)$ is a minimum. The argument that $f''(c) < 0$ implies that $f(c)$ is a max is very similar. $\square$

Again, as we saw just before this proposition, if $f''(c) = 0$, we don't learn anything about what happens at the critical point. This test also doesn't work when $f'(c)$ is undefined because in this case $f''(c)$ is also undefined.

Example. We previously looked at the function $f(x) = 3x^4 - 4x^3 - 12x^2 + 5$ using the first derivative test. We computed that $f'(x) = 12x^3 - 12x^2 - 24x = 12x(x-2)(x+1)$, so the critical points are $x = -1, 0, 2$.

Then $f''(x) = 36x^2 - 24x - 24 = 12(3x^2 - 2x - 2)$. We can compute

$$\begin{aligned} f''(-1) &= 12(3+2-2) = 36 > 0 \\ f''(0) &= -24 < 0 \\ f''(2) &= 12(12-4-2) = 72 > 0 \end{aligned}$$

so by the second derivative test, f has a local maximum at 0 and local minima at -1 and 2.

This was a little faster and easier than the way we original classified the maxima and minima of this function. Typically, when it's relatively straightforward to take another derivative, we should do it! We do have to be aware that the second derivative test may come back inconclusive, though. Here's an example where the second derivative test really isn't very helpful:

Example. Let $f(x) = x^{2/3}(6-x)^{1/3}$. Where does f have relative maxima and minima? Where is it increasing or decreasing?

$$f'(x) = \frac{4-x}{x^{1/3}(6-x)^{2/3}}$$

$$f''(x) = \frac{-8}{x^{4/3}(6-x)^{5/3}}.$$

Then $f'(x) = 0$ when $x = 4$, and $f'(x)$ does not exist when $x = 0$ or $x = 6$, so these are the three critical points.

We can use the second derivative test—or try to. We see that $f''(4) = \frac{-8}{2^{13/3}} = -2^{-4/3} < 0$ so f has a maximum at 4. But at 0 and at 6, the second derivative isn't defined, so the second derivative test isn't useful there.

But we can still use the first derivative test. We get a table:

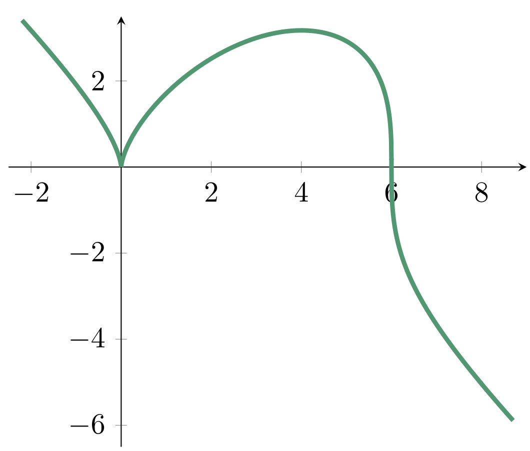
	$4-x$	$x^{-1/3}$	$(6-x)^{-2/3}$	$f'(x)$
$(-\infty, 0)$	$+$	$-$	$+$	$-$
$(0, 4)$	$+$	$+$	$+$	$+$
$(4, 6)$	$-$	$+$	$+$	$-$
$(6, \infty)$	$-$	$+$	$+$	$-$

This tells us that f has a minimum of $f(0) = 0$ at 0 and a maximum of $f(4) = 2^{5/3}$ at 4. It doesn't have a local maximum or minimum at 6.

But now we can do one more thing. Our table tells us that f is increasing on the interval $(0, 4)$, and it's decreasing elsewhere. And further, we can do the same thing for the second derivative. The second derivative is zero, or undefined, at 0 and at 6. So we get

	-8	$x^{-4/3}$	$(6-x)^{-5/3}$	$f'(x)$
$(-\infty, 0)$	$-$	$+$	$+$	$-$
$(0, 6)$	$-$	$+$	$+$	$-$
$(6, \infty)$	$-$	$+$	$-$	$+$

So the function is concave down for $x < 6$ and concave up for $x > 6$. We say that $x = 6$ is a *point of inflection* for this function, where the concavity changes. And we can use this information to sketch an effective graph of the function:



Curve sketching will be the next topic we take up, but before we get there, let's see one more example where the second derivative test makes things a bit easier:

Example. Find and classify the critical points of $f(x) = x + 2\sin(x)$.

We begin by calculating that $f'(x) = 1 + 2\cos(x)$, and this is defined everywhere, meaning the only critical points to worry about are the ones where $f'(c) = 0$. Solving for these, we want

$\cos(x) = -1/2$, which happens whenever x is of the form $2\pi/3 + 2k\pi$ or $4\pi/3 + 2k\pi$ where k is an integer.

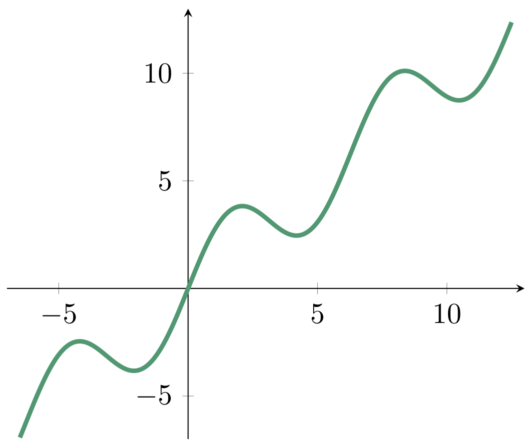
To classify these as minima, maxima, or neither, we notice that taking another derivative is quite easy here: $f''(x) = -2\sin(x)$. Then we have that

$$f''\left(\frac{2\pi}{3} + 2k\pi\right) = -2\sin\left(\frac{2\pi}{3} + 2k\pi\right) = -2\sin\left(\frac{2\pi}{3}\right) = -2\left(\frac{\sqrt{3}}{2}\right) < 0,$$

so all the points $x = 2\pi/3 + 2k\pi$ are maxima and

$$f''\left(\frac{4\pi}{3} + 2k\pi\right) = -2\sin\left(\frac{4\pi}{3} + 2k\pi\right) = -2\sin\left(\frac{4\pi}{3}\right) = -2\left(-\frac{\sqrt{3}}{2}\right) > 0,$$

so all the points $x = 4\pi/3 + 2k\pi$ are minima.



3.4 Sketching graphs of functions

3.5 Applied optimization

4 Integration

4.1 The area problem and the definition of the integral

4.2 Properties of integrals and the first Fundamental Theorem of Calculus

4.3 Antiderivatives and the second Fundamental Theorem of Calculus

4.4 Integration by substitution

4.5 Finding areas

4.6 Applications of the integral

4.7 Finding volumes, solids of revolution

Works consulted

[1] Jay Daigle. Math 1231 spring 2024 course notes. `https://jaydaigle.net/assets/courses/2024-spring/1231/course_notes_1231.pdf` Partially reproduced with permission of the author.

[2] Michael Spivak. Calculus. Cambridge University Press, 2006.

[3] Gilbert Strang and Edwin Herman. Calculus Volume I. OpenStax, 2016.