MATH1070

Calculus Group
Department of Mathematics
University of Pennsylvania

Cover Art by Joanna Raymond Ruiz

Contents

Introduction

This e-textbook, while it could be used in a straight lecture class, was written for a flipped classroom format.

In all but the introductory section (where students may not have had time to read before class) there are many self-check exercises, numbered and offset in red. It is intended that students do these when they come across them during the reading. From the self-check exercises, students can tell what they need to learn and whether they are getting what they should from reading the text. Unless marked with an asterisk, the self-check exercises are at the level where we expect you can figure them out yourself as you do the reading.

Reading before class is mandatory, as is attempting the self-check exercises. In addition to serving as a diagnostic and guide for the student, the self-check exercises let the instructor and the TA know whether students have understood each point as intended and what needs to be gone over in class.

Some conventions we adhere to are as follows. The first time a term is introduced it appears in boldface. This marks it as important and makes it easier to find when looking back. As discussed in Section 1.1, the colon-equal sign is used for defining equalities, reserving the regular equal sign for propositions that could be true or not; this is consistent with conventions in computer science.

The way this book covers content carries an expectation. Any math you know, you should know well enough to

- Explain it to someone else
- Use it to solve an interesting problem
- Recognize when it occurs in an application
- Write a coherent solution that someone else could learn from
- Remember it for many years, or at least significantly beyond the final exam.

Related to this is the emphasis on mathematical modeling. In the old days no one seemed to care about this. More recently, most calculus textbooks address this by including a number of applied examples and problems. We take this a step beyond that by directly addressing things you need to know in order to successfully apply math to physical problems, and by

including problems that are not just a reflection of the mathematical technique just learned, but require thought, organization, choices and recognition of structure. We hope most of them are also interesting.

Some of the topics you will see at the beginning of the course are thought of as high school topics or pre-calculus. The reason they are here is that we find students to be at a disadvantage if they have learned these topics enough to do well on the usual tests but without time to see the connections and subtleties. For this reason, we revisit concepts such as functions and graphing, inequalities, units, proportionality, sequences, limits and continuity. We promise to keep away from the drill and kill versions of these subjects, which you probably already had, but also to give you the instruction and support you need if even these aspects are in need of attention.

Review of basic skills

Very little of the old-sounding material is pure review. Most of it revisits topics while adding depth and connection. Pure review will be limited this section. If the contents of this section are not reasonably familiar to you, then you may have some catching up to do. This may indicate the need for help from the tutoring center, as students are generally expected to know these facts from Algebra II.

Definition 0.1. *When b is a real number and x is a positive integer, the notation b^x means multiply together x copies of b. This is called exponentiation and read as "b to the power x".*

Exponentiation obeys some basic rules. Being able to recall these by trying a simple example is just as good as having them memorized.

Proposition 0.2 (additive law of exponents)**.**

$$b^x \cdot b^y = b^{x+y}\,.$$

Proposition 0.3 (multiplicative law of exponents)**.**

$$(b^x)^y = b^{xy}\,.$$

When $b > 0$, the definition of b^x can be extended to all real values of x.

Definition 0.4 (zero power)**.** *For all $b > 0$, define $b^0 := 1$.*

Definition 0.5 (positive rational powers)**.** *For all real $b > 0$ and integers q, define $b^{1/q} := \sqrt[q]{x}$. For all real $b > 0$ and positive rational $x = p/q$, define $b^x := (\sqrt[q]{x})^p = \sqrt[q]{x^p}$.*

Definition 0.6 (negative powers)**.** *For all real $b > 0$ and positive rational $x = p/q$, define $b^{-x} = 1/b^x$.*

This next definition may not be review, because it involves limits. This motivates our upcoming discussion of limits!

Definition 0.7 (real powers)**.** *For all real $b > 1$ and real x, define $b^x = \lim_{y \to x} b^y$ as y approaches x through rational numbers.*

Because you have not yet seen limits, we include an alternate definition: b^x is the least real number z such that for all rational numbers y,

$$b^y < z \text{ if and only if } y < x\,. \tag{0.1}$$

If this seems overly formal, you can understand it intuitively by realizing that the graph of the function $f(x) := b^x$ over the domain of rationals looks like a smooth curve except the domain is full of holes, and the definition for non-rational x is the one that smoothly fills in the holes. We also remark that we have restricted to the case $b > 1$ so that b^x will be an increasing function of x and we can use a single inequality in (0.1). For $b < 1$ we can either reverse the inequalities or, realizing that $b < 1$ means $b = 1/c$ with $c > 1$, we can just define $(1/c)^x = 1/c^x$.

The logarithm, to a particular base b, is defined to be the inverse function to the function $f(x) := b^x$. Formally,

Definition 0.8 (logarithm to the base b)**.** *For any real $b > 1$, define $\log_b(x)$ to be the unique real number y such that $b^y = x$.*

From the additive and multiplicative rules for exponentiation, we can derive identities for logarithms.

Proposition 0.9 (identities for logarithms)**.**

$$\log_b(xy) = \log_b x + \log_b y \tag{0.2}$$

$$\log_b(x^c) = c \log_b x \tag{0.3}$$

$$\log_b(1/x) = -\log_b x \tag{0.4}$$

$$\log_b x = \log_c x / \log_c b \tag{0.5}$$

Proposition 0.10 (definition of the number e)**.** *There is exactly one real number b for which the slope of the graph of b^x at the point $(0, 1)$ is equal to 1. This number is roughly 2.71828 and is called e.*

Definition 0.11 (exp function and natural log)**.** *Special notation for exponents and logs to the base e are:*

$$\begin{aligned} \exp(x) &:= e^x \\ \ln(x) &:= \log_e(x) \end{aligned}$$

We can take logs (short for "logarithms") to any base, but there are three bases that are most commonly used: $2, e$ and 10. The reason for using e as a base is hinted at in Proposition 0.10, namely that it simplifies many formulas.

The reason for using 2 as a base is that powers of 2 play a big role in computer science and also are conceptually easy. We suggest you memorize at least the first ten of them: $2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, \ldots$. The fact that $2^{10} \approx 1000$ is useful in approximating things. (For example, a kilobyte refers to 1024 bytes, not 1000 bytes.) A notation sometimes used for $\log_2$ is $\lg$.

Finally, the reason for using 10 as a base because we are used to the base-10 numbering system. The size of a number is most obvious to us when we compare it to powers of 10. If a number is given in scientific notation, for example, as 3.124×10^7 we know immediately that it is a little over 31 million. In logarithm facts, the base-10 logarithm of 31 million is 7 plus the base-10 logarithm of 3.1, hence a little under 7.5. The base-10 log of a number gives a direct handle on the size of the number.

Exercise 0.1. *If M is a fifteen digit integer then $\log_{10} M$ is apprioximately what? Give lower and upper bounds: write $C \leq \log_{10} M \leq D$ where C and D are fairly simple numbers and say whether either of these inequalities must be strict ($<$) or not ($\leq$ but could be equal).*

Exercise 0.2. *Write the fact $2^{10} \approx 1000$ as a fact about logarithms.*

Approximations and bounds

This e-textbook is about using math for modeling and coming up with plausible analyses. One of the course goals is number sense. Wikipedia defines this as "an intuitive understanding of numbers, their magnitude, relationships, and how they are affected by operations."

Exercise 0.3. *If you have a model for spread of disease where the number of infections doubles every three days, how long can this go on before the model has to change: A few years? A few months? A few weeks? A few days?*

If some kind of answer began to form in your mind without your stopping to get out a calculator, then you have some of the ingredients of number sense already: perhaps you understand exponential growth, perhaps you can remember about how many people there are in the country or the world, perhaps you are familiar with powers of two and know how they relate to this problem. It is useful to be able to think this way. It's not important whether you use a calculator to answer any given question, but realistically, how often will you stop in casual conversation and whip out a calculator?

In this course, we'll teach you a number of these ingredients: use of logs, converting to powers of ten, tangent line approximation, Taylor polynomials, pairing off positive and negative summands, approximating integrals with sums and *vice versa*. Discussions of these will be brief. The point is to use them when you need them, which turns out to be nearly every lesson.

Today we're going to start with the tangent line approximation. You might think this odd because we haven't taught you calculus yet. Calculus provides a way of computing the slopes of tangent lines to graphs. But conceptually, understanding the tangent line approximation takes knowledge only of algebra and geometry, not calculus. So, we'll preview the idea now, and in fact several more ideas from the course, and then later see how to use calculus to do these analyses more methodically.

Estimating: ladder example

I am hanging wind chimes on my balcony using a ladder 5 meters long. On the highest safe step, my shoulders will be exactly at the top of the ladder, which I need to be at the height of the balcony rail, 4 meters above the ground. Every time I reposition the ladder I scratch the paint, so I'd rather not move it too many times. I need to get my shoulders within a couple of centimeters of the right height in order to drive a nail into the lintel. Where should I put the base of the ladder? The Pythagorean theorem tells me that it should be 3 meters from the wall; see Figure 1. Unfortunately, I didn't measure right, maybe because of the wide hedge at the base of the wall. I am 20 cm too low. Now what?

Solution: Let h be the function representing the height of the ladder as a function of the

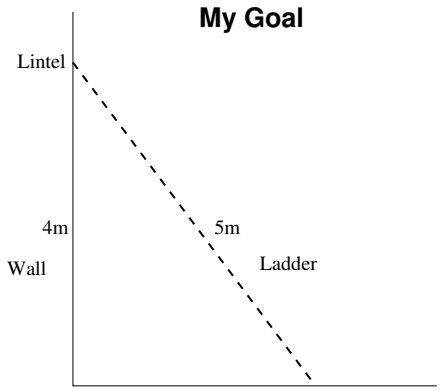


Figure 1

position of the base, in other words, $h(x)$ is the height of the ladder on the wall (in meters) when the base is x meters from the wall. By the Pythagorean Theorem, $h(x) = \sqrt{25 - x^2}$.

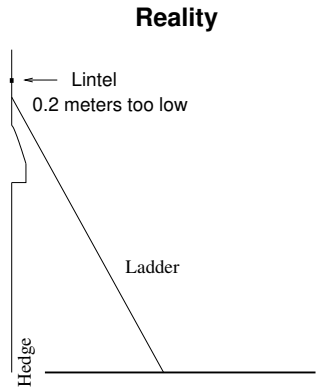


Figure 2

The height I am trying to reach is shown in Figure 2. which has $x = 3$ and $h(x) = 4$. Instead I hit some other point z with $h(z) = 3.8$. Clearly z is too far from the wall. How far do I need to scoot the ladder toward the wall? As you can see in figure 2, due to the balcony and the hedge, it was not feasible to measure either the height of the lintel or the distance I placed the foot of the ladder more accurately.

Figure 3 shows the graph of h and a tangent line to the graph of h at the point $(3, 4)$. The tangent line is a very good approximation to the graph near $(3, 4)$. For values of x between perhaps 2.6 and 3.4, the line is still visually indistinguishable form the graph. If we know

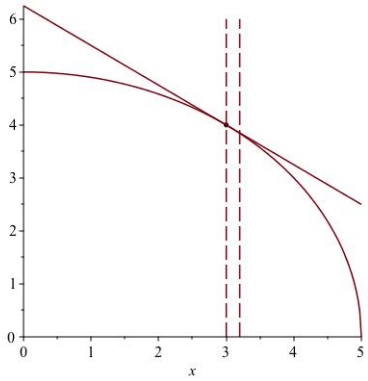
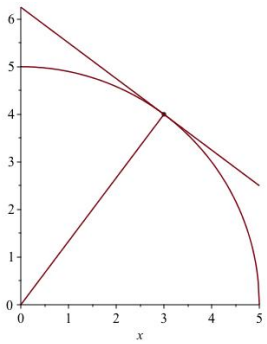


Figure 3

the slope of this line, m, we can write the equation of the line: $(y-4) = m(x-3)$. Because $h(x)$ is very nearly equal to this y (because the curve nearly coincides with the line), we can write $h(x) \approx 4 + m(x-3)$. The wiggly equal sign is not a formal mathematical symbol. Here, it means the two will be close, but has no guarantee of how close, and furthermore, it is only supposed to be close when x is close to 3. This is called an **estimate**. Shortly, we will talk about bounds: estimates that do come with guarantees.

What is the slope of this line? The graph is a quarter-circle. Recall from geometry that any tangent to a circle makes a right angle with the radius. The slope of the radius from $(0,0)$ to $(3,4)$ is $s = 4/3$. The slope of any line making a right angle with this is the negative reciprocal $-1/s = -3/4$. In other words, the slope of the tangent line is $-3/4$. That means to move the tip of the ladder up 0.2 meters, I need the base to be $0.2/(-3/4)$ meters farther from the wall, that is, 0.15 meters closer.



The reason we chose this particular example to demonstrate the tangent line approximation is that we could compute the slope with high school geometry. With calculus, we can do this for pretty much any function we can write down. In fact the word **calculus** when it was invented meant literally "a method of computing".

Bounding

To get an upper bound on $f(x)$ means to find a quantity $U(x)$ that you understand better than $f(x)$ for which you can prove that $U(x) \geq f(x)$. A lower bound is a quantity $L(x)$ that you understand better than $f(x)$ and that you can prove to satisfy $L(x) \leq f(x)$. If you have both a lower and upper bound, then $f(x)$ is stuck for certain in the interval $[L(x), U(x)]$. The smaller the upper bound and the bigger the lower bound, the better, because this traps the value of $f(x)$ in a smaller interval $[L(x), U(x)]$.

While estimating produces statements that are not mathematically well defined, bounding produces inequalities with precise mathematical meaning. Two ways we typically find bounds are as follows.

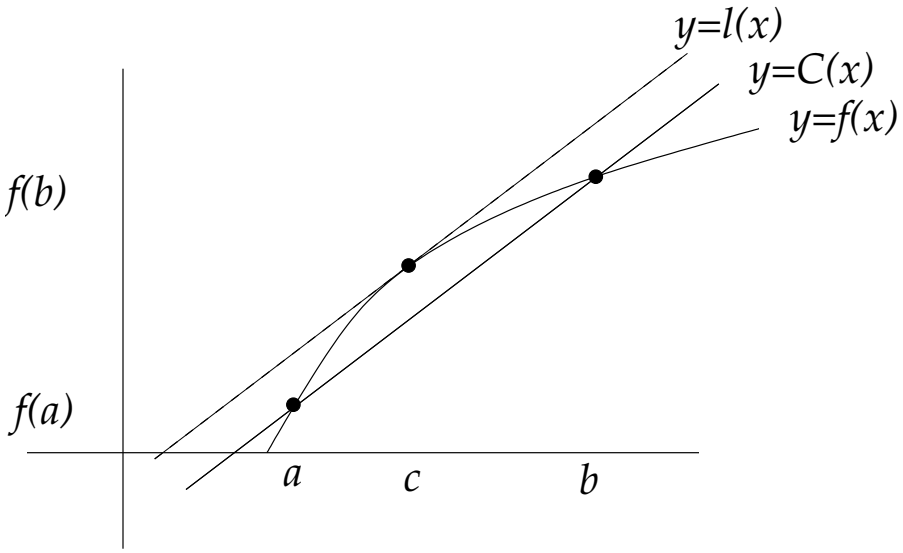
First, if f is increasing then an easy upper bound for $f(x)$ is $f(u)$ for any $u \geq x$ for which we can compute $f(u)$. Similarly an easy lower bound is $f(v)$ for any $v \leq x$ for which we can compute $f(v)$. If f is decreasing, you can swap the roles of u and v in finding upper and lower bounds. There are even stupider bounds that are still useful, such as $f(x) \leq C$ if f is a function that never gets above C. The goal in this case is to pick u and v as close to x as possible while still being able to compute $f(u)$ and $f(v)$.

Example 0.12. Suppose $f(x) = \sin(1)$. The easiest upper and lower bounds are 1 and -1 respectively because sin never goes above 1 or below -1. A better lower bound is 0 because $\sin(x)$ remains positive until $x = \pi/2$ and $1 < \pi/2$. You might in fact recall that one radian is just a bit under $60°$, meaning that $\sin(60°) = \sqrt{3}/2 \approx 0.0866\ldots$ is an upper bound for $\sin(1)$. Computing more carefully, we find that a radian is also less than $58°$. Is $\sin(58°)$ a better upper bound? Probably not, because we don't know how to calculate it, so it's not a quantity we understand better. Of course is we had an old-fashioned table of sines, and all we could remember about one radian is that it is between $57°$ and $58°$, then $\sin(58°)$ would be an excellent upper bound.

Exercise 0.4. *Which is the best of these three choices for a lower bound for $\sqrt{10}$, and why? (a) 3, because we know $3^2 = 9 < 10$; (b) $\sqrt{10} - 0.001$ because this is less than $\sqrt{10}$ by definition, but not by much; (c) 2, because $2^2 = 4 < 10$ and you don't have to think as hard to see this.*

Concavity

A more subtle bound comes when f is known to be concave upward or downward in some region. A **chord** of a graph is a line segment connecting two points on the graph. By definition, a concave upward function lies below its chords and a concave downward function lies above its chords.



In the figure, the function $f(x)$ is concave down, meaning it bends downwards. As long as x is in the interval $[a, b]$, we are guaranteed to have $C(x) \leq f(x)$. Looking at tangent lines instead of chords, if a function is concave down on an interval, then the function always lies below the tangent line. Therefore $l(x)$ is an upper bound for $f(x)$ when $x \in [a, b]$ no matter at which point $c \in [a, b]$ we choose to take the linear approximation. The figure shows the function $y = f(x)$ trapped between the chord and the tangent line over the interval spanned by the chord.

Exercise 0.5. *Did 0.15 meters over- or under-estimate how far we needed to move the base of the ladder?*

In the ladder example, we were lucky that the graph was a familiar geometric shape, a quarter circle, which we know to be convex. We are able to conclude that the tangent line remains above the graph because we know geometrically that the tangent line to a circle touches the circle at one point and otherwise remains outside the circle. Calculus will give us a far more general way to determine concavity.

And now for something completely different: Logarithm Cheat Sheet

These values are accurate[1] to within 1%:

$$\begin{aligned} e &\approx 2.7 \\ \ln(2) &\approx 0.7 \\ \ln(10) &\approx 2.3 \\ \log_{10}(2) &\approx 0.3 \\ \log_{10}(3) &\approx 0.48 \\ e^3 &\approx 20 \end{aligned}$$

Some other useful quantities to with 1%:

$$\begin{aligned} \pi &\approx \frac{22}{7} \\ \sqrt{10} &\approx \pi \\ \sqrt{2} &\approx 1.4 \\ \sqrt{1/2} &\approx 0.7 \\ e^8 &\approx 3000 \end{aligned}$$

Also useful sometimes: $\sqrt{3} = 1.732\ldots$ and $\sqrt{5} = 2.236\ldots$ both to within about 0.003%.

[1] OK so technically $\sqrt{2}$ is about 1.015% greater than 1.4 and 0.7 is about 1.015% less than $\sqrt{1/2}$

1 Variables, functions and graphs

If we count pre-calculus/trigonometry as a pre-requisite, then functions and their graphs are a pre-pre-requisite! But that doesn't mean that you have familiarity with every aspect of these. Recognition of basic types of functions is crucial for being able to use mathematics for modeling and to handle material at the pace and level you will need. So is the ability to go back and forth between analytic expressions for functions and their graphs. So is number sense: knowing approximate values without stopping for a detailed calculuation. So is knowledge of how to use physical units in a math problem. We expect most of these to be unfamiliar to many of you, and have included explanations and some homework; this may be challenging, even though it is on pre-college material. We hope it will be at least somewhat interesting!

In addition, there are some more routine things to discuss up front. In order to have a shared language, we need to agree on notation and terminology. Normally it is a good idea to read everything that is assigned; however if this notation is very familiar, you can probably just answer the self-check questions and skip the reading. We apologize for the length of this preliminary section. When the material becomes harder, the sections will be shorter.

1.1 Notation and terminology

There are several ways to conceive of a function. One is that it is a **rule** that takes an input and gives you an output. This is how most of us think of functions most of the time, but it is not precise (rules are sentences which may be ambiguous or underspecified). For this reason we also need a **formal definition**. A third way to understand functions is via their **graphs**. We now discuss all three of these ways of characterizing a function, beginning with the most formal.

Definition 1.1.

(i) A function is a set of ordered pairs with the property that no two ordered pairs have the same first element.

(ii) The expression $f(x)$ is defined to equal to y if the ordered pair (x, y) is in the set of ordered pairs defining f and undefined otherwise. Informally, $f(x)$ is called the **value** *of the function f evaluated at the* **argument** *x.*

(iii) The **domain** *of f is defined to the set of all first elements of the ordered pairs. The* **range** *of f is defined to be the set of all second elements of the ordered pairs.*

Now let's say the same things verbally. The **domain** of a function is the set of allowed inputs; the **range** is the set of all outputs. We often name functions with letters; f is the typical choice, then g if another is needed, but of course we could name a function anything. While it is common to refer to the function f as $f(x)$, we will try to observe the distinction that f is the function and $f(x)$ is its **value** at the **argument** x, meaning the output when you plug in x as an input. The condition that first coordinates are distinct corresponds to the rule producing an unambiguous answer.

Finally, to describe the function f via its graph, we interpret the ordered pairs as points in the plane, and draw this set as a curve. The condition that first coordinates are distinct corresponds to the so-called **vertical line rule**: any vertical line (vertical lines being sets with a single fixed x-coordinate but all possible y-coordinates) intersects the graph at most once.

In common usage, one might encounter any of the three ways of defining or referring to a function. We don't want to drown in formality, so we usually use something only as formal as needed. Let's look at why we sometimes need formality.

Example 1.2. Suppose we define a function f by $f(x) := x^2 + 2$. Have we formally defined this function? It sounds as if this is the set of ordered pairs

$$\{\ldots, (-2, 6), (-1, 3), (0, 2), (1, 3), (2, 6), \ldots\}.$$

That would be if we meant the domain to be the set of all integers. Maybe instead we meant the domain to be the set of all real numbers. In that case, the "..." in the list is somewhat misleading; we should probably write the set of ordered pairs like this: $\{(x, x^2 + 2) : x \in \mathbb{R}\}$ (we use the notation $\mathbb{R}$ for the real numbers and $\in$ for the "is an element of"). If this function arose in a word problem where $f(x)$ represented the value of some quantity at a time x seconds after the start, maybe it makes sense to allow only nonnegative real numbers as inputs. Formally, this would look like $\{(x, x^2 + 2) : x \text{ is real and nonnegative}\}$, which could also be written $\{(x, x^2 + 2) : x \in [0, \infty)\}$ or $\{(x, x^2 + 2) : x \geq 0\}$, this last version assuming we understood this to mean real numbers at least zero rather than, say, integers at least zero.

Technically, our discussion of the function $x \mapsto x^2 + 2$ referred to three different functions: one whose domain was all integers, one whose domain was all reals, and one whose domain

is all nonnegative reals. You can see they are different functions: even though the defining equation $f(x) := x^2+2$ is the same for all three, they are defined by different sets of ordered pairs. On the other hand, for many purposes, we don't care which of these functions was intended. We can feel free to define the function by $f(x) := x^2 + 2$ without specifying the domain unless and until we get into trouble with the ambiguity in the domain. If we try to answer a question like "How many solutions are there to $f(x) = 3$?" then we will need to be more precise about the domain.

Exercise 1.1. *What are the respective numbers of solutions to $f(x) = 3$ when $f(x) := x^2+2$ and the domain is respectively (a) the integers, (b) the reals, (c) the nonnegative reals?*

In the discussion so far, we have introduced four notations you are probably familiar with, but to be completely explicit, we discuss each briefly.

Maps-to notation. Often we name a function when defining it, then refer to it by name, but we can also refer to it using the "maps-to" symbol $\mapsto$. Thus, $x \mapsto x^2 + 2$ refers to the function that we named f, above. We use this when mentioning a function but rarely when evaluating it at an argument because the notation $(x \mapsto x^2 + 2)(3)$ is an atrocity (but technically equal to 11).

Open and closed interval notation. The interval $[a, b]$ refers to all real numbers x such that $a \leq x \leq b$. When both endpoints are included, this is called a **closed** interval. The interval (a, b) refers to all real numbers x such that $a < x < b$. When both endpoints are excluded, this is called an **open** interval. [Warning: the notation is exactly the same as for an ordered pair! If there is any ambiguity we will try to specify which, for example, "Let (a, b) be the open interval..."] The notations $(a, b]$ and $[a, b)$ are called *half-open* and refer to an interval with one point (the one next to the square bracket) included and one excluded.

Subset notation. To define a subset of some set S, we write $\{x \in S : \cdots\}$ where instead of the three dots we write a property of x that can be true or false. In some books the colon is replaced by a vertical line, the words "such that" or the abbreviation *s.t.* . If the set S is the set of all real numbers it is sometimes omitted. Thus, for example, $\{x : a \leq x < b\}$ refers to the half open interval of real numbers, $[a, b)$.

The defining colon-equal sign. We use $:=$ to mean that the quantity on the left is defined to be the quantity on the right, and a regular equal sign to mean an equation that could hold for some values of the variables and fail for others. Thus, $f(x) := x^2 + 2$ defines a function, whereas $f(x) = x^2 + 2$ is an equation which is true when a given function f,

evaluated at x, has the same value as $x^2 + 2$, and false otherwise.

Exercise 1.2. *Suppose $f(x) := x^2 + 2$. For each of the domains (a)–(c) in Exercise 1.1, write the set of values of x that make the equation $f(x) = 5 - 3x^2$ true. Please simplify your answer(s). Here and throughout, the empty set is denoted by $\emptyset$.*

One final remark about the basic definitions: there is an ambiguity in common usage of the word "range". Sometimes "range" is used to refer to a bigger set than in our definition, namely the set of all things of the type that the function outputs. For example, someone might say that the domain and range of a function $f(x) := x^2 + 2$ is all real numbers. We won't do that here, but you may come across it elsewhere. In this text, technically the range is the set of real numbers that are at least 2.

Exercise 1.3. *What are two formal mathematical ways of writing the set of real numbers that are at least 2, one using set-builder notation and one using interval notation?*

Definition by cases

As we said, the most familiar way of referring to a function is as a rule for converting input to output. Usually the rule is an equation, such as $f(x) := C - x \cdot e^x$, but the rule could be verbal, for example, "Let $f(t)$ be the amount in tons of carbon dioxide emitted in t years." Sometimes we want to talk about functions that are defined by equations, but different ones in different parts of the domain. This is called **definition by cases**. An example from a recent research paper looks like this:

$$f(x) := \begin{cases} -9x & a \leq -3 \\ 2x^2 - 3x & -3 < x < 1 \\ -a^3 & a \geq 1 \end{cases} .$$

A number of useful functions can be defined in this way. For example the absolute value of x, denoted $|x|$, may also be defined in cases:

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

Some remarks on defining by cases:

1. Note that x and $-x$ agree at $x = 0$, so we could have grouped zero with either case. When this happens, writing

$$|x| := \begin{cases} x & x \geq 0 \\ -x & x \leq 0 \end{cases}$$

emphasizes this. If x and $-x$ did not agree at $x = 0$, this would be a badly formed definition.

2. There is a period following the two example definitions but not the one in the first remark. Why? Because well written math follows rules of basic grammar. These rules can be a little different on occasion, but for the most part, you should expect this text to read in complete sentences, to define variables and functions before using them, and when used within sentences, to connect and flow logically, using connecting words like "and", "because", "therefore", and punctuation such as commas and periods.

Exercise 1.4. *Which of the following defines a function whose domain is all real numbers? Explain your reasoning.*

$$f(x) := \begin{cases} x+1 & x > 2 \\ x-1 & x < 2 \end{cases};$$

$$g(x) := \begin{cases} x+1 & x \geq 2 \\ x-1 & x < 2 \end{cases};$$

$$h(x) := \begin{cases} x+1 & x \geq 2 \\ x-1 & x \leq 2 \end{cases}.$$

Free and bound variables

In the defining statement $f(x) := x^2 + 2$, it would define the same function if instead we said $f(u) := u^2 + 2$. It is the same set of order pairs, has the same graph, etc. The variable x (or in the second case, u) is said to be a **bound variable**. The bound variable in this case runs over all values in the domain of f. A variable that is not bound is **free**. For example,

in the definition $f(u) := u^2 + c$, the variable c is free. The definition of the function f depends on the value of c. If $c = 2$, it boils down to the previous definition. If $c = 1$ it is a different function. If c has not been assigned a value, then f is a function whose range is not the real numbers but rather algebraic expressions in the variable c.

Bound variables arise many times throughout this course, in fact throughout math and throughout life! Here is a list of some places bound variables occur in this course, the first two of which you have already seen.

- In the definition of a function
- In the definition of a subset
- In quantifiers
- In limits
- In the definition of a derivative
- In summations
- In the definition of an integral
- In notions of orders of magnitude and asymptotic equivalence
- In Taylor's theorem

A related notion is that of a **quantifier**. Typically we use two quantifiers, **for all** and **there exists**. These two phrases are so important that there are symbols for them. Some people find these intimidating so we won't use them, but in case you encounter them elsewhere, in math they are denoted $\forall$ and $\exists$. A typical use of quantifiers is as follows. A function f is said to be differentiable on an open interval (a, b) if (a, b) is in the domain of f and if, for all $x \in (a, b)$, the derivative $f'(x)$ exists. In this case there was only one quantifier.

Exercise 1.5. *(i) What was the one quantifier? (ii) In the above definition of differentiability, among the variables a, b and x, which are bound and which are free? Intuitively a variable is free if the final answer depends on what value you take for that variable, but bound if you have to consider many values of the variable and put the information together.*

1.2 Useful functions and properties

Here are some more useful special functions. The **greatest integer** function at the **argument** x is denoted $\lfloor x \rfloor$ defined to be the greatest integer y such that $y \leq x$. In other words,

if x is an integer then $\lfloor x \rfloor = x$; if x is positive and not an integer, then $\lfloor x \rfloor$ is the "whole number you get when you write x as a decimal and ignore what comes after the decimal point"; if x is negative and not an integer, it is -1 plus what you get when you ignore the decimals. In older texts, the same function is sometimes denoted $[x]$. This square bracket notation has largely been abandonded in favor of the "floor" notation, because (especially in computer science) we also often want to use the **ceiling** function as well. The ceiling function at the argument x is denoted $\lceil x \rceil$ and is defined to be the least integer y such that $y \geq x$. Informally, $\lfloor x \rfloor$ rounds down to the nearest integer and $\lceil x \rceil$ rounds up.

Exercise 1.6. *What is $\lfloor x \rfloor$ when x is respectively 3, 9.4, $\sqrt{2}$, 0 , -1.5? What is $\lceil x \rceil$?*

Another useful function is the sign function, not to be confused with the sine function! This is defined by

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

Another is the delta function defined by $\delta(x) = 1$ when $x = 0$ and 0 when $x \neq 0$.

Exercise 1.7. *Write the delta functions as a definition by cases.*

We now list certain properties of functions to which we will often refer. A function f is said to be **odd** if $f(-x) = -f(x)$ for all x in the domain of f. It is unclear what is meant if the domain contains x but not $-x$. Similarly an **even** function f is one satisfying $f(-x) = f(x)$.

Exercise 1.8. *For each of these functions, say whether it is odd, even or neither.*

(a) $f(x) := x^2$

(b) $f(x) = 3 - x$

(c) $f(x) = x^3 + x$

(d) $f(x) = \sin x$

(e) $f(x) = \cos x$

A function f is said to be **increasing** if $f(x) \leq f(y)$ for all values of x and y in the domain of f such that $x < y$. Informally, the value of an increasing function gets bigger if the argument gets bigger. If you change the requirement that $f(x) \leq f(y)$ to the strict inequality $f(x) < f(y)$, this defines the notion of **strictly increasing**. **Decreasing** and **strictly decreasing** functions are defined analogously but with one inequality reversed: f

is **decreasing** if $f(x) \geq f(y)$ for all x, y satisfying $x < y$. A (strictly) **monotone** function is one that is either (strictly) increasing or (strictly) decreasing.

We can also say when a function is increasing or decreasing on a part of the domain: f is increasing on the open interval (a, b) if the above inequality holds for all $x, y \in (a, b)$. For any point $c \in (a, b)$, we then also say that f is increasing at c. In other words, to say f is increasing at a point c means there is some $a < c < b$ such that f is increasing on the open interval (a, b).

Exercise 1.9. *Is the sign function strictly increasing, increasing, strictly decreasing, decreasing, or none of the above?*

Exercise 1.10. *Is the delta function monotone?*

1.3 Graphing

As you already know, points in the plane can be labeled by ordered pairs of real numbers. As you also already know, the graph of a function f is the set points in the plane corresponding to the ordered pairs $\{(x, f(x)) : x \in \text{domain of } f\}$.

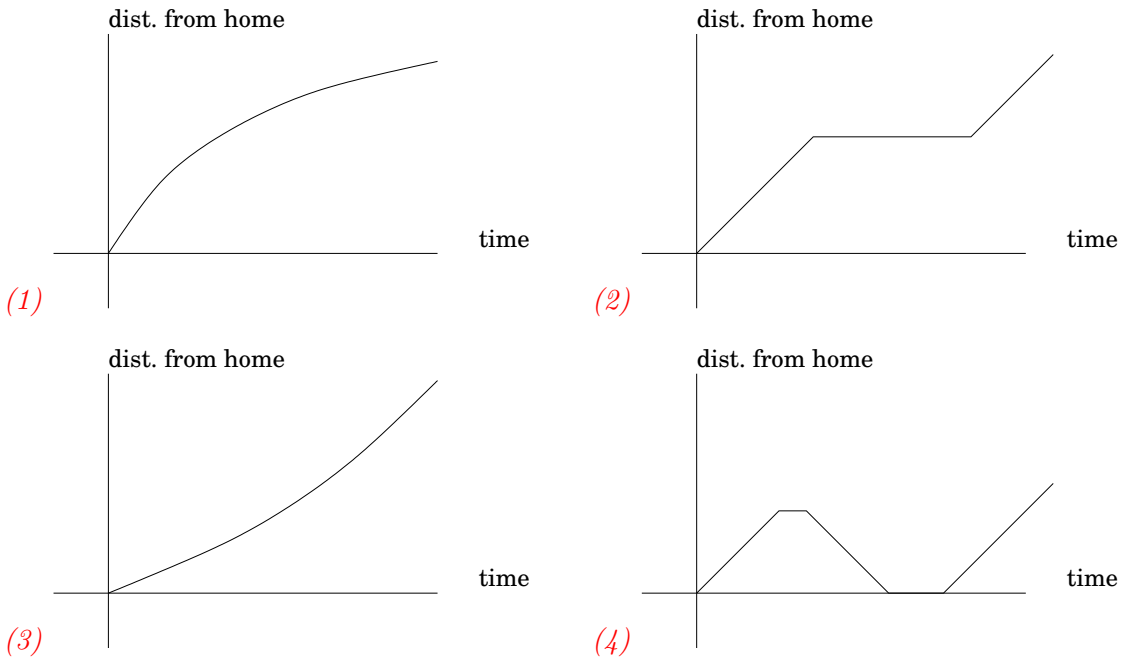
Often the graph of a function is a continuous curve, and can be quickly drawn, conveying essential information about f to the eye much more efficiently than if the reader had to wade through equations or set notation.

Exercise 1.11. *Which of the four graphs, borrowed from Hughes-Hallet et al., best matches each of the following stories?*

(a) I had just left home when I realized I had forgotten my books, so I went back to pick them up.

(b) Things went fine until I had a flat tire.

(c) I started out calmly but sped up when I realized I was going to be late.



Some conventions make graphs even more effective at conveying information. The axes should be labeled (more on that later) but more importantly, marked so that the scale is clear. Rather than just mark where 1 is on the horizontal and vertical axes, it is often helpful to mark any value where something interesting is going on: a discontuity, an asymptote, a local maximum or minimum, or a change of cases for functions defined in cases.

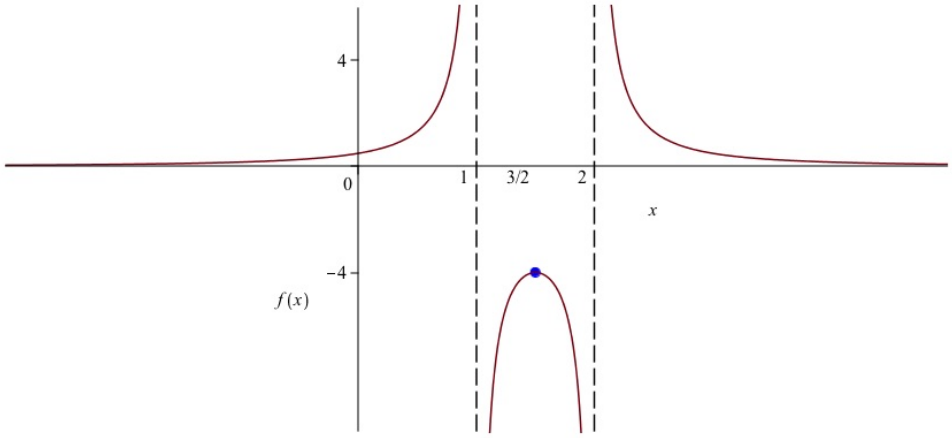


Figure 4: graph of $f(x) := 1/(x^2 - 3x + 2)$

For example, if I graph $x \mapsto 1/(x^2 - 3x + 2)$, I should mark vertical asymptotes (a certain kind of discontinuity) on the x-axis at $x = 1$ and $x = 2$; a dashed vertical line is customary. We should mark a local maximum of -4 (marked on the y-axis) occuring at $x = 3/2$ (marked on the x-axis). When graphing a function on the entire real line, we can't go to infinity and stay in scale, so we either go out of scale or draw a finite portion, large enough to given the idea. Choosing the latter, the resulting picture should look something like the graph in Figure 4. Another way to do this would be to label and mark the point $(3/2, -4)$ on the graph. There is a horizontal asymptote at zero, which we would mark with a dashed horizontal line if it occurred anywhere else, but we don't because it is hidden by the x-axis. If there is a point where an otherwise continuous function fails to be well defined, the convention is to put a small open circle. For example the function $f(x) := x^2/x$ is undefined at zero but is otherwise equal to x; its graph is shown on the left of Figure 5. A solid circle is used to denote a point where the function is defined, as in the graph of the floor function on the right of Figure 5.

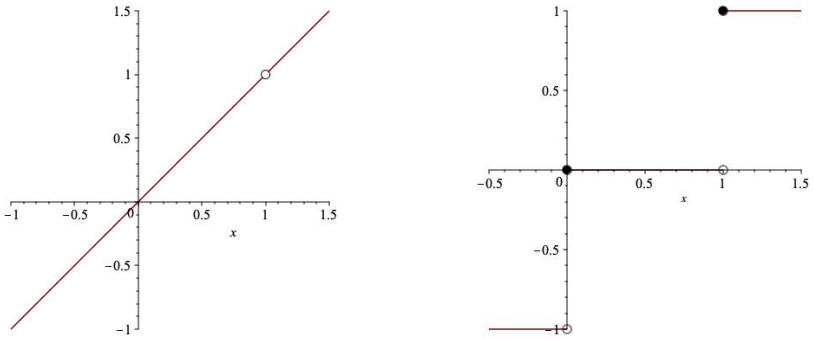


Figure 5: Showing discontinuities in a graph

Here follows a list of tips on graphing an unfamiliar function, call it f. The last three tips on shifting and scaling are ones we have found in the past that many students vaguely recall but get wrong, so please make sure you know them.

(i) Is the domain all real numbers? If not, what is it? If the function has a piecewise definition, try drawing each piece separately.

(ii) Is there an obvious symmetry? If $f(-x) = f(x)$ for all x in the domain, then f is even and there is a symmetry about the y-axis. If $f(-x) = -f(x)$ then f is odd and there is 180-degree rotational symmetry about the origin.

(iii) Are there discontinuities, and if so, where? Are there asymptotes?

(iv) Try values of the function near the discontinuities to get an idea of the shape – these are particularly important places. If the domain includes points on both sides of a discontinuity be sure to test points on each side.

(v) Try computing some easy points. Often $f(0)$ or $f(1)$ is easy to compute. Trig functions are easily evaluated at certain multiples of π.

(vi) Where is f positive?

(vii) Where is f increasing and where is it decreasing? This will be easier once you know some calculus.

(viii) Where is f concave upward versus concave downward? This will be a lot easier once you know some calculus.

(ix) Where are the maxima and minima of f and what are its values there? This will be a lot easier once you know some calculus.

(x) What does f do as x approaches ∞ and $-\infty$?

(xi) Is there a function you understand better than f which is close enough to f that their graphs look similar?

(xii) Is f periodic? Most combinations of trig functions will be periodic.

(xiii) Is the graph of f a shift of a more familiar graph? Graphing $y = f(x) + c$ shifts the graph up by c; this is pretty intuitive; if c is negative the graph shifts downward. Graphing $y = f(x + c)$ shifts the graph left or right by c. If c is positive, the graph shifts *left*.

(xiv) Is the graph of f a rescaling of a more familiar graph? The graph $y = cf(x)$ stretches vertically by a factor of c. When $c \in (-1, 1)$ this is a shrink rather than a stretch.

Exercise 1.12. *What happens when c is negative? Sketch the specific example where $c = -2$ and $f(x) = x^2$ on the domain $[-1, 1]$.*

(xv) The graph of $y = f(cx)$ stretches or shrinks in the horizontal direction. When $c > 1$, it is a shrink. Why? Try sketching $y = \cos x$ and on top of this sketch $y = \cos(2x)$.

Exercise 1.13. *Explain in words why $c \in (0, 1)$ produces a horizontal stretch. What happens when c is negative?*

1.4 Inverse functions

One method of solving the problem of Galileo's experiment involved an inverse function. Let's be explicit about the definitions involved. The inverse function of a function f is the function that answers the question,

> What input do I need to get the given output?

In other words, if g is the inverse function of f then $g(y)$ is whatever value x satisfies $f(x) = y$. If there is more than one answer to this, then f has no inverse function; however, you can usually restrict the domain so there is only one answer. If there is no answer, that's not a problem, it just means that y is not in the domain of g. This happens when y is not in the range of f. Thus, the domain of g is the range of f. Likewise, the range of g is any possible answer to the question above, therefore any x in the domain of f.

Exercise 1.14. *Let $f(x) := \sin x$ on a domain of the form $[-L, L]$, where L is some positive real number. What is the largest value of L such that f is one-to-one and therefore has an inverse?*

The usual notation for the inverse function to f is f^{-1}. This is terrible notation because it is the same as the notation for the -1 power of f, also known as $1/f$. We tried changing the inverse function notation to f^{inv} for the purposes of this class, but then students were confused when they saw f^{-1}. We will stick with the terrible notation, and mention it when confusion might arise.

There is a standard way that the domain is restricted on trig functions so that the inverse function can be defined. For sin and tan it is $[-\pi/2, \pi/2]$. The function cos when restricted to $[-\pi/2, \pi/2]$ is not one-to-one; the standard choice for cos is $[0, \pi]$. These are arbitrary conventions, but are probably built in to your calculator, so we had better adopt them. Also, along with $\sin^{-1}$, $\cos^{-1}$ and $\tan^{-1}$, the conventional names $\arcsin, \arccos$ and $\arctan$ are also used.

Exercise 1.15. *Let f be the squaring function, $f(x) := x^2$. What is the standard name of the inverse function to f, and what choice of domain of f is usually made so that f will be one-to-one?*

Inverse functions occur naturally in mathematical modeling. For example, if $f(t)$ represents how many miles you can walk in t hours, then $f^{-1}(x)$ represents how many hours it takes

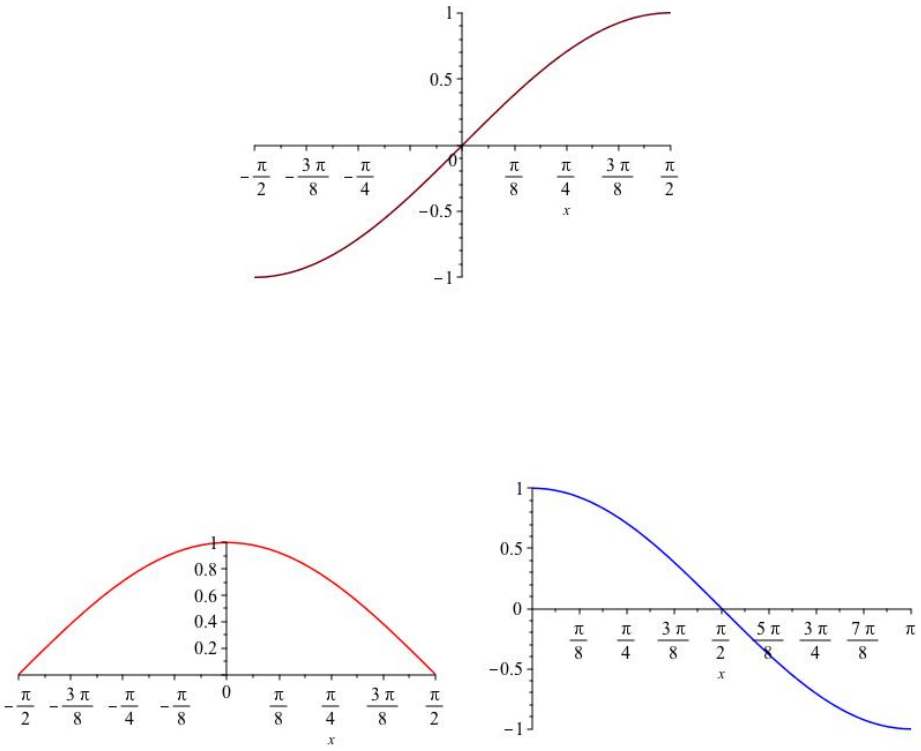


Figure 6: sin is one-to-one on $[-\pi/2, \pi/2]$ (top) but cos is not (left) so we move the window to $[0, \pi]$ (right)

you to walk x miles. Note that in this explanation, x is a bound variable; we could have used any other name, such as t again, only it helps readability if we use names such as t for time and x for distance.

Exercise 1.16. *Define $f(x)$ to be the number of pounds you have to carry when planning a backpacking excursion for x days.*

(a) Give an interpretation for $f^{-1}(v)$

(b) Give interpretations for $f^{-1}(v) + f^{-1}(w)$ and $f^{-1}(v + w)$.

(c) Which do you think would be greater?

How does the graph of an inverse function relate to the graph of a function? The roles of x and y have switched. When the first and second coordinate of an ordered pair are switched, the point reflects across the diagonal line $y = x$. Thus, the graph of the inverse function is the original graph (on the appropriate domain) reflected across the diagonal. The blue curve in Figure 7 is the plot of $f(x) := x^3 - 3x$ from $x = 1$ to $x = 3$, an interval on which f is one-to-one. The red curve shows f^{-1} on the corresponding interval $[-2, 18]$.

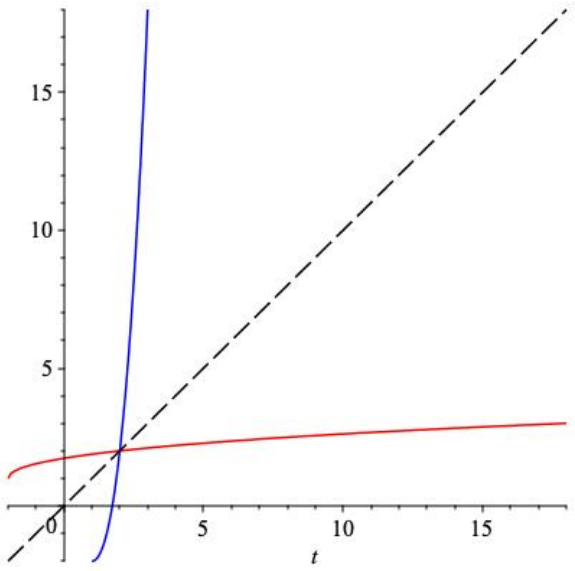


Figure 7: The function $x \mapsto x^3 - 3x$ on $[1, 3]$ and its inverse

Exercise 1.17. *Why is* $[-2, 18]$ *the "corresponding interval"?*

2 Units, proportionality and mathematical modeling

2.1 Physical units and formulas

One skill most students need practice with is writing formulas for functions given by verbal descriptions. Try this multiple choice question before going on.

Exercise 2.1. *Knowing that an inch is 2.54 centimeters, if $f(x)$ is the mass of a bug x centimeters long, what function represents the mass of a bug x inches long?*

(a) $2.54f(x)$

(b) $f(x)/2.54$

(c) $f(2.54x)$

(d) $f(x/2.54)$

It helps to think about all such problems in units. Although inches are bigger than centimeters by a factor of 2.54, numbers giving lengths in inches are *less than* numbers giving lengths in centimeters by exactly this same factor. Writing this in units prevents you from making a mistake. The quantity 1 inch is the same as the quantity 2.54 centimeters, so their quotient in either order is the number 1 (unitless). We can multiply by 1 without changing something. Thus,

$$x \text{ in} \times \frac{2.54 \text{ cm}}{1 \text{ in}} = 2.54x \text{ cm}\,.$$

This shows that replacing x by $2.54x$ converts the measurement, and therefore (c) is the correct answer. Here are some more helpful facts about units.

1. You can't add or subtract quantities unless they have the same units. That would be like adding apples and oranges!
2. Multiplying (resp. dividing) quantities multiplies (resp. divides) the units.
3. Taking a power raises the units to that power. For example, if x is in units of length, say centimeters, then $3x^2$ will have units of area, in this case square centimeters. Most functions other than powers require unitless quantites for their input. For example, in a

formula $y = e^{***}$ the quantity *** must be unitless. The same is true of logarithms and trig functions: their arguments are always unitless[2].

4. Units tell you how a quantity transforms under scale changes. For example, a square inch is 2.54^2 times as big as a square centimeter.

Exercise 2.2. *Suppose a pear growing on a tree doubles in length over the course of two weeks. By what factor does its volume increase?*

Often what we can easily tell about a function is that it is proportional to some combination of other quantities, where the **constant of proportionality** may or may not be known, or may vary from one version of the problem to another. Constants of proportionality have units, which may be computed from the fact that both sides of an equation must have the same units.

Example 2.1. If the monetization of a social networking app is proportional to the square of the number of subscribers (this representing perhaps the amount of messaging going on) then one might write $M = kN^2$ where M is monetization, N is number of subscribers and k is the constant of proportionality. You should always give units for such constants. They can be deduced from the units of everything else. The units of N are people and the units of M are dollars, so k is in dollars per square person. You can write the constant as k $\frac{\$}{\text{person}^2}$.

To say A is **inversely proportional** to B means that A is proportional to $1/B$. If a quantity A is proportional to both quantities B and C, which can vary independently, then A must be proportional to $B \cdot C$, so $A = k\,B\,C$ for some constant of proportionality, k.

Example 2.2. If the expected profit on a home sale is proportional to the assessed value of the home and inversely proportional to the number of days it has been on the market, we could capture that relation as $P = kV/T$ where P is profit in dollars, V is assessed price in dollars, T is number of days on the market, and k is a constant of proportionality.

Exercise 2.3. *What are the units of k in Example 2.2?*

Warning: Sometimes in mathematical modeling, an equation represents an empirical law, which is a rough fit to some function. For example, if it is observed that the blood volume of small mammals is roughly proportional to the 2.65 power of the mammal's length, sensible

[2]assuming we consider a radian to be physically unitless

units will not be assignable to the proportionality constant k in the formula $BV = kL^{2.65}$. In this case we just have to live with the fact that k has units involving fractional powers of length that won't make much sense outside of this context.

An important point when writing up your work: You don't just write $M = kN^2$ without stating the interpretations of the three variables. Also, there would not usually be a $:=$ here, because you are not defining the function $M(N) := kN^2$ as much as you are saying that two observed quantities M and N vary together in a way that satisfies the equation $M = kN^2$. There isn't a clear line here, but the style of the definition can be important in conveying to the reader what's going on.

Example 2.3. The present value under constant discounting is given by $V(t) = V_0 e^{-\alpha t}$ where V_0 is the initial value and α is the discount rate. What are the units of α? They have to be inverse time units because αt must be unitless. A typical discount rate is 2% per year. You could say that as "0.02 inverse years." We hope that by the end of the semester, the notion of an inverse year is somewhat intuitive.

Exercise 2.4. *Write a formula expressing the statement that risk of viral infection in an enclosed space is proportional to the square of the number of people and inversely proportional to the cube root of the volume. Be sure to give the units of the constant of proportionality.*

Often quantities are measured as proportions. For example, the proportional increase in sales is the change in sales divided by sales. In an equation: the proportional increase in S is $\Delta S/S$. Here, ΔS is the difference between the new and old values of S. You can subtract because both have the same units (sales), so ΔS has units of sales as well. That makes the proportional increase unitless. In fact proportions are always unitless.

Percentage increases are always unitless. In fact they are proportional increases multiplied by 100. Thus if the proportional increase is 0.183, the percentage increase is 18.3%. In this class we aren't going to be picky about proportion versus percentage. If you say the percentage increase is 0.183 or the proportional change is 18.3%, everyone will know exactly what you mean. But you may as well be precise.

Exercise 2.5. *The proportional increase in an animal's weight during the first week of life is observed to be exponential in the percentage of a certain protein in the blood at birth. Do the units make sense or not?*

Units behave predictably under differentiation and integration as well. We will refer back to this when we define the relevant concepts, but you may as well see a preview now. The

derivative $(d/dx)f$ has units of f divided by units of x. You can see this easily on the graph in Figure 8 because the derivative is a limit of rise over run, where rise has units of f and run has units of x. The integral $\int f(x)\,dx$ has units of f times units of x. Again you can see it from a picture (Figure 9), because the integral is an area under a graph where the y-axis has units of f and the x-axis has units of, well, x.

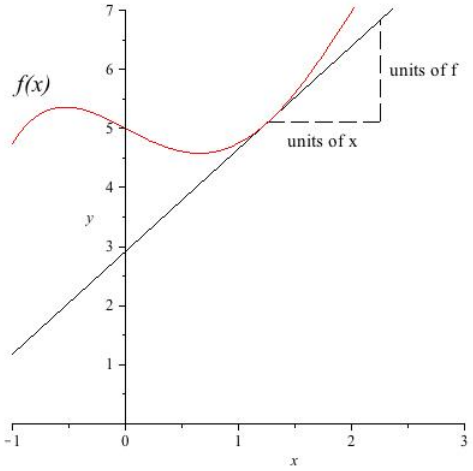


Figure 8: units of the derivative

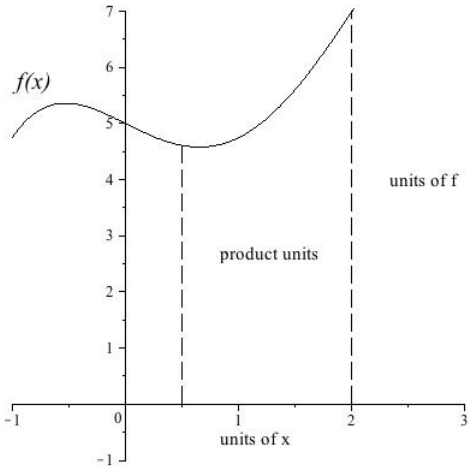


Figure 9: units of the integral

2.2 Modeling

Mathematical modeling means writing down mathematics corresponding to a given physical scenario, along with equations or other relations that could be expected to hold, at least approximately or under further assumptions.

Unpacking this, we see a number of features. First of all one must define mathematical objects in the model: variables, sets, functions, equations and so forth. Secondly, one must give **interpretations** of everything in the model. An interpretation tells what physical quantity is associated with each of the constants and variables and what relation is meant by each function. Physical quantities include units, so this part always involves stating units. Note: the interpretation tells how the math relates to the scenario; it is not itself mathematical. Thirdly, often one needs to add hypotheses about the scenario. These say the circumstances under which would you expect the mathematics to be correct for the model. This hypotheses are also physical, not mathematical. Lastly, if there are questions given in the scenario, it is necessary to say what part of the mathematics answers the question(s). After this, what is left is a math problem: solve for the quantities that answer the questions.

In the following example, we have underlined parts of the modeling exercise that reflect the outline we have given, such as naming of variables, interpretation, units and hypotheses.

Example 2.4.

Scenario: Galileo observes that objects falling a short time seem to fall a distance that was proportional to the square of the time and independent of the object: 4 feet for an object falling half a second, 9 feet after three quarters of a second, 16 feet after one second, and so forth. Galileo decides to measure the Tower of Pisa by dropping a stone from the top of the tower and measuring the time it takes for him to hear it hit the ground. Make a model for this and use it to estimate the elapsed time Galileo measured between dropping the stone and hearing the sound.

Model: Let $f(t)$ be the distance in feet that an object falls in t seconds, starting from rest. The wording of the scenario tells us that $f(t) = c\,t^2$ where c has units of feet per seconds squared. This assumes we set $t = 0$ at the time of release and measure distance from the point of release. We are asked to determine t such that $f(t) = h$, where h is the height of the Tower of Pisa. Equivalently, we need to find $f^{-1}(h)$. We assume that the model is accurate. What that means in this case is that we can ignore things such as air resistance and the time lag for

the sound of impact to get back to Galileo's ear.

Solution: We look up the height of the Tower of Pisa to find that $h = 186$ feet. We solve for c given Galileo's data for small distances and find that $c = 16$ (for example: $f(1/2) = c(1/2)^2 = 4$ implies $c = 16$). We can solve directly for $16t^2 = 186$ or we can compute the inverse function to f yielding $f^{-1}(x) = \sqrt{x/16}$ and substitute 186 for x. Either way we get $t = \sqrt{186/16} \approx 3.40$. It may sound pedantic, but probably we should justify our choice of the positive square root by saying the whole experiment only covers time after the release, that is, $t \geq 0$.

Were the hypotheses warranted? Many objects would be slowed by air resistance over such a distance. Probably Galileo would have had to drop something like a rock in order for the fall not to have been significantly slowed. Looking up the speed of sound, it would take an extra 1/4 second to register the sound. Probably Galileo could measure time to within greater accuracy that 1/4 second, so this hypothesis is definitely shaky.

Squares and Powers of 2 Cheat Sheet

If you know the powers of 2 you can do the same thing with $\log_2$ that you can do with $\log_{10}$. Because I indeed am a Geek, I have listed the first few powers of 2 and am suggesting you be at least somewhat familiar with them. By the way, you should also recognize the first twenty squares:

$$1, 4, 9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169, 196, 225, 256, 289, 324, 361, 400\,.$$

No kidding, when you come across one of these numbers under a radical, you know immediately it can be factored out. Here are the powers of 2.

$$\begin{aligned}
2^0 &= 1 \\
2^1 &= 2 \\
2^2 &= 4 \\
2^3 &= 8 \\
2^4 &= 16 \\
2^5 &= 32 \\
2^6 &= 64 \\
2^7 &= 128 \\
2^8 &= 256 \\
2^9 &= 512 \\
2^{10} &= 1,024 \\
2^{11} &= 2,048 \\
2^{12} &= 4,096 \\
2^{13} &= 8,192 \\
2^{14} &= 16,384 \\
2^{15} &= 32,768 \\
2^{16} &= 65,536 \\
2^{20} &\approx 1,000,000 \\
2^{30} &\approx 1,000,000,000 \\
2^{100} &\approx 10^{30}
\end{aligned}$$

2.3 Exponential and logarithmic relationships

The log cheatsheet is there to encourage you to use logs for quick computations. The squares and powers of two are just for fun (OK it was written by geeks). We're going to take a quick break from concepts to get the hang of computing with logs.

Example 2.5. What is the probability of getting all sixes when rolling 10 six-sided dice? It's 1 in 6^{10} but how big is that? If we use base-10 logs, we see that $\log_{10}(6^{10}) = 10\log_{10} 6 = 10(\log_{10}(2) + \log_{10}(3)) \approx 10(.78) = 7.8$. So the number we're looking for is approximately $10^{7.8}$ which is $10^7 \times 10^{0.8}$ or $10,000,000$ times a shade over $10^{.78}$, this latter quantity being very close to 6 according to the one-digit logs you computed. So we're looking at a little over sixty million to one odds against.

Exercise 2.6. *Roughly how big is 5^{11}? Just one significant digit is fine.*

These are not just random examples, it is always the best way to get a quick idea of the size of a large power. When the base is 10 we already know how many digits is has, but when the base is something else, we quickly compute $\log_{10}(b^a) = a \cdot \log_{10}(b)$.

Example 2.6. Why is the value $\ln(10) \approx 2.3$ on your log cheatsheet so important? It converts back and forth between natural and base-10 logs. Remember, $\log_{10} x = \ln x / \ln 10$. Thus the constant $\ln 10$ is an important conversion constant that just happens to be closer than it looks (the actual value is $2.302\ldots$). So for example,

$$e^8 \approx 10^{8/2.3} \approx 10^{3.5} = 1000 \times 10^{0.5} = 1000\sqrt{10} \approx 3,000\,.$$

Exercise 2.7. *Estimate $2.7^{2.3}$ using the log cheatsheet, then use a calculator to find a more accurate decimal approximation.*

Exercise 2.8. *A certain astronomical computation yields the number $\exp(24)$. How many digits will this be? (Meaning, how many digits before the decimal point.)*

Recall in the definition of e, the slope of the graph e^x at $(0,1)$ is 1, therefore the tangent line approximation is $e^x \approx 1 + x$. In case you didn't do practice problem #2, you should know that this approximation is very good when $x < 0.1$. Let's see what this means for doing typical interest computations. Suppose, for example, your company grows in value by 6% each year for 20 years. By what factor C does the value increase over this time? The answer is 1.06^{20}, but about how big is that? For a quick answer, take logs. Using the fact

that $\ln 1.06 \approx 0.06$, we see that $\ln C = \ln(1.06^{20}) \approx 20 \times 0.06 = 1.2$. We'd rather have this in base ten, so we compute $\log_{10} C = \ln C / \ln 10 \approx \ln C / 2.3 \approx 1.2/2.3 \approx 0.5$, maybe a little bigger like 0.52 or so. Looking at the log cheatsheet shows this means C should be between 3 and 4, somewhat closer to 3. In fact to two significant figures, the growth factor is 3.2.

Exercise 2.9. *Historical economists look at real (inflation-adjusted) growth rates over periods of a century or more. If the real annual growth rate averages 2%, what should be the growth factor over the century and a half from 1870 to 2020?*

The multiplicative frame

If you ask someone to state a relationship between the numbers 20 and 30, the most common answer is that 30 is ten more than 20. A more fundamental answer is that 30 is 50% more, or equivalently that 30 is three halves of 20. The section on proportionality is designed to emphasize multipicative thinking over additive thinking. Additive thinking is more common only because we find it computationally easier to add than to multiply. Saying that multiplicative thinking is more fundamental is not a precise mathematical statement, so there's no way to prove it. One reason to believe it is that the statement remains the same no matter what units you use (as long as the 30 and the 20 are in the same units).

Exercise 2.10. *A small city has 40,000 households. To organize an emergency response system, the city wants to organize groups of households on a scale "halfway between the individual household and entire city scale." What size of groups of households best fulfills this?*

Exponentials and logarithms are built to express multiplicative facts. In fact the additive laws of exponentiation and logarithms basically convert multiplicative facts to additive facts, thereby converting the more fundamental fact to the type you can compute more easily.

Much of what you learn on topic of exponential and logarithmic relationships insights such as this one:

> If you observe that $\ln x$ has increased by about 0.7, what does this mean about the increase that has occurred in x?

A tip about setting up equations representing functional relationships: when a quantity has different values at different times such as "before" and "after", using one variable to

represent both quantities can lead to mess and confusion. Better to use different names such as x_1 and x_2, or x_{init} and x_{final}, or possibly x and x', etc. Using this idea on the question above sets up an equation like this: $\ln x_2 \approx \ln x_1 + 0.7$. From here, exponentiating leads to

$$x_2 \approx e^{\ln x_1 + 0.7} = x_1 \cdot e^{0.7} \approx 2x_1 \,.$$

So, if you observe $\ln x$ increasing by about 0.7, you will know that x had approximately doubled. This is what it means that logarithms transfer multiplicative scales to additive ones. A multiplicative relation such as doubling transfers to an additive relation, namely addition of about 0.7.

Exercise 2.11. *When x triples, what happens to the base-ten log of x? What about the natural log of x?*

Exercise 2.12. *(*) If $\ln x$ has tripled, what has happened to x?*

One more thing to keep in mind about logarithms and exponentials is that they do not scale with units. If I change the units of x from inches to centimeters, and if $y = e^x$, then in the new units $y' = e^{2.54x'} = y^{2.54}$. The new exponential appears to be the old one to the 2.54 power. What does that even mean? It is a tipoff that x should not be exponentiated: anything other than a unitless constant is likely to be meaningless when exponentiated. The same is true for logarithms and trig functions.

If $\log x$ increases at a constant additive rate, then x increases at a constant multiplicative rate. What does this mean?

If a quantity Q increases at a constant additive rate, it means that if you wait one unit of time, Q always increases by the same additive amount. In fact, between any two times s and t the increase will be $c(t - s)$.

Exercise 2.13. *What are the units of c in this case?*

If a quantity Q increases at a constant *multiplicative* rate, it means waiting one unit of time always multiples Q by the same amount, and in general, between times s and t, the factor by which Q increases will be c^{t-s} where c is the factor by which Q increases in one unit of time.

Exercise 2.14. *What are the units of c in this case?*

To get back to the question of what it means about logs relating additive to multiplicative growth, if $\log x = a + bt$ (constant additive growth over time) then $x = e^{a+bt} = e^a e^{bt} = AB^t$

where $A = e^a$ and $B = e^b$. This is constant multiplicative growth.

Constant multiplicative growth rates occur in a lot of applications. This is also called *exponential growth* because the formula for a quantity growing multiplicatively is Ae^{bt} (also e^{a+bt} or AB^t). When $b < 0$, it is called exponential *decay* or *decrease*.

Here are a few examples. Equilibrating: if an item is hotter or colder than its environment then the temperature difference between the object and its environment, as a function of time, decreases exponentially (here, in Ae^{bt}, the coefficient b is negative). Money accumulating (fixed rate) interest grows exponentially. So, unfortunately does debt (just put a minus sign on the money). Population tends to grow this way (again unfortunately, in most cases). Radioactive substances decay exponentially. So does the portion of DNA remaining unmutated. Present value analyses, under a fixed discount rate, imply exponential decrease of the present value for revenue at future times. Time series data for which the correlations decay exponentially are common.

If we get to assume a nice clean exponential model, and can observe at more than one time point, then exponential growth/decay models are nearly as easy to solve as linear growth models (a highlight of eighth grade math). You should learn this both conceptually and as a mindless skill.

Example 2.7. A viral infection is spreading exponentially through the community. On the first day that the outbreak had a name, there were 25 infections. A week later there were 40 infections. How many infections will there be in another two weeks? When will the number of infections reach 200,000, which is the size of the entire local population?

Solution #1 (plug in logs): Let $N(t)$ denote the number of infections after t weeks. Our model is $N(t) = Ae^{bt}$. The given information is that plugging in $t = 0$ and $t = 1$ give $N = 25$ and $N = 40$ respectively. Because $e^0 = 1$, we have $25 = A$, while $40 = Ae^b$. This gives $e^b = 40/25 = 8/5$, hence $b = \ln(8/5)$. In another two weeks we will have $t = 3$, so

$$N(3) = 25e^{3\ln(8/5)}\,.$$

When $N = 200,000$ we have $25e^{t\log(8/5)} = 200,000$ hence

$$e^{t\ln(8/5)} = \frac{200,000}{25} = 8,000 \text{ hence } t = \frac{\ln 8000}{\ln(8/5)}\,.$$

Solution #2 (growth factor): If we use the growth factor B in the equation AB^t instead of the exponential constant b in Ae^{bt} we may get away without logs. In a week the increase

was from 25 to 40, a factor of 8/5 so clearly $B = 8/5$. Thus $N = 25(8/5)^t$. In three weeks we have $N(3) = 25(8/5)^3 = 512/5 = 102.5$. Evidently the expression $25e^{3\ln(8/5)}$ can be simplified! The time needed to get to 200,000, a growth factor of 8000, is t such that $(8/5)^t = 8000$. This is, by definition $\log_{8/5} 8000$, which is equal to $\log_b 8000/\log_b(8/5)$. The previous answer was a special case of this in base e, but the ratio of two logs is the same in any base. Using base ten, for example, we get approximately $3.9/(0.9 - 0.7) = 19.5$. So it should take between nineteen and twenty weeks to saturate the city.

Exercise 2.15. *Is exponential growth a more realistic model when a small portion of the population is infected or when a large portion is infected?*

3 Limits

You might not think limits would show up in a calculus course oriented toward application. Wrong! There are a lot of reasons why you need to understand the basics of limits. You should know these reasons, so here they are.

1. You have already seen they show up in the definition of powers and logarithms when the exponent is not rational.
2. The definition of derivative (instantaneous rate of change) is a limit.
3. The number e is defined by a limit.
4. Continuous compounding is a limit.
5. Limits are needed to understand improper integrals, such as the integrals of probability densities.
6. Infinite series, which we will discuss briefly, require limits.
7. Discussing relative sizes of functions is really about limits.

3.1 Definitions of limit

You should learn to understand limits in four ways:

Intuitive
Pictorial
Formal
Computational

Intuitive: The limit as $x \to a$ of $f(x)$ is the numerical value (if any) that $f(x)$ gets close to when x gets close to (but does not equal) a. This is denoted $\lim_{x\to a} f(x)$. If we only let x approach a from one side, say from the right, we get the one-sided limit $\lim_{x\to a^+} f(x)$.

Please observe the syntax: If I tell you a function f and a value a then the expression $\lim_{x\to a} f(x)$ takes on a numerical value or "undefined". The variable x is a bound variable; it does not have a value in the expression and does not appear in the answer; it stands for

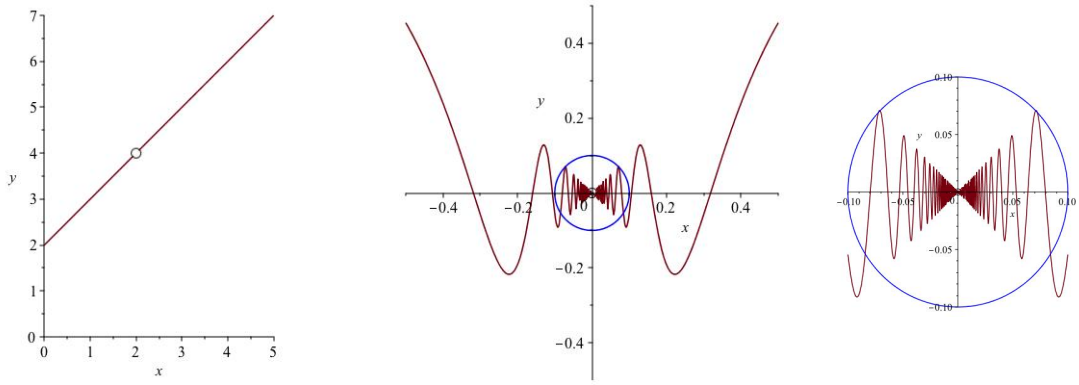


Figure 10: Left: $f(x) = x+2$ except that f is undefined at $x = 2$; Center: a wiggly function near zero; Right: zooming in on the wiggly function at zero

a continuum of possible values approaching a. The variable a is free and does show up in the answer; for example $\lim_{x\to a} x^2$ is equal to a^2.

Pictorial: If the graph of f appears to zero in on a point (a, b) as the x-coordinate gets closer to a, then b is the limit, even if the actual point (a, b) is not on the graph. For example, suppose $f(x) = \dfrac{x^2-4}{x-2}$. Canceling the factor of $x-2$ from top and bottom, you can see this is equal to $x+2$, except when $x = 2$ because then you get zero divided by zero. Functions like this are not just made up for this problem. They occur naturally when solving simple differential equations, where indeed something different might happen if $x = 2$. The graph of f has a hole in it, which we usually depict as an open circle, as in the left side of Figure 10. The value of $\lim_{x\to 2} f(x)$ is 2, even though f is undefined precisely at 2.

In this example the function f behaved very nicely everywhere except 2, growing steadily at a linear rate. The center figure shows the somewhat less well behaved function $g(x) := x\sin(1/x)$. This function is undefined at zero. As x approaches zero, the function wiggles back and forth an infinite number of times, but the wiggles are smaller and smaller. Intuitively, the value of the function g seems to approach zero as x approaches zero. Pictorially we see this too: zooming in on $x = 0$ in the right-hand figure, corroborates that $g(x)$ approaches zero.

Exercise 3.1. *Sketch the function* $f(x) := \begin{cases} e^{-x} & x \geq 0 \\ 0 & x < 0 \end{cases}$ *; see Exercise 3.2, upcoming. Does*

the limit $\lim_{x\to 0} f(x)$ *exist?*

We can take limits at infinity as well as at a finite number. The limit as $x \to \infty$ is particularly easy visually: if $f(x)$ gets close to a number C as $x \to \infty$ then f will have a horizontal asymptote[3] at height C. Thus $3 + \dfrac{1}{x}$, $3e^{-x}$ and $3 + \dfrac{\sin x}{x}$ all have limit 3 as $x \to \infty$, as shown in Figure 11

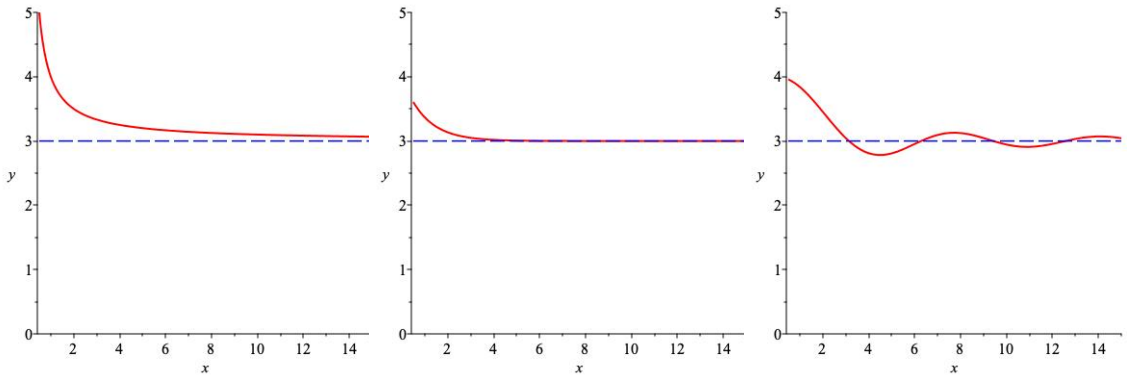


Figure 11: Three functions all having limiting value 3 as $x \to +\infty$

Formal: The precise definition of a limit is a little unexpected if you've never seen it before. We don't define the value of $\lim_{x\to a} f(x)$. Instead, we define when the statement $\lim_{x\to a} f(x) = L$ is true. It can be true for at most one value L. If there is such an L, we call this the limit. If there is no L, we say the limit does not exist. When asked for the value of $\lim_{x\to a} f(x)$, you should answer with either a real number, or "DNE", for "does not exist". We won't have to spend a lot of time on the formal definition. You should see and grasp it at least once. Use of the Greek letters ε and δ for the bound variables is a strong tradition.

Definition 3.1. *If f is a function whose domain includes an open interval containing the real number a, we say that* $\lim_{x\to a} f(x) = L$ *if and only if the following statement is true.*

> *For any positive real number ε (think of this as acceptable tolerance in the y value) there is a corresponding positive real δ (think of this as guaranteed accuracy in the x-value) such that for any x other than a in the interval $[a-\delta, a+\delta]$, $f(x)$ is guaranteed to be in the interval $[L-\varepsilon, L+\varepsilon]$.*

[3] A horizontal line can be an asymptote for f even if f crosses back and forth over the line; we will see a formal definition soon in Definition 3.6.

In symbols, the logical implication that must hold is:

$$0 < |x - a| < \delta \Longrightarrow |f(x) - L| < \varepsilon\,.$$

Remark. Loosely speaking, you can think of ε as an acceptable error tolerance and δ as how tightly you control the input. The limit statement says, you can meet even the pickiest error tolerance provided you can tune the input sufficiently well. Why is this a difficult definition? Chiefly because of the quantifiers. The logical form of the condition that must hold is: *For all* $\varepsilon > 0$ *there exists* $\delta > 0$ *such that for all* $x \in [a - \delta, a + \delta]$, $\cdots$. This has three alternating quantifiers (for all... there exists... such that for all...) as well as an if-then statement after all this. Experience shows that most people can easily grasp one quantifier "for all" or "there exists", but that two is tricky: "for all ε there exists a δ . . .". A three quantifier statement usually takes mathematical training to unravel.

Some people find it easier to conceive of the formal definition as a game. Alice is trying to show it's true. Bob is trying to show it's false. Alice says to Bob, no matter what ε you give me, I can find a δ to make the implication true. (The implication is that all x-values fitting into Alice's δ-interval will give values of $f(x)$ inside Bob's interval.) Now they play the game: Bob tries to come up with a value of ε so small as to thwart Alice. Then Alice has to say her δ. If she can always do so (assuming Bob has not made a blunder in overlooking the right choice of ε) she wins and the limit is L. If not (unless Alice has overlooked a δ that would have worked), Bob has won and the limit is not L.

3.2 Variations

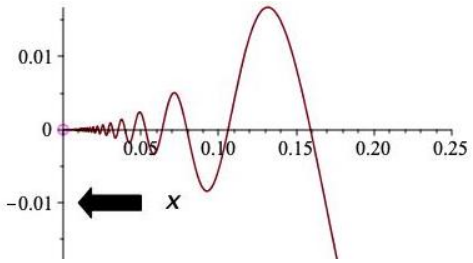
Before introducing computational apparatus for limits, we need to finish the definitions by defining some variations: one-sided limits, limits at infinity and "limits of infinity" (which are in quotes because technically they are not limits at all).

One-sided limits

Change the definition so that $f(x)$ is only required to approach L when $x \to a$ if x is greater than a. We say x "approaches a from the right," thinking of a number line. If the value of $f(x)$ approaches L when x approaches a from the right, we say that the limit from the right of $f(x)$ at $x = a$ is L, and denote this $\lim_{x \to a^+} f(x) = L$. If we require $f(x)$ to approach L

when x approaches a but only for those x that are less than a, this is called having a limit from the left and is denoted $\lim_{x\to a^-} f(x) = L$.

Remark.
Just like wind directions (North wind, South wind, etc.), one-sided limits are named for the direction they come from, not the direction x is moving. Thus, $\lim_{x\to 0^+}$ is evaluated by letting x approach zero from the positive direction, as shown to the right.



Exercise 3.2. *The lifetime of a light bulb is often modeled as a random variable*[4] *with density* $f(x) = ce^{-cx}$ *when* $x \geq 0$ *and* $f(x) = 0$ *when* $x < 0$ *(light bulbs cannot have negative lifetimes). Here* c *is some positive constant. What are* $\lim_{x\to 0^+} f(x)$ *and* $\lim_{x\to 0^-} f(x)$*?*

Both kinds of one-sided limits require something less stringent, so the statement $\lim_{x\to a} f(x) = L$ automatically implies both $\lim_{x\to a^+} f(x) = L$ and $\lim_{x\to a^-} f(x) = L$. Likewise, if $f(x)$ is forced to approach L when x approaches a from the right, but also when x approaches a from the left, then this covers all x, and the (unrestricted) limit will be L. If you want, you can summarize this as a theorem – wait, no it's too puny, let's make it a proposition. We won't be referring to this too often, but here it is.

Proposition 3.2. *For every function* f *and real numbers* a *and* L*,*

$$\lim_{x\to a} f(x) = L \text{ if and only if } \lim_{x\to a^+} f(x) = L \text{ and } \lim_{x\to a^-} f(x) = L\,.$$

In words, a limiting value for a function exists at a point if and only if the two one-sided limits exist are equal.

Exercise 3.3. *Suppose* f *is a function satisfying* $\lim_{x\to 4^-} f(x) = 2$ *and* $\lim_{x\to 4^+} f(x) = 1$*.*

(i) Sketch a graph of such a function.

(ii) What is $\lim_{x\to 4} f(x)$*?*

Example 3.3 (one-sided limits)**.** Let $f(x) = \lfloor x \rfloor$, the greatest integer function. Let's evaluate the one-sided limits and two-sided limit at a couple of values. First, take $a = \pi$, you

[4] You haven't studied probability densities yet, but all that matters here is the function f.

know, the irrational number beginning 3.14 . . .. If we just look near this value, say between 3.1 and 3.2, it is completely flat: a constant function, taking the value 3 everywhere. So of course the limit at $x = \pi$ will also be 3. This is the same by words or pictures; see Figure 12. By the formal definition, no matter what ε is chosen, you can take $\delta = 0.1$, say,

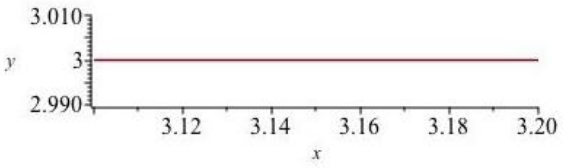


Figure 12: an interval where the greatest integer function is constant

and $f(x)$ will be within ε of 3 because it will be exactly 3. So the limit is 3, hence so are both one-sided limits as in the picture just above.

Now take x to be an integer, say $a = 5$. The limit from the right looks like it did before, with $f(x)$ taking the value 5 for every sufficiently close x (here sufficient means within 1) greater than 5. On the other hand, when x is close to 5 but less than 5, we will have $f(x) = 4$, as in the picture below. Thus,

$$\begin{aligned} \lim_{x \to 5^+} f(x) &= 5 \\ \lim_{x \to 5^-} f(x) &= 4 \\ \lim_{x \to 5} f(x) &= \text{DNE}\,. \end{aligned}$$

The two-sided limit does not exist because the two one-sided limits are unequal; see Figure 13.

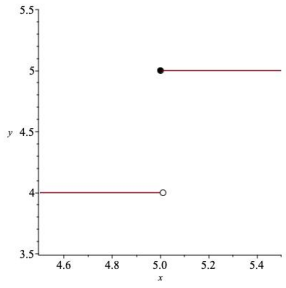


Figure 13: an interval where the greatest integer function is discontinuous

Exercise 3.4. *Let $f(x) = sgn(x)$, the sign function. Use the verbal, pictorial or formal definition, as you please, to give values of these limits.*

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

How about if we take the absolute value: is $\lim_{x\to 0} |sgn(x)|$ any different?

Limits at infinity

You have already seen the pictorial and verbal version of a limit at infinity. Here is the formal definition. It repeats a lot of the definition of a limit at $x = a$. The only difference is that instead of having to come up with an interval $[a-\delta, a+\delta]$ guaranteeing $f(x)$ is within ε of the limit, you have to come up with an "interval near infinity". This turns out to mean an interval $[M, \infty)$. In other words, there must be a real number M guaranteeing $f(x)$ is with ε of L when $x > M$.

Remark. Informally, "close to infinity" turns into "sufficiently large". In the tolerance/accuracy analogy, getting $f(x)$ to be close to L to within the acceptable tolerance will result from guaranteed largeness of the input rather than guaranteed closeness to a.

Definition 3.4. *We say that $\lim_{x\to\infty} f(x) = L$ if and only if L is a real number and:*

> *For any positive real number ε (think of this as acceptable tolerance in the y value) there is a corresponding real M (think of this as guaranteed minimum value for x) such that for any x greater than M, $f(x)$ is guaranteed to be in the interval $[L-\varepsilon, L+\varepsilon]$.*

In symbols, the logical implication that must hold is:

$$x > M \Longrightarrow |f(x) - L| < \varepsilon\,.$$

If a real number L exists satisfying this, we write $\lim_{x\to\infty} f(x) = L$. Sometimes to be completely unambiguous, we put in a plus sign: $\lim_{x\to+\infty} f(x) = L$.

Exercise 3.5. *True or false?*

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

Limits at $-\infty$ are defined exactly the same except for a single inequality that is reversed. Now the implication that must hold is that for some (possibly very negative) M,

$$x < M \Longrightarrow |f(x) - L| < \varepsilon\,.$$

When this holds, we write $\lim_{x\to-\infty} f(x) = L$. When no such L exists, we write $\lim_{x\to-\infty} f(x) = DNE$ or just $\lim_{x\to-\infty} f(x)$ DNE.

Example 3.5. Let $f(x) := \frac{x}{\sqrt{1+x^2}}$. Because $\sqrt{1+x^2}$ is a little bigger than $|x|$ but almost the same when x or $-x$ is large, this function satisfies

$$\begin{aligned} \lim_{x\to\infty} f(x) &= 1 \\ \lim_{x\to-\infty} f(x) &= -1 \end{aligned}$$

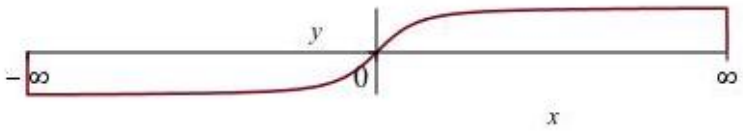


Figure 14: graph of $x/\sqrt{1+x^2}$

The graph of this function is shown in Figure 14. It has horizontal asymptotes at 1 and -1. This suggests how to define a horizontal asymptote.

Definition 3.6. *A function f or its graph is said to have a* **horizontal asymptote** *at height b if $\lim_{x\to\infty} f(x) = b$ or $\lim_{x\to-\infty} f(x) = b$.*

Exercise 3.6.

(i) Sketch a graph of a function f for which $\lim_{x\to-\infty} f(x)$ exists but $\lim_{x\to+\infty} f(x)$ does not.

(ii) Give a formula defining a function $g(x) := \cdots$ such that $\lim_{x\to-\infty} g(x)$ exists but $\lim_{x\to+\infty} g(x)$ does not.

(iii) Which of these two things was easier to do?

"Limits" of infinity

Consider the function $f(x) = 1/x^2$, defined for all real numbers except zero. What happens to $f(x)$ as $x \to 0$? By our definitions, $\lim_{x\to 0} 1/x^2$ DNE. But we can see that $f(x)$ "goes to infinity". Because infinity is not a number, the limit technically does not exist. However, it is useful to classify DNE limits as ones where the function approaches ∞ (or $-\infty$) versus ones where there is no consistent behavior.

Remark. This time, instead of staying within a tolerance of ε in the output, we make the output sufficiently large (greater than any given N) or small. We do this by guaranteeing δ accuracy in the input (for limits as $x \to a$) or by making the input sufficiently large or small (limits as $x \to \pm\infty$).

Formally, this turns into the following definition.

Definition 3.7. *If f is a function and a is a real number, we say that $\lim_{x\to a} f(x) = +\infty$ if for every real N there is a $\delta > 0$ such that $0 < |x - a| < \delta$ implies $f(x) > N$.*

Again, if we reverse the last inequality to require that $f(x) < N$ (and N can be a very negative number) we get the definition for a limit of negative infinity. Please remember these are all subcases of limits that don't exist! If you show that a limit is infinity, you have shown that the limit does not exist (and you have specified a particular reason it doesn't exist).

Example 3.8. Let's check that $\lim_{x\to 0} 1/x^2 = +\infty$. Given a positive real number N, how can we ensure $f(x) > N$? Answer: for positive numbers, f is decreasing and $f(x) = N$ precisely when $x = 1/\sqrt{N}$. Therefore, if we keep x positive but less than $1/\sqrt{N}$ then $f(x)$ will be greater than N. We have just shown that $\lim_{x\to 0^+} 1/x^2 = +\infty$. Similarly, when x is negative, if we keep x in the interval $(-1/\sqrt{N}, 0)$ we ensure $1/x^2 > N$. So $\lim_{x\to 0^-}$ is also $+\infty$. Both one-sided limits are $+\infty$, therefore

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

Don't forget, it follows from the limit being $+\infty$ that

$$\lim_{x\to 0} \frac{1}{x^2} \text{does not exist.}$$

For one-sided limits and limits at infinity, the DNE case also includes a case where the limit would be said to be infinity. Stating all these would be repetitive. Try one, to make sure you agree it's straightforward.

Exercise 3.7. *Write a formal definition for the statement* $\lim_{x\to a^+} f(x) = -\infty$.

Exercise 3.8. *Consider the function* $1/x$. *Write one-sided infinite limit statements for* $\lim_{x\to 0^+} 1/x$ *and* $\lim_{x\to 0^-} 1/x$.

Limit of a sequence

A special case of limits at infinity is when the domain of f is the natural numbers. When f is only defined at the arguments $1, 2, 3, \ldots$, it is more usual to think of it as a sequence $b_1, b_2, b_3, \ldots$, where $b_k := f(k)$. The definition of a limit at infinity can be applied directly, resulting in the definition of the limit of a sequence.

Definition 3.9 (limit of a sequence)**.** *Given a sequence* $\{b_n\}$ *and a real number* L *we say* $\lim_{n\to\infty} b_n = L$ *if and only if for all* $\varepsilon > 0$ *there is an* M *such that* $|b_n - L| < \varepsilon$ *for every* $n > M$.

Remark. Often we use letters such as n or k to denote integers and x or t to denote real numbers. Therefore, by context, $\lim_{n\to\infty} 1/n$ denotes the limit of a sequence while $\lim_{t\to\infty} 1/t$ denotes the limit at infinity of a function. Formally we should clarify and not count on the name of a variable to signify anything! But because the two definitions agree, often we don't bother.

Exercise 3.9. *Evaluate these three limits of sequences.*

(i) $\lim_{n\to\infty}(-1)^n$

(ii) $\lim_{n\to\infty}(1/2)^n$

(iii) $\lim_{n\to\infty} 2^n$

Pictorially, if a sequence has a limit L, then for every pair of parallel horizontal lines, however narrow, enclosing the height L, the sequence must eventually stay between them. This is shown in Figure 15.

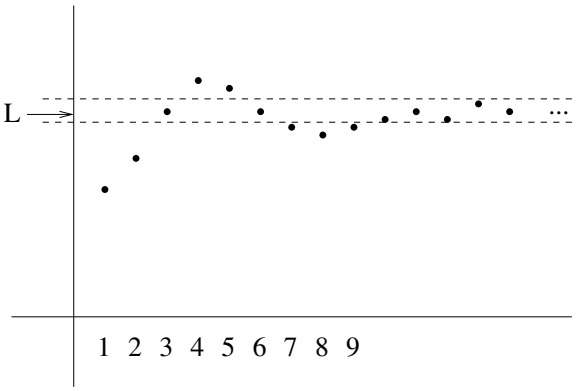


Figure 15: For these two parallel lines, once $k > 9$, the height b_k is between the lines

As you will see, Propositions 3.12 and 3.15 give ways to determine limits of more complicated functions once you understand limits of some basic functions. Here is another piece of logic that can help do the same thing. You will prove it in your homework.

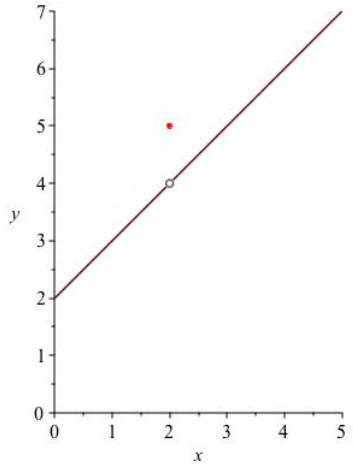
Theorem 3.10 (sandwiching)**.** *Let a be a real number or $\pm\infty$ and let f, g and h be functions satisfying $f(x) \leq g(x) \leq h(x)$ for every x. If $\lim_{x\to a} f(x) = L$ and $\lim_{x\to a} h(x) = L$ then also $\lim_{x\to a} g(x) = L$. If we know only that $\lim_{x\to a^+} h(x) = \lim_{x\to a_-} h(x) = L$ then we can conclude $\lim_{x\to a^+} g(X) = L$, and same for limits from the left.*

The same fact is true of sequences: if $a_n \leq b_n \leq c_n$ for these three sequences and the first and last sequence converge to the same limit L, then so does the middle one. We will not do anything with this now, but will get back to this fact in a week or two. The next exercise brushes up on the logical syntax of limits.

Exercise 3.10. *Evaluate $\lim_{x\to 4} cx$. Evaluate $\lim_{t\to a} bt$. In each of these two cases, say which variables (any letter appearing in the expression other than letters spelling "lim") are free and which are bound. Did your answers involve only the free variables?*

3.3 Continuity

A function f is said to be continuous at the value a if the limit exists and is equal to the function value, in other words, if $\lim_{x\to a} f(x) = f(a)$. Intuitively, this means the limit at a exists and there is no hole: the function is actually defined at a and wasn't given some weird other value. To illustrate what we mean, to the right is a picture of a function that is discontinuous at $x = 2$ even though $\lim_{x\to 2} f(x)$ exists and so does $f(2)$, because the values don't agree.



Exercise 3.11. *Is* $\lim_{x\to a} f(x) - f(a) = 0$ *the same as* f *being continuous at* a*? Explain why or why not.*

Continuity on regions

A function is said to be continuous on an open interval (a, b) if it is defined and continuous at every point of (a, b). A function is said to be continuous on an closed interval (a, b) if it is defined and continuous at every point of (a, b), with only one-sided contintuity required at a^+ and b_-. A function f is said to be just plain continuous if it is continuous on the whole real line. Note: these definitions can have unintended consequences if the domain is strange; mostly our domains will be intervals or all real numbers.

Exercise 3.12. *Which of the basic trig functions* sin, cos *and* tan *are continuous on* $(0, 2\pi)$*? You don't need to prove your answer, just to have an intuitive justification in mind.*

Before going on to use the notion of continuity to help us compute limits, we will state one famous result which will seem either stupid and obvious or deep and tricky.

Theorem 3.11 (Intermediate value theorem)**.** *Let* f *be a continuous function defined on the closed interval* $[a, b]$ *and suppose that* y *is any value between the values* $f(a)$ *and* $f(b)$*. Then there is some number* c *in the interval* $[a, b]$ *satisfying* $f(c) = y$*.*

This says, basically, a continuous function can't get from one value to another without hitting everything in between. The theorem is most often used when there is a number we can only define this way. For example, let $f(x) := e^x/x$, which is an increasing function on the half-line $[1, \infty)$. We want to say "let c be the value for which $f(c) = 3$." How do we know there is one? Well, $f(1) = e$, which is less than 3, and $f(3) \approx 6.695$ which is greater than 3. So there must be an argument between 1 and 3 where f takes value 3. There can be only one because f is strictly increasing (you can prove after another two sections).

3.4 Computing limits

Computing a limit by verifying the formal definition is a real pain. There is computational apparatus that allows us to compute limits of many functions once we know limits of a few simple ones. One approach we have seen in textbooks is to give a list of rules that work. It looks something like this.

Proposition 3.12. *If* $\lim_{x\to a} f(x) = L$ *and* c *is a real number then*

$$\begin{aligned} \lim_{x\to a} cf(x) &= cL \\ \lim_{x\to a} f(x)^c &= L^c \text{ provided } L > 0 \end{aligned}$$

Example 3.13. Suppose f is a polynomial: $f(x) = b_n x^n + \cdots + b_1 x + b_0$. What is $\lim_{x\to a} f(x)$? We hope you think this is a really boring example. Of course, the polynomial is continuous (picture in your mind the graph of a polynomial) and $\lim_{x\to a} f(x) = f(a)$. It is an example of the Proposition for these reasons: (1) we can evaluate the limit at a of the monomial x^k as a^k (the second conclusion); (2) we can evaluate the limit at a of each monomial $b_k x^k$ as $b_k a^k$ by applying the first conclusion with $c = b_k$ and $f(x) = x^k$; (3) we can sum the whole thing to evaluate the limit at a of f by the third conclusion - wait, there is no third conclusion, it's the first conclusion of Proposition 3.15. This fact comes up a lot, so we record it as a proposition.

Proposition 3.14. *Polynomials are continuous. The limit of a polynomial* f *at* a *is always given by* $f(a)$.

Exercise 3.13. *Evaluate* $\lim_{x\to -1} 3x^2 + 2x + 1$.

Proposition 3.15. *If f and g are functions and a, K, L are real numbers with $\lim_{x\to a} f(x) = K$ and $\lim_{x\to a} g(x) = L$, then*

$$\begin{aligned}
\lim_{x\to a} f(x) + g(x) &= K + L \\
\lim_{x\to a} f(x) - g(x) &= K - L \\
\lim_{x\to a} f(x) \cdot g(x) &= K \cdot L \\
\lim_{x\to a} \frac{f(x)}{g(x)} &= \frac{K}{L} \text{ provided } L \neq 0
\end{aligned}$$

Exercise 3.14. *Use Proposition 3.15 to evaluate two of these three limits. For the third, can you find a way to evaluate it?*

(a) $\lim_{x\to 1} \ln x - \sqrt{x}$

(b) $\lim_{x\to 0} x \sin x$ *(see Exercise 3.12)*

(c) $\lim_{x\to 3}(x^2 - 9)/(x - 3)$

So that Propositions 3.12 and 3.15 don't look like arbitrary rules from out of nowhere, you should realize they can be proved, and in fact follow from one basic theorem.

Theorem 3.16 (composition with a continuous function)**.** *If the function f has a limit L at $x = a$ and the function H is continuous at L then $H \circ f$ will have the limit $H(L)$ at $x = a$. Formally,*

$$\lim_{x\to a} f(x) = L \text{ implies } \lim_{x\to a} H(f(x)) = H(L) \text{ provided } H \text{ is continuous at } L\,.$$

Why do the two propositions 3.12 and 3.15 follow from this principle? Let $H(x)$ be the continuous function cx. Then $H \circ f$ is $cf(x)$ and we recover the first conclusion of Proposition 3.12. Setting $H(x) := x^c$ recovers the second conclusion.

Exercise 3.15. *A related fact about limits is computation by change of variables. Suppose g is a function such that $\lim_{x\to 0} g(x) = 3$. What is $\lim_{x\to 0} g(2x)$? This question will be discussed further. For now, give a short answer and try to explain in words.*

Some more techniques and tricks

This course is more about using limits than it is about computational technique, but you should at least see some of the standard techniques for cases that go beyond what's in Propositions 3.12 and 3.15.

Suppose you need to evaluate $\lim_{x\to a} f(x)/g(x)$. If both f and g have nonzero limits at a, say L and M, then Proposition 3.15 tells you $\lim_{x\to a} f(x)/g(x) = L/M$. In fact if $L = 0$ but $M \neq 0$, this still works. If $M = 0$ but $L \neq 0$, then the question of evaluating $\lim_{x\to a} f(x)/g(x)$ also has an easy answer.

Exercise 3.16. *What is the easy answer?*

The remaining case, when $L = M = 0$, can be enigmatic. Calculus provides one solution you will see in a few weeks (L'Hôpital's rule), but you can often solve this with algebra. If you can factor out $(x - a)$ from both f and g, you may get a simpler expression for which at least one of the functions has a nonzero limit.

Example 3.17. What is $\displaystyle\lim_{x\to 5} \frac{x^2 - 25}{x^2 - 5x}$?

Both numerator and denominator are continuous functions with values of zero (hence limits of zero) at 5. That suggests dividing top and bottom by $x-5$, resulting in $\displaystyle\lim_{x\to 5} \frac{x+5}{x}$. Both numerator and denominator are continuous functions so we can just evaluate and get $10/5$ so the answer is 2.

Sometimes you have to do a little algebra to simplify. Here's an example of one of the most common simplification tricks.

Example 3.18. What is $\displaystyle\lim_{x\to 0} \frac{\sqrt{x+1} - 1}{x}$?

Multiplying and dividing by the so-called conjugate expression, where a sum is turned into a difference or vice versa, gives

$$\begin{aligned}
\lim_{x\to 0} \frac{\sqrt{x+1} - 1}{x} &= \lim_{x\to 0} \frac{\sqrt{x+1} - 1}{x} \frac{\sqrt{x+1} + 1}{\sqrt{x+1} + 1} \\
&= \lim_{x\to 0} \frac{x}{x(\sqrt{x+1} + 1)} \\
&= \lim_{x\to 0} \frac{1}{\sqrt{x+1} + 1}\,.
\end{aligned}$$

The numerator and denominator are continuous at $x = 0$ with nonzero limits of 1 and 2 respectively, so the limit is equal to $1/2$.

This algebra trick occurs so commonly throughout mathematics that you should always think about conjugate radicals every time you see an expression with a square root added to or subtracted from something!

Further tricks can wait until you've learned some more background. Although limits are needed to define derivatives, you can then use derivatives to evaluate more limits (L'Hôpital's rule). Similarly, limits are used to define orders of growth, which can then be used to evaluate more limits.

4 Derivatives

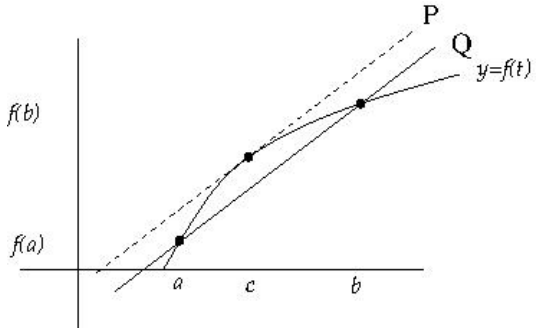
4.1 Concept of the derivative

It's easy to define your average speed for a trip: take the number of miles, divide by the number of hours, and there's your average speed in miles per hour. If you journey at constant speed, then that's also your speed at every moment of the trip. Most of us do not travel at constant speed. What is your speed then? How do you define it? How do you measure it? How do you compute it if you know some equation for your position at time t?

The concept of instantaneous speed is subtle. It is what spurred the invention of calculus over a few decades near the year 1700. It is a very general notion. Average speed is distance traveled per total time. Instantaneous speed is some instantaneous version.

Exercise 4.1.

The figure at the right shows distance traveled ($f(t)$ against time (t). The slope of line P can be interpreted as what in terms of speed? What about the slope of line Q?



If you replace "distance traveled" by "production price" and "time elapsed" by "units produced" you get the notions of average production cost per unit; marginal cost per unit is the instantaneous version. The list of applications is endless. Mathematically, they are all the same: if f is a function and x_0 and x_1 are starting and ending arguments for f, then the average change in f over the interval is $(f(x_1) - f(x_0))/(x_1 - x_0)$; the instantaneous rate of change of f with respect to x is called the **derivative** of f with respect to x and denoted $f'(x)$.

Exercise 4.2. *Suppose $f(x) = mx + b$. What is $f'(x)$?*

In this section we will see how to understand f' both physically and mathematically. We will continue to use instantaneous speed as a running example of the physical concept, and instantaneous rate of change of $f(x)$ with respect to x as the corresponding mathematical concept.

Important remark: we can take the slope of the function f at any point. Taking it at x gives a value we call $f'(x)$. That means that f' is a function: give it an argument x and it will produce the slope of f at that point. It will be helpful to keep in mind that the derivative operator takes as input functions f and produces as output their derivatives f'. Operator is a fancy word for a function whose input and output are functions rather than numbers. Taking derivatives is a **linear** operator. This is captured in Propositions 5.1 - 5.3 below.

Exercise 4.3. *Suppose you replace "distance traveled" by "elevation of trail" and "time elapsed" by "distance hiked". What would be the physical interpretation of the instantaneous rate?*

4.2 Definitions

Most functions we use in mathematical modeling have unique tangent lines at most points. The slope of the tangent line to the graph of f at the point $(x, f(x))$ seems like one reasonable definition of $f'(x)$. In rare cases, such as you have already seen, we can use geometry to prove there is exactly one line tangent to the graph of f at a point and compute the slope.

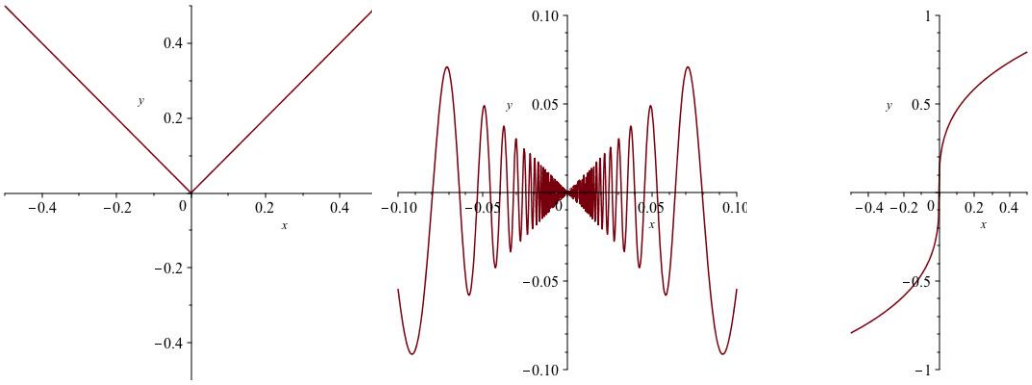


Figure 16: graphs of $|x|$, $x\sin(1/x)$ and $\sqrt[3]{x}$

Unfortunately, there are not many functions for which the graph is a well known geometric object. In most cases we can't use geometry to conclude that there is a tangent line, that there is only one tangent line, or what the slope of this line is, if indeed there is exactly one. Keeping this in mind, we will use limits to come up with a definition that works for most functions and, when it does not work, as in the examples in Figure 16, gives an indication

of why. In cases when it does not work, in fact we would probably agree that there is no good way to make sense of the instantaneous slope.

Exercise 4.4. *The graphs of $|x|$, $x\sin(1/x)$ and $\sqrt[3]{x}$ are shown in Figure 16. All contain the point $(0,0)$ provided we add zero to the domain of the second function and define the function to be zero there. In each case, say whether there is one, none, or more than one tangent line to the graph at $(0,0)$. In which of these cases do you think there is a well defined slope of the tangent $(0,0)$?*

We can take average slopes over any interval we want. The slope over the interval $[a,b]$ is the slope of the **secant line** passing through $(a,f(a))$ and $(b,f(b))$. This is also called the **difference quotient** of f at the arguments a and b. What happens when one endpoint of the interval is x and the other is very close to x? Pictorially, it looks the slope get very close to the slope of the tangent line at $(x,f(x))$. Figure 17 shows an example where $a = 1/2$ and secant lines (blue) are drawn through various values of b. These appear to converge to the tangent line at $(1/2, f(1/2))$ which is black and dashed.

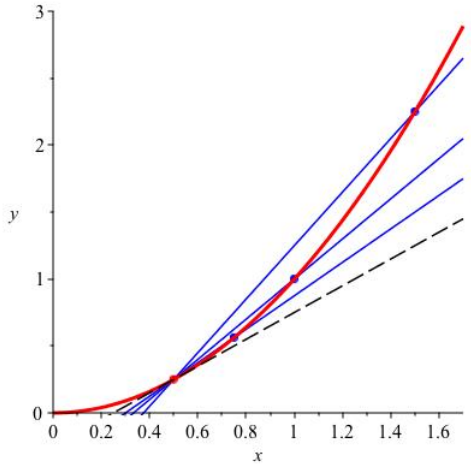


Figure 17: three secant lines approaching a tangent line

Definition using limits

The derivative is a mathematical definition meant to compute the slope of the tangent line at a. Definition 4.1, however, only talks about limits of slopes of secants, not of tangents. Do you think these two notions will always coincide? There isn't a right answer to this.

Definition 4.1. *Let f be a function whose domain contains an interval around the point a. Define*

$$f'(a) := \lim_{b \to a} \frac{f(b) - f(a)}{b - a} \tag{4.1}$$

if the limit exists, and say that $f'(a)$ is undefined if the limit does not exist. Because we want to emphasize that $b - a$ is going to zero, we often define $h := b - a$ and rewrite the definition as

$$f'(a) := \lim_{h \to 0} \frac{f(a + h) - f(a)}{h} \,. \tag{4.2}$$

The two definitions (4.1) *and* (4.2) *are algebraically equivalent.*

Exercise 4.5.

(i) In (4.1)*, which variables are free and which are bound?*

(ii) In Figure 17 What values of a and b are being illustrated?

(iii) Suppose a student complains that Figure 17 illustrates a limit of the form $\lim_{b \to a^+}$, not $\lim_{b \to a}$. What could you add to the picture to address her concerns?

Example 4.2. Let $f(x) = x^2$. Let's see the definition to try to compute $f'(1)$. By definition, this is

$$\lim_{b \to 1} \frac{f(b) - f(1)}{b - 1} \,.$$

Evaluating the numerator, gives

$$\lim_{b \to 1} \frac{b^2 - 1}{b - 1} = \lim_{b \to 1} b + 1 = 2 \,.$$

The first equality is true because we can cancel the factors of $b - 1$ (remember, the limit looks at values of b near 1 but not equal to 1). The second equality is true because we can evaluate the limit of the polynomial $b + 1$ at $a = 1$ by plugging in 1 for b (Proposition 3.14).

Exercise 4.6. *Let $f(x) = x^2 + 5$. Compute $f'(3)$ directly from the definition, as we did in the previous example (show your work: you can upload a pdf, write in text in using a lot of parentheses, or use the Canvas equation editor).*

Notation

We already agreed to use a prime after the function name as one way to denote a derivative. Thus the derivative of f is f', the derivative of g is g', the derivative of Γ is Γ', etc. We may need to refer to the derivative of a function when it has not been given a name. One could imagine something like the notation $(cx)'$ for the derivative of the function "multiply by c", or perhaps the more precise[5] $(x \mapsto cx)'$

To avoid ambiguity, we use the notation $\frac{df}{dx}$ for the derivative of f with respect to x. This is better than f' when there is more that one variable that could be differentiated. You can also write this as $\frac{d}{dx}f$ when f is a big long cumbersome expression, for example,

$$\frac{d\left(\frac{e^{x^2-1}\sin x}{1+x}\right)}{dx} \qquad \text{is the same as} \qquad \frac{d}{dx}\left(\frac{e^{x^2-1}\sin x}{1+x}\right) \quad .$$

Then there is the question of how to write $f'(a)$, the value of the function f' at argument a, in this notation. Should we write $\frac{d\,f(a)}{dx}$ or $\frac{df}{dx}(a)$? The second is better, for example, $\frac{d(x^3-3x+1)}{dx}(a)$, because the first looks like you are differentiating a constant. Another common way of writing this is $\left.\frac{d\,(x^3-3x+1)}{dx}\right|_{x=a}$.

Exercise 4.7. *Suppose the number of feet an object has fallen after t seconds is given by $16t^2 + ct$ where c is its initial downward velocity[6]. Write an expression for the downward instantaneous speed of the object after s seconds. Please don't compute any derivatives, just write an expression in some notation involving a derivative.*

Further interpretations: error propagation and marginal effect

You have seen examples in which derivatives represent speed. More generally, the derivative of a function of time represents the rate of change of the quantity per time. Here are some other things derivatives commonly represent.

Suppose you have a formula $f(x)$ involving a quantity x that is measured, but with measurement error. Then $f'(x)$ tells you how much error you get in f per amount of error in measuring x.

[5] More precise because it is distinguished $(c \mapsto cx)'$ in which x is the free variable, c is the (bound) variable, and the function is "multiply by x".

[6] This is in fact true when air resitance is ignored and the earth's gravitational constant is approximated.

Example 4.3. A 4×8 foot board is cut parallel to the long side to obtain a 3×8 board. The accuracy of the cut is $1/4$ inch. What is the accuracy of the area, in square feet? Writing $A = \ell \times w$ and differentiating gives $dA/dw = \ell = 8$ feet in our case. Therefore, the error in area (in square feet) is 8 feet times the measurement error in the width (in linear feet). Plugging in a measurement error of $1/4$ inch, which equals $1/48$ feet, we see the area is accurate to within $8 \text{ ft} \times \frac{1}{48} \text{ ft} = \frac{1}{6} \text{ ft}^2$.

The symbol Δ is the upper case Greek letter Delta an often used to denote change in a quantity or error in a measurement.

Exercise 4.8. *Let Δx denote the possible error in x, and Δf denote the possible resulting in $f(x)$. Write a formula for these quantities in terms of the derivative of f.*

Another interpretation is the marginal effect of the variable on the function. For example, if $f(x)$ represents the cost of producing x barrels of refined oil, then $f'(x)$ is the marginal cost of production of more oil. Unless f is linear, this will depend on x. The marginal cost of further production usually depends on the present level of production.

4.3 First and second derivatives, and sketching

Knowledge of the deriviatve can help you sketch a function more accurately. The very first practice problem asked you to incorporate slope information into a sketch. Sketching is as much an art as a science, but there are methodical ways to use information about the function and its derivatives.

To begin with, knowing where the derivative is positive and negative determines whether it is sloping up or down as you move right. In other words, the sign of the derivative indicates whether the function is increasing or decreasing. Where the sign of the derivative changes from positive to negative as you move right, the function changes from increasing to decreasing. That means someone hiking on the graph of the function from left to right has been walking upwards and now begins to walk downward; see Figure 18.

Exercise 4.9. *What does the hiker's landscape look like if f' is negative to the left of the value $x = a$ and positive to the right?*

Because transitions in the sign of f' correspond to hilltops and valley floors, finding values of x that are maxima and minima for $f(x)$ involves finding values of x for which $f'(x) = 0$.

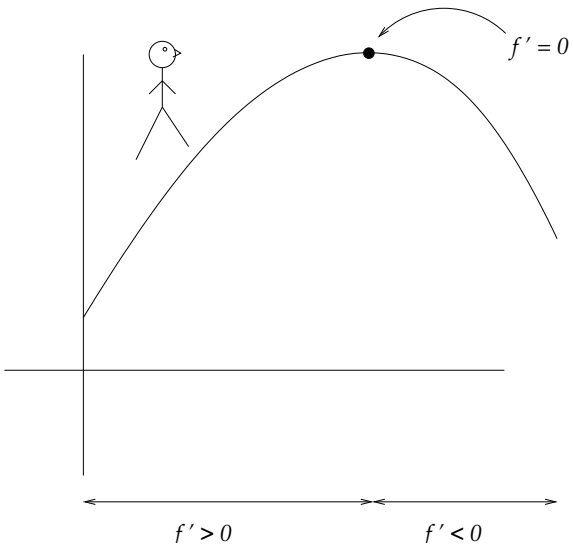


Figure 18: The hilltop, where the function changes from increasing to decreasing, occurs exactly where $f' = 0$.

We discuss this at greater length in Chapter 7. For the purposes of sketching, the moral of the story is: know where f' is positive and where it is negative, and use this to depict a function that is increasing and decreasing in the right places.

Exercise 4.10. *Sketch a function f such that f' is positive when $x < 1$, dips to zero at $x = 1$, is positive again until $x = 3$, is zero at $x = 3$ and is negative to the right of that. See if you can also make f have a unique zero at $x = -2$.*

The second derivative

A Japanese proverb says, "The other side also has another side." The function f has a derivative. This is also a function. Therefore, *The derivative also has a derivative.* Not quite so poetic, but very useful for sketching functions. It is called the second derivative, denoted f'', or $\dfrac{d^2 f}{dx^2}$. The sign of the derivative says where a function is increasing or decreasing, therefore the *sign* of f'' indicates where the *slope* f' is *increasing* or *decreasing*. We use italics here as a visual reminder that there are a number of levels (original function, first derivative, second derivative) and attributes (positive/negative, increasing/decreasing) and it's easy to get mixed up what corresponds to what.

Remark. The placement of the 2 in the numerator of $\frac{d^2 f}{dx^2}$ may seem strange, but it reflects something important: (d/dx) is a differential operator, and (d^2/dx^2) is the result of applying this operator twice. This becomes important in later courses such as Math 114.

Exercise 4.11. *Let $g(x) := x^2$. Compute $g''(x)$.*

Of course, not every function is differentiable, and not every derivative is itself differentiable, so f'' may not exist even if f' exists.

We have talked informally about functions that are concave up or down. It is time to give a definition. In fact we give two definitions, one algebraic and one pictorial. The pictorial one is in fact more general because it works when f' does not exist. When f' exists on (a, b), then the two definitions agree.

Definition 4.4 (concavity)**.**
When the function f' exists and is increasing, we say that f is concave upward.
When the function f' exists and is decreasing, we say that f is concave downward.

Definition 4.5 (concavity: pictorial definition)**.** *If (a, b) is an open interval in the domain of f and if for every pair of numbers $x, y \in (a, b)$ the graph of f on (a, b) lies below the line segment connecting $(x, f(x))$ to $(y, f(y))$, we say that f is concave upward on (a, b).*

Exercise 4.12. *If f'' exists and is positive, can you conclude anything about concavity of f? How about if f'' exists and is negative?*

To summarize, if f'' exists on (a, b) then the sign of f'' determines the concavity of f. If f'' doesn't exist or you can't compute it, use Definition 4.4 or 4.5.

Points of inflection

We never formally defined a tangent line. One definition would be "A line that touches a graph of a function at precisely one point and stays on one side of the graph other than this." Here are four ways this definition may fail to capture what some people think a tangent line should be. For each example, please say whether you think the given line ought to count as a tangent line.

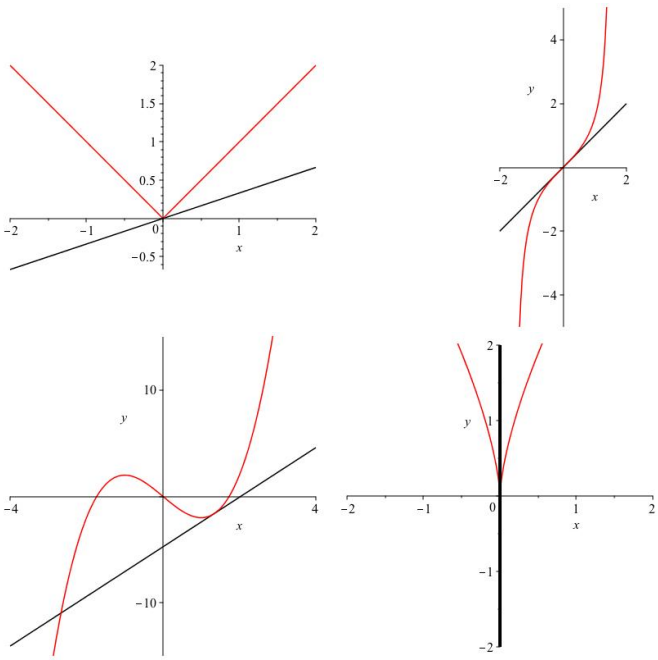
Exercise 4.13.

(a) Graph of $y = |x|$; any line with absolute slope less than 1

(b) Graph of $y = \tan x$, line of slope 1 through the origin

(c) Graph of a cubic with a tangent line that interesects the cubic elsewhere

(d) Graph with a vertical cusp



As you can see, the intuitive definition of tangent line is subject to unanticipated judgment calls. This motivates a more formal definition.

Definition 4.6 (tangent line)**.** *If f is differentiable at a, the tangent line to f at a is defined to be the line $(y - f(a)) = m(x - a)$ where $m = f'(a)$.*

Exercise 4.14. *Is the point $(a, f(a))$ always on this line? Explain why or why not.*

One confusing case is when the second derivative is zero. What happens to the concavity at such a point? Often it switches from up to down or *vice versa*. Wherever concavity switches is called a **point of inflection**. The geometric concept of an inflection point does not require calculus, though the notion seems not to have been discussed much before the advent of calculus.

Exercise 4.15. *Which of the figures in Exercise 4.13 shows a point of inflection?*

Exercise 4.16.

(i) Sketch a graph of the sine function.

(ii) Mark the intervals where sine increases and those where it decreases.

(iii) On the same graph sketch the cosine function.

(iv) *The derivative of* sin *is* cos*; what does this imply about the values of cosine on the marked intervals?*

(v) *Where are the points of inflection for sine and what happens to the cosine at those arguments?*

5 Computing derivatives

There are a lot of rules for computing derivatives that are relatively easy to remember and use. These rules are theorems – they can all be derived from the definition via limits and some computation. You will get familiar enough with these rules that you will happily use them without thinking. The structure of this chapter is backwards: we give you nearly all the rules right away, then give arguments for some of them, postponing some of the arguments until we have developed a few more tools. We do this because calculus is so much more fun when you know enough do a few computations!

5.1 Rules for computing derivatives

The rules have two forms. Some just tell you the derivative of a particular function like $\sin x$ or a class of functions like b^x. Others are rules for combining and transforming. They tell you, if you know f' and g', what the derivatives are of $f+g$, fg, $f \circ g$, and so forth.

The combining rules

Proposition 5.1 (sum rule)**.** *Let f and g be differentiable functions. Then $(f+g)' = f'+g'$.*

Proposition 5.2 (difference rule)**.** *Let f and g be differentiable functions. Then $(f-g)' = f'-g'$.*

Proposition 5.3 (multiplication by a constant)**.** *Let f be a differentiable function and c be a constant. Then $(cf)' = cf'$.*

Exercise 5.1. *Using the three propositions above, as well as examples you've worked out earlier, compute the derivative of $x - 3\sqrt{x}$.*

Proposition 5.4 (product rule)**.** *Let f and g be differentiable functions. Then $(fg)' = f'g + g'f$.*

Proposition 5.5 (quotient rule)**.** *Let f and g be differentiable functions. Then for any x such that $g(x) \neq 0$,*

$$\frac{d}{dx}\frac{f(x)}{g(x)} = \frac{gf' - fg'}{g^2},$$

all functions on the right-hand side evaluated at x.

Proposition 5.6 (chain rule)**.** *Let f and g be differentiable functions. Let a be a real number inside an open interval in the domain of g such that $g(a)$ is inside an open interval in the domain of f. Then*

$$\frac{d}{dx} f(g(x))\Big|_{x=a} = \left(\frac{df}{dx}\Big|_{x=g(a)} \right) \left(\frac{dg}{dx}\Big|_{x=a} \right) .$$

We can write this more compactly as

$$(f \circ g)'(x) = f'(g(x))g'(x)\,;$$

the longer version can help unravel any confusion.

A collection of rules for particular functions

We list a few that are either obvious from the definition or are ones you've worked out already.

Proposition 5.7 (easy cases)**.** *Let c be any real constant. Then,*

$$\begin{aligned} \frac{d}{dx}\, c &= 0 \\ \frac{d}{dx}\, cx &= c \\ \frac{d}{dx}\, x^2 &= 2x \\ \frac{d}{dx}\, \sqrt{x} &= \frac{-1}{2\sqrt{x}} \qquad \text{for } x > 0\,. \end{aligned}$$

the same for all x.

Exercise 5.2. *Which functions f have the property that f' is a constant function? Sketch the graph of f in the case that f' is the constant function $1/2$.*

Proposition 5.8 (powers and transcendental functions)**.** *In the following list, if no restric-*

tions are given on x, then the statement holds for all real x.

1. $\frac{d}{dx}\, x^n = nx^{n-1}$ *when n is a positive integer*
2. $\frac{d}{dx}\, x^r = rx^{r-1}$ *when $x \neq 0$ and r is any nonzero real number*
3. $\frac{d}{dx}\, e^x = e^x$
4. $\frac{d}{dx}\, a^x = a^x \cdot \ln a$ *for $a > 0$ and all real x*
5. $\frac{d}{dx}\, \ln x = \frac{1}{x}$ *for $x > 0$*
6. $\frac{d}{dx}\, \sin x = \cos x$
7. $\frac{d}{dx}\, \cos x = -\sin x$
8. $\frac{d}{dx}\, \tan x = \sec^2 x$ *when this is finite*
9. $\frac{d}{dx}\, \arcsin x = \frac{1}{\sqrt{1-x^2}}$
10. $\frac{d}{dx}\, \arccos x = \frac{-1}{\sqrt{1-x^2}}$
11. $\frac{d}{dx}\, \arctan x = \frac{1}{1+x^2}$

Exercise 5.3. *Use rule # 4 to compute the slope of the function $f(x) := a^x$ at $x = 0$. For which a is this clope equal to 1? Is this consistent with Proposition 0.10?*

Exercise 5.4. *Let $f(x) := x^{-1}$ and $g(x) := x^3$. This exercise takes you step by step through a test of the product rule.*

(i) What is f'?

(ii) What is g'?

(iii) what is $(f')(g')$?

(iv) What does the product rule give you for $(fg)'$?

(v) What do you get for $(fg)'$ by first multiplying, then using rule #1 from Proposition 5.8 (the power rule)?

You are probably pretty experienced at taking apart algebraic expressions into sums and differences of products and quotients of simpler expressions. Here are some more exercises to check that you can do this and then apply the differentiation rules above.

Exercise 5.5. *Use the sum, difference, product and quotient rules, along with derivatives given in Proposition 5.8 to evaluate $f'(x)$ in each of these cases.*

(i) $f(x) := x^3 e^x$

(ii) $f(x) := \frac{1}{x^{2.5}}$

(iii) $f(x) := x \ln x - x$

(iv) $f(x) := x \arcsin x$

Taking apart algebraic expressions into compositions of functions, as is needed for the chain rule, can be a little trickier.

Example 5.9. In order to differentiate $(1+x^2)^{1/3}$ you need to recognize this as a composition $f(g(x))$ with $f(x) = x^{1/3}$ and $g(x) = 1+x^2$. The chain rule tells us that the derivative of $(1+x^2)^{1/3}$ at $x = a$ will be given by

$$\left(\frac{d}{dx} x^{1/3} \Big|_{x=1+a^2} \right) \left(\frac{d}{dx} (1+x^2) \Big|_{x=a} \right) . \tag{5.1}$$

The derivative of $x^{1/3}$ is $(1/3)x^{-2/3}$ be the power rule (the second identity in Proposition 5.8); the derivative of $1+x^2$ is $0+2x = 2x$ by the sum rule and the power rule. This shows (5.1) to equal

$$\left(\frac{1}{3} x^{-2/3} \Big|_{x=1+a^2} \right) (2x|_{x=a}) = \frac{1}{3}(1+a^2)^{-2/3}(2a) .$$

The next few exercises check on your understanding of the chain rule. The first two tell you how to choose f and g. The last two do not.

Exercise 5.6. *Let $f(x) = e^x$ and $g(x) = -x$. Use the chain rule to evaluate the derivative of e^{-x}.*

Exercise 5.7. *Let $f(x) = \sqrt{x}$ and $g(x) = 1 + x^2$. Use the chain rule to evaluate the derivative of $\sqrt{1+x^2}$*

Exercise 5.8. *Evaluate $h'(x)$ where $h := \ln(1+x^2)$. To do so, first state a choice of functions f and g such that $h(x) = f(g(x))$. Then use the chain rule.*

Exercise 5.9. *Evaluate $h'(x)$ where $h := e^{-x^2/2}$. To do so, first state a choice of functions f and g such that $h(x) = f(g(x))$. Then use the chain rule.*

5.2 Arguments and proofs

Proofs are for convincing others, as well as for deciding whether you know something for sure, in all cases. The next two exercises ask for opinions on whether or not a proof is needed. There's no right answer, but we expect you to give a good sense of why or why not.

Exercise 5.10 (sum rule - obvious or not?)**.** *The sum rule, in an applied setting, says something like this. Suppose Dick's net worth at time t, call it $f(t)$, is increasing at a certain rate, and Jane's, call it $g(t)$, is increasing at another rate. Then their joint fortune (they are married) is increasing at a rate that is the sum of the two individual rates. Stated in these terms, is the sum rule obvious or does it require proof?*

Exercise 5.11. *In applied terms, suppose $f(t)$ is the length in meters of a turtle that is t days old and $g(t) = 3.3f(t)$ is the length in feet. Then $g'(t)$, the rate of increase of length in feet per day, should be 3.3 times $f'(t)$, the rate of increase in meters per day. Obvious or not?*

In case some of you answered that it was not obvious, here is a mathematical proof. In most of the upcoming proofs, we need to use the definition of the derivative as a limit of difference quotients We don't need to use the ε-δ definition of limit, just known facts about limits.

Proof of sum law: Let $h = f + g$. By definition

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

The difference quotient on the right-hand side simplifies to $\dfrac{f(x)-f(a)}{x-a} + \dfrac{g(x)-g(a)}{x-a}$. This is a sum of two things. The limit of the sum is the sum of limits, therefore

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

As you can see, the logic broke this down into small steps, justified by facts we have accumulated. The proof didn't add a whole lot to our understanding, although it does help to nail down the fact that this holds whenever $f'(a)$ and $g'(a)$ exist, without exceptions for when one of them is zero, or undefined for values other than a, or anything like that.

We'll ask you to do one of these on your own, then not bother you with proofs of things that are borderline obvious.

Exercise 5.12. *Prove Proposition 5.3. It's pretty similar to the proof for the sum rule but a little easier.*

A close up look at the product rule

We mentioned earlier what units a derivative has, but never discussed why. Now is a good time. Taking the limit of an expression gives something with the same units. The derivative is the limit of a difference quotient $(f(x+h) - f(x))/h$. The numerator is the difference between two things with the same units, namely the units of the value of f. The denominator has units of the argument of f. So the difference quotient has units of the value of f divided by the argument of f. For example, if $f(t)$ is distance traveled in the time t, then f' has units of distance per time (such as MPH).

Why is $(fg)'$ not equal to $f'g'$? There are many reasons, one of which is the units. In an application, the values of f and g might have different units, but if both are being differentiated with respect to x then they must have the same input units. The units of $(fg)'$ are, as we have just seen, units of f times units of g divided by units of x, the argument. Unfortunately $f'g'$ has the units of f/x times the units of g/x, so one too many units of x in the denominator.

We now present three arguments for the product rule. When we're done, we'll take a poll of which is most convincing.

Intuitive proof: If f is a constant, so all the change in the product fg comes from changes in g, then we have seen $(fg)' = f \cdot g'$. If g is a constant, then similarly, $(fg)' = gf'$. In reality, both are changing, so the rate of change of the area is the sum of these two individual rates.

Picture proof: Suppose $f(t)$ is the length in meters of a growing rectangular blob at time t seconds, and $g(t)$ is its width. How fast is the area growing at time t?

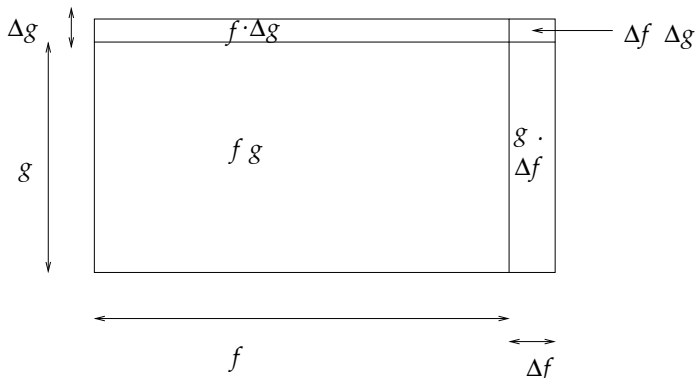


Figure 19: Pictorial proof of the product rule

Figure 19 shows the classical pictorial argument. When time increases by a small quantity Δt, both f and g increase by small quantities, which we respectively call Δf and Δg, and the area increases by $f\Delta g$ plus $g\Delta f$ plus $(\Delta f)(\Delta g)$. We know that Δf is approximately $f'(t)\Delta t$, because in the limit as $\Delta t \to 0$, the ratio $\Delta f/\Delta t$ converges to $f'(t)$. Similarly, $\Delta g \approx g'(t)\Delta t$. From the picture, you can see that $\Delta(fg) = f\Delta g + g\Delta f + (\Delta g)(\Delta f)$. So

$$\frac{\Delta fg}{\Delta t} = f\frac{\Delta g}{\Delta t} + g\frac{\Delta f}{\Delta t} + \frac{(\Delta f)(\Delta g)}{\Delta t}\,.$$

Taking limits on the right hand side as $\Delta t \to 0$ gives $f'g + g'f + \lim_{\Delta t\to 0}(\Delta f)(\Delta g)/\Delta t$. This last limit should be zero. Why? Say $f'(t) = a$ and $g'(t) = b$. Then $\Delta f \approx a\Delta t$ and $\Delta g \approx b\Delta t$, so

$$\lim_{\Delta t\to 0}\frac{(\Delta f)(\Delta g)}{\Delta t} \approx \lim_{\Delta t\to 0}\frac{f'(t)(\Delta t)g'(t)(\Delta t)}{\Delta t} = \lim_{\Delta t\to 0} f'(t)g'(t)(\Delta t)$$

which is zero.

Aside. *We could have called δt something like h, in keeping with the notation in the definition of derivative. We have purposely used different notation here to get you used to seeing multiple different looks. All are common in textbooks. The different notations affect your brain slightly differently. The Δf and Δt notation is designed to make you think of a physical quantity changing as another physical quantity changes. The notation $f(x+h)$ is designed to make you think of a mathematical function with an argument x increased by a small amount h. Both are important frames in which to think.*

FORMAL PROOF: The simplest algebraic proof of the product rule is a bit more "out of the

blue" because it relies on this trick:

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

and hence

$$\frac{f(x+h)g(x+h) - f(x)g(x)}{h} = f(x+h)\frac{g(x+h)-g(x)}{h} + g(x)\frac{f(x+h)-f(x)}{h}\,.$$

The trick was, we added and subtracted $f(x+h)g(x)$ in order to be able to separate the original difference quotient into two pieces, both of which look a function times a simpler difference quotient. Taking limits and using the fact that limits of sums are sums of limits, and the same for products, gives

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

Exercise 5.13. *Because f and g are differentiable, they are continuous. The formal proof above uses that fact that one of the two is continuous at x but does not use continuity of the other. Which continuity fact is needed and where is it used?*

A physics proof of the derivative of the sine function

Suppose a toy car is moving around a circular track of radius one meter, so that its speed is constant 1 meter per second; the coordinates of the point are $x = \cos t, y = \sin t$. By definition of radian, its angle with respect to the horizontal increases at a rate of one radian per second. The northward (y-direction) speed is the derivative of $\sin t$. Suppose at time x a gate opens up and the car stops turning to stay on the track and coasts straight onward at its present speed of 1. Its northward speed during the time $[x, x+1]$ is the derivative of the sine function at time x. To evaluate this, we just have to check how far northward the car went from time x to $x+1$. This is just analytic geometry. The car goes one unit tangent to the circle during this time interval from the point $(\cos x, \sin x)$ (B in Figure 20) to the point $(\cos t - \sin t, \sin t + \cos t)$ (A in the figure). Therefore the derivative of sin is cos. For free, we also get (by looking at the x coordinate) that the derivative of cos is $-\sin$.

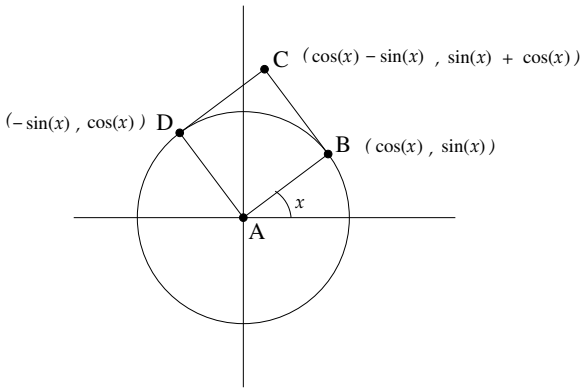


Figure 20
ABCD is a square of side 1 tangent to the unit circle as shown.
At time x the car is at point B, making angle x with the x-axis.
From time x to time $x+1$ the car travels in a straight line to C.

The chain rule

The easiest way to make sense of the chain rule is in terms of related rates. Think of x, u and y as physical quantities related by rules. If you change x, it changes u. The specific rule is $u = g(x)$. If you change u it changes y. The specific rule is $y = f(u)$.

$$x \xrightarrow{\quad g \quad} u \xrightarrow{\quad f \quad} y$$

Aside. *Suppose x is time, u is how many liters of air you breathed in that time and y is how much CO_2 you produced. If you breathe six liters per minute, and you produce $1/20$ liter CO_2 for every liter of air you breathe in, what is your rate of production of CO_2? This simple word problem, which most of you would solve without much thought, turns into the chain rule if your respiration rate or rate of CO_2 production per breath is no longer constant and refer instead to the present instantaneous rates.*

What does this mean quantitatively? The rate of change of u with respect to x is $g'(x)$. This is illustrated on the left side of Figure 21, where the infinitesimal changes dx and du are depicted. The slope of the hypotenuse of the small triangle is $g'(x)$, where in the diagram, the value of x is roughly $1/2$. On the right side of the figure, we see that this small

change in u leads to a proportionate small change in y. The ratio, dy/du is equal to $f'(u)$. One question remains: at what value of u is this ratio evaluated? In the figure, it appears $u \approx 1/8$. More precisely, if we originally took x to be $1/2$, the u value will be $f(1/2)$. In other words, the value from the u-axis (vertical in the first graph) is copied to the second graph (where the u-axis is now the horizontal axis). In other words, f' is evaluated at u, which is $g(x)$. Thus $dy/dx = du/dx \cdot dy/du|_{u=g(x)}$.

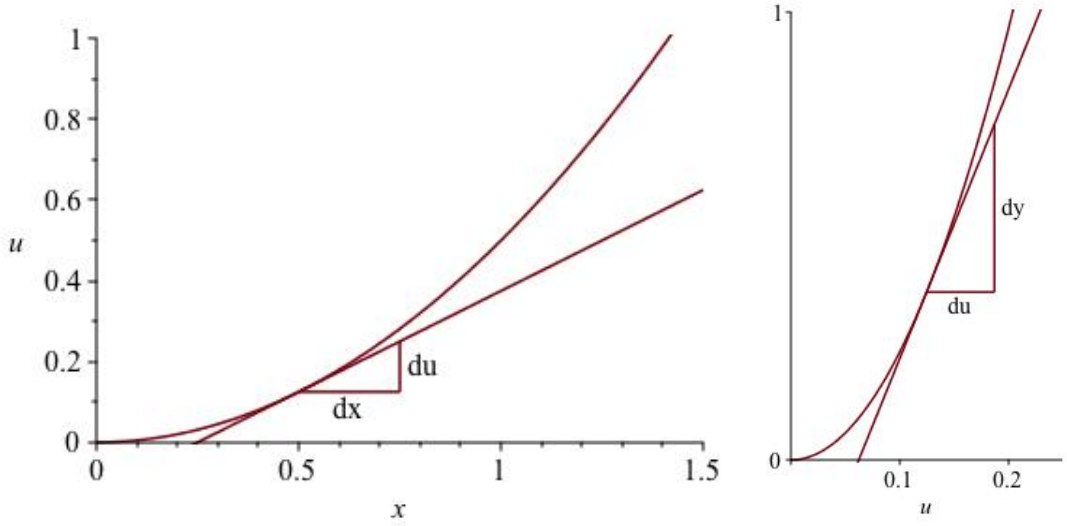


Figure 21: x affects u, which in turn affects y

If we want to make this into a formal proof, we might start by writing

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

If $g(a+h)$ could be replaced by the tangent line approximation $g(a)+hg'(a)$ then the proof would finish easily: letting $\varepsilon := hg'(a)$,

$$\lim_{h\to 0} \frac{f(g(a)+hg'(a))-f(g(a))}{h} = \lim_{\varepsilon\to 0} \frac{f(g(a)+\varepsilon)-f(g(a))}{(\varepsilon/g'(a)} = g'(a)f'(g(a))\,.$$

It is indeed true that the tangent line approximation is close enough to g itself to make this work, but proving that takes a trickier argument than we want to go into here.

Aside. *The first equality above used the a change of variables between ε and h in the limit. We hope you made a note of this following the class discussion of Exercise 3.15.*

6 Asymptotic analysis and L'Hôpital's rule

6.1 Indeterminate forms

This is an optional (but fun) intro to "the infinity rules". Recall from Chapter 3 that limits (two-sided, one-sided, or a limit at $\pm\infty$) that evaluate to DNE may broken into three categories, limits of $+\infty$, limits of $-\infty$ or no limit not even an infinite one, which we will write as UND for "undefined".

Jake is trying to evaluate $\lim_{x\to\infty} \frac{1}{x}$. He says that plugging in ∞ for x you get $1/\infty$ which is 0. Jen is trying to evaluate $\lim_{x\to 0} \frac{\ln x}{\sqrt{x}}$. She says that plugging in ∞ for x you get ∞/∞ which is 1. If you're gut feeling is that Jake is right and Jen is wrong, then you have good instincts. Jake's logic is correct because every time $\lim_{x\to\infty} f(x) = 1$ and $\lim_{x\to\infty} g(x) = \infty$, it follows that $\lim_{x\to\infty} f(x)/g(x) = 0$. Jen's problem contains an *indeterminate form*, meaning that when both $\lim_{x\to\infty} f(x) = \infty$ and $\lim_{x\to\infty} g(x) = \infty$, there are multiple possible values for $\lim_{x\to\infty} f(x)/g(x)$, including any positive real number, ∞, or UND.

When you learn about complex numbers, they seem in one sense like make-believe but in another sense like ordinary math because they obey clear rules. Learning about infinity is different. The word is in the vocabulary of most children, but no one knows the rules! Is infinity part of math? Part of philsophy? Science fiction? It turns out infinity does obey some very clear rules, as long as you decide to define it as a limit. (Trust mathematicians to take the fun out of it!)

Suppose, in addition to the real numbers, we include the numbers $+\infty, -\infty$ and UND. These are the possible limits a function can have. The goal is to create combining rules for limits under the basic operations: addition, subtraction, multiplication, division and taking powers. One rule is that once something is undefined, it stays that way. Limits that DNE could turn out to be $\pm\infty$ rather than UND, but once a limit gets classified as UND, nothing can be inferred about what you get when you add it to something, multiply it, etc. Thus, UND + 3, UND $-\infty$, $-\infty\cdot$ UND and UND / UND are all undefined[7]

[7] Occasionally this classifies a limit as undefined when there is a value, but that's OK as long as we understand UND to mean that our combining rules alone don't determine the value. Example: $2/0 =$ UND but if you know the particular function with a limit of zero is always positive then the limit is actually $+\infty$. Similarly, if $\lim_{x\to\infty} f(x) =$ UND then also $\lim_{x\to\infty} f(x)/f(x) =$ UND by this rule, not the obvious limit, 1.

We want a theory that makes Jake's limit 0 and Jen's limit UND. Let a and b be *extended real numbers*, that is they are either real numbers, or $+\infty$ or $-\infty$. We don't bother with UND because we already agreed that once either a or b is UND then any combination of them is UND. Since you're reading this optional section for your own edification, stop before you read the next definition and think how this theory would work.

Definition 6.1 (operations with infinity)**.** *If a, b and L are extended real numbers, we say that $a + b = L$ if for every extended real number c and every pair of functions f and g such that $\lim_{x\to c} f(x) = a$ and $\lim_{x\to c} g(x) = b$, it is true that $\lim_{x\to c} f(x) + g(x) = L$. If no such extended real L exists we say that $a + b$ is UND and call this an* **indeterminate form***. Extending this definition with any other binary operation in place of addition gives definitions as well for subtraction, multiplication, division, to the power, etc.*

Example 6.2 ($2 + 2 = 4$)**.** Let's be sure we haven't destroyed anything we already knew! We'll check it on "2+2=4", often used for something everyone knows[8]. Checking this is not completely trivial (!) but it follows from Proposition 3.15.

Example 6.3 ($1/\infty$ (Jake's example))**.** To check that $1/\infty = 0$ we need to show that $\lim_{x\to c} f(x) = 1$ and $\lim_{x\to c} g(x) = \infty$ imply that $\lim_{x\to c} f(x)/g(x) = 0$. Briefly, if f is getting near 1 and g is getting very large, you can see that f/g must be getting very small, i.e., close to 0. If you are curious, your TA or instructor can supply the formal proof you'd see in an honors calculus class.

Example 6.4 (∞/∞ (Jen's example))**.** Take $c = infty$ and $f(x) = g(x) = x$. Then obviously $\lim_{x\to\infty} x/x = 1$. On the other hand, changing $f(x)$ to $2x$, we get $\lim_{x\to\infty} f(x)/g(x) = 2$, or if we take $f(x) = x^2$ and $g(x) = x$ we get $\lim_{x\to\infty} f(x)/g(x) = \infty$. Because many different limits are possible, ∞/∞ is undefined. Jen may or may not be right that $lim_{x\to\infty} \ln x/\sqrt{x} = 1$, but the argument that $\infty/\infty = 1$ is bogus.

Exercise 6.1. *Using Definition 6.1 as a guide, say what you think the value (possibly $\pm\infty$ or UND) is for each of these three expressions. You don't need a proof, just a guess.*

(i) $4/\infty$

(ii) 1^{∞}

(iii) $3^{-\infty}$

[8]This famous identity is used as a test for brainwashing in George Orwell's classic novel *1984*.

6.2 L'Hôpital's rule

The previous section is optional because you don't ever NEED to know whether a form is indeterminate. L'Hôpital's rule allows us resolve indeterminate forms in some cases. The hypotheses involve particular indeterminate forms such as $0/0$, but you don't need the infinity rules to use it. Rather, the infinity rules give you an alternate way to evaluate limits when the expression is NOT really an indeterminate form.

In other words, L'Hôpital's rule can determine a limit of an expression such as $f+g$ or f/g or f^g, etc., when this limit is not determined just by knowing the limit of f and the limit of g, so if you do have an indeterminate form, L'Hôpital's rule is often your best option. The basic version of L'Hôpital's rule involves just the one indeterminate form $0/0$.

Theorem 6.5 (L'Hôpital's rule, first version[9])**.** *Let f and g be functions differentiable on an interval containing the point a, except possibly at the point a, where f and g are not required to be defined. Suppose f and g both have limit zero at a and suppose g' is nonzero on the interval. If $\lim_{x\to a} f'(x)/g'(x) = L$ for some finite L, then the limit $\lim_{x\to a} f(x)/g(x)$ exists and is equal to L.*

Example 6.6. L'Hôpital's rule computes $\lim_{x\to 0} \sin(x)/x$ much more easily than in the video from a few weeks ago. Let $f(x) = \sin x$, $g(x) = x$ and $a = 0$ and observe that the continuous functions f and g both vanish at zero, hence $\lim_{x\to 0} f(x) = \lim_{x\to 0} g(x) = 0$. Therefore,

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

You might wonder, when we first evaluated this limit, why did we do it the hard way? Remember, we did not and will not prove L'Hôpital's rule . For this reason it's good to see some things that can be done without it.

Exercise 6.2. *Use L'Hôpital's rule to evaluate the following limits. Please state what are f, g, a, f' and g', as well as the value of the limit.*

[9] L'Hôpital's rule uses derivatives to compute limits. You might object that this is circular because limits are used to define derivatives. It is not circular, because in each case, we use facts we already know to compute ones we don't. We should probalby avoid using L'Hôpital's rule to prove general theorems about derivatives, given that we are not going to prove L'Hôpital's rule and don't know what theorems about derivatives it relied on. But it's safe to use L'Hôpital's rule to evaluate individual derivatives. We promise no individual derivative was used in the proof of L'Hôpital's rule.

(a) $\lim_{x\to 1} \dfrac{e^x - 1}{x}$

(b) $\lim_{x\to 10} \dfrac{\sqrt[3]{x} - \sqrt[3]{10}}{\sqrt{x} - \sqrt{10}}$

There are two common mistakes in applying L'Hôpital's rule. One is trying to use it the other way around. If f/g has a limit at a, that doesn't mean f'/g' does, or that these even exist. The other is to try to use it when f or g has a nonzero limit at a. For example, if $\lim_{x\to a} f(x) = 5$ and $\lim_{x\to a} g(x) = 3$ then $\lim_{x\to a} f(x)/g(x) = 5/3$ (the nonzero quotient rule) and is probably not equal to $\lim_{x\to a} f'(x)/g'(x)$.

Exercise 6.3. *Which (possibly several, possibly none) of these uses of L'Hôpital's rule are valid (hypotheses are satisfied and conclusion is correctly applied)?*

(i) $\lim_{x\to 3} \dfrac{x^2 - 10}{x - 3} = \lim_{x\to 3} \dfrac{2x}{1} = 6.$

(ii) $\lim_{x\to 2} \dfrac{x^2 - 4}{x - 2} = \lim_{x\to 2} \dfrac{2x}{1} = 4.$

(iii) $\lim_{x\to\infty} \dfrac{6 - e^{-x}}{3 - e^{-2x}} = 2$ *and the respective derivatives on top and bottom are* e^{-x} *and* $2e^{-2x}$*, therefore* $\lim_{x\to\infty} \dfrac{e^{-x}}{2e^{-2x}} = 2.$

More general versions

If the hypotheses hold only from one side, for example $\lim_{x\to a^+} f(x) = \lim_{x\to a^+} g(x) = 0$, then the conclusion still holds on that side: if $\lim_{x\to a^+} f'(x)/g'(x) = L$ then $\lim_{x\to a^+} f(x)/g(x) = L$. Also, the limit can be taken at $\pm\infty$ and nothing changes.

Proposition 6.7 (Improved L'Hôpital's rule)**.**

(i) Suppose f and g are differentiable on an open interval (a, b), with f and g both having limit zero at a. Suppose that $g' \neq 0$ on (a, b) and $\lim_{x\to a^+} f'(x)/g'(x) = L$. Then $\lim_{x\to a^+} f(x)/g(x) = L$.

(ii) Suppose f and g are differentiable on an open interval (b, a) with f and g both having limit zero at a. Suppose that $g' \neq 0$ on (b, a) and $\lim_{x\to a^-} f'(x)/g'(x) = L$. Then $\lim_{x\to a^-} f(x)/g(x) = L$.

(iii) Suppose f and g are differentiable on an open interval (b, ∞) with f and g both having limit zero at infinity. Suppose that $g' \neq 0$ on (b, ∞) and and $\lim_{x\to\infty} f'(x)/g'(x) = L$. Then $\lim_{x\to\infty} f(x)/g(x) = L$. The same holds for limits at $-\infty$, replacing the interval with $(-\infty, b)$.

Exercise 6.4. *Which of these would you use to evaluate the limit at zero of $\ln(1+x)/\sqrt{x}$, and what is the limit?*

Turning other indeterminate forms into $0/0$

The case $0 \cdot \infty$

Suppose $\lim_{x\to a} f(x) = 0$ and $\lim_{x\to a} g(x) = \infty$. How can we compute $\lim_{x\to a} f(x) \cdot g(x)$? We know that $\lim_{x\to a} 1/g(x) = 1/\infty = 0$. Therefore, an easy trick is to replace multiplication by g with division by $1/g$. Letting h denote $1/g$, we have

$$\lim_{x\to a} f(x)g(x) = \lim_{x\to a} \frac{f(x)}{h(x)}$$

which is the correct form for L'Hôpital's rule.

Example 6.8. What is $\lim_{x\to 0^+} x \cot x$? Letting $f(x) = x$ and $g(x) = \cot x$ we see this has the form $0 \cdot \infty$. Letting $h(x) = 1/g(x) = \tan x$ we see that

$$\lim_{x\to 0^+} x \cot x = \lim_{x\to 0^+} \frac{x}{\tan x} = \lim_{x\to 0^+} \frac{x}{\sin x} \cdot \cos x\,.$$

The limit at 0 of $x/\sin x$ is 1 and the limit of the continuous function $\cos x$ is $\cos(0) = 1$, therefore the answer is $1 \cdot 1 = 1$.

The case ∞/∞

You could invert both f and g, writing $\dfrac{f(x)}{g(x)}$ as $\dfrac{1/g(x)}{1/f(x)}$. There is a reasonable chance that L'Hôpital's rule can be applied to this. There is also another version of L'Hôpital's rule specifically for this case.

Theorem 6.9 (L'Hôpital's rule for ∞/∞)**.** *If both f and g tend to ∞ or $-\infty$ as $x \to a$, then*

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}$$

whenever the right-hand side is a real number or $\pm\infty$.

Exercise 6.5. *Compute* $\lim_{x\to\infty} \dfrac{x}{e^x}$.

The cases 1^∞, 0^0 and ∞^0

The idea with indeterminate powers is to take the log, compute the limit, then exponentiate. The reason this works is that e^x is a continous function. Theorem 3.16 says that if $\lim_{x\to a} h(x) = L$ then $\lim_{x\to a} e^{h(x)} = e^L$.

The way we will use this when evaluating something of the form $\lim_{x\to a} f(x)^{g(x)}$ is to take logarithms. Algebra tells us $\ln f(x)^{g(x)} = g(x)\ln f(x)$. If we can evaluate $\lim_{x\to a} g(x)\ln f(x) = L$ then we can exponentiate to get $\lim_{x\to a} f(x)^{g(x)} = e^L$.

Exercise 6.6. *What is* $\lim_{x\to\infty} x^{1/\ln(x)}$*? RP says: "Maybe I'm warped, but I think this one is cute: surprising and easier than it looks."*

Example 6.10 (continuous compounding)**.** Suppose you have a million dollars earning a 12% annual interest rate for one year. You might thing after a year you will have 1.12 million dollars. But no, things are better than that. The bank compounds your interest for you. They realize you could have cashed out after half a year with 1.06 million and reinvested for another half year, giving you 1.1236 million, which doesn't seem so different but is actually 3600 dollars more. You could play this game more frequently, dividing the year into n periods and earning $12\%/n$ interest n times, so your one million becomes $(1+0.12/n)^n$ million.

With computerized trading, you could make the period of time a second, or even a microsecond. Does this enable you to claim an unboundd amount of money after one year? To answer that, let's compute the amount you would get if you compounded *continuously*, namely $\lim_{n\to\infty}(1+0.12/n)^n$. Taking logs gives $\ln(1+0.12/n)^n = n\ln(1+0.12/n)$. Changing to the variable $x := 1/n$,

$$\begin{aligned}
\lim_{n\to\infty} n\ln(1+0.12/n) &= \lim_{x\to 0}\frac{\ln(1+0.12x)}{x} \\
&= \lim_{x\to\infty}\frac{(d/dx)\ln(1+0.12x)}{(d/dx)x} \text{ (L'Hôpital's rule)} \\
&= \lim_{x\to\infty}\frac{0.12/(1+0.12/x)}{1} \text{ (use the chain rule)} \qquad = 0.12\,.
\end{aligned}$$

Therefore, $\lim_{n\to\infty}(1+0.12/n)^n = e^{0.12} \approx 1.12749685$ million dollars. That's better than the \$120,000 you earn without compounding, or the \$3,600 more than that you earn compounding once, but it's not infinite, it's just another \$3896.85 better.

Exercise 6.7. *What is* $\lim_{t\to 0}(1+t)^{1/t}$*? This limit is sometimes used to define the famous constant named after Euler.*

Because interest is quoted in both continuous and annualized rates, we need to agree on terminology to distunguish between these. Our terminology is reasonably consistent with industry usage, however you should be warned that real world usage can vary quite a bit. For us an **interest rate** always refers to a continuous exponential growth rate, r, quoted either as a real number in units of inverse time, or a percentage R so that $R = R/100$. For example, an interest rate of $r = 0.07$ annually (which is in units of inverse time because "annually" means "per year") is the same as an annual interest are of $R = 7\%$ and corresponds to the way your money would grow in a savings account that offered 7% interest compounded continuously. As we have seen in Example 6.10, this corresonds to a one-year **growth factor** of $e^{0.07}$.

The growth factor for t years instead of one year is easily seen to be e^{rt}; thus money growing at a constant continuous interest rate is an example of exponential growth.

Exercise 6.8. *Verify this by dividing the* t *years into* tn *intervals of size* $1/n$ *years (as was done in Example 6.10 with* $t = 1$*) and computing* $\lim_{n\to\infty}(1+r/n)^{tn}$*.*

There's also a name for the annual yield, which is how much the interest looks like if you receive it in a lump sum at the end of a year. For example, an interest rate of $r = 7\%$ gives a one-year growth factor of $e^{0.07}$, leading to an annual yield of $e^{0.07} - 1$, which is a little over 0.0725. Multiplying by 100 to write this as a percentage we say that the **APY** which stands for **Annual Percentage Yield** is a little over 7.25%. Because consumers find the idea of annual yields easier to understand, banks have now for decades been required to quote interests rates in terms of the APY.

Letting r be the interest rate, so $R = 100R$ is the percentage interest rate, with g denoting the growth factor and $y = g - 1$ the annual yield, we can solve for any of these in terms of

any other to obtain

$$\begin{aligned} g &= e^r && (6.1)\\ y &= e^r - 1 && (6.2)\\ r &= \ln g && (6.3)\\ r &= \ln(1+y) && (6.4) \end{aligned}$$

If you prefer things in percentages, the APY for example, in terms of the percentage interest rate R, would be given by $\text{APY} = (100e^{R/100} - 1)$.

Exercise 6.9. *What is the inverse function for this, that writes R in terms of the APY?*

Repeated use of L'Hôpital's rule

Sometimes when trying to evaluate $\lim_{x\to a} f(x)/g(x)$ you find that $\lim_{x\to a} f'(x)/g'(x)$ appears a bit simpler, but you still can't tell what it is. You might try L'Hôpital's rule twice. If $f'(x)$ and $g'(x)$ tend to zero as $x \to a$ (if they don't, you can probably tell what the limit is), then you can use f' in place of f and g' in place of g in L'Hôpital's rule. If you can evaluate the limit of $f'(x)/g'(x)$ then this must be the limit of $f(x)/g(x)$. You can often do a little better if you simplify $f'(x)/g'(x)$ to get a new numerator and denominator whose derivatives will be less messy.

Example 6.11. Repeated L'Hôpital's rule makes another limit that was formerly painful into a piece of cake: $\lim_{x\to 0}(1-\cos x)/x^2$. Both numerator and denominator are zero at zero, so we apply L'Hôpital's rule to see that the limit is equal to $\lim_{x\to 0} \sin x/(2x)$. You can probably remember what this is, but in case not, one more application of L'Hôpital's rule shows it to be equal to $\lim_{x\to 0} \cos x/2 = \cos(0)/2 = 1/2$.

Exercise 6.10. *Compute* $\displaystyle\lim_{x\to\infty} \frac{x^3}{e^x}$.

6.3 Orders of growth at infinity

Often in mathematical modeling, one hears statements such as "This model produces a much smaller growth rate than the other model, as time gets large." This statement sounds

vague: how much is "much smaller" and what are "large times"? In this section we will give a precise meaning to statements such as this one.

Why are we spending our time making a science out of vague statements? Answer: (1) people really think this way, and it clarifies your thinking to make these thoughts precise; (2) a lot of theorems can be stated with these as hypotheses; (3) knowing the science of orders of growth helps to fulfill the Number Sense mandate because you can easily fit an unfamiliar function into the right place in the hierarchy of more familiar functions.

We focus on two notions in particular: when one function is **much** bigger/smaller/closer than another, and when two functions are **asymptotically equal**.

Mostly we will be comparing functions of x as $x \to \infty$. Let f and g be positive functions.

(*i*) We say the function f is **asymptotic to** the function g, short for "asymptotically equal to", if

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1\,.$$

This is denoted $f \sim g$.

(*ii*) The function f is said to be **much** smaller than g, or to grow "much more slowly" if

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0\,.$$

This is denoted $f \ll g$. Typically this notation is used only when g is positive.

Example 6.12. Is it true that $x^2 + 3x$ is asymptotically equivalent to x^2? Intuitively it should be true because $3x$ is a lot smaller than x^2 when x is large (in fact, it is *much smaller*) so adding it to x^2 should be negligible. We check that

$$\lim_{x\to\infty} \frac{x^2+3x}{x^2} = \lim_{x\to\infty} 1 + \frac{3}{x} = 1\,,$$

therefore indeed $x^2 + 3x \sim x^2$.

Exercise 6.11. *True or false:*

(i) $x \sim x + 1$

(ii) $e^x \sim e^{x+1}$

(iii) $\ln x \ll x$

Example 6.13. Let's compare two powers, say x^3 and $x^{3.1}$. Are they asymptotically equivalent or does one grow much faster? Taking the limit at infinity we see that $\lim_{x\to\infty} x^3/x^{3.1} = \lim_{x\to\infty} x^{-0.1} = 0$. Therefore, $x^3 \ll x^{3.1}$. This is shown on the left side of Figure 22.

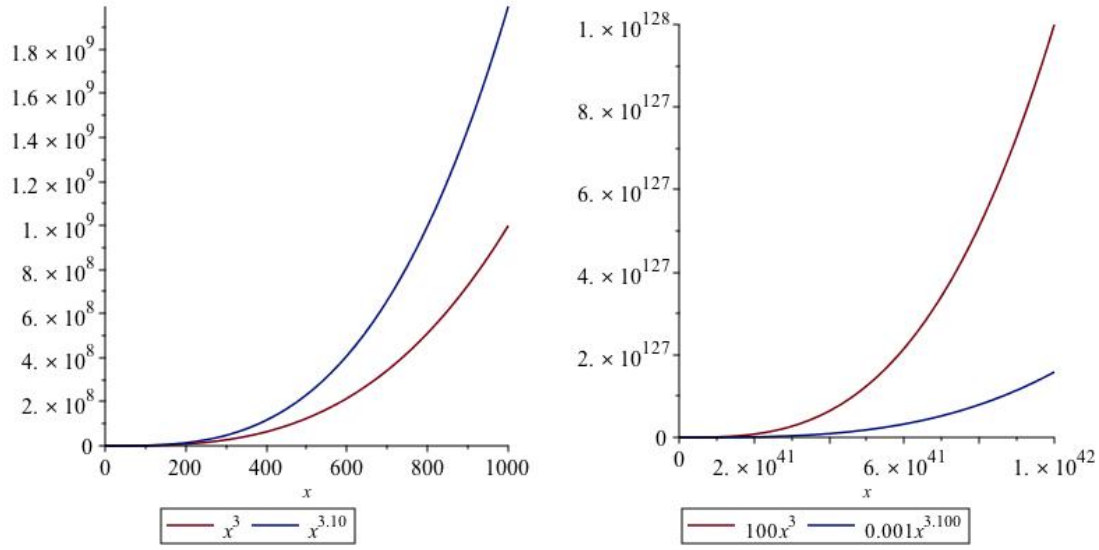


Figure 22: Comparing various multiples of x^3 and $x^{3.1}$

What about comparing $100x^3$ with $0.001x^{3.1}$? The plot on the right side of Figure 22 appears to show that $100x^3$ remains much greater than $0.001x^{3.1}$, at least beyond a duodecillion (look it up). Doing the math gives

$$\lim_{x\to\infty} \frac{100x^3}{0.001x^{3.1}} = \lim_{x\to\infty} 100000x^{-0.1} = 0\,.$$

Therefore, again, $100x^3 \ll 0.001x^{3.1}$. Whether or not you care what happens beyond 10^{42} depends on the application, but the math is pretty clear: if $a < b$, then $Kx^a \ll Lx^b$ for any positive constants K and L.

Discussion

This is a general rule: the function $g(x) + h(x)$ will be asymptotic to $g(x)$ exactly when $h(x) \ll g(x)$. Why? Because $(g(x) + h(x))/g(x)$ and $h(x)/g(x)$ differ by precisely 1. It

follows that if $g(x) + h(x) \sim g(x)$ then

$$1 = \lim_{x\to\infty} \frac{g(x)+h(x)}{g(x)} = \lim_{x\to\infty} 1+\frac{h(x)}{g(x)} = 1+\lim_{x\to\infty} \frac{h(x)}{g(x)} \qquad \text{hence} \qquad \lim_{x\to\infty} \frac{h(x)}{g(x)} = 0\,,$$

or in other words, $h \ll g$. The chain of identies runs backward as well: $g + h \sim g$ if and only if $h \ll g$.

Another principle is that if $f \sim g$ and $h \sim \ell$ then $f \cdot h \sim g \cdot \ell$. This is literally just the product rule for limits. The same is true for nonzero quotients, for the same reason.

Example 6.14. We know $x + 1/x \sim x$ and $2 - e^{-x} \sim 2$, therefore

$$\frac{x+1/x}{2-e^{-x}} \sim \frac{x}{2}\,.$$

These two facts give important techniques for estimating. They allow you to clear away irrelevant terms: in any sum, every term that is much less than one of the others can be eliminated and the result will be asymptotic to what it was before. You can keep going with products and quotients.

Example 6.15. Find a nice function asymptotically equal to $\sqrt{x^2+1}$. The notion of "nice" is subjective; here it means a function you're comfortable with, can easily estimate, and so forth.

> Because $1 \ll x^2$ we can ignore the 1 and get $\sqrt{x^2}$ which is equal to x for all positive x. Therefore, $\sqrt{1+x^2} \sim x$.

Beware though, if $f \sim g$ and $h \sim \ell$, it does not follow that $f + h \sim g + \ell$ or $f - h \sim g - \ell$. Why? See the following self-check exercise.

Exercise 6.12. *Let $f(x) = x + 1$ and $g(x) = x + 1/x$. Let $h(x) = x$. Evaluate the truth or falsity of these claims, then say in words what went wrong with the proposed "subtraction principle for asymptotic equivalence."*

(i) $f \sim g$

(ii) $h \sim h$

(iii) $f - h \sim g - h$

It should be obvious that the relation $\sim$ is symmetric: $f \sim g$ if and only if $g \sim f$. Formally,

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1 \iff \lim_{x\to\infty} \frac{g(x)}{f(x)} = 1$$

because one is the reciprocal of the other. On the other hand, the relation $f \ll g$ is anti-symmetric: it is not possible that both $f \ll g$ and $g \ll f$.

It is good to have an understanding of the relative sizes of common functions. Here is a summary of some basic facts from today's lesson, practice problems and homework problems.

Proposition 6.16.

1. *Positive powers all go to infinity but at different rates, with the higher power growing faster.*
2. *Exponentials grow at different rates and every exponential grows faster than every power.*
3. *Logarithms grow so slowly that any power of* $\ln x$ *is less than any positive power of* x*.*

6.4 Comparisons elsewhere and orders of closeness

Everything we have discussed in this section has referred to limits at infinity. Also, all our examples have been of functions getting large, not small, at infinity. But we could equally have talked about functions such as $1/x$ and $1/x^2$, both of which go to zero at infinity. It probably won't surprise you to learn that $1/x^2$ is much smaller than $1/x$ at infinity.

Exercise 6.13. *Use the definitions to verify that* $1/x^2 \ll 1/x$*.*

These same notions may be applied elsewhere simply by taking a limit as $x \to a$ instead of as $x \to \infty$. The question then becomes: is one function much smaller than the other as the argument approaches a? In this case it is more common that both functions are going to zero than that both functions are going to infinity, though both cases do arise. Remember: at a itself, the ratio of f to g might be $0/0$ or ∞/∞, which of course is meaningless, and can be made precise only by taking a limit as x approaches a.

The notation, unfortunately, is not built to reflect whether $a = \infty$ or some other number. So we will have to spell out or understand by context whether the limits in the definitions of $\ll$ and $\sim$ are intended to occur at infinity or some other specificed location, a.

Example 6.17. Let's compare x and x^2 at $x = 0$. At infinity, we know $x \ll x^2$. At zero, both go to zero but at possibly different rates. Have a look at Figure 23. You can see that x has a postive slope whereas x^2 has a horizontal tangent at zero. Therefore, $x^2 \ll x$ as $x \to 0^+$. You can see it from Figure 23 or from L'Hôpital:

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

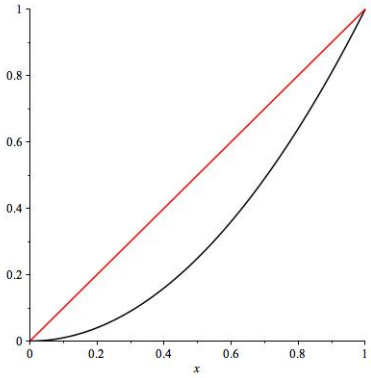


Figure 23: Comparing x (red) and x^2 (black) at $x = 0$

Example 6.18. What about x^2 and x^4 near zero? Both have slope zero. By eye, x^4 is a lot flatter. Maybe $x^4 \ll x^2$ near zero. It is not clearly settled by the picture (do you agree? see Figure 24), but the limit is easy to compute.

Exercise 6.14. *Compute the limit needed to settle the previous answer.*

Here is a less obvious example, still with powers.

Example 6.19. Let's compare $\sqrt{x}$ and $\sqrt[3]{x}$ near zero. See Figure 25. Is one of these functions much smaller than the other as $x \to 0^+$? Here, the picture is pretty far from giving a definitive answer!

We try evaluating the ratio: $f(x)/g(x) = x^{1/2}/x^{1/3} = x^{1/2-1/3} = x^{1/6}$. Therefore,

$$\lim_{x\to 0^+} \frac{f(x)}{g(x)} = \lim_{x\to 0^+} x^{1/6} = 0$$

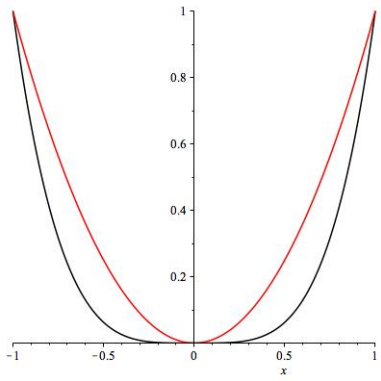


Figure 24: Comparing x^2 (red) and x^4 (black) at $x = 0$

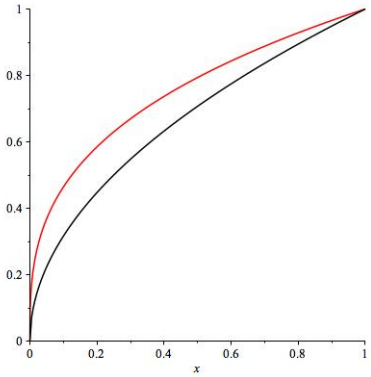


Figure 25: Comparing $\sqrt{x}$ (black) and $\sqrt[3]{x}$ (red) at $x = 0$

and indeed $x^{1/2} \ll x^{1/3}$. Intuitively, the square root of x and the cube root of x both go to zero as x goes to zero, but the cube root goes to zero a lot slower (that is, it remains bigger for longer).

Exercise 6.15. *Let a, b, K, L be positive constants with $a < b$. Determine which of Kx^a or Lx^b is much greater than the other at $x = 0$, if either.*

Suppose f and g are two nice functions, both of which are supposed to be approximations to some more complicated function H near the argument a. The question of whether $f - H \ll g - H$, or $g - H \ll f - H$, or neither as $x \to a$ is particularly important because it tells us whether one of the two functions f and g is a much better approximation to H than is the other. We will be visiting this question shortly in the context of the tangent

line approximation, and again later in the context of Taylor polynomial approximations.

"For sufficiently large x"

Often when discussing comparisons at infinity we use the term "for sufficiently large x". That means that something is true for every value of x greater than some number M (you don't necessarily know what M is). For example, is it true that $f \ll g$ implies $f < g$? No, but it implies $f(x) < g(x)$ for sufficiently large x. Any limit at infinity depends only on what happens for sufficiently large x.

Example 6.20. We have seen that $\ln x \ll \sqrt{x-5}$. It is not true that $\ln 6 < \sqrt{6-5}$ (the corresponding values are about 1.8 and 1) and it is certainly not true that $\ln 1 < \sqrt{1-5}$ because the latter is not even defined. But we can be certain that $\ln x < \sqrt{x-5}$ for sufficiently large x. The crossover point is between 10 and 11.

7 Optimization

7.1 Definitions of Minima and Maxima, and their existence

Many of you have seen max-min problems before. If not, pay attention! Finding the maximum or minimum of a function is one of the crowning achievements of calculus. This occurs in business (maximize profit), medicine (minimize mortality), mechanical engineering (what is the maximum load?), economics (maximize utility), population genetics (maximize selective advantage), actuarial science (minimize risk), and further applications in every field that uses mathematical models and methods.

The following definitions give precise meaning to notions you have probably already seen. Some vocabulary may be new but none of it is rocket science.

Definition 7.1.

- *A point $x \in [a,b]$ such that $f(x) \leq f(y)$ for all $y \in [a,b]$ is called a* **minimum** (Plural: minima)*. This is also called a* **global** *or* **absolute minimum on** $[a,b]$*.*
- *A point $x \in [a,b]$ such that $f(x) \geq f(y)$ for all $y \in [a,b]$ is called a* **maximum** (Plural: maxima)*. This is also called a* **global** *or* **absolute maximum on** $[a,b]$*.*
- *The word for a something that is a minimum or maximum is* **extremum** *or* **extreme value** (Plural: extrema)*.*
- *A* **local minimum** *is a value x such that $f(x) \leq f(y)$ for all y in some open interval I containing x, which could be a lot smaller than the whole interval (a,b). The terms* **local maximum** *and* **local extremum** *are defined analogously.*
- *A* **critical point** *is a point where f' is zero or undefined.*

A subtle but important piece of vocabulary distinguishes between the **location** of the extremum (the value of x) and the **value** of the extremum, namely $f(x)$ where x is the location. When we refer to "the maximum" without saying "location" or "value" it is assumed we mean the value. Both are important though, as can be seen through these examples.

- If I want to build a building to house my flying squirrels, I need to know what the maximum height they're capable of flying is, but I don't really care when they get to that height.
- If I need to build a window which admits the most possible light, what I care about is how to set the dimensions (an input), but the amount of light actually let in (in lumens, say) isn't really needed.
- If I'm running a widget factory and I want to know what production level will maximize my profit, the input where the maximum occurs (a number of widgets per hour) is important, but for fiscal planning I also need to know what that maximum (a number of dollars) actually is.

Before we start looking for extrema, it might occur to you to question whether they exist.

Exercise 7.1.

(i) Find a discontinuous function defined on the interval $[-2, 1]$ with no absolute maximum nor minimum on that interval.

(ii) Find a continuous function on $(-2, 1)$ with no absolute maximum nor minimum on that interval.

Now that you have seen some scenarios where functions have no absolute extrema on an interval, here is a theorem guaranteeing the opposite.

Aside. *Like the Intermediate Value Theorem, this theorem requires mathematical analysis to prove; we would say it was obvious were it not for the counterexamples we paraded by you in the self-check exercises!*

Theorem 7.2. *Let f be a continuous function on the closed interval $[a, b]$. Then f has at least one absolute minimum on $[a, b]$ and at least one absolute maximum on $[a, b]$.*

Exercise 7.2. *What hypothesis of Theorem 7.2 is violated in each part of Exercise 7.1?*

7.2 The role of calculus

Theorem 7.3 (Fermat's Theorem)**.** *Suppose a function f has a minimum at a point c in some open interval I. If f is differentiable at c then $f'(c) = 0$.*

This result should seem very credible an intuitive level. If $f'(c) > 0$ then moving to the left from c to $c - \varepsilon$ should produce a greater value of f. Likewise, if $f'(c) < 0$ then moving to the right should produce a greater value. This is the most intuitive justification we could write down, though not exactly airtight.

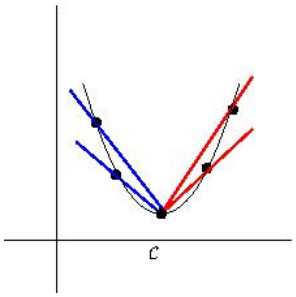


Figure 26: difference quotients between c and points to the right (red) are positive; those to the left (blue) are negative

Here is a more airtight argument. Because f is differentiable at c, the one-sided derivatives exist and are equal. The derivative from the right is $\lim_{x \to c^+} \dfrac{f(x) - f(c)}{x - c}$; because c is a minimum, both top and bottom of this fraction are positive (the numerator could be zero). The limit of nonnegative numbers is nonnegative, hence $f'(c_+) \geq 0$; see Figure 26. Similarly, $f'(c_-)$ is a limit in which each term is nonpositive, thus $f'(c_-) \leq 0$. For these to be equal, both must equal zero. This finishes the proof.

Exercise 7.3. *Suppose f is differentiable on $[a, b]$ (derivatives at the endpoints are one-sided). If the minimum of f on this interval occurs at the left endpoint, can you conclude that the one-sided derivative there is zero? Explain.*

Let's get the logic straight. It is of the form `minimum` $\Rightarrow$ $f' = 0$. The converse is not necessarily true: $f' = 0$ $\Rightarrow$ `minimum`. Nevertheless, everyone's favorite procedure for finding minima is to set f' equal to zero. Why does this work, or rather, when does this work? From Theorem 7.2, if f is defined and continuous on a closed interval $[a, b]$, then indeed f has to have a minimum somewhere on $[a, b]$. We can find it by using Theorem 7.3 to rule out where it's not: if $a < c < b$ and $f'(c) \neq 0$, then definitely the minimum does not occur at c. Where can it be then? What's left is the point a, the point b, every point where f' is zero, and every point where f' does not exist. An identical argument shows the same is true for the maximum. Summing up:

Theorem 7.4. *Suppose f is continuous on $[a, b]$ and differentiable everywhere on (a, b)*

except for a finite number of points $c_1, \ldots, c_k$. Then the minimum value of f on $[a, b]$ occurs at one or more of the points $\{a, b, c_1, \ldots, c_k,$ anywhere $f' = 0\}$, and nowhere else. The maximum also occurs at one or more of these points and nowhere else.

Exercise 7.4. *Let $f(x) := |x|$ and let $[a, b]$ be the interval $[-2, 2]$. Does the theorem say f must have a minimum on this interval? If so, what does the theorem say about where the minimum must be? Answer the same question for the maximum of f on $[a, b]$.*

Being differentiable except for a number of points is sometimes called being **piecewise differentiable**, because the function is differentiable in pieces, the pieces being the intervals $(c_0, c_1), (c_2, c_2), \ldots, (c_{k-1}, c_k)$.

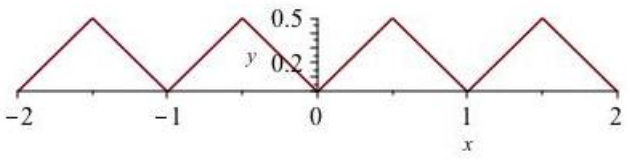


Figure 27

Exercise 7.5. *Let f be the "sawtooth" function shown in Figure 27, defined by letting $f(x)$ be the distance from x to the nearest integer, either $\lfloor x \rfloor$ or $\lceil x \rceil$.*

Is f piecewise differentiable on $[-2, 2]$? If so, give a value of k and $c_0, \ldots, c_k$ that show this to be true. If not, say why not.

You can write Theorem 7.4 as a procedure if you want. Even if you're looking only for the minimum or only for the maximum, the procedure is the same so it will find both.

Procedure 7.5 (finding extreme values)**.**

(1) Make sure f is continuous on $[a, b]$; if not, abort procedure.

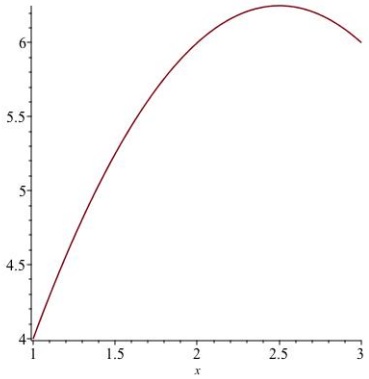
(2) Write down all $x \in (a, b)$ where $f'(x) = 0$.

(3) Add to these all $x \in [a, b]$ where $f'(x)$ DNE.

(4) Add to the list the points a and b.

(5) For every point x on the list, compute $f(x)$; the one(s) of these where f is greatest will be the maxima; the one(s) where f is least will be the minima.

Example 7.6. Find the maximum of $f(x) := 5x - x^2$ on the interval $[1, 3]$; see the figure at the right. Computing $f'(x) = 5 - 2x$ and setting it equal to zero we see that $f'(x) = 0$ precisely when $x = 2\frac{1}{2}$. There are no points where f is undefined, so our list consists of just the one point plus the two endpoints: $\{1, 2\frac{1}{2}, 3\}$. Checking the values of f there produces $4, 6\frac{1}{4}, 6$. The maximum is the greatest of these, occurring at $x = 2\frac{1}{2}$.



Exercise 7.6. *Find the maximum and minimum of $x^3 - x^2 - 2x$ on the interval $[-1, 3]$.*

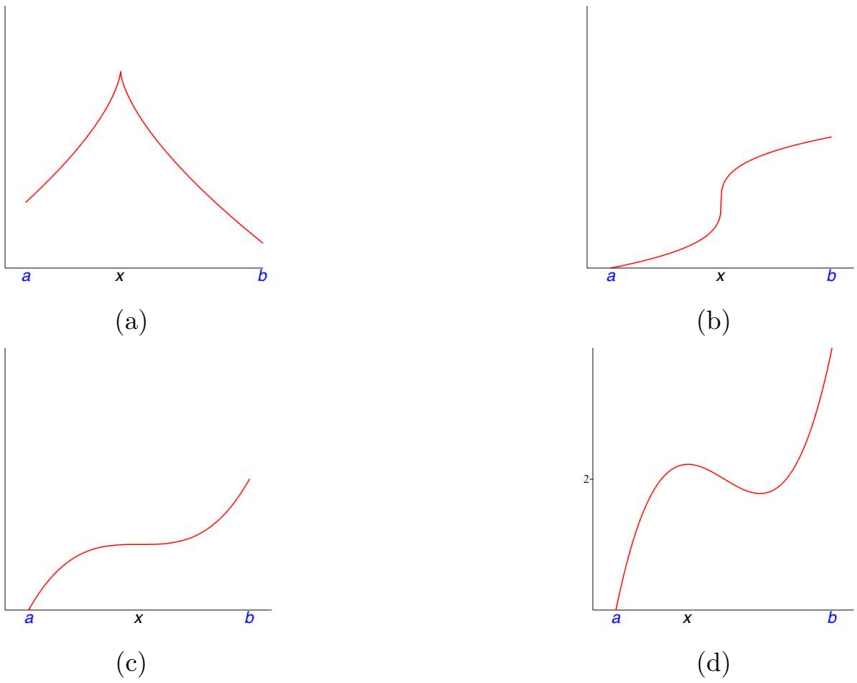


Figure 28

Exercise 7.7. *Here are some other things you may find when you use Procedure 7.5. Match each of these verbal descriptions to the role of x in one of the four pictures in Figure 28. Then state which picture has an endpoint that is not a global extremum.*

- *f has a local extreme value at x but not a global one*

- *$f'(x) = 0$ but f is neither a local minimum nor a local maximum*
- *$f'(x)$ is undefined but x is in fact a maximum or minimum*
- *$f'(x)$ is undefined and f is not an extremum*

Example 7.7 (interval is not closed, function has no minimum)**.** Let $f(x) = x$ and consider the half-closed interval $(0, 1]$. In this case we have a continuous function but not a closed interval. This example represents a scenario where you make a donation in bitcoin to enter a virtual tourist attraction and you want to spend as little as possible. You have 1 bitcoin, so that's the maximum you can donate; donations can be any positive real number but zero is not allowed. The minimum of x on $(0, 1]$ does not exist: there is no smallest positive real number. The interpretation is clear: no matter how little you donate, you could have donated less. Mathematically, this clarifies the need for a closed interval in Theorem 7.2.

Exercise 7.8.

(i) True or false: the function e^{-x} has a global minimum on the whole real line?

(ii) True or false: the function xe^{-x} has a global minimum on the nonnegative half-line $[0, \infty)$?

Second derivatives

Recall that wherever f has a second derivative, if $f'' \neq 0$ then the sign of f'' determines the concavity of f. If $f''(x) > 0$ then f is concave upward and if $f''(x) < 0$ then f is concave downward. At a point where $f' = 0$, if we know the concavity, we know whether f has a local maximum or local minimum.

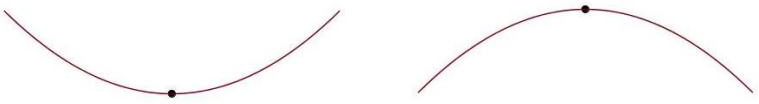


Figure 29: a critical point where $f'' < 0$ (left) and where $f'' > 0$ (right)

Example 7.8. What are the extrema of the function $f(x) := x^2 + 1/x$ on the interval $(0, 2)$? The only critical point is where $f'(x) = 2x - 1/x^2 = 0$, hence $x = \sqrt[3]{1/2}$. Here, $f''(x) = 2 + 2/x^3 > 0$ therefore this is a local minimum. There are not any local maxima. This means f has no global maximum on $(0, 2)$. It may have a global minimum, and indeed, Figure 30 shows that $x = \sqrt[3]{2}$ is a local minimum. We will discuss further tools for arguing whether a local extremum on a non-closed interval is a global extremum.

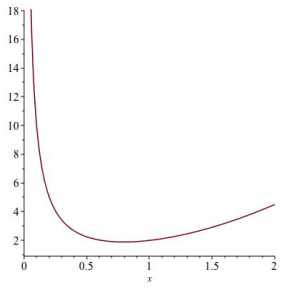


Figure 30: the function $x^2 + 1/x$ on the interval $(0, 2)$

Remark. If the second derivative vanishes along with the first, you won't know any more than you did already.

Applications

Finding extrema is part of a subject called *optimization*. The prototypical application is that you control a parameter x and are would like to maximize some **objective function**, f, which is perhaps how large you can build something, or perhaps revenue minus cost.

Example 7.9. The logistic equation models growth rate per unit time, call it R, of a population as $R(x) = Cx(A - x)$. Here C is a constant of proportionality, x is the present population, and A is a theoretical limit on the population size supported by the habitat. At what size is the population growing the fastest?

We need to find the maximum of $R(x) := Cx(A - x)$ on $[0, A]$. The reason for restricting to this interval is that we are told the population size is constrained to be at most A, and of course it has to be nonnegative. Computing $R'(x) = C(A - 2x)$, we find $R' = 0$ for a single value, $x = A/2$. Checking the endpoints, we find R is zero at both. Therefore the maximum value occurs at $x = A/2$.

Exercise 7.9. *What are the units of x, R, A and C?*

Example 7.10. Suppose the cost of supplying a station is proportional to the distance from the station to the nearest port, and the cost of the land for the station is inversely proportional to the distance to the nearest port. Adding together these costs, what is the least expensive distance at which to put the station?

Letting x be the distance to the nearest port and $f(x)$ be the cost, we are told that $f(x) = ax + b/x$ where a and b are unspecified constants. The value of $f(x)$ is defined for every positive x and f is continuous on $(0, \infty)$. We seek the global minimum of f on $(0, \infty)$. We are not guaranteed there is a minimum. When we solve for $f'(x) = 0$ we find

$$0 = f'(x) = a - \frac{b}{x^2} \qquad \text{hence} \qquad x = \sqrt{\frac{b}{a}}\,.$$

At this value, $f(x) = a\sqrt{\frac{b}{a}} + b/\sqrt{\frac{b}{a}} = 2\sqrt{ab}$. Checking what happens near 0 and ∞, we find $\lim_{x\to 0} f(x) = \infty$ and $\lim_{x\to\infty} f(x) = \infty$. Therefore, there is a minimum value, which we have determined to be $\sqrt{ab}$ occuring at $x = \sqrt{\frac{a}{b}}$.

You might have noticed there are two free variables in this example, the unspecified constants a and b. It's worth observing that everything interesting in the problem depends only on the ratio b/a. One might check whether this makes sense from the units. The units of a are in dollars per distance. The units of b are dollars per inverse distance, so dollars times distance. Dividing and simplifying, we see that b/a has units of distance, which corroborates that $x = \sqrt{b/a}$ is a reasonable solution for the location of the minimum, since this really is a "location" as measure in distance to the nearest port.

Example 7.11. The functions $x^\gamma e^{-x}$, for $x \geq 0$, arise in probability modeling. They are called Gamma densities. We will return to these in Section 12.3. For now, we would like to understand the shape of these functions. An example with $m = 5$ is shown in Figure 31. The place where one is mostly to find the random variable is where the maximum of the density occurs. Where does the maximum of $f(x) := x^5 e^{-x}$ occur? We know that the value is zero at $x = 0$ and positive everywhere else. We also know $\lim_{x\to\infty} f(x) = 0$. This means there must be a maximum at some positive finite x. The function f is differentiable for all positive x, therefore the maximum can only occur where $f' = 0$. Solving

$$0 = f'(x) = 5x^4 e^{-x} - x^5 e^{-x}\,.$$

Factoring out $x^4 e^{-x}$, we see that $x = 5$. Therefore, the maximum occurs at $x = 5$.

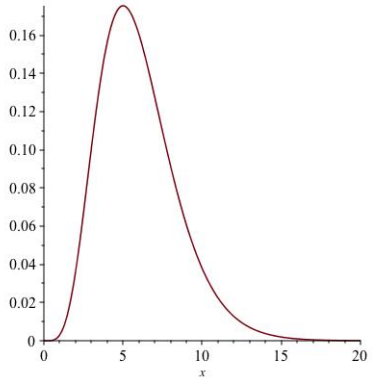


Figure 31: Gamma-5 density

Exercise 7.10. *Why does a limit of zero at infinity imply that f must have a maximum at some positive, finite x? Any convincing argument is fine here.*

Example 7.12. Let h be the height of a member of a carnivore species. In this simple model, the food gathering capability of an individual is given by kh^2 while its daily food needs are given by ch^3.

(a) Why?

(b) What are the units of c and k?

(c) To maximize food gathering ability minus food needs, how tall should members of this species be?

(a) We can only make educated guesses about the reason the equations in the model have this form. If an animal's speed is proportional to its height then the model stipulates territory is proportional to the square of this. Perhaps territory is the area that can be reached in a given amount of time such as an hour or a day. As to why food needs would be proportional to volume, one might imagine that sustaining and nourishing tissue requires nutrients proportional to volume.

(b) Units of c are food per length^3 and units of k are food per length^2. For example, if food is measured in kilograms and length in meters, then food per length^3 would be kg/m^3; however one might measure food in other ways such as calories, or numbers of a particular animal of prey, etc.

(c) The objective function we want to maximize is $kh^2 - ch^3$. Having been told no limitations on size, we assume h can be any positive real number, though we may have to retract that if the optimum turns out to have unrealistic scale. Differentiating $f(h) := kh^2 - ch^3$ with respect to h yields $2kh - 3ch^2$ and setting equal to zero gives the two solutions 0 and $x_* := (2k)/(3c)$. This indeed has units of length. Clearly $f(0) = 0$. The value of the objective function at x_* is $4k^3/(27c^2)$, which is positive. Therefore the maximum of f on $[0, \infty)$ is either $4k^3/(27c^2)$ achieved at $h = (2k)/(3c)$ or there is no maximum because the function can get arbitrarily large as $h \to \infty$. At infinity, $f(h) \sim -ch^3$ because $kh^2 \ll ch^3$ as $h \to \infty$. Therefore, h has a maximum at a positive location, whose value is $4k^3/(27c^2)$

Exercise 7.11 (optional, it's a bit of a computation, though not hard)**.** *Continuing the previous example, suppose that for lions $k = 0.001$ gazelles per square meter, and $c = 0.0004$ gazelles per cubic meter. What length of lion maximizes its excess food gathering ability, and how many gazelle carcasses per day will be left over for the other lions in the pride?*

7.3 Applications

There's no new material in this section, just some typical applications of optimization using differential calculus.

A geometric optimization problem

Example 7.13. We're going to build a window in the shape of a rectangle topped by an equilateral triangle. We want to make a window which lets in the most light – that is, with the greatest possible area. In order to build the window, we have to use wood trim. We have 16 feet of wood trim to build the window with.

Such a window has two dimensions: the width and the height of the rectangle. The rectangular portion has area and the triangular portion has area. So the total area is given by

$$A(w, h) = wh + \frac{\sqrt{3}}{4}w^2 \,.$$

We also need to take into account the fact that our supplies are limited. Two pieces of trim with length h and four of length w add up to $2h + 4w$ which we can set equal to 16 because

if they add up to less we would increase h to takeup the slack, obviously giving us more light. Thus $h = 8 - 2w$ and we can plug in to get

$$A(w) = w(8 - 2w) + \frac{\sqrt{3}}{4}w^2 \,.$$

Clearly w can't be less than zero or greater than 4, so we are left with the calculus problem of maximizing $A(w)$ over $w \in [0, 4]$. There's only one critical point, when $A'(w) = 8 - 4\,w + 1/2\,w\sqrt{3} = 0$, whose solution is $w_* = \dfrac{16}{8 - \sqrt{3}} \approx 2.55$ feet. We are also interested in the value of the maximal area which is $A(w_*) = 64/(8 - \sqrt{3}) \approx 10.21$ square feet.

Optimization in business

Consider a company whose main business is producing and selling sneakers. In real life it's very complicated, taking into account things such as labor costs, transportation, import/export taxes, management costs, durable equipment versus expendable supplies, and so forth. But one can get a handle on basic decision making with a simplified model, taking into account only a few variables, as follows.

Let p be the selling price of a pair of sneakers. This may seem like an odd choice for the lone independent variable in such a model until one realizes that the retail price is the one thing the company completely controls. According to economic theory, the demand $N(p)$ for the sneakers will be a function of the price; this is pretty credible. The equation $P(p) = N(p)(p - U(p))$ represents the fact that the profit P is found by multiplying the number of pairs of sneakers sold times the difference between the price p and the production cost $U(p)$ per pair. One might also write this as revenue minus cost, where revenue is your gross sales $pN(p)$ and the production cost $U(p)N(p)$ is the unit cost times the number of units.

The big simplification in this model involves the supposition that $N(p)$ and $U(p)$ are knowable and furthermore have simple formulas. In fact, a lot might be inferred about $N(p)$ might be available from marketing data and known demographics. In the region where the maximum of P occurs, N may indeed be approximated by a simple formula. In the case of the unit production cost, $U(p)$ might be difficult to know because of the huge basket of things it includes taxes, management costs, excess inventory, etc., and while it is a function of p (because it is a function of N and N is a function of p) it is unlikely to satisfy a nice formula or be mathematically tractable.

Example 7.14. Suppose that $U(p)$ is constant: no matter how many sneakers you make, the marginal cost of producing one more is the same amount, say c dollars. Suppose that $N(p)$ obeys some power law, $N(p) = bp^{-\alpha}$. Thus,

$$P(p) = bp^{-\alpha}(p - c)\,.$$

Can we determine the best price to set? Looking for critical points we find

$$P'(p) = bp^{-\alpha} - \alpha bp^{-\alpha-1}(p - c)\,.$$

Setting this equal to zero we factor out $bp^{-\alpha-1}$ and find that $0 = p - \alpha(p - c)$, and solving for p gives $p = c\dfrac{\alpha}{\alpha - 1}$.

This is a good chance to practice asking questions. Before you read on, please stop and think about what questions you should be asking. When fractional exponents are involved, units are often nonsensical, so let's not go there. What about the signs: is α positive or negative, and does the formula make sense? It seems the way we set things up, α should be positive so that the demand can decrease to zero, not increase to infinity, as the price rises. Something must be messed up when $\alpha = 1$, but what and why? In fact, something is messed up when $\alpha \leq 1$: the critical point is a minimum rather than a maximum. In fact when $\alpha < 1$, say $1/2$ for example, the model is nonsensical. You can price the sneakers at a trillion dollars, sell only $1/1,000,000$ of a pair, and make a million dollars. The nonsense is that there's no good interpretation of selling a small fraction of one pair of sneakers. The same issue arises in principle when $\alpha > 1$, say $\alpha = 2$, only it doesn't matter, because if you sell a trillionth of a pair for a million dollars per pair, almost no money (or sneakers) changes hands. It's OK to model $N(p)$ as a continuous variable when small values of N correspond to irrelevant parts of the scenerio but not when they the small values of N coresond to ridiculously huge transactions.

Next question: say $\alpha > 1$; do things make sense now? The best price point is the cost, c, multiplied by $\alpha/(\alpha - 1)$. It's a good sign that $\alpha/(\alpha - 1) > 1$; it means you are setting the price above cost. Notice that as $\alpha \to \infty$, $\alpha/(\alpha-1)$ goes to 1 from above. You might interpret that as saying that when consumers are very cost-sensitive (large α), then you shouldn't set the price much above your actual cost. What about the constant of proportionality b? It doesn't appear at all in the formula for the best price point. The profit you make at this price point will be proportional to b but the price point doesn't change with b. Whether this makes intuitive sense is up to you. It kind of does to me. This is an oversimplied model, to be sure, but seems to be getting at some real phenomena.

8 Further topics in differential calculus

Calculus has been around for 300 years. The applications and techniques don't all fit nicely into chapter length categories. Here, we tie up some loose ends and mention a few things we think you shouldn't miss.

8.1 Differentiating inverse functions

This section pays back a debt by addressing those functions in Proposition 5.8 whose derivations we have not yet discussed: powers, exponentials, logarithms and inverse trig functions. To clarify our terminology, the reason x^a is called a power, while x^a is called an exponential, is that we are differentiating with respect to x, while a plays the role of a constant.

For positive integer powers x^n there are many ways of computing the derivative. One is by expanding it out:

$$(x+h)^n - x^n = nhx^{n-1} + \binom{n}{2} h^2 x^{n-2} + \cdots + nh^{n-1}x + h^n \,.$$

Dividing by h and taking the limit as $h \to 0$ shows that the derivative of x^n is nx^{n-1}. Another way is to prove it by induction, using the product rule to get from $(d/dx)x^n = nx^{n-1}$ to $(d/dx)x^{n+1} = (n+1)x^n$.

For negative integer powers you can use the quotient rule, writing $x^{-n} = 1/x^n$ and using the known derivative for positive integer values of n. For rational powers, it is easiest after proving a combining rule that tells us how to compute the derivative of the inverse function f^{-1} if we know the derivative of f. The derivation is a quick use of the chain rule.

Proposition 8.1.

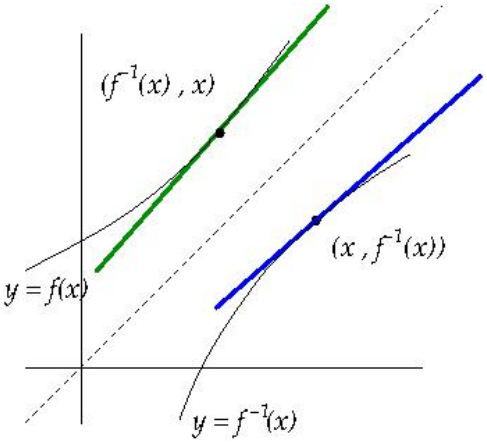
$$\frac{d}{dx} f^{-1}(x) = \frac{1}{f'(f^{-1}(x))} \,. \tag{8.1}$$

Usual Proof: By definition $f(f^{-1}(x)) = x$. Takingthe derivative of both sides,

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

and dividing both sides by $f'(f^{-1}(x))$ yields the result.

One of the instructors called this proof "efficient but unenlightning." In case you feel the same way, here is a pictorial proof.



Graphs of f and f^{-1} (in black) are reflections of each other across the diagonal line $y = x$ (dashed). The tangent to $y = f^{-1}(x)$ at $(x, f^{-1}(x))$ (blue) is the reflection of the tangent to the line $y = f(x)$ at $(f^{-1}(x), x)$ (green). The green line has slope $f'(f^{-1}(x))$, therefore its reflection, the blue line, has slope reciprocal to this, namely $1/f'(f^{-1}(x))$.

Exercise 8.1. *Suppose f has input units of people and output units of money. Do a unit analysis of equation (8.1): what are the units of each side, and are they the same?*

Example. *Square root is the inverse function to squaring. Using Proposition 8.1 quickly computes the derivitive of the square root. Letting $f(x) = x^2$ in Proposition 8.1, and using $f'(x) = 2x$, the conclusion becomes*

$$\frac{d}{dx}\sqrt{x} = \frac{1}{2 \cdot \sqrt{x}}\,.$$

Exercise 8.2. *Use a similar method to compute $\dfrac{d}{dx}\sqrt[3]{x}$. Show your work.*

Similarly, this allows us to show $(d/dx)x^{1/n} = (1/n)x^{1/n-1}$. Using the chain rule, because $x^{k/n} = (x^{1/n})^k$, we can then compute $(d/dx)x^{k/n}$ for any nonzero integers k and n. So now we have verified that the derivative of x^r is rx^{r-1} for all rational numbers r.

At the end of the section we will finish this argument by handling the case of exponents that are not rational numbers.

Inverse trig functions

We've already computed the derivatives of the basic trig functions (parts 6, 7 and 8). What remains are the inverse trig functions. Use the inverse function rule, obviously! For example,

if $f(x) := \sin x$ then the derivative of arcsin is computed by

$$\frac{d}{dx} \arcsin x = \frac{1}{\cos(\arcsin x)} .$$

Some of you may recognize the identity $\cos(\arcsin y) = \sqrt{1-y^2}$. In case not, it's an easy piece of geometry. For any $y \in [-1, 1]$, $\arcsin y$ is a value between $-\pi/2$ and $\pi/2$, denoted by θ in Figure 32. In the figure, the measure of BC is $|y|$ and the measure of AC is $\cos\theta = \cos(\arcsin y)$, and the Pythagorean theorem shows what we want, namely $\cos \arcsin y = \sqrt{1-y^2}$.

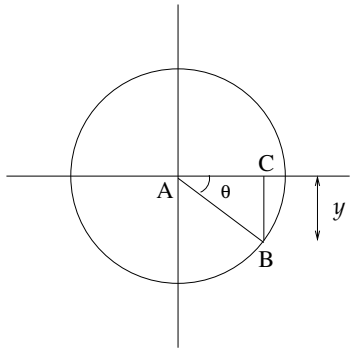


Figure 32

8.2 Related rates

Sometimes two quantities vary with time and one is a function of the other. In this case, the rate of change of one quantity determines the rate of change of the other. In old style textbooks, this was a major topic even though there isn't all that much to say. We think it is more proportionate to illustrate with one example, give you a few practice problems and call it a day.

Example 8.2. Suppose the volume of a balloon increases as a function of time. The radius, being a function of the volume, will therefore increase at a different rate. Writing $R = f(V)$ and $V = g(t)$, we have $R = f(g(t))$. Therefore by the chain rule,

$$\frac{dR}{dt} = \frac{dR}{dV}\frac{dV}{dt} .$$

This notation hides where each derivative is evaluated but the meaning is clear. Letting primes denote time derivatives, $R' = V' \cdot dR/dV$.

The rate of increase of radius and the rate of increase of volume are therefore called **related rates**. Knowing one always gives you the other, provided you know the present volume and can compute dR/dV. For a spherical balloon, $V = (4\pi/3)R^3$, therefore $R = \sqrt[3]{3V/(4\pi)} = (3/(4\pi))^{1/3}V^{1/3}$ and we can compute $dR/dV = (1/3) \cdot (3/(4\pi))^{1/3}V^{-2/3}$. In other words, if the present volume is V, then the rate the radius is growing in, say, cm/sec, is equal to $\sqrt[3]{3/(4\pi)}/3$ times the rate the volume is growing in cm^3/sec divided by the two thirds power of the volume.

Exercise 8.3. *A conical tank (picture included so you know that conical tanks aren't just a fiction of calculus) has radius $0.8h$ at height h from the bottom.*

(i) What is the volume of the interior of the tank up to height h? Write this as $V = f(h)$ for some function f. You can find this in Wikipedia if you don't know.

(ii) Write an equation relating dV/dt to dh/dt.

(iii) If the tank is emptying at a rate of 2 liters per minute (a liter is 1000 cubic centimters), and the tank is currently filled to a height of h, how quickly is the height decreasing?

(iv) The units of h were not given. Did you choose units? Does this affect the answer?

8.3 Exponentials revisited

Recall that e was defined to be the positive real number such that e^x has slope 1 at $x = 0$. In other words, by definition,

$$\lim_{h\to 0} \frac{e^h - 1}{h} = 1\,.$$

From this we can compute the derivative of e^x at any point. Let $f(x) := e^x$. Then

$$f'(x) = \lim_{h\to 0} \frac{e^{x+h} - e^x}{h} = \lim_{h\to 0} e^x \frac{e^h - 1}{h} = e^x \lim_{h\to 0} \frac{e^h - 1}{h} = e^x\,.$$

Next, for some $a > 0$, let $f(x) := a^x = (e^{\ln a})^x = e^{x \ln a}$. The chain gives

$$f'(x) = e^{x \ln a} \frac{d}{dx}(x \ln a) = a^x \ln a \,.$$

At this point, we have derived parts 1, 3 and 4 of Proposition 5.8. For Part 5, letting $f(x) := e^x$ so $f^{-1}(x) = \ln x$, we use the inverse function rule Proposition 8.1 and $(d/dx)e^x = e^x$ to obtain

$$\frac{d}{dx} \ln x = \frac{1}{e^{\ln x}} = \frac{1}{x} \,.$$

Paying back a debt, this finishes off Part 2 of Proposition 5.8. For any real r and positive x, let $f(x) := x^r = e^{r \ln x}$ and use the chain rule to obtain

$$f'(x) = e^{r \ln x} \frac{d}{dx}(r \ln x) = x^r \frac{r}{x} = r x^{r-1} \,.$$

Differential equations

The course after this one studies differential equations. This semester we get only a tiny preview of this subject. A differential equation arises when you have a function that is unknown and your information about it includes something about the derivative. The simplest example is when you know the derivative outright, for example $f'(t) = 5 + 4t$. Integral calculus then produces a formula for f. In this case, because you are familiar with derivatives of polynonmials, you can probably recognize the solution $f(t) = 5t + 2t^2$. There are other possible solutions, all differing by a constant, for example $f(t) = 1 + 5t + 2t^2$. The general solution is $f(t) = c + 5t + 2t^2$ where c can be any constant. Further information is required to figure out c; if you know even a single value of f, such as $f(7) = -2$, you can solve for c.

The differential equation we will study here is the next simplest one: $f'(t) = kf(t)$. This is more subtle because the derivative is not given outright but rather is related to the function itself (of course f represents the same function on both sides of the equation). This method of solution of this equation is similar to the previous example. You can solve it because you can recall a function that behaves this way, namely the function $f(t) := e^{kt}$. That is the simplest looking solution but there are others. The most general solution is $f(t) = A\, e^{kt}$ where A can be any constant. When you study methodical solutions to differential equations you will be able to prove that these are the only solutions. In the present course, we won't discuss the problem at that level but you are free to assume this is true: if $f'(t) = k\, f(t)$

for all t then $f(t) = A\,e^{kt}$ for some real number A. Note that the constant k is not like the constant A: the constant k is part of the equation you were given altering it will make the function no longer a solution to the equation.

Because $f' = kf$ is such a basic equation, it occurs pretty commonly in applications. For this reason it pays to study the functions Ae^{kt} in detail. When $k > 0$ this represents exponential growth. Some things that behave this way under the right circumstances are populations, money (both assests and debt), epidemics, adoption of new technology, and pyramid schemes. In all of these cases, it's easy to argue that the rate of increase should, to a first approximation, be proportional to the present size; in other words, $f' = kf$.

Aside. *One such argument goes like this: dividing the money (or population, or infection, etc.) into small units, each unit produces the same net growth independently of the others and independently of how much time has passed. Therefore, the growth should be proportional to the number of units presently existing. Whether this constant of proportionality k should be independent of time versus being a function $k(t)$ is not clear from this argument and would need to be addressed separately in any justification of the model.*

When $k = -\ell < 0$, this is called exponential decay. The classic example of exponential decay is a radioactive material breaking down through alpha or beta decay. Other things that decay exponentially under the right circumstances are temperature difference, correlations in time series data and valuations of future goods. These examples were mentioned briefly in Section 2.3. Calculus gives a a reason to believe why exponential growth and decay are plausible models for these physical phenomena. It is because the underlying mechanisms force f' to be proportional to f.

Exercise 8.4. *Suppose the underlying mechanisms force f' to be proportional to $f - c$ rather than f. Write down a guess as to what would the differential equation look like.*

Time constants

Suppose $f(t) := Ae^{kt}$ where t is in units of time and f is a quantity in some units we will just refer to as "units of f". Recall from the introduction to units early in the course that the exponent kt is required to be unitless if the expression is to make physical sense. That means the constant k has to have units of inverse time. Such constants are called **time constants**.

At first these can be difficult to make physical sense of. We understand the quantity 0.02 days, but what is the physical significance of the quantity 0.02 inverse days? Most directly it means that if t is the reciprocal, namely 50 days, then $kt = 1$ (unitless) and the quantity Ae^{kt} is $A \cdot e$, a factor of e greater than it was at the start (because at the start, $Ae^{k\cdot 0} = A$.

Exercise 8.5. *In March, 2020, the U.S. COVID-19 epidemic was increasing exponentially with a time constant of 1.4 inverse weeks. By roughly what factor did the number of total cases increase each week in March?*

If k is negative, then f represents exponential decay. For example if $k = -0.02$ inverse days, then after 50 days, the function will have decreased by a factor of e.

Which is bigger, an inverse second or an inverse minute. Minutes of course are much longer than second: one minute equals 60 seconds. On paper one can convert between inverse time units as well. For example,

$$1\text{sec}^{-1} = \frac{1}{\text{sec}} \cdot \frac{60\text{sec}}{1\text{minute}} = \frac{60}{\text{minute}} = 60\text{min}^{-1}$$

so one inverse second is 60 inverse minutes. To make this a little more intuitive, think of one inverse second as 1/sec which we might write say aloud as "one per second". The phrases "one per second" and "sixty per minute" should sound believably the same.

Consider a quantity that is decaying exponentially. As a function of time, the quantity is represented as a function $f(t) := Ae^{-kt}$. Such a quantity is said to have a **half-life**. Regardless of how much of the quantity there is originally, the time until half remains is always the same. It's too bad the concept wasn't first conceived as eth-life, the time it takes to reduce by a factor of e, because that is clearly the time for kt to become -1, in other words the reciprocal of k (it's a good thing that k has inverse time units so its reciprocal is a time). No matter, if instead of $kt = -1$ we say $kt = -\ln 2 \approx 0.7$, then e^{-kt} will be $1/2$. So the half-life is just $(\ln 2)$ times the eth-life, that is, $(\ln 2)/k$.

Exercise 8.6. *Polonium-210 is a radioactive substance and decays to lead with a half-life of about 138 days. What is the present rate of decay of a sample of 5 micrograms of Polonimum-210? Please give units.*

8.4 Tangent line estimates and bounds using calculus

Let's sum up what we already know about the tangent line approximation, this time in the language of calculus. If f is a function differentiable in an open interval I containing a, then the tangent approximation to $f(x)$ at a is the function

$$L(x) := f(a) + (x - a)f'(a)\,.$$

Exercise 8.7. *Compute the tangent line approximation for* $f(x) := \sqrt[3]{1+x}$ *near* $x = 0$. *What quick estimate does this give of* $\sqrt[3]{1.06}$*? Please check this against a numerical computation on your computer and say how close the quick estimate was.*

If f is twice differentiable in I and $f'' \geq 0$ on I then $L(x) \leq f(x)$ for all $x \in I$, that is, the tangent line approximation is a lower bound for the actual value. Reversing the inequality to $f'' \leq 0$ reverses the conclusion to $L(x) \geq f(x)$. Making the inequality strict makes the conclusion strict, except at a where f and L always agree; see Figure 33.

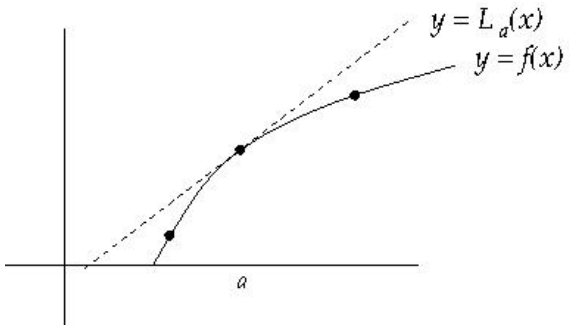


Figure 33: $f'' < 0$ on the interval shown, hence for any a, $L_a(x) \geq f(x)$ with equality only at $x = a$.

Exercise 8.8. *Compute the tangent line approximation to* $\sin(\pi/5)$ *at any nearby point* a *where you know the value of* $\sin(a)$. *Write the result as an algebraic expression involving* π *and say whether this is an upper bound, lower bound or neither.*

We have said before that $f(x) \approx L(x)$ when x is near a. How close are these two? At the end of the course you will see that L is just the first in a series of estimates $P_1, P_2, \cdots$ that approximate f better and better. These are the Taylor polynomials, the first being linear, the second quadratic, and so on, the n^{th} one having degree n. Each one is the best approximation for a polynomial of its degree. How good an approximation are they? The

degree n Taylor polynomial differs from f at x by a term on the order[10] of $(x-a)^{n+1}$. Because the tangent line approximation $L = P_1$ is the first, it differs from f by on the order of $(x-a)^2$, meaning possibly $2(x-a)^2$ or $10(x-a)^2$ but not anything $\gg (x-a)^2$ as $x \to a$.

When talking about orders of magnitude of functions near a, recall that $(x-a)^2 \ll |x-a|$, in other words the difference between f and L at any x is much less than the distance that x is from a. The above facts about Taylor polynomials are a preview. We won't discuss them more now, but instead will focus only on P_1, which is also denoted L. This proposition is weaker than what we just told you about how close the tangent line approximation is, but has the virtue of being easy to prove.

Proposition 8.3. *The tangent line approximation is better than linear, meaning that*

$$|L(x) - f(x)| \ll |x-a| \qquad \text{as} \qquad x \to a\,.$$

You can see this algebraically. By definition of $\ll$, we need to check that

$$\lim_{x\to a} \left| \frac{L(x)-f(x)}{x-a} \right| = 0\,.$$

This follows from

$$\lim_{x\to a} \frac{L(x)-f(x)}{x-a} = 0$$

by composition with the absolute value function, which is continuous. We evaluate this:

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

Exercise 8.9. *Using a calculator, compute the difference between the cube root of* 1.06 *and your tangent line estimate in Exercise 8.7. Does this corroborate Proposition 8.3? Does it corroborate the assertion that $|P_1 - f|$ should be on the scale of $|x-a|^2$? In each case say why or why not.*

[10] We have not formally introduced the phrase "on the order of" but what we mean here is that the first quantity should not be much more than the second: it should not be true that $x - a \ll |P_n(x) - f(x)|$.

The mean value theorem

In class we will discuss the following theorem. Please read it now to see whether it makes intuitive sense to you. The hypotheses will be filled in after the class discussion centered on the counterexamples in Figure 34.

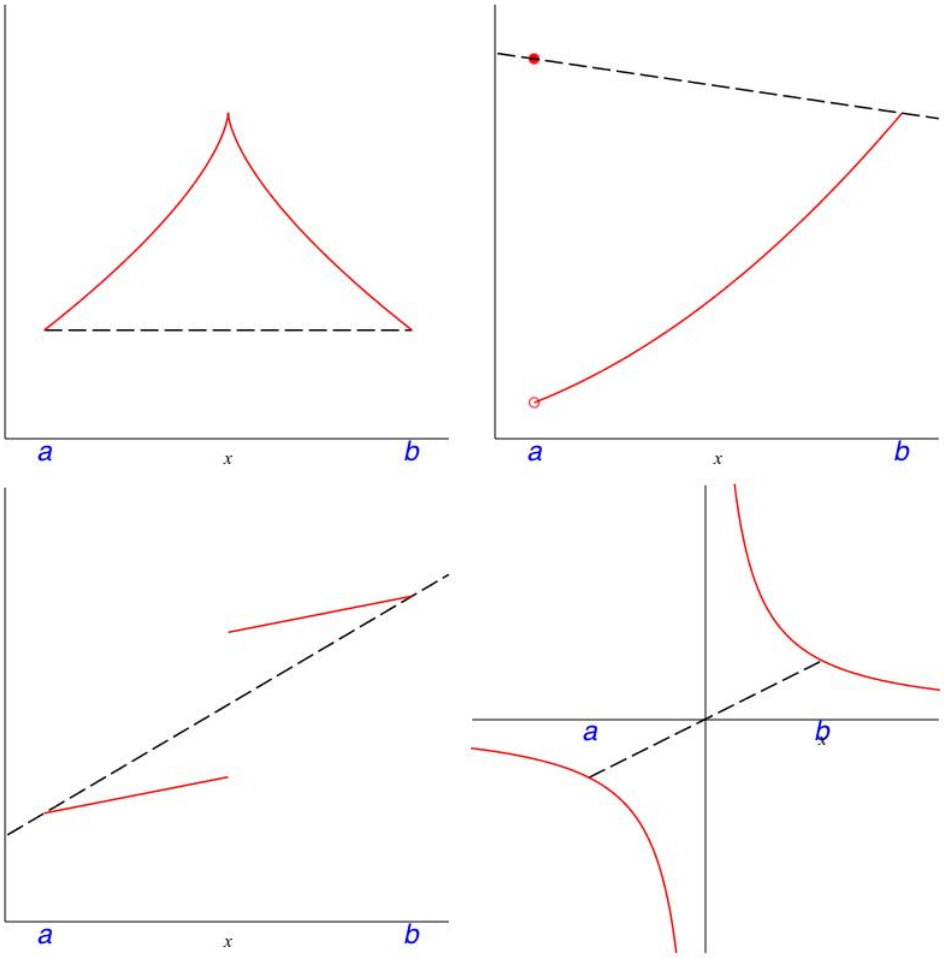


Figure 34: In each case the dashed line illustrates the average slope $\dfrac{f(b)-f(a)}{b-a}$

Theorem 8.4 (Mean value theorem)**.** *Let f be a function and $a < b$ be real numbers. Assuming some hypotheses* _____*, there must be a number $c \in (a, b)$ where the slope of f is equal to the average slope over (a, b), that is,*

$$f'(c) = \frac{f(b) - f(a)}{b - a}\,. \tag{8.2}$$

Example 8.5. Let $f(x)$ be the position (mile marker) of a PA Turnpike driver at time x. Suppose the driver entered the Turnpike at Mile 75 (New Stanton) at 4pm and exited at Mile 328 (Valley Forge) at 7pm. What does the Mean Value Theorem tell you in this case? The average slope of f over interval [4pm,7pm] is the difference quotient $(f(7) - f(4))/(7 - 4) = (325 - 75)/3 = 84\frac{1}{3}$. Thus there is some time c between 4pm and 7pm that $f'(c) = 84\frac{1}{3}$ MPH, in other words, that this driver was traveling at a speed of $84\frac{1}{3}$ MPH. Bonus question: can the Mean Value Theorem be used in court by Law Enforcement? It has been ruled in some states that this is legal evidence of the car having violated a speed limit, but not that the particular driver has done so.

Exercise 8.10. *Let $f(x) := 1/x$ and let $a < b$ be positive real numbers. What, explicitly in terms of a and b, is the number c guaranteed by the Mean value theorem?*

9 Summation

9.1 Sequences

On page 51 we briefly discussed sequences. When working with sums and the "Sigma" notation for summations, you need to be able to write formulas for sequences you understand intuitively. For example, if you want to write the sequence $7, 9, 11, 13, \ldots$ in the notation $\{b_n : n \geq 1\}$, so that $b_1 = 7$, $b_2 = 9$ and so on, one choice would be to say,

> "Let $\{b_n : n \geq 1\}$ be the sequence defined by $b_n := 5 + 2n$."

The subscript n is called the **index**[11] (plural: **indices**). Indexing can begin at any natural number. In this case, as is most common, we began at $n = 1$. Defining $\{b_n : n \geq 3\}$ by $b_n := 1 + 2n$ yields the same sequence: $7, 9, 11, 13, \ldots$. Secondly, the informal notation $7, 9, 11, 13, \ldots$ is not mathematically precise because it assumes we all agree exactly what the pattern is. Producing a formula for the n^{th} term removes any ambiguity. A formula is often necessary if you want to sum the sequence or to use it to define other sequences. This section considers some common types of sequences and gives you some practice writing a formula for the general term.

Exercise 9.1. *Write a formula for the general term of the "place value" sequence* $1, 10, 100, 1000, 10000, \ldots$. *You can choose any letter for the indexing variable (we chose* n *above), the sequence name (we chose* b *above) and the starting index (we chose 1 at first, then 3 for constrast). Whatever you choose, write the definition in a full sentence, similar to the quoted sentence above.*

Definition 9.1. *A sequence is called* **arithmetic** *(adjective, accent on the third syllable, to rhyme with "alpahbetic") if the difference between successive terms is constant.*

Our example sequence $7, 9, 11, 13, \ldots$ is an arithmetic sequence with common difference 2. It is particularly easy to write a formula for the general term of an arithmetic sequence if you start indexing at zero. The n^{th} term is the zeroth term plus n copies of the common

[11] We don't absolutely need new notation. A sequence could be thought of as a function $n \mapsto b_n$ from the natural numbers to the real numbers, but the notation is useful because it sets us up to imagine that we will be looking at the numbers $b_1, b_2, \ldots$ rather than the relationship between n and b_n. In fact we define sequences to be the same if the numbers are the same, even if the indices are different.

difference. In notation, if the common difference is d and the sequence is $\{a_k : k \geq 0\}$, this means $a_k = a_0 + kd$. Setting $a_0 = 7$ and $d = 2$ gives $a_k = 7 + 2k$ for the sequence $7, 9, 11, 13, \ldots$.

Exercise 9.2. *Which of these sequences appear to be arithmetic sequences?*

(i) $9, -11, 13, -15, \ldots$

(ii) $\sin(1), \sin(3), \sin(5), \sin(7), \ldots$

(iii) $30, 27, 24, 21, \ldots$

(iv) the sequence defined for $n \geq 0$ *by* $b_n := 1/(5 + 2n)$

(v) the sequence defined for $n \geq 0$ *by* $b_n := 14 - n/2$

Definition 9.2. *A sequence is called* **geometric** *if the ratio between successive terms is constant. In other words, if the sequence is* $\{u_j\}$*, then the ratio* u_{j+1}/u_j *has some common value* r *for all* j*.*

For example, the sequence $10, 20, 40, 80, 160, \ldots$ is geometric with common ratio 2.

Exercise 9.3. *Write a formula for the general term of this geometric sequence.*

Sequences with alternating signs appear often enough that it's a good idea to know a way to write their general term. The key to being able to write such sequences is to notice that $(-1)^n$ bounces back and forth between $+1$ and -1. The odd terms are negative, so starting with $n = 1$ (or 3 or 5, etc.) starts with -1 whereas starting with 0 (or 2, -2, etc.) starts with $+1$. You can incorporate this in a sum as a multiplicative factor and it will change the sign of every second term. Thus for example, to write the sequence $1, -2, 3, -4 \ldots$ you can write $(-1)^{n+1} \cdot n$. Note that we used $(-1)^{n+1}$ rather than $(-1)^n$ so that the term corresponding to $n = 1$ was positive rather than negative.

When the sum has a pattern that takes a couple of steps to repeat, the greatest integer function can be useful. For example, $1, 1, 1, 2, 2, 2, 3, 3, 3, \ldots$ can be written as $a_n := \left\lfloor \dfrac{n+2}{3} \right\rfloor$ for $n \geq 1$. Actually, it comes out a little more simply if you index starting from zero: $a_n := \left\lfloor \dfrac{n}{3} \right\rfloor + 1$ for $n \geq 0$.

Definitions by cases work for sequences just the way they do for functions. Suppose you want to define a sequence with an opposite sign on every third term, such as $-1, -1, 1, -1 -$

$1, 1, \ldots$. You can do this by cases as follows.

$$a_n := \begin{cases} -1 & n \text{ is not a multiple of 3} \\ 1 & n \text{ is a multiple of 3} \end{cases}.$$

Plenty of sequences don't fit any of these molds. Writing a formula for the general term is a matter of trying an expression, seeing if it works, then if not, tinkering to get it right.

9.2 Finite series

Let's talk for a minute about a notation you have likely seen before. It is called the "Sigma" notation because Σ is a capital Greek Sigma. The notation involves an indexing variable which runs between a lower limit and an upper limit. The lower and upper limits are required to be integers[12]. If the indexing variable is n, the lower limit is L, the upper limit is U and the general term is b_n, the summation looks like $\sum_{n=L}^{U} b_n$. What this means is to add together all the values of b_n starting with $n = L$ and ending with $n = U$.

Example 9.3. $\sum_{n=1}^{5} 2^{-n}$ represents the sum $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32}$.

The summand, as you can see is usually a function of the indexing variable; otherwise, the summand would not change from term to term. There may be other variables, for example $\sum_{k=3}^{6} kx$ evaluates to $3x + 4x + 5x + 6x$, which is equal to $18x$. Note that this other variable x persists when the sum is evaluated. It is a free variable. On the other hand, the index of summation, k in this case, is a bound variable. It runs over a set of values (in this case 3 to 6) and does not appear in the final value.

Exercise 9.4. *In the sum* $\sum_{k=1}^{n} \frac{C}{k}$*, which of the variables* k, n *and* C *are free and which are bound?*

When a sequence is summed, it is called a **series** (the plural is also "series").

[12] This is unlike computer science, where a loop counter can increment by any number

Example 9.4. The sum $\sum_{n=5}^{19} \frac{3}{n-2}$ represents a series with 15 terms because there are 15 integers in the range from 5 to 19. Informally, we might write this sum by writing the first few terms and the last term, with dots in between (traditionally the dots are centered for series, as opposed to at the bottom of the line for sequences). Thus we would write $\frac{3}{3} + \frac{3}{4} + \cdots + \frac{3}{17}$, assuming this conveyed enough information for the reader to understand the precise sum. Of course there is no reason why the index should go from 5 to 19. There have to be fifteen terms, but why not write the sum with the index going from 1 to 15? Then it would look like

$$\sum_{n=1}^{15} \frac{3}{n+2}\,.$$

Another natural choice is to let the index run from 0 to 14:

$$\sum_{n=0}^{14} \frac{3}{n+3}\,.$$

All three of these formulas represent the exact same sum.

Exercise 9.5. *Write a summation that sums the integers from 1 to 100 for which the lower limit is -5.*

9.3 Some series you can explicitly sum

The series in Example 9.4 sums to a rational number. According to Excel it is equal to 23763863/4084080. There isn't any really nice formula for this sum in terms of the values 5 and 19 and the function $n \mapsto 3/(n-2)$. In fact most series don't have nice summation formulas. Arithmetic and geometric series are exceptions. Because they are common and the formulas are simple and useful, we include them in this course.

Arithmetic series

Here's an example of how to sum an arithmetic series, which generalizes easily to summing any arithmetic series. This particular example is a well known piece of mathematical folklore (google "Gauss child sum").

Example 9.5. Problem: Sum the numbers from 1 to 100. Solution: Pair the numbers starting from both ends: 1 pairs with 100, 2 pairs with 99, and so forth, ending at 50 paired with 51. There are 50 pairs each summing to 101, so the sum is $50 \times 101 = 5050$.

Another way to get the same formula is only slightly different.

Example 9.6. Evaluate $\sum_{n=13}^{29} n$. There are 17 terms and the average is 21, which can be computed by averaging the first and last terms: $(13 + 29)/2 = 21$. Therefore, the sum is equal to $17 \times 21 = 357$.

Exercise 9.6. *Suppose we want to sum the arithmetic series $\sum_{k=L}^{U} a+kd$. We have already seen that every arithmetic series can be written this way, so this exercise solves the problem of summing every arithmetic series (yet is easy enough to put in an exercise!).*

(i) How many terms are there in this series?

(ii) Pairing from both ends, what is the common sum of each pair?

(iii) If the number of terms is even, what is the formula for the sum?

(iv) If the number of terms is odd, what is the formula?

Advice: this is more general than our usual exercise. You might find it easier to try a few examples with numbers before doing the exercise with algebraic expressions.

Geometric series

The standard trick for summing geometric series is to notice that the sum and r times the sum are very similar. It is easiest to explain with an example.

Example 9.7. Evaluate $\sum_{n=1}^{10} 7 \cdot 4^{n-1}$.

To do this we let S denote the value of the sum. We then evaluate $S - 4S$ (because $r = 4$).

I have written this out so you can see the cancellation better.

$$\begin{aligned} S - 4S &= 7 + 28 + 112 + \cdots + 7 \cdot 4^9 \\ &\quad - (28 + 112 + \cdots + 7 \cdot 4^9 + 7 \cdot 4^{10}) . \end{aligned}$$

Thus,

$$(1-4)S = 7 - 7 \cdot 4^{10} .$$

From this we easily get $S = (7 - 7 \cdot 4^{10})/(1-4) = 7(4^{10} - 1)/3 = 2446675$.

Exercise 9.7. *The chance that it takes precisely n rolls of a standard die in order to roll your first 6 is $(5/6)^{n-1}(1/6)$. Sum 10 terms of a geometric series to find the chance that you first see a 6 by the time of your tenth roll.*

General case: Evaluate $\displaystyle\sum_{n=1}^{M} A \cdot r^{n-1}$.

Letting S denote the sum we have $S - rS = A - Ar^M$ and therefore

$$S = A \, \frac{1 - r^M}{1 - r} .$$

When A and r are positive, all the terms are positive, hence the sum is positive as well. When $r < 1$ this is very evident from the formula. When $r > 1$ it is true as well, but easier to see multiplying top and bottom by -1 so as to get $A(r^M - 1)/(r - 1)$. When $r = 1$ this quotient is undefined, however the sum is very easy: M copies of A sum to $A \cdot M$.

9.4 Infinite series

No discussion of series would be satisfied if it didn't answer the question, "Is $0.9999\ldots$ (repeating) actually equal to 1?" As you can probably guess, it is a matter of definition. However, there is a standard definition, and therefore we can in fact supply an answer (see below).

Definition 9.8.

$$\sum_{n=L}^{\infty} b_n := \lim_{M \to \infty} \sum_{n=L}^{M} b_n .$$

This definition might require a bit of unpacking. First of all, the colon-equal is right: the symbol $\sum_{n=L}^{\infty} b_n$ on the left is not already defined, and we are *defining* it to be the value on the right. So what we are saying is that the sum of an infinite series is the limit of a certain sequence, called the **sequence of partial sums**.

Example 9.9. How does this definition apply to the so-called **harmonic series**, $\sum_{n=1}^{\infty} 1/n$? It says that this infinite sum is equal to the limit of the sequence $\{H_M\}$, where H_M is the **harmonic number** $\sum_{n=1}^{M} 1/n$. The harmonic numbers H_M are said to be the **partial sums** of the harmonic series. Interpreting the infinite sum in this way doesn't tell us whether the limit is defined, or if so, what it is, it just tells us that if we can evaluate the limit $\lim_{M\to\infty} H_M$, this is *by definition* the sum of the harmonic series. If the limit is undefined, then the sum of the harmonic series is undefined.

Exercise 9.8. *The* **alternating harmonic series** *is the series* $1 - 1/2 + 1/3 - 1/4 + \cdots$.

1. *Write this as an infinite summation.*
2. *Write the value of this infinite sum as a limit.*
3. *State your guess as to whether this limit is defined; if so, estimate (unscientifically) what it is; if not, say whether or not you think the limit is* ∞ *or* $-\infty$.

Because we know how to sum finite geometric series, we can sum infinite geometric series.

Example 9.10. Problem: evaluate $1 + 1/2 + 1/4 + 1/8 \cdots$. Solution: this is the infinite sequence $\sum_{n=0}^{\infty} (1/2)^n$. The value is the limit of the partial sums $S_M := \sum_{n=0}^{M} (1/2)^n$. Evaluating these finite sums gives

$$S_M = \frac{1 - (1/2)^{M+1}}{1 - 1/2} = 2 - \frac{1}{2^M}\,.$$

The infinite sum is then $\lim_{M\to\infty} 2 - (1/2)^M$ which is clearly equal to 2.

Exercise 9.9. *Write* $0.9999\ldots$ *(repeating) as an infinite geometric series, then evaluate it to see if it is really equal to 1.*

9.5 Financial applications

Consider a mortgage loan (loan for a house) or car loan, at a annual interest rate r. Typically payments on these are made monthly, which we will take to be every $1/12$ of a year instead

of counting days (most car loans in fact assume this). Recall from Exercise 6.8 that the one-month growth factor (the factor by which your debt grows each month) is $e^{r/12}$. That's only if you don't pay off the loan. Actually, these loans are typically configured so you pay a fixed amount every month until the loan is paid off in an integer number of months (usually, in fact, an integer number of years). To agree on some notation, let r be the annual interest rate, P be the **principal**, that is the initial debt, and let M be the monthly payment.

In order to deal successfully with used car sales people, it's helpful to understand how these determine your balance over the successive months. The key relation is to understand what happens from one month to the next. We will discuss this, then leave the rest of the balance sheet computation for in-class discussion and homework. To determine your debt after a month, just take your initial debt P, multiply by the factor $e^{r/12}$ for the growth of the debt over the first month, and subtract the amount you just paid off, namely M. We can write this as $P_1 = e^{r/12}P_0 - M$. It holds equally from any month to the next: $P_{n+1} = e^{r/12}P_n - M$, where P_n is your debt after n months.

How about your retirement account? Say while you're working, you put M dollars every month into an interest bearing account. How much do you have after n months? It's the same formula, with an opposite sign because you're adding to your balance, not subtracting.

Exercise 9.10. *Write a formula for your retirement balance after $n+1$ months, P_{n+1}, in terms of your balance P_n after n months.*

A guaranteed rate annuity works similarly. By the time you retire you have put P dollars into an account. (How did this happen? See Exercise 9.10.) You hand this over to a company who guarantees you a certain APY every year, call it Y. Each year you also withdraw a fixed amount to live on, call it M.

Exercise 9.11. *Write a formula for P_{n+1} in terms of P_n, Y and M.*

The University of Pennsylvania's endowment works something like this. The balance increases by roughly 5% each year due to the growth of the investments and new donations. Meanwhile, during the year, the university spends roughly 3.4% of the present endowment. Unlike the formula for growth of a retirement fund or reduction of debt, this one is only approximate because the actual return varies. Nevertheless, it is useful for forecasting. Let E_n denote the size of the endowment after n years.

Exercise 9.12. *What is the relation of E_n to E_{n+1}? In what way does this formula differ from the other three (loan, retirement account, annuity)?*

10 Integrals

10.1 Area

Integrals compute many things, the most fundamental of these being area. The definition of area is more subtle than one might think. Most people's understanding of area is based on a physical concept of how much two-dimensional space is taken up. For example, if you have to paint an irregular flat shape, how much paint does it take?

Looking back at the treatment of area in the pre-college math curriculum, you can see the steps toward a mathematical definition. First, for rectangles with integer sides a and b, you can count the number of 1×1 squares needed to make the rectangle, leading to the area formula $A = a \times b$. From the physical point of view this is a formula, but from the mathematical point of view it is a definition, extended later to non-integer side lengths. Areas of triangles are not studied until much later. For right triangles with sides a and b and hypotenuse c, the area is shown to be equal $ab/2$ by showing that two of these fit together to make an $a \times b$ rectangle. This invokes a new principle: areas of congruent figures are equal. To compute the area of a parallelogram or trapezoid, the **dissection** principle is invoked: cutting up and rearranging the pieces of a figure preserves the area. These principles, all of which make intuitive and physical sense, are illustrated in Figure 35.

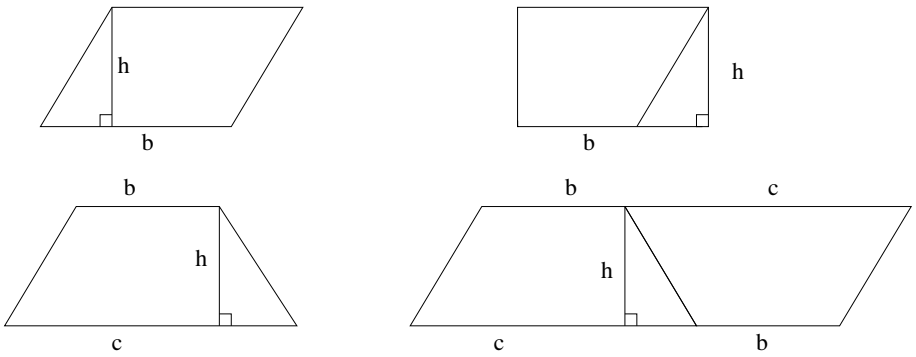


Figure 35: identifying congruent pieces of a dissection to evaluate areas of parallelograms and trapezoids

Exercise 10.1. *Write a sentence for each of the two rows explaining how it proves an area formula. What is being asserted to have the same area as what, why is this true, and what is the conclusion?*

The area of a circle is introduced, usually without much explanation. Do you know why the area of a circle of radius r is equal to πr^2? One common explanation is that areas of similar figures are related by a scaling principle. Recalling that area has units of squared length, it makes sense that scaling a figure by λ should scale the area by λ^2. All circles are similar; if follows that the area of a circle should be Kr^2 for some constant K. We can name this constant π but that leaves a nagging question unanswered. Scaling also shows that the circumference of a circle should be proportional to the radius, therefore $C = K'r$ for some other constant K'. This turns out to be 2π. But why should K' be double K? An argument involving dissections and limits is shown in Figure 36.

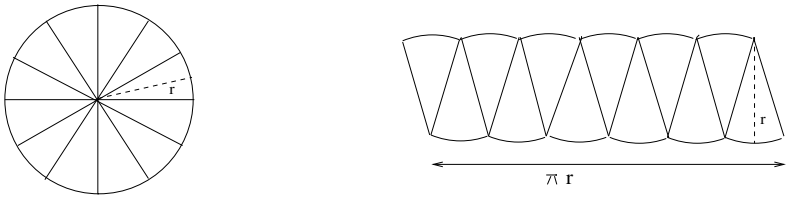


Figure 36: a limit of dissections relates the constants for circumference and area

Exercise 10.2. *In Figure 36 the measure πr on the right refers to the total curved length of the bottom. We have not defined limits of shapes, but intuitively, what is the limiting shape on the right and what are its dimensions?*

Once limits are brought into the discussion, there is a way to define areas of much more general shapes. The idea is this: put as many non-overlapping squares of some small side length ε as you can inside the shape. These cover an area less than the area of the shape, but if ε is small, it seems credible that the area is getting close to the area of the shape. If the limit as $\varepsilon \to 0^+$ exists, this should be the area of the shape. Similarly, you could completely cover the shape with squares of side ε if you are willing to cover a slightly too big region. When ε is small, you don't cover too much extra area. The limit as $\varepsilon \to 0^+$ should also be the area of the shape. To make a long story short (you can hear the full story in Math 360), there are many shapes for which it is possible to prove that these two limits exist and are equal. For these shapes we can *define* area to be this common limiting value. This mathematical definition captures our existing physical intuition and is also consistent with the principles we already adpoted: congrunce, scaling and dissection.

With this build up, we will mathematically define area for a certain restricted class of shapes. The class of shapes we start with will be the class of shapes that are rectangular on three sides but whose top is described by an arbitrary continuous function. More precisely, let $a < b$ be real numbers and let f be positive and continuous on the closed interval $[a, b]$. We will define the area of the region R bounded on the left by the vertical line $x = a$, on the right by the vertical line $x = b$, on the bottom by the x-axis (the line $y = 0$), and on the top by the graph of f (the curve $y = f(x)$). This region is shown in Figure 37.

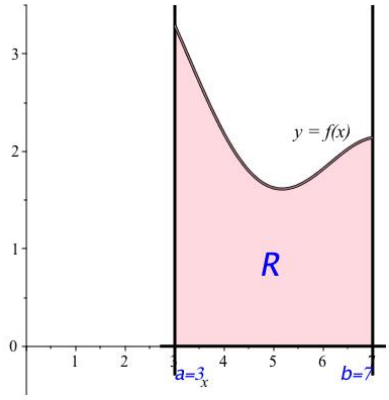


Figure 37: region between the x-axis and the graph of a function

10.2 Riemann sums and the definite integral

We now define the lower and upper Riemann sums with n rectangles for a function f on an interval $[a, b]$. If you prefer a picture, refer to Figure 38.

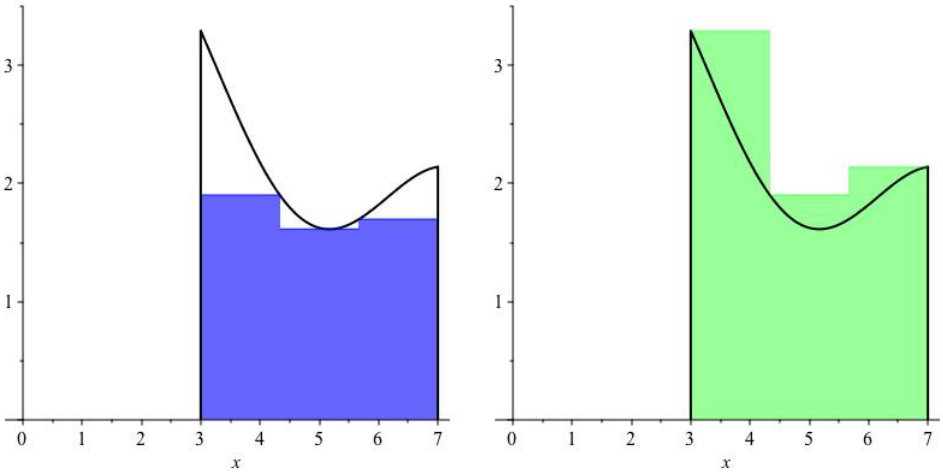


Figure 38: lower and upper Riemann sums

Exercise 10.3. *In Figure 38, what value of n was used?*

Definition 10.1. *Let f be a nonnegative continuous function on an interval $[a, b]$ and let n be a positive integer. Let $I_1, \ldots, I_n$ denote the intervals you get when you divide $[a, b]$ into n equally sized intervals. For each interval I_k, let y_k be the minimum value of f on I_k and let R_k be the rectangle with base I_k on the x-axis and height y_k. The* **lower Riemann sum** *for f on $[a, b]$ with n rectangles is the sum of the areas of the rectangles R_k, for $1 \leq k \leq n$. The* **upper Riemman sum** *is defined similarly, with the maximum value instead of the minimum value on each interval.*

Exercise 10.4. *What are the endpoints of the interval I_2 in Figure 38? What is the approximate value of y_2?*

Example 10.2. We are not given precise values for the function f in Figure 38, but we can estimate from the graph. The rectangles each have width $4/3$. The respective heights for the lower Riemann sum appear to be roughly $1.9, 1.6$ and 1.7, making the lower Riemann sum equal to $(4/3)1.9 + (4/3)1.6 + (4/3)1.7 = (4/3)5.2 \approx 6.93$. The upper Rieman sum is computed from rectangles with approximate heights $3.3, 1.9$ and 2.15, leading to a total area of $(4/3)7.35 = 9.8$.

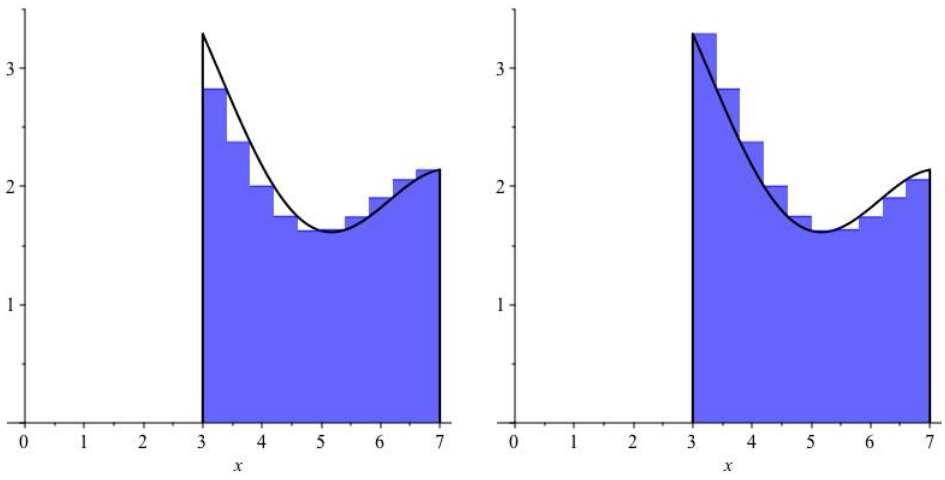


Figure 39: left and right Riemann sums

The **left Riemann sums** and **right Riemann sums** are defined similarly, except that instead of using the minimum or maximum values of the function on each sub-interval the left Riemann sums uses the value at the left endpoint of each interval I_k, while the right Riemann sum uses the value at the right endpoint of each sub-interval I_k. Examples are shown in Figure 39.

Exercise 10.5. *Is the left Riemann sum on the left of Figure 39 or on the right?*

The upper and lower Riemann sums give upper and lower bounds on the area of the figure. The left and right Riemann sums are neither upper nor lower bounds for the area, but they are sandwiched in between the lower and upper Riemann sums, so they also converge to the area. They are useful because always choosing the left endpoint (or always choosing the right endpoint) leads to a simpler formula.

Exercise 10.6. *Write a summation formula for the left Riemann sum for f on $[a, b]$ with 10 rectangles. It should have 10 terms and look like this:* $\sum_{n=1}^{10} \cdots$.

The values of the lower and upper Riemann sums in Figure 38 are approximately 6.9 and 9.8. These are not very close to each other, leaving considerable uncertainty about the true area. Replacing by the left (say) Riemann sums, we can program the sum into a computing device and compute for much greater values of n. If we increase n from 3 to 10, as in Figure 39, we find the Riemann sums come out to approximately 8.48 and 8.02 – somewhat better. These are not necessarily bounds: the true value could be greater than both, or less than both, or in between. Replacing n by 50 gives 8.28. This is again not a bound, however the following theorem guarantees that as $n \to \infty$, this will converge to the area.

Theorem 10.3. *The upper Riemann sums for any continuous function f on any closed interval $[a, b]$ converge as $n \to \infty$. The lower Riemann sums converge to the same value. It follows that you can let $y_k = f(x_k)$ for* any *$x_k \in I_k$ and the sums of rectangle areas will still converge to this common limit.*

Definition 10.4. *The common limit in Theorem 10.3 is called the* **definite integral of f from a to b** *and is denoted* $\int_a^b f(x)\,dx$.

Exercise 10.7. *Let f be the constant function c. How far apart are the lower and upper Riemann sums for $\int_3^9 c\,dx$ (pick any value of n)? What does that tell you about the definite integral $\int_a^b c\,dx$?*

Remark. The variable x is a bound variable; the notation $\int_a^b f(u)\,du$ would represent the same thing. Also, as in the notation for derivatives, you shouldn't try to interpret what the symbol du means on its own. It evokes the width of an infinitesimal rectangle, but you can't always count on it to behave nicely in equations.

Exercise 10.8. *From this construction and theorem, you can deduce some identities for integrals. Simplify these definite integrals in terms of more basic ones.*

(i) $\int_a^b f(x)\,dx + \int_b^c f(x)\,dx$

(ii) $\int_a^b 3 + 10f(x)\,dx$

10.3 Interpretations of the integral

Area is the most visually obvious interpretation but there are many others. If material (or charge, or mass, etc.) is spread out unevenly over an interval, the **density** at any point is the amount of material per length near that point. It has units of material divided by length. The total amount of material in the interval is gotten by summing how the amount of material over small intervals. When the interval is small enough, we can estimate the amount of material as $f(x)$ times the length of the interval where x is any point in the interval. This is not exact because f generally will still vary over the interval, but not by much when x is small. The limit as the interval lengths go to zero will be $\int_a^b f(x)\,dx$ and will represent the total material.

Example 10.5. A 3-inch blade of grass is covered in mold. The amount of mold decreases up the blade because it is killed by sunlight. The density of mold per inch is $1000e^{-x/3}$ spores per inch at height x inches from the ground. The total number of spores on the blade of grass is given by $\int_0^3 1000e^{-x/3}\,dx$.

Exercise 10.9. *Why did we use 0 and 3 for the limits of integration in Example 10.5?*

Integrals can also be used to give averages. For a finite collection, the average is defined to be the total divided by the number you added to get the total. Averages over an interval are defined similarly.

Definition 10.6 (average over an interval)**.** *The average of a quantity varying over an interval* $[a, b]$ *according to a function* f *is defined to be* $\dfrac{1}{b-a}\displaystyle\int_a^b f(x)\,dx$.

Example 10.7. Suppose the temperature over a day is $f(t)$ degrees Celsius t hours after midnight. The average temperature over the day is then $\dfrac{1}{24}\displaystyle\int_0^{24} f(T)\,dt$.

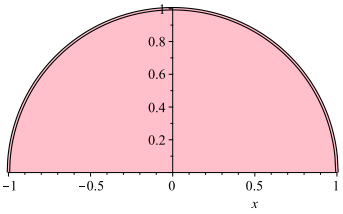
Exercise 10.10. *Suppose $f(x)$ is some constant c on the interval $[a, b]$. Intuitively, what is the average of f on $[a, b]$? Compute the average value of f on $[a, b]$ directly from the definitions and check that it is what you expected.*

An integral is a limit of a sum of rectangle areas. The units are therefore the same units as the rectangle areas. The rectangles live on a graph where the x-axis has units of the argument variable and the y-axis has units of the function. Therefore the rectangle units, hence the integral units, are units of the argument times units of the function. In the grass example, the function was density (spores per inch) and the argument was inches, therefore the integral had units of spores. It is a good thing that this agrees with our interpretation of the integral as the total number of spores. In the temperature example, f has units of temperature and t is in units of time, so the integral of f has units of temperature times time. This sounds like a strange unit but it's not unheard of. Severity of cold spells is measured, for example, in heating degree-days. The average is the integral divided by the time, so it is in units of temperature. Of course: the average temperature should be a temperature!

In physics there are countless things represented by integrals. One is the **moment**. Suppose mass is spread out along $[a, b]$ with density f (you know what that means now, right?). Integrate f and you get the total mass. If instead you compute $\int_a^b x\, f(x)\, dx$ you get the **moment of inertia**, which tells you how much the weight counts when balancing (imagine a teeter-totter pivoting on the origin), or how much torque is needed to produce a given angular acceleration.

In probability theory, random quantities can be discrete or continuous. If the random quantity X is discrete it means that there is a sset of values $x_1, x_2, \ldots$ such that probabilities for $X = x_k$ sum to 1. This could be a finite sum or the sum of an infinite sequence (you now know the definition of an infinite sum, right?). For continuous quantities, you need integrals. The probabilities for finding X to take various values are spread continuously over an interval (possibly an infinite interval such as the whole real line). There will be a **probability density function** f such that the probability of finding X in a given interval $[a, b]$ will be $\int_a^b f(x)\, dx$. We will say more about this in Section 12, after we have defined integrals where one or both of the limits of integration can be infinite.

Going back to the area interpretation, you may ask what about more general shapes? It turns out you don't really need straight sides. The vertical walls on the left and right sides of the regions in Figures 37 and 38 can disappear. For example, letting $f(x) = \sqrt{1-x^2}$ and $[a, b] = [-1, 1]$ produces the upper half of a circle.



So far we have required f to be a nonnegative function. What if f is negative? Let Δ_k denote the width of I_k. The most useful definition turns out to be that the integral is still the limit of sums of the quantities $\sum_k f(x_k) \cdot \Delta_k$ but we must interpret this as a new concept, called **signed area** rather than area. We won't worry too much about signed area; it just means we need to keep track of whether f is positive or negative before we know whether $\int_a^b f(x)\,dx$ computes area or its negative. Figure 40 shows a function which is positive on $[0, 0.42]$ and negative on the interval $[0.42, 1]$. The integral $\int_0^1 f(x)\,dx$ will be slightly negative because it adds a positive area A_1 to a negative signed area A_2.

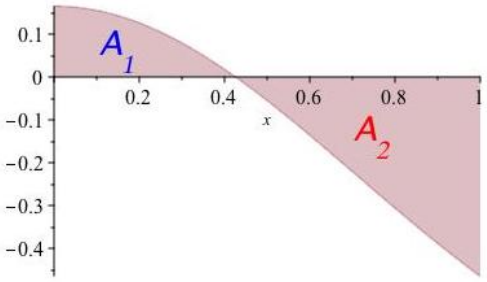


Figure 40: the signed area A_2 will be negative because it is below the x-axis

Another useful definition along the same lines switches the upper and lower limits.

Definition 10.8. *If $a < b$ then $\int_b^a f(x)\,dx$ is defined to equal $-\int_a^b f(x)\,dx$.*

Suppose f and g are functions such that $f \geq g$ on $[a, b]$. One interpretation of $\int_a^b [f(x) - g(x)]\,dx$ is that it is the area of the shape with upper boundary f and lower boundary g. We started out computing areas of a very specific set of shapes, looking like three sides of a rectangle and a possibly curved upper boundary. Using the idea of upper and lower boundaries we can use integrals to give the area of a much greater variety of shapes.

Exercise 10.11. *On a coordinate axis, draw a heart shape (you know, the classic Valentine's heart). Then draw in values a and b on the x-axis and graphs of functions f and g such that the area of the heart is computed by $\int_a^b [f(x) - g(x)]\,dx$.*

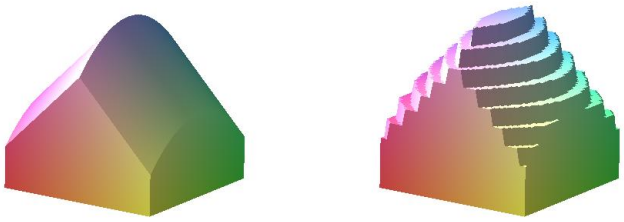


Figure 41: a solid volume (left) cut into slabs (right)

The examples of densities of quantities spread out along a line is somewhat limited. When quantities spread out, usually they spread over a region in a plane or in three dimensions. The next calculus course covers multivariable integration. Still, there are some higher dimensional things you can do with ordinary integrals. One of these is to compute a volume of an object if you know the area of its cross-sections. Dividing the object into n very thin slabs, the volume of the k^{th} one is roughly the thickness Δ_k times the cross-sectional area of the k^{th} slab, call it A_k; see Figure 41.

The limit of $\sum_{k=1}^n A_k \Delta_k$ should give the volume. Line up the slabs so that the x-axis goes perpendicular to the slabs. This limit looks awfully similar to the limit of $\sum_{k=1}^n f(x_k)\Delta_k$ where x_k is any point on the x-axis inside the k^{th} slab and f is the function telling the cross-sectional area at every x-value. Therefore, the volume is computed by $\int_a^b f(x)\,dx$ where a and b are the x-values at the first and last slab respectively.

Example 10.9 (area of a pyramid)**.** We write an integral for area of a pyramid whose base is a square of side length s and whose height is h. It corresponds best to the description above if we orient it so the height is measured along the x-direction with the apex at the origin. See Figure 42. The cross-section is a square with side increasing linearly from 0 to s as x increases from 0 to h. Thus, the side length is given by $\ell(x) = (s/h)x$, hence the cross-sectional area is given by $f(x) = (s/h)^2x^2$ between $x = 0$ and $x = h$. The volume is therefore given by $\int_0^h (s/h)^2x^2\,dx$. When you learn to compute integrals, this will be a pretty easy one.

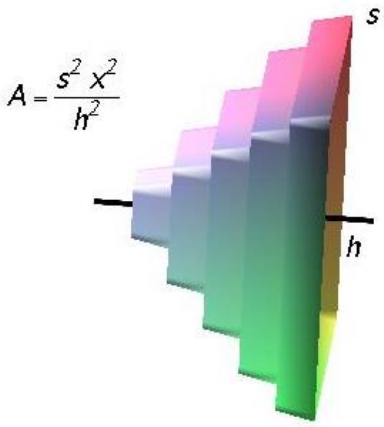


Figure 42: a pyramid, cut into slabs along the x-direction

10.4 The fundamental theorem of calculus

The reason we make such a fuss over integrals is that they can often be exactly computed. To see how this works, we look at the **indefinite** integral. Replacing the upper limit on the integral by a variable yields a function of that variable. To say this in another way, we may consider $\int_a^b f(x)\,dx$ as a function of the free variables a and b (it can't be a function of x because x is a bound variable). Let a remain a constant but consider b to be a variable. We then have a function, $b \mapsto \int_a^b f(x)\,dx$. Denote this function by G, in other words $G(b) := \int_a^b f(x)\,dx$.

Example 10.10. Let $f(x) := 3x$ and $a = 0$. Then $G(b) := \int_0^b 3x\,dx$. Definite integrals

compute area, hence $G(b)$ is the area of the triangle with vertices at the origin, $(b, 0)$ and $(b, 3b)$. The triangle area formula gives $G(b) = (3/2)b^2$.

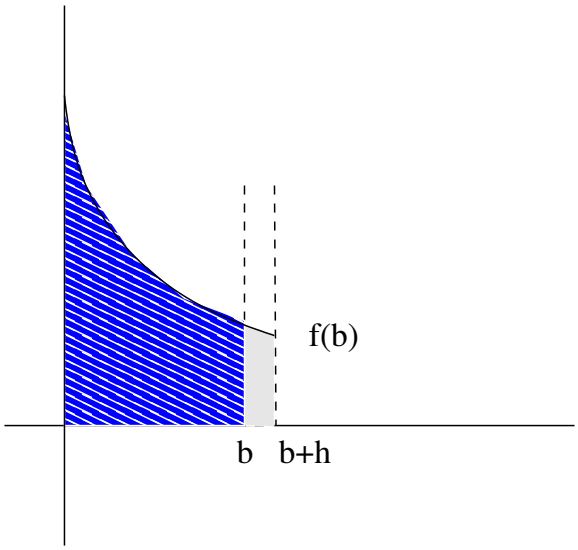
For fun (we have a warped sense of fun), compute G'. That's an easy one: $G'(b) = 3b$. Note that this is the integrand of the original integral, with the free variable b in place of the bound variable x. This is not a coincidence, as the following theorem asserts.

Theorem 10.11 (Fundamental Theorem of Calculus)**.** *Let f be a continuous function on an interval $[a, c]$. For $b \in (a, c)$, let $G(b) := \int_a^b f(x)\,dx$. Then $G'(b) = f(b)$.*

Sketch of Proof: The derivative from the right is given by

$$G'(b^+) = \lim_{h \to 0^+} \frac{G(b+h) - G(b)}{h}\,.$$

When h is small, the value of $G(b + h)$ is very well approximated by $G(b) + hf(b)$; in the picture at the right, $G(b)$ is the blue area and $G(b+h)$ is the blue area plus the shaded black and white area. Plugging this in gives $\dfrac{G(b) + hf(b) - G(b)}{h} = f(b)$. To turn this into a proof, you need to use continuity of f to show that the error replacing $G(b + h)$ by $G(b) + hf(b)$ is $\ll h$, so the approximation does not affect the limit. You already know enough to understand the argument, but in the interest of time, the details are left to a course in mathematical analysis.



Anti-derivatives

The Fundamental Theorem of Calculus says we can evaluate integrals of f if we know a funcdtion G whose derivative is f. That motivates the next definition.

Definition 10.12. *An* **anti-derivative** *of a function f is any function G such that $G' = f$.*

How do we find anti-derivatives? The next chapter is entirely about computing these. Like rules for differentiation, rules for anti-differentiation start from a collection known results. For derivatives, we obtained these from the definition by computing limits. For anti-derivatives, we will get these simply by remembering some basic derivatives. The simple rule yielding the derivative of a polynomial may be run backwards. So for example the monomial ax^m has anti-derivative $\frac{a}{m+1}x^{m+1}$. We can sum these, obtaining the anti-derivative of any polynomial: an anti-derivative of $\sum_{k=0}^{m} a_k x^k$ is given by $\sum_{k=0}^{m} \frac{a_k}{k+1}x^{k+1}$. In fact this works for negative or fracdtional powers, as long as the power is not -1.

Exercise 10.12.
(i) Why can't the power be -1?
(ii) Compute an anti-derivative of $x^2 - 5x + 6$.
(iii) Compute a different anti-derivative of $x^2 - 5x + 6$.

We say "an anti-derivative" rather than "the anti-derivative" because there is more than one. The functions G and $G + c$, where c is a constant, have the same derivative, so one is an anti-derivative of f if the other is. This is the only way anti-derivatives can differ[13]. Once you know the value of the anti-derivative at any point, it is easy to reconstruct the correct anti-derivative as an integral, as in the following example.

Example 10.13. Suppose G is an anti-derivative of f and $G(3) = 7$. We will look for an anti-derivative of the form $G(b) = c + \int_a^b f(x)\,dx$. To write G as an integral with a variable upper limit, begin by choosing the constant for the lower limit. The most convenient choice is 3, because we are supposed to know the value of G at 3. The function $b \mapsto \int_3^b f(x)\,dx$ is zero at 3, so we will need to add 7. We therefore choose

$$G(b) = 7 + \int_3^b f(x)\,dx\,.$$

For concreteness, let's see how this works with the example from above: $f(x) = 3x$. Then $G(b) = 7 + \int_3^b 3x\,dx$. We already computed $\int_0^b 3x\,dx = \frac{3}{2}b^2$ and similarly $\int_0^3 3x\,dx = (3/2)3^2 = 27/2$. Subtracting, $\int_3^b 3x\,dx = (3/2)b^2 - 27/2$. Thus the anti-derivative we are looking for is $7 + (3/2)b^2 - 27/2 = (3/2)b^2 - 13/2$.

This example shows a general principle, which we record as a proposition.

Proposition 10.14 (computing definite integrals with anti-derivatives)**.**
The definite integral $\int_a^b f(x)\,dx$ is equal to $G(b) - G(a)$, also denoted $G|_a^b$, when G is any

[13]This follows from some technical analysis which we won't be doing.

anti-derivative of f.

Note: this implies that $H(b) - H(a) = G(b) - G(a)$ when H is any other anti-derivative of f. In other words, differences of an anti-derivative at a specified pair of points do not depend on which particular anti-derivative was chosen.

Exercise 10.13. *Compute $\int_1^6 x^2 - 5x + 6\,dx$.*

10.5 Estimating sums via integrals

We have seen integrals interpreted as areas and volumes, totals and averages, moments, and probabilities. One more use of an integral is to estimate a sum. In a way this is the reverse of the definition, which tells you that an integral is estimated by Riemann sums, in fact is a limit of such sums. Going the other way, if we have a sum, we can write an integral for which it is a Riemann sum. We may then expect the integral to be a good approximation for the sum. This will be easier when we know how to compute more integrals, but there are plenty we can already compute. We illustrate with a long example. It starts with the fact that the derivative of $\ln x$ is $1/x$. This means that an anti-derivative of $1/x$ is $\ln x$.

Example 10.15 (harmonic sum estimated by an integral)**.** Problem: estimate the 100^{th} harmonic number $1 + 1/2 + 1/3 + \cdots + 1/100$. To solve this, we may as well estimate $H_n := \sum_{k=1}^{n} 1/k$ for any positive integer n. Summing $1/n$ looks a lot like integrating $1/x$. In fact, suppose we write a Riemann sum for $\int_1^n 1/x\,dx$ that has precisely $n-1$ rectangles. Then the intervals I_k are just the intervals $[1,2], [2,3], \ldots, [n-1, n]$. Even better, we can make areas of the rectangles be exactly the same as in the sum. We just need to use the upper Riemann sum: $1 + 1/2 + \cdots + 1/(n-1)$; see the left-hand side of Figure 43 for a picture of this when $n = 9$.

We have shown that H_{n-1} is an upper Riemann sum for $\int_1^n 1/x\,dx$. By Proposition 10.14 the integral is the difference of anti-derivatives:

$$\int_1^n \frac{1}{x}\,dx = \ln n - \ln 1 = \ln n\,.$$

Therefore, we have shown the bound $H_{n-1} \geq \ln n$. In particular, choosing $n = 101$, we see that $H_{100} \geq \ln 101 \approx 4.615$.

Is this an upper bound or a lower bound? It depends on your point of view. If we were trying to figure out the integral up to 100, H_{100} would be an upper bound on the value. But

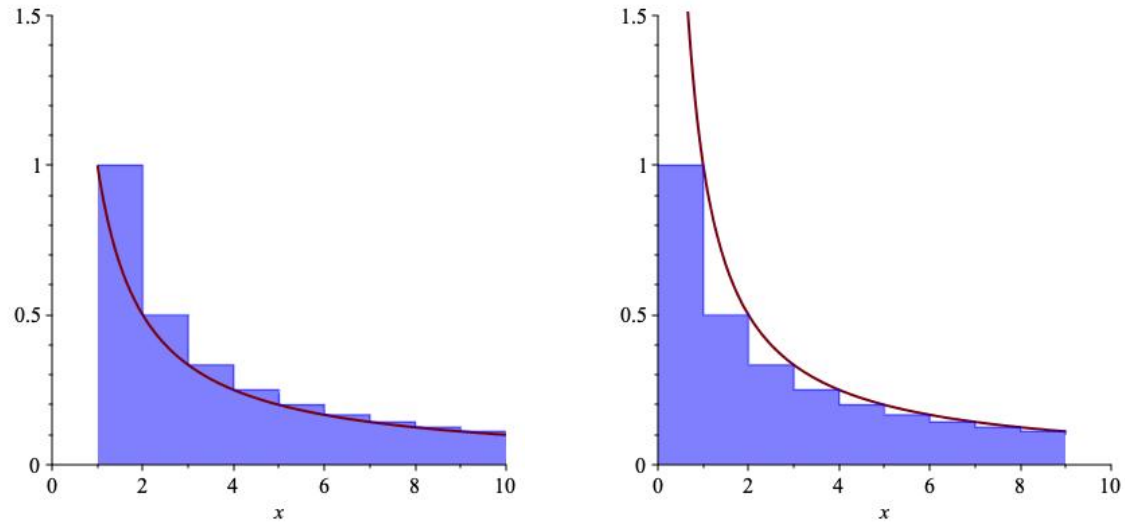


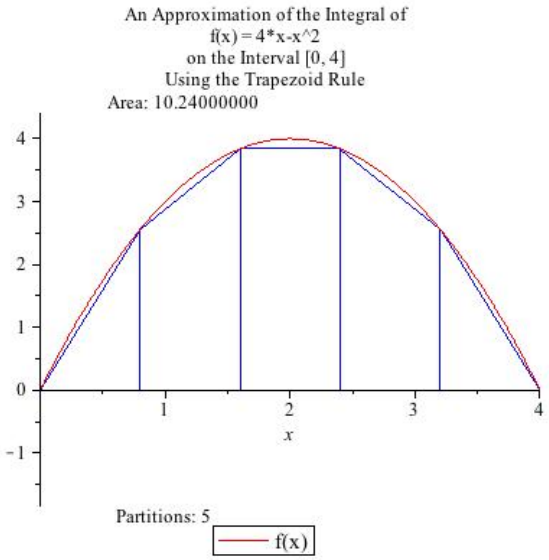
Figure 43: representing the harmonic sum as upper and lower Riemann sums

in this case we know the integral and are trying to estimate H_{100}. The integral provides a lower bound, in this case 4.615.

What about an upper bound on H_{100}. The obvious thing is to see if we can make the same sum be a lower Riemann sum. Watch what happens when you try to do this. Take the graph, shift all the rectangles one unit left, and voilà! (See the right-hand side of Figure 43.) This shows that H_{100} is a lower Riemann sum for a slightly different integral, namely $\displaystyle\int_0^{100} \frac{1}{x}\,dx$. Alas, this is not an integral we can do because $1/x$ is not continuous at $x = 0$. In fact, when we study improper integrals, we will see this evaluates to $+\infty$. Sure, we get the upper bound $H_{100} \leq \infty$, but that is hardly useful. All is not lost, however, if we use some common sense. The same picture shows that an upper bound for the harmonic sum starting at 2 instead of 1 is

$$\sum_{k=2}^{100} \frac{1}{k} \leq \int_1^{100} \frac{1}{x}\,dx = \ln 100 \approx 4.605\,.$$

So, adding back the 1, we see that $H_{100} \leq 1 + \ln 100 \approx 5.605$. This is about as good as we can do with the techniques we have so far: $4.615 \leq H_{100} \leq 5.605$. For the record, $H_{100} \approx 5.1874$.



Trapezoidal approximation

Sometimes it can be frustrating using Riemann sums because a lot of calculation doesn't get you all that good an approximation. You can see a lot of "white space" between the function f and the horzontal lines at the top of the rectangles that make up the upper or lower Riemann sum. If instead you let the rectangle become a right trapezoid, with both its top-left and top-right corner on the graph $y = f(x)$, then you get what is known as the **trapezoidal approximation**. The figure shows a trapezoidal approximation of an integral $\int_0^4 f(x)\,dx$ with five trapezoids. Note that the first and last trapezoid are degenerate, that is, one of the vertical sides has length zero and the trapezoid is actually a right triangle. It is perfectly legitimate for one or more of the trapezoids to be degenerate.

Because the tops of the slices are allowed to slant, they remain much closer to the graph $y = f(x)$ than do the Riemann sums. Because the area of a right trapezoid is the average of the areas of the two rectangles whose heights are the value of f at the two endpoints, it is easy to compute the trapezoidal approximation: it is just the average of the left-Riemann sum and the right-Riemann sum corresponding to the same partition into vertical strips.

Example 10.16. Compute the trapezoidal approximation for $\int_1^2 \frac{1}{1+x^2}$ with 10 trapezoids. Averaging the left and right Riemann sums always gives a sum containing the $n-1$ common terms plus half the first term for the left Riemann sum and the last term for the right

Riemann sum. In this case one gets

$$\frac{1}{2}\frac{f(1)}{10} + \frac{1}{2}\frac{f(0)}{10} + \sum_{1}^{9} \frac{1}{10} f\left(1 + \frac{j}{10}\right) .$$

The outcome of trapezoidal approximation in general can be summarized as, "Sum the values of f along a regular grid of x-values, counting endpoints as half, and multiply by the spacing between consecutive points."

The trapezoidal estimate is usually much closer than the upper or lower estimate, though it has the drawback of being neither an upper nor a lower bound. However, if you know the function to be concave upward then the trapezoidal estimate is an upper bound. Similarly if $f'' < 0$ on the interval then the trapezoidal estimate is an lower bound. In the figure, f is concave downward and the trapezoidal estimate is indeed a lower bound.

Example 10.17. The function $1/(1+x^3)$ is concave upward on $[1, 2]$ (compute and see that the second derivative is a positive quantity divided by $(1+x^3)^3$) so the trapezoidal estimate should be not only very close but an upper bound. Indeed, the trapezoidal estimate is the average of the upper and lower previously computed and is equal to 0.25485... which is indeed just slightly higher than the true value of 0.25425....

Aside. *Just as Riemann sums estimate by strips with constant height (degree zero) and trapezoids estimate by strips whose height is a linear function, you could imagine using higher degree polynomials (because you can still compute their areas exactly).* Simpson's rule*, for example, uses quadratic functions. It gets very good results! We won't discuss it here but you might want to ask your instructor about higher degree polynomial approximations, which can be programmed without too much difficulty into a computation package or even a spreadsheet.*

11 Computing integrals

All continuous functions have anti-derivatives, but not all of the anti-derivatives have names. For example, the definite integral $\int_3^8 \frac{1}{\ln x}\,dx$ is a well defined quantity; indeed $\int_a^b \frac{1}{\ln x}\,dx$ is well defined for any $b > a > 1$, but the function $b \mapsto \int_a^b (1/\ln x)\,dx$ is not equal to any combination of named functions such as powers, logs, exponentials and trig functions. The same is true of the normal (bell curve) density function e^{-x^2}, or $\sqrt{\sin x}$ or $\sqrt{1-4x^2}/\sqrt{1-x^2}$. The prevalence of functions like this is the reason we need good numeric approximations to integrals. In the remainder of this section we concentrate on anti-derivatives for which reasonably nice exact expressions exist.

11.1 Remembering and guessing

Computing derivatives, as you saw in Chapter 5, rests on combination rules and working out some basic cases. For anti-derivatives the same is true, with" working out" replaced by "remembering". In other words, if you remember what the derivative of f is, then you know how to compute an anti-derivative of f'. This is how we computed anti-derivatives for polynomials, for example. The strategy is then: (1) list the derivatives we already know, organized in a way that allows us to query what function goes with a given derivative; and (2) give combining rules for anti-derivatives. This gives the following proposition. Note that in each case, remembering allows us to identify just one of the antiderivatives; we trust you can compute the others from that.

Notation: we use an integral sign without upper and lower limits to denote the antiderivative: e.g., $\int (3x^2+1)\,dx$ is equal to x^3+x, plus any constant. We usually write this as x^3+x+C. By custom, we don't change the variable. In previous sections, for example, we were careful to write $\int_0^b (3x^2+1)\,dx$ as a function of b, namely b^3+b. But when writing the indefinite integral we tend to write $\int (3x^2+1)\,dx = x^3+x+C$, not b^3+b+C. This is because it's shorthand for

The indefinite integral of the function $x \mapsto 3x^2+1$ is any function $x \mapsto x^3+x+C$.

The variable x is bound, so the choice of letter does not affect the meaning.

Proposition 11.1. *The following basic anti-derivatives are computed by reversing Proposition 5.8.*

(i) $\displaystyle\int x^m = \frac{1}{m}x^{m-1} + C$ *as long as* $m \neq 0$.

(ii) $\displaystyle\int \frac{1}{x}\,dx = \ln x + C$

(iii) $\displaystyle\int \cos x\,dx = \sin x + C$

(iv) $\displaystyle\int \sin x\,dx = -\cos x + C$

(v) $\displaystyle\int \sec^2 x = \tan x + C$

(vi) $\displaystyle\int e^x\,dx = e^x + C$

(vii) $\displaystyle\int \frac{1}{\sqrt{1-x^2}}\,dx = \arcsin x + C$

(viii) $\displaystyle\int \frac{1}{1+x^2}\,dx = \arctan x + C$

Exercise 11.1. *Use Proposition 11.1 to compute this definite integral:* $\displaystyle\int_0^1 \frac{1}{1+x^2}\,dx$. *You will also need Proposition 10.14, which you should get used to using without even thinking of it as an extra step.*

The derivative of a sum or difference is the sum or difference of the derivatives. The derivative of $c \cdot f$ is c times the derivative of f for any real constant c. This leads immediately to the following proposition.

Proposition 11.2 (linearity of the anti-derivative)**.**

$$\int [f(x) + g(x)]\,dx = \int f(x)\,dx + \int g(x)\,dx\,; \tag{11.1}$$

$$\int c \cdot f(x)\,dx = c \cdot \int f(x)\,dx\,. \tag{11.2}$$

PROOF: Let F be an anti-derivative of f and G be an anti-derivative of G. Then $(F+G)' = F' + G' = f + g$ therefore $(F + G)$ is an antiderivative of $f + g$, proving (11.1).

Exercise 11.2. *The proof of the second statement of Proposition 11.2 is even shorter. See if you can write it down.*

The word "anti-derivative" is a mouthful and so is the verb form "anti-differntiate". Because computing integrals comes down to anti-differentiation, common practice is to use the verb **integrate** in place of "anti-differentiate". We also call an anti-derivative an "integral". Propositions 11.1 and 11.2 allow us to compute some more integrals.

Example 11.3. Compute the integral of $\dfrac{a\cos x + b/\cos x}{\cos x}$. Simplifying,

$$\frac{a\cos x + b/\cos x}{\cos x} = a + b\sec^2 x\,.$$

Therefore

$$\begin{aligned}\int \frac{a\cos x + b/\cos x}{\cos x}\,dx &= \int [a + b\sec^2 x]\,dx \\ &= \int a\,dx + b\int \sec^2(x)\,dx \\ &= ax + \tan x + C\,.\end{aligned}$$

Exercise 11.3. *One of your classmates argues this is wrong: $\int a\,dx = ax+C$ and $\int \sec^2(x)\,dx = \tan x + C$, therefore the answer should be $ax + \tan x + 2C$. Explain what is going on: is the original answer is right, or the new answer, or both?*

Example 11.3 should worry you. Does it seem a bit contrived? The expression $\dfrac{a\cos x + b/\cos x}{\cos x}$ just happens to simplify into two expressions covered by the list of cases in Proposition 11.1. If that seems like a piece of luck, it is. With only Propositions 11.1 and 11.2 you won't get very far. The next two sections give two rules for combining integrands that will greatly increase your ability to integrate. Keep in mind though, that in some sense you are still lucky whenever you can compute an analytic expression for an anti-derivative: many anti-derivatives have no nice formula.

11.2 Integration by parts

The sum rule for derivatives is simple enough that it leads directly to (11.1), which is an identical rule for anti-derivatives. There is also a product rule, but it does not lead directly

to an identical rule for anti-derivatives. That's because it is not symmetric. The derivative of fg is not $f'g'$ but rather $f'g + g'f$. Therefore, if we want to run it backwards, we get

$$\int [f'(x)g(x) + g'(x)f(x)]\,dx = f(x)g(x) + C\,. \tag{11.3}$$

The problem is, this doesn't tell us how to integrate a product such as $f'g'$, but rather $f'g + g'f$. This is great if someone asks us to compute the anti-derivative of $f'g + g'f$, but this is rare, harder to spot, and does not answer the question as to the anti-derivative of the product.

The best we can do is to exploit (11.3) as much as we can. This leads to the following proposition.

Proposition 11.4 (integration by parts)**.** *Let u and v be differentiable functions. Suppose $u'v$ is known to have anti-derivative G. Then $v'u$ has anti-derivative $uv - G$. In a single equation,*

$$\int v'u\,dx = uv - \int u'v\,dx\,. \tag{11.4}$$

Proof: This is just the product rule run in reverse: $(uv)' = u'v + v'u$, therefore

$$(uv - G)' = u'v + v'u - G' = v'u\,.$$

The way this works in practice is that when integrating an expression, you try to identify the expression as $v'u$ for some functions u and v. Then you check whether you already know the anti-derivative to $u'v$. If so, you subtract this from uv and you are done. Sometimes there are several possible ways to do this, in which case you may have to try them all until you find one that works.

Example 11.5. Use integration by parts to integrate xe^x. Oviously this decomposes as a product of x and e^x. One of these should be v' and the other should be u. Let's try setting

$$\begin{aligned} v' &= x\,; \\ u &= e^x\,. \end{aligned}$$

At first this goes smoothly: the expression we chose for v has a known anti-derivative and the one we chose for u has a known derivative, therefore we can find v and u':

$$\begin{aligned} v &= \frac{x^2}{2} + C\,; \\ u' &= e^x\,. \end{aligned}$$

Unfortunately the next step doesn't work: $u'v = e^x(x^2/2 + C)$, which is not something whose anti-derivative we recognize no matter what choice we make for the constant C.

Back to the drawing board. Let's try switching it:

$$\begin{aligned} v' &= e^x\,; \\ u &= x\,. \end{aligned}$$

Again it goes smoothly at first: the expression we chose for v has a known integral e^x and the one we chose for u has a known derivative 1, therefore

$$\begin{aligned} v &= e^x + C\,; \\ u' &= 1\,. \end{aligned}$$

Now we're in better shape. Choose $C = 0$ (usually this works if anything does). Then $u'v = e^x$, for which an integral is known, namely e^x. Therefore,

$$\int xe^x\,dx = \int uv'\,dx = uv - \int u'v\,dx = xe^x - \int e^x\,dx = xe^x - e^x + C\,.$$

We did a long-winded example to show you the process of trial and error and to show how each step works. What would have happened if we chose a different value of C? It turns out it always works exactly as well.

Exercise 11.4. *Complete the computation in the previous example, choosing $C = 7$ instead of $C = 0$, to see that it works out the same after some cancellation. [Bonus question: can you see why this cancellation always happens?]*

It usually takes several worked examples and a lot of practice before integration by parts feels natural. Because "a lot of practice" means different things to different people, we include only a few mandatory self-check and howework problems, putting a greater number online for those who want to practice.

Example 11.6. Compute the definite integral $\int_0^{2\pi} x\sin(x)\,dx$. We start with the indefinite integral, which we compute by parts. Based on what happened with xe^x, let's decide to start with the choice $u = x, v' = \sin x$. Then $v = -\cos x$ and $u' = 1$, which yields

$$\int x\sin(x)\,dx = -x\cos x - \int(-\cos x)\,dx = -x\cos x - (-\sin x) = \sin x - x\cos x + C\,.$$

Evaluating the definite integral (notice we chose $C = 0$),

$$\begin{aligned}\int_0^{2\pi} x\sin(x)\,dx &= [\sin x - x\cos(x)]_{x=2\pi} - [\sin x - x\cos(x)]_{x=0} \\ &= [\sin(2\pi) - 2\pi\cos(2\pi)] - [\sin(0) - 0\cdot\cos(0)] \\ &= -2\pi\,.\end{aligned}$$

Exercise 11.5. *Evaluate* $\int_0^{\pi} x\cos(x)\,dx$.

Here are a few more tips to help you use integration by parts. Also, you should see a notational variation that is common in textbooks and on the web. Instead of $\int v'u\,dx = uv - \int u'v\,dx$, people sometimes write

$$\int u\,dv = uv - \int v\,du\,.$$

Because u and v are functions of x, you can think of $du := u'(x)\,dx$ and $dv := v'(x)\,dx$, whereby this form of the identity comes out to exactly the same thing as (11.4).

Repeated integration by parts

Sometimes integration by parts doesn't quite get you to an expression $u'v$ that you know how to evaluate, but it gets you closer, so that repeating the integration by parts solves the problem.

Example 11.7. Compute $\int x^2e^x$. Letting $v' = e^x$ and $u = x^2$ gives

$$\int x^2e^x\,dx = x^2e^x - \int 2xe^x\,dx\,.$$

That last expression isn't covered by Proposition 11.1 but we just saw (take out the constant factor 2) that it can be done by parts and integrates to $2(xe^x - e^x) = 2(x-1)e^x$. Therefore,

$$\int x^2e^x\,dx = x^2e^x - 2(x-1)e^x = (x^2 - 2x + 2)e^x\,.$$

It should be apparent you can integrate $p(x)e^x$ this way for any polynomial p. Some textbooks have a separate algorithm for this called **tabular integration**. We won't be teaching that, but you can google it if you ever need the anti-derivative of $p(x)e^x$ where

$p(x)$ has degree more than, say, 3 (doing it by hand gets longer and more complicated as the degree of p grows). To see how this will go, try the following exercise, which is about as much as we would ever ask you to do by hand.

Exercise 11.6. *Compute $\int x^3 e^x$. Double check afterword by differentiating your answer.*

Don't forget v' could be 1

You can always decompose any expression as itself times 1. In the langauge of $v\,du$ and $u\,dv$, that says $\int f(x)\,dx$ can always be thought of as $u\,dv$ where $u(x) = f(x)$ and $dv = dx$, that is, $v' = 1$. This only sometimes works but it's good to know.

Example 11.8. Compute $\int \ln(x)\,dx$. There's only one term to decompose so we pretty much have to use the $dv = dx$ trick. Setting $u(x) = \ln x$ and $dv = dx$, gives (recalling that the derivatve of $\ln x$ is $1/x$),

$$\int \ln(x)\,dx = (\ln x)(x) - \int x \cdot \frac{1}{x}\,dx = x\ln x - \int 1 \cdot dx = x\ln x - x + C\,.$$

This is a good one to memorize - it's very useful to recall quickly how to integrate the natural log.

11.3 Substitution

Integration by parts is what you get from reversing the product rule. Reversing the chain rule is called **substitution**. You can probably guess what it says. The chain rule says $(d/dx)f(g(x)) = f'(g(x))g'(x)$. Therefore, we need a rule to tell us that $\int f'(g(x))g'(x)\,dx = f(g(x))$. This gives the simplest form of the substitution method.

Theorem 11.9. *Suppose g is differentiable on an interval (a, b) and let I (which will also be a closed interval) be the range of g. Suppose h is differentiable on I. Then*

$$\int h'(g(x))g'(x)\,dx = h(g(x)) + C\,.$$

Writing f for h', this becomes

$$\int f(g(x))g'(x)\,dx = \left(\int f\right) \circ g \tag{11.5}$$

where the identity $f = h'$ allows us to write the indefinite integral $\int f$ in place of h on the right. This second form is sometimes clearer because we often arrive at the form $f(g(x))g'(x)$ before we have identified the antiderivative of f, hence it makes sense for the right-hand side to leave $\int f$ unevaluated.

Example 11.10. We compute the integral of $\dfrac{(\ln x)^2}{x}$. The numerator $(\ln x)^2$ looks like a composition $f(g(x))$ where $f(x) = x^2$ and $g(x) = \ln x$. We are in luck because $g'(x) = 1/x$ so there is alread a g' sitting there. The expression to be integrated looks like $f(g(x))g'(x)$, so applying (11.5),

$$\int \frac{(\ln x)^2}{x}\,dx = \left(\int x^2\right) \circ \ln\,.$$

The indefinite integral of x^2 is $x^3/3$, so the final answer is that the indefinite integral of $(\ln x)^2/x$ is $(\ln x)^3/3 + C$.

Exercise 11.7. *Use the substitution method to evaluate* $\int (2x)e^{x^2}\,dx$.

The substitution rule is very often stated in the language of science, with a variable u, thought of as a physical quantity related to the variable x via $u = g(x)$. For this reason the substitution method is commonly referred to as "u-substitution", a name which is a little silly only because it ties the method to a particular variable name u when of course you could choose any name. Instead of a theorem, this version is usually described as a procedure.

1. Change variables from x to u (hence the common name "u-substitution")
2. Keep track of the relation between dx and du
3. If you chose correctly you can now do the u-integral
4. When you're done, substitute back for x

Again, we let g be the function relating u to x via $u = g(x)$, and again you need hypotheses, namely the ones stated in Theorem 11.9). Then $du = g'(x)\,dx$. Usually you don't do this kind of substitution unless there will be an $g'(x)\,dx$ term waiting which you can then turn into du. Also, you don't do this unless the rest of the occurences of x can also be turned into u. If g has an inverse function, you can do this by substituting $g^{-1}(u)$ for x everywhere. Now when you reach the fourth step, it's easier because you can just plug in $u = g(x)$ to get things back in terms of x.

This notation gives a particularly nice simplification when $u = x + c$ for some constant c. Replacing x by $x + c$ is called a **translation**. In the first unit of the course, we discussed what this does to the graph. It is a very natural change of variables, corresponding to a different starting point for a parametrization.

Example 11.11 (translation)**.** Compute the indefinite integral of $\sqrt{x+6}$. Let $u = x + 6$. Then $du = dx$. Integrating the $1/2$ power (one of the basic facts in Proposition 11.1),

$$\int \sqrt{x+6}\,dx = \int \sqrt{u}\,du = \frac{2}{3}u^{3/2} = \frac{2}{3}(x+1)^{3/2} + C\,.$$

The moral of this story is that you can "read off" integrals of translations. For example, knowing $\int \cos x\,dx = \sin x$ allows you to read off $\int \cos(x - \pi/4)\,dx = \sin(x - \pi/4)$. Don't let this example fool you into thinking it works this way for functions other than translations. Thinking that $\int \cos(\sqrt{x})\,dx = \sin(\sqrt{x}) + C$ is wrong; it is the calculus equivalent of the algebra mistake $(a+b)^2 = a^2 + b^2$.

Here's an example of u-substitution with something other than a translation.

Example 11.12. Compute $\displaystyle\int \sin^n x\,\cos x\,dx$.

Solution: substitute $u = \sin x$ and $du = \cos x\,dx$. This turns the integral into $\int u^n\,du$ which is easily valuated as $u^{n+1}/(n+1) + C$. Now plug back in $u = \sin x$ and you get the answer

$$\frac{\sin^{n+1} x}{n+1} + C\,.$$

You might think to worry whether the substitution had the right domain and range, was one to one, etc., but you don't need to. When computing an indefinite integral you are computing an anti-derivative and the proof of correctness is whether the derivative is what you started with. You can easily check that the derivative of $\sin^{n+1} x/(n+1)$ is $\sin^n x \cos x$.

After a translation, the next simplest substitution is a **dilation**, where $u(x) = cx$ for some nonzero real number c. This is the other case in which substitution always succeeds: if you can integrate $f(x)$ you can always integrate $f(cx)$. We leave it to you to work this out, first in an example, then in the general case.

Exercise 11.8.

(i) Use substitution to integrate $\cos(5x)$*.*

(ii) Suppose you know the anti-derivative for f; say $f = h'$. Use subsitution to work out the general formula for $\int f(cx)\,dx$.

When evaluating a definite integral you can compute the indefinite integral as above and then evaluate. A second option is to change variables, including the limit of integration, and then never change back.

Example 11.13. Compute $\displaystyle\int_1^2 \frac{x}{x^2+1}\,dx$.
If we let $u = x^2+1$ then $du = 2x\,dx$, so the integrand becomes $(1/2)\,du/u$. If x goes from 1 to 2 then u goes from 2 to 5, thus the integral becomes

$$\int_2^5 \frac{1}{2}\frac{du}{u} = \frac{1}{2}(\ln 5 - \ln 2)\,.$$

Of course you can get the same answer in the usual way: the indefinite integral is $(1/2)\ln u$; we substitute back and get $(1/2)\ln(x^2+1)$. Now we evaluate at 2 and 1 instead of 5 and 2, but the result is the same: $(1/2)(\ln 5 - \ln 2)$.

Backwards substitution

There are times when the best substitution is of the form $x = g(u)$ rather than $u = g(x)$. No matter what f and g are, the substitution $x = g(u)$, $dx = g'(u)\,du$ always leads to a new integral, it's just hard to choose g in a way that makes the new integral simpler than the old one. It turns out there are some integrals, not apparently involving trig functions, where substituting $x = g(u)$ for some trig function g will magically unlock a dead end. Knowing tricks for dealing with a wide class of anti-derivative extractions is not the aim of this course, therefore we will not be featuring this method in the text. If you're interested in seeing one of these, try googling "integrate sqrt(1-xˆ2)".

Looking it up

Math is about understanding relations of a precise nature, about abstraction, and about making models of physical phenomena. It is also about building a library of computational tricks, but that's only a small part of math, and it's somewhat time-consuming. We have taught you what we think it is reasonable for you to know and remember – to have in

your quick-access library. For all the other integrals currently known to mankind, there are lookup tables. The following integral table is stolen from a popular calculus book. Use it as a handy reference, as needed.

1. $\int k\,dx = kx + C$ (any number k)

2. $\int x^n\,dx = \frac{x^{n+1}}{n+1} + C$ $(n \neq -1)$

3. $\int \frac{dx}{x} = \ln|x| + C$

4. $\int e^x\,dx = e^x + C$

5. $\int a^x\,dx = \frac{a^x}{\ln a} + C$ $(a > 0,\ a \neq 1)$

6. $\int \sin x\,dx = -\cos x + C$

7. $\int \cos x\,dx = \sin x + C$

8. $\int \sec^2 x\,dx = \tan x + C$

9. $\int \csc^2 x\,dx = -\cot x + C$

10. $\int \sec x \tan x\,dx = \sec x + C$

11. $\int \csc x \cot x\,dx = -\csc x + C$

12. $\int \tan x\,dx = \ln|\sec x| + C$

13. $\int \cot x\,dx = \ln|\sin x| + C$

14. $\int \sec x\,dx = \ln|\sec x + \tan x| + C$

15. $\int \csc x\,dx = -\ln|\csc x + \cot x| + C$

16. $\int \sinh x\,dx = \cosh x + C$

17. $\int \cosh x\,dx = \sinh x + C$

18. $\int \frac{dx}{\sqrt{a^2 - x^2}} = \sin^{-1}\left(\frac{x}{a}\right) + C$

19. $\int \frac{dx}{a^2 + x^2} = \frac{1}{a}\tan^{-1}\left(\frac{x}{a}\right) + C$

20. $\int \frac{dx}{x\sqrt{x^2 - a^2}} = \frac{1}{a}\sec^{-1}\left|\frac{x}{a}\right| + C$

21. $\int \frac{dx}{\sqrt{a^2 + x^2}} = \sinh^{-1}\left(\frac{x}{a}\right) + C$ $(a > 0)$

22. $\int \frac{dx}{\sqrt{x^2 - a^2}} = \cosh^{-1}\left(\frac{x}{a}\right) + C$ $(x > a > 0)$

12 Integrals over the whole real line

12.1 Definitions

The situation when integrating out to infinity is similar to the situation with infinite sums. Because there is no already assigned meaning for summing infinitely many things, we **defined** this as a limit, which in each case needs to be evaluated:

$$\sum_{k=1}^{\infty} a_k \qquad := \qquad \lim_{M\to\infty} \sum_{k=1}^{M} a_k \,.$$

It is the same when one tries to integrate over the whole real line. We define such integrals by integrating over a bigger and bigger piece and taking the limit. In fact the definition is even pickier than that. We only let one of the limits of integration go to zero at a time. Consider first an integral over a half-line $[a, \infty)$.

Definition 12.1 (one-sided intergal to infinity)**.** *Let a be a real number and let f be a continuous function on the infinite interval $[a, \infty)$. We define*

$$\int_a^{\infty} f(x)\,dx := \lim_{M\to\infty} \int_0^M f(x)\,dx \,. \tag{12.1}$$

One-sided infinite integrals $(-\infty, b]$ are defined similarly:

$$\int_{-\infty}^{b} f(x)\,dx := \lim_{M\to-\infty} \int_M^b f(x)\,dx \,.$$

Exercise 12.1. *Write down the defining limit for $\displaystyle\int_{-\infty}^{3} e^x\,dx$ and evaluate the limit.*

We remark that you can often substitute ∞ into the antiderivative and subtract: $\displaystyle\int_1^{\infty} dx/x^2 = (-1/x)|_1^{\infty} = 0 - (1) = 1$. If the value of $-1/(\infty)$ were not obvious, you would need limits.

Aside. *When you say $-1/\infty = 0$, recalling Definition 6.1, you are really saying $\lim_{M\to\infty} -1/M = 0$ and then quickly evaluating that limit in your head.*

If we want both limits to be infinite then we require the two parts to be defined separately.

Definition 12.2 (two-sided integral to infinity)**.** *Let a be a real number and Let f be a continuous function on the whole real line. Pick a real number c and define*

$$\int_{-\infty}^{\infty} f(x)\,dx := \int_{-\infty}^{c} f(x)\,dx + \int_{c}^{\infty} f(x)\,dx\,. \tag{12.2}$$

If either of these two limits is undefined, the whole integral is said not to exist.

Example 12.3. What is $\displaystyle\int_{-\infty}^{\infty} \frac{x}{x^2+1}$? Choosing $c = 0$, we see it is the sum of two one-sided infinite integrals $\int_0^\infty x/(x^2+1)\,dx + \int_{-\infty}^0 x/(x^2+1)\,dx$. Going back to the definition replaces each one-sided infinite integral by a limit:

$$\lim_{M\to\infty} \int_0^M \frac{x}{x^2+1}\,dx + \lim_{M\to\infty} \int_{-M}^0 \frac{x}{x^2+1}\,dx\,.$$

It looks as if this limit is coing to come out to be zero because $x/(x^2+1)$ is an odd function. Integrating from $-M$ to M will produce exactly zero, therefore

$$\lim_{M\to\infty} \int_{-M}^{M} \frac{x}{x^2+1}\,dx = \lim_{M\to\infty} 0 = 0\,. \tag{12.3}$$

Be careful! The definition says not to evaluate (12.3) but rather to evaluate the two one-sided integrals separately and sum them. We will come back to finish this example later.

At this point you should be bothered by three questions.

1. What is c? Does it matter? How do you pick it?
2. If we get $-\infty + \infty$, shouldn't that possibly be something other than "undefined"?
3. Why do we have to split it up in the first place?

The answer to the first question is, pick c to be anything, you'll always get the same answer. This is important because otherwise, what we wrote isn't really a definition. The reason the integral does not depend on c is that if one changes c from, say, 3 to 4, then the first of the two integrals loses a piece: $\int_3^4 f(x)\,dx$. But the second integral gains this same piece, so the sum is unchanged. This is true even if one or both pieces is infinite. Adding or subtracting the finite quantity $\int_3^4 f(x)\,dx$ won't change that.

The answer to the second question is yes, sometimes you can be more specific. The one-sided integral to infinity is a limit. Cases where a finite limit does not exist can be resolved into limits of ∞ or $-\infty$, along with the remaining cases where no limit exists even allowing for infinite limits. Because integrals over the whole real line are sums of one-sided (possibly infinite) limits, the rules for infinity from Sections 3.2 and 6.2 can be applied. In other words, integrals over the whole real line are the sum of two one-sided limits; we can add real numbers and $\pm\infty$ according to the rules in Definition 6.1: $\infty + \infty = \infty$ (and analogously with $-\infty$), $\infty + a = \infty$ when a is real (and analogously with $-\infty$), $\infty - \infty = UND$, $UND +$ anything $= UND$, and so on.

The third question is also a matter of definition. The reason we make the choice to do it this way is illustrated by the integral of the sign function

$$f(x) = \operatorname{sign}(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$$

On one hand, $\int_{-M}^{M} f(x)\,dx$ is always zero, because the postive and negative parts exactly cancel. On the other hand, $\int_{M}^{\infty} f(x)\,dx$ and $\int_{-\infty}^{M} f(x)\,dx$ are always undefined. Do we want the answer for the whole integral $\int_{-\infty}^{\infty} f(x)\,dx$ to be undefined or zero? There is no intrinsically correct choice here but it is a lot safer to have it undefined. If it has a value, one could make a case for values other than zero by centering the integral somewhere else, as in the following exercise.

Exercise 12.2. *What is* $\lim_{M\to\infty} \int_{7-M}^{7+M} \operatorname{sign}(x)\,dx$*?*

Example 12.4. The function $\sin(x)/x$ is not defined at $x = 0$ but you might recall it does have a limit at 0, namely $\lim_{x\to 0} \sin(x)/x = 1$. Therefore the function

$$\operatorname{sinc}(x) := \begin{cases} \sin(x)/x & x \neq 0 \\ 1 & x = 0 \end{cases}$$

is a continuous function on the whole real line. Its graph is shown in Figure 44. To write down a limit that defines this integral, we first choose any c. Choosing $c = 0$ makes things symmetric. The integral is then defined as the sum of two integrals, $\int_{-\infty}^{0} \operatorname{sinc}(x)\,dx + \int_{0}^{\infty} \operatorname{sinc}(x)\,dx$. Going back to the definition of one-sided integrals as limits, this sum of integrals is equal to

$$\lim_{M\to-\infty} \int_{M}^{0} \operatorname{sinc}(x)\,dx + \lim_{M\to\infty} \int_{0}^{M} \operatorname{sinc}(x)\,dx\,.$$

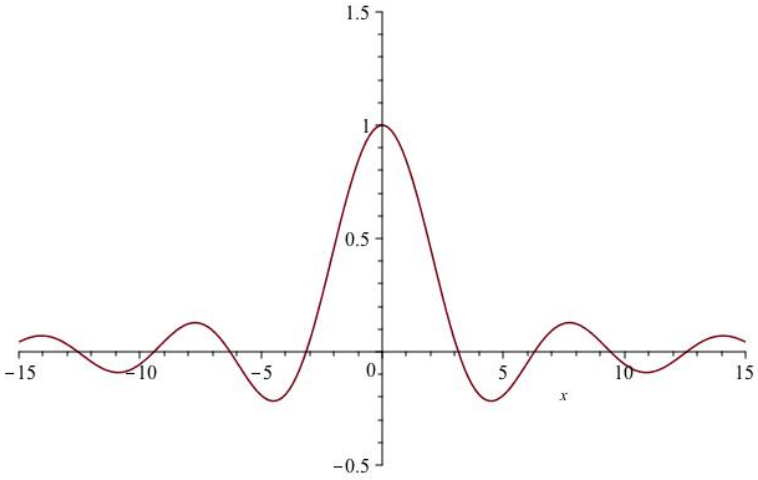


Figure 44: graph of the function sinc

It is not obvious whether these limits exist. One thing is easy to discern: because sinc is an even function, the two limits have the same value (whether finite or not). We can safely say:

$$\int_{-\infty}^{\infty} \operatorname{sinc}(x)\,dx = 2 \cdot \lim_{M\to\infty} \int_{0}^{M} \operatorname{sinc}(x)\,dx\,.$$

Exercise 12.3. *Evaluate $\int_{-\infty}^{\infty} x\,dx$ by writing down the definition via limits and then evaluating.*

12.2 Convergence

The central question of this section is: how do we tell whether a limit such as $\int_b^\infty f(x)\,dx$ exists. If so, we would like to evaluate it if possible, and estimate it otherwise. When discussing convergence you should realize that $\int_a^\infty f(x)\,dx$ either diverges for all values of a or converges for all values of a as long as f is defined and continuous on $[a,\infty)$. For this reason, we use the notation $\int^\infty f(x)\,dx$ or, to be really blunt, $\displaystyle\int_{\text{who cares}}^{\infty} f(x)\,dx$.

Exercise 12.4. *Explain the "you should realize" comment in a concrete context by stating a reason why $\displaystyle\int_2^\infty e^{-3\ln(\ln x)}\,dx$ converges if and only if $\displaystyle\int_6^\infty e^{-3\ln(lnx)}\,dx$ converges. Hint: remember the questions we said should bother you, "What is c? Does it matter?"*

Case 1: you know how to compute the definite integral

Suppose $\int_b^M f(x)\,dx$ is something for which you know how to compute an explicit formula. The formula will have M in it. You have to evaluate the limit as $M \to \infty$. How do you do that? There is no one way, but that's why we studied limits before. Apply what you know. What about b, do you have to take a limit in b as well? I hope you already knew the answer to that. In this definition, b is any fixed number. You don't take a limit.

These special cases will become theorems once you have worked them out.

Name of test	Type of integral	Condition for convergence
	$\int_b^\infty e^{kx}\,dx$	
power test	$\int_b^\infty x^p\,dx$	
	$\int_b^\infty \frac{(\ln x)^q}{x}\,dx$	

You will work out these cases in class: write each as a limit, evaluate the limit, state whether it converges, which will depend on the value of the parameter, k, p or q. Go ahead and pencil them in once you've done this. The second of these especially, is worth remembering because it is not obvious until you do the computation where the break should be between convergence and not.

Exercise 12.5. *Work out the first special case: for what real k does the integral converge?*

Case 2: you don't know how to compute the integral

In this case you can't even get to the point of having a difficult limit to evaluate. So probably you can't evaluate the improper integral. But you can and should still try to answer whether the integral has a finite value versus being undefined. This is where comparison tests come in. You build up a library of cases where you do know the answer and then, for the rest of functions, you try to compare them to functions in your library.

Sometimes a comparison is informative, sometimes it isn't. Suppose that f and g are positive functions and $f(x) \le g(x)$ for all x. Consider several pieces of information you might have about these functions.

	Comparison tests	
(a)	$\int_b^\infty f(x)\,dx$ converges to a finite value L.	conclusion:
(b)	$\int_b^\infty f(x)\,dx$ does not converge.	conclusion:
(c)	$\int_b^\infty g(x)\,dx$ converges to a finite value L.	conclusion:
(d)	$\int_b^\infty g(x)\,dx$ does not converge.	conclusion:

In which cases can you conclude something about the other function? We are doing this in class. Once you have the answer, either by working it out yourself or from the class discussion, please pencil it in here so you'll have it for later reference.

Exercise 12.6. *Suppose you want to show that* $\displaystyle\int_1^\infty \frac{3+sin(x)}{x^2}\,dx$ *converges. Which pair of facts allows you to do this?*

(a) $\frac{3+\sin x}{x^2} \ge \frac{2}{x^2}$ *and* $\int^\infty \frac{2}{x^2}\,dx$ *converges*

(b) $\frac{3+\sin x}{x^2} \le \frac{4}{x^2}$ *and* $\int^\infty \frac{4}{x^2}\,dx$ *does not converge*

(c) $\frac{3+\sin x}{x^2} \le \frac{4}{x^2}$ *and* $\int^\infty \frac{4}{x^2}\,dx$ *converges*

(d) $\frac{3+\sin x}{x^2} \le \frac{4}{x^2}$ *and* $\int^\infty \frac{2}{x^2}\,dx$ *does not converge*

Asymptotic comparison tests

Here are two key ideas that help your comparison tests work more of the time, based on the fact that the question "convergence or not?" is not sensitive to certain things.

(1) Multiplying by a constant does not change whether an integral converges. That's because if $\lim_{M\to\infty} \int_b^M f(x)\,dx$ converges to the finite constant L then $\lim_{M\to\infty} \int_b^M Kf(x)\,dx$ converges to the finite constant KL.

Exercise 12.7. *Does* $\displaystyle\int^\infty \frac{10}{x}\,dx$ *converge or not? In either case, give a reason why. If it converges, say to what. If it does not converge, is the value* ∞ *or* $-\infty$ *or is it truly undefined?*

(2) It doesn't matter if $f(x) \le g(x)$ for every single x as long as the inequality is true *for sufficiently large* x. For example, if $f(x) \le g(x)$ once $x \ge 100$, then you can apply the comparison test to compare $\int_b^\infty f(x)\,dx$ to $\int_b^\infty g(x)\,dx$ as long as $b \ge 100$. But even if not, once you compare $\int_{100}^\infty f(x)\,dx$ to $\int_{100}^\infty g(x)\,dx$, then adding the finite quantity $\int_b^{100} f(x)\,dx$ or $\int_b^{100} g(s)\,dx$ will not change whether either of these converges.

Putting these two ideas together leads to the conclusion that if $f(x) \le Kg(x)$ from some point onward and $\int_b^\infty g(x)\,dx$ converges, then so does $\int_b^\infty f(x)\,dx$. The theorem we just proved is:

Theorem 12.5 (asymptotic comparison)**.** *If f and g are positive functions on some interval (b,∞) and if there are some constants M and K such that*

$$f(x) \le Kg(x) \text{ for all } x \ge M \tag{12.4}$$

then convergence of the integral $\int_b^\infty g(x)\,dx$ implies convergence of the integral $\int_b^\infty f(x)\,dx$.

In particular, if $f(x) \ll g(x)$ as $x \to \infty$ then (12.4) holds, hence convergence of the integral $\int_b^\infty g(x)\,dx$ implies convergence of the integral $\int_b^\infty f(x)\,dx$.

Exercise 12.8. *Let $f(x) := 3x^3/(x-17)$ and $g(x) := x^2$. Is it true that $f(x) \le Kg(x)$ from some point onward? Explain.*

Example 12.6 (power times negative exponential)**.** Does $\displaystyle\int_1^\infty x^8 e^{-x}\,dx$ converge? One way to do this is by computing the integral exactly. This takes eight integrations by parts, and is probably too messy unless you figured out how to do "tabular" integration (optional when you learned integration by parts). In any case, there's an easier way if you only want to know whether it converges, but not to what.

We claim that $x^8 e^{-x} \ll e^{-(1/2)x}$ (you could use $e^{-\beta x}$ in this argument for any $\beta \in (0,1)$). It follows from the asymptotic comparison test that convergence of $\int_1^\infty e^{-(1/2)x}$ implies convergence of $\int x^8 e^{-x}\,dx$. We check the claim by evaluating

$$\lim_{x\to\infty} \frac{x^8 e^{-x}}{e^{-(1/2)x}} = \lim_{x\to\infty} \frac{x^8}{e^{(1/2)x}} = 0$$

because we know the power x^8 is much less than the exponential $e^{(1/2)x}$.

Exercise 12.9. *Does $\displaystyle\int_{18}^\infty \frac{x^3}{x-17} e^{-x}\,dx$ converge? [You can use the reuslt of Exercise 12.8.]*

A particular case of Theorem 12.5 is when $f(x) \sim g(x)$. When two functions are asymptotically equivalent, then each can be upper bounded by a constant multiple of the other, hence we have the following proposition.

Proposition 12.7. *If f and g are positive functions and $f \sim g$ then $\int^{\infty} f(x)\,dx$ converges if and only if $\int^{\infty} g(x)\,dx$ converges.*

Example 12.8.

(*i*) Does $\displaystyle\int_1^{\infty} \frac{dx}{x^2+3x}$ converge?

ANSWER: We can use comparison test (c) here: $\dfrac{1}{x_2+3x} \leq \dfrac{1}{x^2}$ and we know $\displaystyle\int_1^{\infty} \frac{dx}{x^2}$ converges, hence so does $\displaystyle\int_1^{\infty} \frac{dx}{x^2+3x}$.

(*ii*) Does $\displaystyle\int_4^{\infty} \frac{dx}{x^2-3x}$ converge?

ANSWER: Now the inequatlity goes the other way, so we are in case (c) of the comparison test and we cannot conclude anything from direct comparison. However, we also know $\dfrac{1}{x^2-3x} \sim \dfrac{1}{x^2}$ as $x \to \infty$, therefore we can conclude convergence again by Proposition 12.7.

Did you wonder about the lower limit of 4 in part (*ii*)? That wasn't just randomly added so you'd be more flexible about the lower limits of integrals to infinity. It was put there to ensure that f was continuous; note the discontinuity at $x = 3$.

Exercise 12.10. *Find a simple function g such that $(3x+\cos(x))/x^3 \sim g(x)$ as $x \to \infty$. Then determine whether $\displaystyle\int_1^{\infty} \frac{3x+\cos x}{x^3}\,dx$ converges.*

12.3 Probability densities

Students have varied backgrounds when it comes to probability. A few have taken courses in probability. Most have seen a little probability theory in high school. Some have never studied anything to do with probability. Because of the varied backgrounds, we take a couple of paragraphs to discuss the key concepts.

The first thing students usually learn is **discrete probability**, where the random variables take values in a finite set, with given probabilities for each outcome. That's because this

can be studied with middle school mathematics. For example, rolling two 6-sided dice leads to 36 possible outcomes, each equally likely; this in turn leads to 11 possible outcomes for the sum of the two dice, with probabilities ranging from 1/36 for 2 and 12 to 6/36 for 7. All questions about rolls of finitely many dice can be answered with careful analysis and basic arithmetic.

Random variables whose values are spread over all real numbers, or a real interval, require calculus to define and study. These are called **continuous** random variables, and are the topic of this section.

Philosophically, a real-valued random variable X is a quantity that has a value equal to some real number, but will have a different value each time some kind of experiment is run. It is unpredictable, therefore we cannot answer the question "What is the value of X?" but only "What is the probability that the value of X lies in the set A?" For example, suppose we throw a dart at a 12 foot wide wall, from a long enough distance and with poor enough aim that it is as likely to hit any region as any other (if we miss completely, we get another try). Say the random variable X is the distance (in feet) from the left edge of the wall. We can ask for the probability that $X \leq 2$, that is that the dart lands within two feet of the left edge.

Exercise 12.11. *What should this probability be? Forget about calculus, just use intuition.*

For discrete random variables you answer this type of question by summing the probability that X is equal to y for every y in the set A. For continuous random variables, the probability of being equal to any one real number is zero. In the example with the dart, the probability that it lands exactly $\sqrt{3}$ feet from the left edge (or 1 foot, or 1/3 of a foot, or any other real number of feet) is zero. The only way to get a nonzero probability is to consider an entire interval of values. Thus the most basic questions we ask about X are: what is the probability that $X \in [a, b]$, where $a < b$ are fixed real numbers. These probabilities will be governed by a **probability density**, which is a nonnegative function telling how likely it is for X to be in an interval centered at any given real number.

Definition 12.9 (probabilitiy densities)**.**

1. *A probability density is a nonnegative function f such that $\int_{-\infty}^{\infty} f(x)\,dx = 1$.*
2. *A random variable X is said to have probability density f if the probability of finding X in any interval $[a, b]$ is equal to $\int_a^b f(t)\,dt$.*
3. *We denote the probability of finding X in $[a, b]$ by $\mathbb{P}(X \in [a, b])$.*

Exercise 12.12. *Why do we require f to integrate to 1?*

Sometimes f is defined only on an interval $[a, b]$ and not on the whole real line. The interpretation is that the random variable X takes values only in $[a, b]$. Probabilities for X are then defined by integrating in sub-intervals of $[a, b]$. Often one extends the definition of f to all real numbers by making it zero off of $[a, b]$. This may result in f being discontinuous but its definite integrals are still defined.

Example 12.10. The standard exponential random variable has density e^{-x} on $[0, \infty)$. If X has this density, what is $\mathbb{P}(X \in [-1, 1])$? This is the same as $\mathbb{P}(X \in [0, 1])$, because X cannot be negative. We compute it by $\displaystyle\int_0^1 e^{-x}\,dx = e^{-x}\big|_0^1 = 1 - e^{-1}$. As a quick reality check we observe that the quantity $1 - \frac{1}{e}$ is indeed between zero and one, therefore it makes sense for this to be a probability.

Exercise 12.13. *Write the statement $X \geq m$ as a statement about X being in a (possibly infinite) interval. Letting f be the probability density of X, write an integral computing $P(X \geq m)$.*

Often the model dictates the form of the function f but not a multiplicative constant.

Example 12.11. For example, if we know that $f(x)$ should be of the form Cx^{-3} on $[1, \infty)$ then we would need to find the right constant C to make this a probability density. The function f has to integrate to 1, meaning we have to solve

$$\int_1^\infty Cx^{-3}\,dx = 1$$

for C. Solving this results in $C = 2$, therefore the density of f is $2/x^3$ on $[1, \infty)$.

Exercise 12.14. *Suppose X has density proportional to $\cos(x)$ on the interval $[-\pi/2, \pi/2]$. What value of C makes $C \cos x$ a probability density on this interval?*

Several important quantities associated with a probability distribution are the mean, the variance, the standard deviation and the median. Again, a couple of paragraphs don't do justice to these ideas, but we hope they explain the concepts at least a little and make the math seem more motivated and relevant.

Probably the simplest concept intuitively is the median. This is the 50^{th} percentile of the distribution.

Definition 12.12. *The* **median** *of a random variable X having probability density f is the real number m such that*

$$\mathbb{P}(X > m) = \mathbb{P}(X < m) = \frac{1}{2}\,. \tag{12.5}$$

Exercise 12.15. *Write (12.5) as an equation with an integral in it.*

Definition 12.13.

1. *If X has probability density f, the* **mean** *or* **expectation** *of X (the two terms are synonyms) is the quantity $\mathbb{E}X := \int_{-\infty}^{\infty} x\, f(x)\, dx$. A variable commonly used for the mean of a distribution is μ.*
2. *If X has probability density f and mean μ, the* **variance** *of X is the quantity*

$$\mathrm{Var}\,(X) := \int_{-\infty}^{\infty} (x-\mu)^2\, f(x)\, dx\,.$$

3. *The* **standard deviation** *of X is the quantity*

$$\sigma := \sqrt{\mathrm{Var}\,(X)}\,.$$

To understand these intuitively, you might recall what happens when rolling a die. Each of the six numbers comes up about $1/6$ of the time, so in a large number N of dice rolls you will get about $N/6$ of each of the six outcomes. The average will therefore be

$$\frac{1}{N}\left[(N/6)\cdot 1 + (N/6)\cdot 2 + (N/6)\cdot 3 + (N/6)\cdot 4 + (N/6)\cdot 5 + (N/6)\cdot 6\right]\,.$$

We can write this in summation notation as

$$\sum_{j=1}^{6} j\cdot \mathbb{P}(X=j)\,.$$

Exercise 12.16. *A carnival game that costs a dollar to play gives you a quarter for each spot on a roll of a die (e.g., 75 cents if you roll a 3). When you have spent N dollars, about how many quarters will you have received?*

When instead there are infinitely many possible outcomes spread over an interval, the sum is replaced by an integral

$$\int_{-\infty}^{\infty} x\cdot f(x)\, dx\,.$$

A famous theorem in probability theory, called the Strong Law of Large Numbers, says that this still computes the long term average: the long term average of independent draws from a distribution with probability density function f will converge to $\int x \cdot f(x)\,dx$.

Exercise 12.17. *The random variable X has probability density $2x$ on $[0,1]$. If you sample a million times and take the average of the samples, roughly what will you get?*

It is more difficult to understand why the variance has the precise definition it does, but it is easy to see that the formula produces bigger values when the random variable X tends to be farther from its mean value μ. The standard deviation is another measure of dispersion. To see why it might be more physically relevant, consider the units.

Probabilities such as $\mathbb{P}(X \in [a,b])$ can be considered to be unitless because they represent ratios of like things: frequency of occurrences within the interval $[a,b]$ divided by frequency of all occurrences. Probability densities, integrated against the variable x (which may have units of length, time, etc.) give probabilities. Therefore, probability densities have units of "probability per unit x-value", or in other words, inverse units to the independent variable.

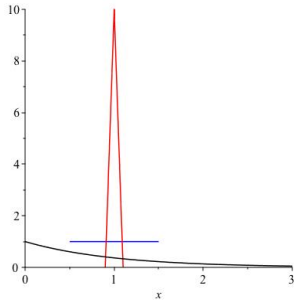
The units of the mean are units of $\int xf(x)\,dx$, which is units of f times x^2; but f has units of inverse x, so the mean has units of x. This makes sense because the mean represents a point on the x-axis. Similarly, the variance has units of x^2. It is hard to see what the variance represents physically. The standard deviation, however, has units of x. Therefore, it is a measure of dispersion having the same units as the mean. It represents a distance on the x-axis which is some kind of average discrepancy from the mean[14].

Exercise 12.18. *Here are three probability densities with mean 1. Rank them in order from greatest to least standard deviation. You don't have to compute precisely unless you want to; just state an answer and justify it intuitively. The three densities are graphed to the right.*

(a) $f(x) := 1$ *on* $[1/2, 3/2]$ *(blue)*

(b) $f(x) := 10 - 100|x-1|$ *on* $[0.9, 1.1]$ *(red)*

(c) $f(x) := e^{-x}$ *on* $[0,\infty]$ *(black)*



[14] To be precise, a root-mean-square discrepancy.

Some common probability densities

There are a zillion different functions commonly used for probability densities. Three of the most common are named in this section: the exponential, the uniform, and the normal. These are common in probability for reasons analogous to why exponential behavior is common in evolving systems. They come from simple properties.

The uniform, as the name applies, arises when a random quantity is uniformly likely to be anywhere in an interval. It is often used as an "uninformed" model when all you know is that a quantity has to be somewhere in a fixed interval. The normal arises when many small independent contributions are summed. It is often used to model observational error. The exponential is the so-called memoryless distribution. It arises when the probability of finding X in the next small interval, given that you haven't already found it, is always constant.

All three of these are parametrized families of distributions. Once values are picked for the parameters you get a particular distribution. This section concludes by giving definitions of eac h and discuss typical applications.

The exponential distribution

The exponential distribution has a parameter μ which can be any positive real number. Its density is $(1/\mu)e^{-x/\mu}$ on the positive half-line $[0, \infty)$. This is obviously the same as the density Ce^{-Cx} (just take $C = 1/\mu$) but we use the parameter μ rather than C because a quick computation shows that the mean of the distribution is equal to μ.

Exercise 12.19. *Integrate by parts with $u = x$ and $dv = \mu^{-1}e^{-x/\mu}$ to show that the mean of the exponential with parameter μ is μ. Don't forget to write integrals to ∞ as limits.*

The exponential distribution has a very important "memoryless" propoerty. If X has an exponential density with any parameter and is interpreted as a waiting time, then once you know it didn't happen by a certain time t, the amount of further time it will take to happen has the same distribution as X had originally. It doesn't get any more or any less likely to happen in the the interval $[t, t + 1]$ than it was originally to happen in the interval $[0, 1]$.

The median of the exponential distribution with mean μ is also easy to compute. Solving $\int_0^M \mu^{-1}e^{-x/\mu}\,dx = 1/2$ gives $M = \mu \cdot \ln 2$. When X is a random waiting time, the interpreta-

tion is that it is equally likely to occur before $\ln 2$ times its mean as after. Because $\ln 2 \approx 0.7$, the median is significantly less than the mean. When modeling with exponentials, it is good to remember it produces values that are unbounded but always positive.

Any of you who have studied radioactive decay know that each atom acts randomly and independently of the others, decaying at a random time with an exponential distribution. The fraction remaining after time t is the same as the probability that each individual remains undecayed at time t, namely $e^{-t/\mu}$, so another interpretation for the median is the **half-life**: the time at which only half the original substance remains. Other examples are the life span of an organism that faces environmental hazards but does not age, or time for an electronic component to fail (they don't seem to age either).

The uniform distribution

The uniform distribution on the interval $[a, b]$ is the probability density whose density is a constant on this interval: the constant will be $1/(b-a)$. This is often thought of the least informative distribution if you know that the the quantity must be between the values a and b. The mean and median are both $(a+b)/2$.

Aside. *The uniform distribution is less common in nature than the exponential or normal. On the other hand, if you ask a computer to generate a random number in some range, it will pick from the uniform distribution unless you program it otherwise.*

Exercise 12.20. *Use calculus to prove that a constant function C on an interval $[a, b]$ is a probability density if and only if $C = 1/(b-a)$.*

Example 12.14. In your orienteering class you are taken to a far away location and spun around blindfolded when you arrive. When the blindfold comes off, you are facing at a random compass angle (usually measured clockwise from due north). It would be reasonable to model this as a uniform random variable from the interval $[0, 360]$ in units of degrees.

Exercise 12.21. *The mean and median are both $180°$. Why are these not meaningful measures of the center of the distribution in this case?*

The normal distribution

The normal density with mean μ and standard deviation σ is the density

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} .$$

The **standard normal** is the one with $\mu = 0$ and $\sigma = 1$. There is a very cool mathematical reason for this formula, which we will not go into. When a random variable is the result of summing a bunch of smaller random variables all acting independently, the result is usually well approximated by a normal. It is possible (though very tricky) to show that the definite integral of this density over the whole real line is in fact 1 (in other words, that we have chosen the right constant to make it a probability density).

Annoyingly, there is no nice antiderivative, so no way in general of computing the probability of finding a normal between specified values a and b. Because the normal is so important in statistical applications, they made up a notation for the indefinite integral in the case $\mu = 0, \sigma = 1$, using the capital Greek letter Phi:

$$\Phi(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx .$$

So now you can say that the probability of finding a standard normal between a and b is exactly $\Phi(b) - \Phi(a)$. In the old, pre-computer days, they published tables of values of Φ. It was reasonably efficient to do this because you can get the antiderivative F of any other normal from the one for the standard normal by a linear substition: $F(x) = \Phi((x - \mu)/\sigma)$.

13 Taylor polynomials

13.1 Approximating functions by polynomials

Polynomials are simpler than most other functions. This leads to the idea of approximating a complicated function by a polynomial. Taylor realized that this is possible provided there is an "easy" point at which you know how to compute the function and its derivatives. Given a function $f(x)$ and a value a, we will define for each degree n a polynomial $P_n(x)$ which is the "best n^{th} degree polynomial approximation to $f(x)$ near $x = a$."

It pays to start very simply. A zero-degree polynomial is a constant. What is the best constant approximation to $f(x)$ near $x = a$? Clearly, the constant $f(a)$. What is the best linear approximation? We already know this, and have given it the notation $L(x)$. It is the tangent line to the graph of $f(x)$ at $x = a$ and its equation is $L(x) = f(a) + f'(a)(x - a)$. So now we know that

$$\begin{aligned} P_0(x) &= f(a) \\ P_1(x) &= f(a) + f'(a)(x - a) \end{aligned}$$

We illustrate this pictorially as follows.

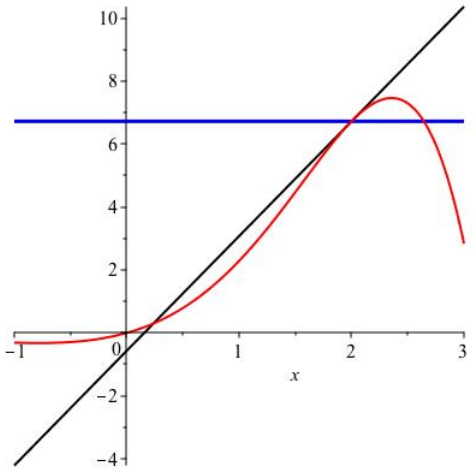


Figure 45: A function (red), its constant (blue), and linear (black) approximations at $x = 2$

Figure 45 shows the graph of a function f along with its zeroth and first degree Taylor polynomials at $x = 2$. The zeroth degree polynomial is the flat line and the first degree Taylor polynomial is the tangent line. To refresh your memory on how well these approximate $f(x)$ near $x = a$, you might want to look back at Proposition 8.3 and Exercise 8.9.

Exercise 13.1. *Suppose $f'(a) \neq 0$, which is true in Figure 45, for example, at $a = 2$. Multiple choice question: How good an approximation is P_0 near $x = a$?*

(i) $f(x) - P_0(x) \sim a$

(ii) $f(x) - P_0(x) \sim x - a$

(iii) $f(x) - P_0(x) \sim f'(a) \cdot (x - a)$

(iv) $f(x) - P_0(x) \sim f'(a) \cdot (x - a)^2$

ALTERNATE VERSION: Let $f(x) := x^2$ and let $a = 2$. How good an approximation is $P_0(x) := 4$ to $f(x)$ as $x \to 2$?

(i) $f(x) - P_0(x) \sim a$, *in other words,* $x^2 - 4 \sim 2$

(ii) $f(x) - P_0(x) \sim x - a$, *in other words,* $x^2 - 4 \sim x - 2$

(iii) $f(x) - P_0(x) \sim f'(a) \cdot (x - a)$, *in other words,* $x^2 - 4 \sim 4(x - 2)$

(iv) $f(x) - P_0(x) \sim f'(a) \cdot (x - a)^2$, *in other words,* $x^2 - 4 \sim 4(x - a)^2$

To figure out the best degree-n polynomial approximation for all n, the one idea you need is that the polynomial P_n should match all the derivatives of f up through the first n (the zeroth being the value of f itself). Let's check we've already done this for P_0 and P_1. Check: P_0 was chosen to match the function value at a. Check: P_1 matches the first derivative because $P_1(x)$ is a line; it has the same derivative everywhere, $f'(a)$, chosen to match the derivative of f at the point a.

Figure 46 shows P_3, P_4 and P_5 at $x = 2$ for the same function, with P_5 shown in long dashes, P_4 in shorter dashes and P_3 in dots. As n grows, notice how P_n beceoms a better approximation and stays close to f (shown in red) for longer.

Proposition 8.3 showed that $|P_1 - f| \ll |x - a|$ near $x = a$ and Exercise 8.9 gave evidence that in fact $|P_1 - f|$ was on the scale of $|x - a|^2$, at least for a particular example. In the coming sections we will see that this is true in general, and in fact that $|P_n - f|$ is of the scale $|x - a|^{n+1}$ near $x = a$. This is one of the main motivations for studying Taylor approximations.

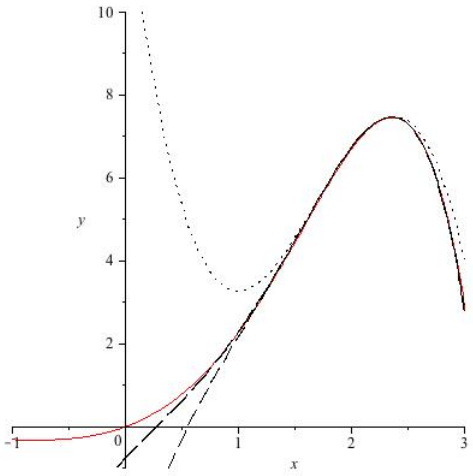


Figure 46: Successive Taylor approximations P_3 (dots), P_4 (short dashes), P_5 (long dashes) and f in red

Exercise 13.2. *The Taylor series for $1/x$ near $x = 1$ happens to obey the approximation $|P_n(x) - f(x)| \approx |x - a|^{n+1}$ very closely. About how many digits after the decimal point would the approximation $P_6(1.01)$ capture of the true value of $1/1.01$?*

13.2 Taylor's formula

There is a formula for computing P_n. It's easiest to see what's going on when computing Taylor polynomials near $x = 0$. The algebra for these is enough simpler that these Taylor polynomials have a different name. A Taylor polynomial near $x = 0$ is called a **MacLaurin** polynomial.

The formula for Taylor and MacLaurin polynomials uses some possibly unfamiliar notation: $f^{(k)}$ refers to the k^{th} derivative of the function f. This is better than f', f'', etc., because we can use it in a formula as k varies. In this notation, $f^{(0)}$ denotes f itself.

Proposition 13.1 (MacLaurin's formula)**.** *Let f be a function that is n times differentiable on an interval containing 0. The polynomial P_n whose 0^{th} through k^{th} derivatives match those of f is given by the formula*

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!} x^k \,. \tag{13.1}$$

Exercise 13.3. *Use formula* (13.1) *to compute* P_4 *near* $x = 0$ *for the function* $f(x) = \cos(x)$.

The reason it's easy to check MacLaurin's formula is that $(d/dx)^j x^k$ is a simple computation. When $j > k$, you get zero. When $j = k$ you get the constant $k!$. When $j < k$ you get $k\,(k-1)\cdots(k-j+1)x^{k-j}$ which may seem messy but at the value $x = 0$ is is zero.

Exercise 13.4. *What is the* 6^{th} *derivative evaluated at* $x = 0$ *of the polynomial* $10 + 11x + 12x^2 + 13x^3 + 14x^4 + 15x^5 + 16x^6 + 17x^7 + 18x^8$ *?*

Proof of MacLaurin's formula: Observe that P_n as defined by (13.1) is indeed a polynomial of degree at most n. Let's check that the j^{th} derivative of P_n matches the j^{th} derivative of f at the value $x = 0$ for each j from 0 to n. Taking the j^{th} derivative of each term and evaluating at $x = 0$ gives 0 for each term except the term $k = j$, which contributes $k!$ times $f^{(k)}(0)/k!$. This is equal to $f^{(k)}(0)$, therefore we have matched the j^{th} derivative of f at zero.

Taylor polynomials do the same thing as MacLaurin polynomials except at a point $x = a$ where a is not necessarily zero. The resulting polynomial P_n is a polynomial in x of degree n, so you could write it as $\sum_{k=0}^{n} b_k x^k$. However, it is much easier to check that the derivatives match those of f at the point a if you write it instead as a sum $\sum_{k=0}^{n} b_k (x-a)^k$. This is still a polynomial of degree at most n, now written in a way that makes it easier to evaluate repeated derivatives at the point a. In fact the same argument proves the following more general formula.

Proposition 13.2 (Taylor's formula)**.** *Let* a *be any real number and let* f *be a function that can be differentiated at least* n *times at the point* a*. The* **Taylor polynomial** *for* f *of order* n *about the point* a *is the polynomial* $P_n(x)$ *defined by*

$$P_n(x) := \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k\,. \tag{13.2}$$

Exercise 13.5. *Identify the free and the bound variables on the right-hand side of* (13.2)*. Do all the free variables appear on the left? What does that tell you about the notation* $P_n(x)$*?*

Remember to read this sort of thing slowly. Here is roughly the thought process you should go through when seeing MacLaurins's formula and Taylor's formula for the first time.

- It looks as if P_n is a polynomial in the variable x with $n+1$ terms.
- Really the polyonomial depends on both n and a. It should really be called $P_{n,a}(x)$.
- Taylor's formula generalizes MacLaurin's formula because when $a = 0$, the quantity $x - a$ is just x.
- The coefficients are the derivatives of f at zero divided by successive factorials.
- Hey, what's zero factorial? Oh, it's defined to be 1. Who knew?
- Hey, what's the zeroth derivative $f^{(0)}(a)$? Oh, it's just $f(a)$.
- The degree of $P_n(x)$ will be n unless the coefficient on the highest power $(x-a)^n$ is zero, in which case the degree will be less.

Next you should try a simple example.

Example 13.3. Let $f(x) := x$, with $n = 3$ and $a = 2$. The value of $f(a)$ is 2 and the first three derivatives of $f(x)$ are constants: 1, 0, 0. Therefore

$$P_3(x) = 2 + 1 \cdot (x-2) + \frac{0}{2!}(x-2)^2 + \frac{0}{3!}(x-2)^3 .$$

In other words, $P_3(x) = x$. Obviously P_4, P_5 and so on will also be x. Maybe this example was too trivial. But it does point out a fact: if f is a polynomial of degree d then the terms of the Taylor polynomial beyond degree d vanish because the derivatives of f vanish. In fact, $P_n(x) = f(x)$ for all $n \geq d$. When $a = 0$ the Taylor polynomials for $n < d$ are also pretty simple:

Proposition 13.4. *If $f(x) = \sum_{k=0}^{d} a_k x^k$ is a degree-d polynomial, then $P_n(x) = f(x)$ for $n \geq d$, while for $n < d$, $P_n(x) = \sum_{k=0}^{n} a_k x^k$.*

Exercise 13.6. *What are The Maclaurin polynomials P_0, P_1, P_2, P_3 and P_4 for $f(x) := x^2$?*

Example 13.5. $f(x) = e^x$, $n = 3$ and $a = 0$. We list the function and its derivatives out to the third one.

k	$f^{(k)}(x)$	$f^{(k)}(a)$	$\frac{f^{(k)}(a)}{k!}(x-a)^k$
0	e^x	1	1
1	e^x	1	x
2	e^x	1	$\frac{x^2}{2}$
3	e^x	1	$\frac{x^3}{6}$

Summing the last column we find that the cubic Maclaurin polynomial is given by $P_3(x) = 1 + x + x^2/2 + x^3/6$.

Example 13.6. Let $f(x) = \ln\sqrt{x}$ and expand around $a = 1$. We'll do the first two terms this time.

k	$f^{(k)}(x)$	$f^{(k)}(a)$	$\frac{f^{(k)}(a)}{k!}(x-a)^k$
0	$\ln\sqrt{x}$	0	0
1	$\frac{1}{2x}$	$\frac{1}{2}$	$\frac{1}{2}(x-1)$
2	$\frac{-1}{2x^2}$	$-\frac{1}{2}$	$-\frac{1}{4}(x-1)^2$

Summing the last column we find that $P_2(x) = \dfrac{x-1}{2} - \dfrac{(x-1)^2}{4}$. If you don't have a computing device and you need a quick estimate $\ln\sqrt{1.4}$, this is one you can do in your head (really!).

13.3 Computing Taylor polynomials

You can always compute a Taylor polynomial using the formula. But sometimes the derivatives get messy and you can save time and mistakes by building up from pieces. Taylor polynomials follow the usual rules for addition, multiplication and composition. If f and g have Taylor polynonmials P and Q of order n then $f + g$ has Taylor polynomial $P + Q$. This is easy to see because the derivative is just the sum of the derivatives. Furthermore, the order n Taylor polynomial for fg is $P \cdot Q$ (ignore terms of order higher than n). This is

because the product rule for the derivative of fg looks exactly like the rule for multiplying polynomials. I won't present a proof here but you can feel free to use this fact.

Some vocabulary: polynomials of degrees 0, 1, 2, 3 and 4 are called, respectively, **constant**, **linear**, **quadratic**, **cubic** and **quartic**. Thus, for example, when we refer to the "cubic Taylor polynomial", we mean P_3. Don't confuse this with "the Taylor polynomial that has 3 terms"! That could be a quadratic Taylor polynomial or it could be a Taylor polynomial of very high degree that only has three nonzero terms.

Example 13.7. What is the cubic Maclaurin polynomial for $e^x \sin x$? The respective cubic Taylor polynomials are $1 + x + x^2/2 + x^3/6$ and $x - x^3/6$. Multiplying these and ignoring terms with a power beyond 3 we get

$$P_3(x) = x\left(1 + x + \frac{x^2}{2}\right) - \frac{x^3}{6} \cdot 1 = x + x^2 + \frac{x^3}{3}\,.$$

Exercise 13.7. *What is the cubic Maclaurin polynomial for $(1 + x^2)e^x$?*

You can do the same thing with division, assuming you learned polynomial long division (this is useful? Who knew!). If you have Taylor series around a point a other than zero, you will be dealing with polynomials in $(x - a)$ rather than in x.

Example 13.8. To compute the quadratic Taylor polynomial for $\ln\sqrt{x}/(1+2x)$ at $x = 1$, rather than having to differentiate this twice, look back at Example 13.6 to see that the quadratic Taylor polynomial for the numerator is $0 + \frac{1}{2}(x-1) - \frac{1}{4}(x-1)^2$. For the denominator it is $3 + 2 \cdot (x-1) + 0 \cdot (x-1)^2$. We have written itthis way so you can see the coefficient sequences: $0, 1/2, -1/4$ in the numerator and $3, 2, 0$ in the denominator. Synthetic polynomial long division then gives a coefficient sequence of $0, 1/6, -7/36$, so

$$P_2(x) = \frac{1}{6}(x-1) - \frac{7}{36}(x-1)^2\,.$$

```
                0   1/6   −7/36
            ┌──────────────────
3   2   │ 0   1/2   − 1/4
          0     0
          ─────
                1/2   −1/4
                1/2   1/3
                ─────────
                      −7/12
```

Exercise 13.8. *What is the cubic Maclaurin polynomial for $e^x/(1 + x^2)$?*

Perhaps the most useful manipulation is composition. I will illustrate this by example. The Maclaurin polynomial for e^{x^2} is obtained by plugging in x^2 for x in the Maclaurin polynomial for $e^x : 1 + (x^2) + (x^2)^2/2! + \cdots$.

One last trick arises when computing the Taylor series for a function defined as an integral. Suppose $f(x) = \int_b^x g(t)\,dt$. Then $f'(x) = g(x)$ so if you know g and its derivatives, you know the derivatives of f. If g has no nice indefinite integral, then you don't know the value of f itself, except at one point, namely $f(b) = 0$. Therefore, a Taylor series at b is the most common choice for a function defined as $\int_b^x$ of another function.

Example 13.9. Suppose $f(x) = \int_1^x \sqrt{1+t^3}\,dt$. The Taylor series can be computed about the point $a = 1$. From $f'(x) = \sqrt{1+x^3}$, $f''(x) = 3x^2/(2\sqrt{1+x^3})$ we get

$$f(1) = 0, \quad f'(1) = \sqrt{2}, \quad f''(1) = 3/(2\sqrt{2})$$

and therefore $P_2(x) = \sqrt{2}(x-1) + \dfrac{3}{4\sqrt{2}}(x-1)^2$.

13.4 Approximating with Taylor polynomials

The next section gives precise statements about how closely Taylor polynomials approximate function values. For now, we will take this on faith and see how to use Taylor polynomials.

Example 13.10. What's a good approximation to $e^{0.05}$? The Maclaurin polynomial will provide a very accurate estimate with only a few terms. The linear approximation, 1.05, is already not bad. The quadratic approximation is

$$1 + 0.05 + (1/2)(0.05)^2 = 1 + 0.05 + 0.00125 = 1.05125\,.$$

The true value is $1.05127\ldots$ so the quadratic approximation is quite good!

Taylor series are sometimes useful in approximating integrals when you can't do the integral: you approximate the integrand by a Taylor polynomial, then integrate precisely (polynomial anti-deriviatives are easy to calculate).

Example 13.11. Is it easier to approximate $\int_0^{1/2} \cos(\pi x^2)\,dx$ via trapezoidal approximation or Taylor integration?

The Taylor approach starts by computing some P_n at some point in the interval. The midpoint 1/4 would probably give the greatest accuracy but computations would be messier.

Instead take $a = 0$. There, P_4 is easily computed by substituting πx^2 for x in the Maclaurin polynomial for cosine. Which one? $P_2(x) = 1 - x^2/2$ is good enough to get all terms up to degree 4 after the substitution: plugging in πx^2 for x gives

$$P_4(x) = 1 - \frac{\pi^2}{2}x^4 \,.$$

Integrating,

$$\int_0^{1/2} P_4(x) = \left(x - \frac{\pi^2}{10}x^5\right)\Big|_0^{1/2} \,.$$

This comes out to $1/2 - \pi^2/320 \approx 0.46916$ which is accurate to within 0.001. The trapezoidal appoximation gives roughly 0.464907 which is off by four times as much.

Exercise 13.9. *Estimate $\int_{-1}^{1} e^{-x^2}\,dx$ by integrating the quadratic Taylor polynomial exactly. How close do you get to the numerical answer of 1.49?*

13.5 Taylor's theorem with remainder

Aside. *This last section is ambitious. Given the circumstances of having to adapt to an online format, it's likely you won't get to it. If you do get through this section, you will have absorbed a good dose of mathematical reasoning. You will probably be a lot better prepared for further study in math than many students who place into higher courses.*

The central question for this section is, how good an approximation to f is P_n? We will give a rough answer and then a more precise one.

Rough answer: $P_n(x) - f(x) \sim K(x-a)^{n+1}$ near $x = a$. For example, the linear approximation P_1 is off from the actual value by a quadratic quantity $K(x-a)^2$. If x differs from a by about 0.1 then $P_1(x)$ will differ from $f(x)$ by something like 0.01 (we are being rough here and pretending $K = 1$). If x agrees with a to four decimal places, then $P_1(x)$ should agree with $f(x)$ to about eight places. Similarly, the quadratic approximation P_2 differs from f by a multiple of $(x-a)^3$, and so on.

You can skip the justification of this answer, but we thought we'd include the derivation for those who want it because it's just an application of L'Hôpital's rule. Once you guess that $P_n(x) - f(x) \sim K(x-a)^n$, you can verify it by starting with the equation

$$\lim_{x\to 0} \frac{f(x) - P_n(x)}{(x-a)^{n+1}} \,,$$

and repeatedly applying L'Hôpital's rule until the denominator is not zero at $x = a$. Because the derivatives of f and P_n at zero match through order n, it takes at least $n+1$ derivatives to get something nonzero, at which point the denominator has become the nonzero constant $(n+1)!$. The limit is therefore $f^{(n+1)}(a)/(n+1)!$, which may or may not be zero but is surely finite.

We know the Taylor polynomial is an order $(x-a)^{n+1}$ approximation but there is a constant K in the expression which could be huge. What about actual bounds can we obtain on $f(x) - P_n(x)$? These are given by Answer # 2, which is called Taylor's Theorem with Remainder.

Theorem 13.12 (Taylor's Theorem with Remainder)**.** *Let f be a function with $n+1$ continuous derivatives on and interval $[a, x]$ or $[x, a]$ and let P_n be the order n Taylor polynomial for f about the point a. Then*

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

for some c between a and x. This is illustrated in Figure 47.

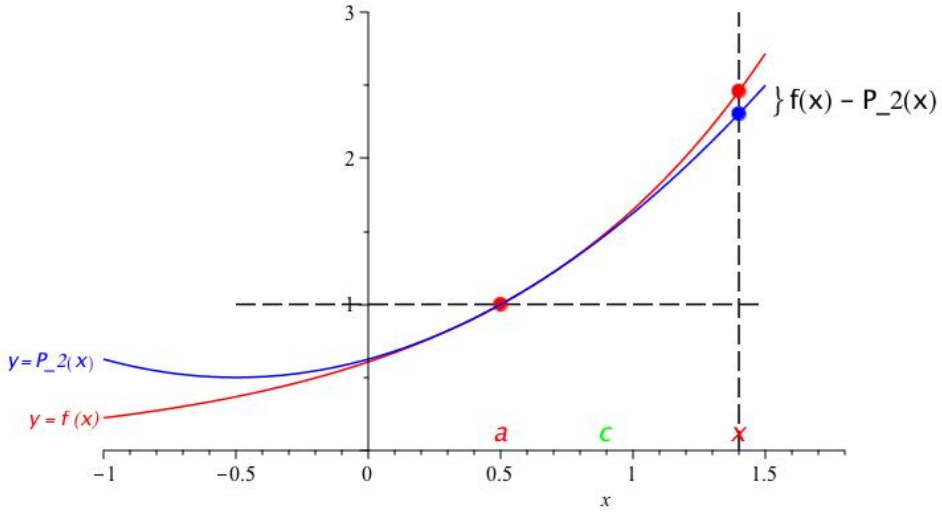


Figure 47: the difference $f(x)-P_2(x)$ is equal to $(x-a)^3$ times $f^{(3)}(c)/3!$ for some c between a and x

The theorem tells us that the constant k in the rough answer is $f^{(n+1)}(c)/(n+1)!$ for this unknown c. This is at first a little mysterious and difficult to use, which is why we'll be

doing some practice. The exact value of c will depend on a, x, n and f and will not be known. However, it will always be between a and x.

Example 13.13. Suppose $f(x) := \sqrt{x}$, $a = 9$ and $n = 1$. Observing that $f(9) = 3$ and $f'(9) = 1/(2\sqrt{9}) = 1/6$, we see that the tangent line approximation $P_1(x)$ is equal to $3 + (x - 9)/6$. What can we infer about the value of $\sqrt{10}$ from this? With $x = 10$, Theorem 13.12 tells us that

$$\sqrt{10} - \frac{3 + (10 - 9)}{6} = \frac{f''(c)}{2!}(10 - 9)^2$$

for some c between 9 and 10. Using $f''(x) = -(1/4)x^{-3/2}$, this simplifies to

$$\sqrt{10} - 3\frac{1}{6} = -\frac{1}{8}c^{-3/2}\,.$$

Exercise 13.10. *Suppose $f(x) := e^{-x}$, $a = 0$ and $n = 1$. What does Theorem 13.12 say about $f(0.4)$?*

In Example 13.13, we still don't know which number between 9 and 10 is the actual c. Often we can get a good idea of the error by examining the possible values of the right-hand side a little more closely. Frequently, for example the sign of $f^{(n+1)}$ does not change and is known to us. Also frequently it is greatest at the point a where we can compute everything exactly. For example, if $f^{(n+1)}$ is known to be positive on the interval $[a, x]$, and known to be greater at a than at larger values, we can conclude that $P_n(x) - f(x)$ is between 0 and $(x - a)^{n+1} \cdot f^{(n+1)}(a)/(n + 1)!$. Here is a very similar example, except that $f^{(n+1)}$ is known to be negative.

Example (13.13 continued)**.** *We don't know which number between 9 and 10 is c, but examining values of $c^{-3/2}$ when c is between 9 and 10, we see that they are all positive, with a maximum of $9^{-3/2} = 1/27$. This is pretty small, which is nice for us because it implies that the error $\sqrt{10} - 3\frac{1}{6}$ is a negative number whose magnitude is no greater than $1/(8 \cdot 27)$ which is less than .005 because 8 times 27 is more than 200. Evidently $3\frac{1}{6}$ is a very good approximation to $\sqrt{10}$.*

Exercise 13.11. *Use the same technique to say how good an approximation $3\frac{1}{12}$ is to $\sqrt{9\frac{1}{2}}$.*

Things don't always work out so nicely. It is pretty common that you know the sign of $f^{(n+1)}$, and almost always you can compute $f^{(n+1)}(x)$ precisely at $x = a$, but it is only moderately likely that its magnitude will be maximized at $x = a$.

Exercise 13.12. *Why do you usually know the exact value of $f^{(n+1)}(a)$?*

Here is an example of what you can do when you don't know the maximum magnitude of $f^{(n+1)}(x)$ on $[a, x]$.

Example 13.14. Let $f(x) = e^x, a = 0$ and $n = 2$. Because $f^{(n)}(e^x) = e^x$ for all n, and $e^0 = 1$, we see that $f^{(n)}(0) = 1$ for all n, and in particular that

$$P_2(x) = 1 + x + \frac{x^2}{2}\,.$$

Let's use P_2 to estimate $e^{0.4}$. This is just like Exercise 13.10 except with e^x instead of e^{-x}. First, $P_2(0.4) = 1 + 0.4 + (0.4)^2/2 = 1.48$ precisely. Plugging in $f'''(x) = e^x$, Theorem 13.12 tells us that

$$e^{0.4} - 1.48 = \frac{f'''(c)}{3!}(0.4)^3 \approx 0.021e^c$$

for some $c \in [0, 0.4]$. We can see the maximum of the right-hand side is attained at $c = 0.4$ rather than $c = 0$. The value of f''' there is $e^{0.4}$ which happens to be the quantity we are going to a lot of trouble to estimate. So of course we don't already know what it is. The trick is to use any crude upper bound. For example, e is less than 3 and 0.4 is less than $1/2$, so $e^{0.4} < \sqrt{3}$, which we happen to know to be approximately 1.732. If we didn't know this, we could use $e < 4$ instead of $e < 3$, leading to $e^{0.4} < 4^{0.5} = 2$. That's pretty rock solid. So then the error, which is known to be positive, is less than $2 \cdot 0.021 = 0.042$ and we have $1.48 < e^{0.4} < 1.522$. The true value to three deicmals is 1.492.