Math 1080, Fall 2023, E-textbook

Contents

Introduction

This e-textbook was written at a time of uncertainty about in-person versus online instruction. It was written to be used flexibly. For an in-person course, it provides a number of activities for a flipped classroom and self-check exercises so that students can tell what they need to learn and whether they are getting what they should from reaching the text.

In each section, there are self-check exercises, numbered and offset in red. It is intended that students do these when they come across them during the reading. Typically, doing the reading before class is mandatory and so are the self-check questions. The purpose of the self-check questions is purely diagnostic. Students need to know whether they have understood each point as intended and instructors need to know what needs to be gone over in class.

Self-check exercises vary a bit in difficulty. Self-check exercises which are more challenging or more open-ended are marked with a +, or sometimes more than one + if they're particularly challenging.

Some other 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 below, 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.

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.

(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.

Domains and ranges of functions can any sets – they don't even have to be set of numbers. In calculus courses, usually the domain and range are subsets of the real numbers. Later in the course we will talk about multivariable functions where the domain is $\mathbb{R}^n$, the set

of ordered n-tuples $(x_1, \ldots, x_n)$ of real numbers. For the remainder of this introduction, however, we will stick to functions of one real variable.

We write $f : A \to B$ as shorthand for the statement that the domain of f is A and the outputs take values in B, that is, the range is (not necessarily all of) B. The last sentence in the previous paragraph can be rephrased as, "For the remainder of this introduciton, we will stick with functions $f : \mathbb{R} \to \mathbb{R}$."

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 0.1. 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 "$\ldots$" 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 thought 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 0.1. *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 $a \le x \le b$. When both endpoints are included, this is called a **closed** interval. The interval (a, b) refers to all real numbers x such that $a \le x \le b$. 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.

Product notation for intervals and other sets. The notation $[a, b] \times [c, d]$ refers to the rectangle in the plane consisting of points (x, y) such that $x \in [a, b]$ and $y \in [c, d]$. More generally (though we won't use this), the **Cartesian product** of any two sets $S \times T$ is defined as the set of ordered pairs (x, y) such that $x \in S$ and $y \in T$. This notation extends to more than two variables, for example $[0, 1] \times [0, 1] \times [0, 1]$ is the unit cube in $\mathbb{R}^3$.

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, $\{x : a \le x < b\}$ 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 0.2. *Suppose $f(x) := x^2 + 2$. What values of x make the equation $f(x) = 5 - 3x^2$ true? Please simplify you 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, they might say that the domain and range of a function $f(x) := x^2 + 2$ is all real numbers. It's fine to define the domain to be all real numbers, but then technically the range is the set of real numbers that are at least 2.

Exercise 0.3. *What are two formal mathematical ways of writing the set of real numbers that are at least 2, one using set 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 0.4. *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**[1]. The bound variable in this case runs over all values in the domain of f. It has no particular value. A variable that is not bound is **free**. For example, in the definition $f(u) := u^2 + c$, the variable c is free.

---

[1]The word "bound" here is the past tense of "bind" and has nothing to do with lower and upper bounds.

9

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. Changing the value of a bound variable doesn't make sense, let alone change the value of the expression.

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 calculus, the first two of which you have already seen in this introduction.

- 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.

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 0.5. *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

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

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

Exercise 0.6. *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 0.7. *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 $x < y$.

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

Physical units

This multiple choice problems seems to trip up far more people than it should. Give it a try before reading on.

Exercise 0.9. *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) $2.54 f(x)$

(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 0.10. *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 0.2. 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 $k \ \dfrac{\$}{\text{person}^2}$.

---

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

1 Integration in more than one variable

One of the main topics of this course is the very large subject of calculus of functions of several variables. Before we tackle this, you should know the applications that motivate this topic. Here is a list of some key applications.

1. Totals of quantities spread out over an area.

2. Joint probabilities for more than one random variable: what is the probability that a pair of random variables (X, Y) is in a certain set of possible values?

3. Optimization: how do I find the maximum or minimum of a function $f(x, y, z)$ on a region R? For a concrete example, if I have a limit L on how much I can spend on production and advertising in total, and my revenue will be some function $f(p, a)$, then how much should I invest in production and how much in advertising? I will need to maximize $f(p, a)$ over the region $R := \{(p, a) : p + a \leq L\}$.

4. Marginal rates, for example the marginal rate of increase of revenue per increase in spending on production from the *status quo*.

When dealing with these sorts of questions, the functions and their notation can start to seem difficult and abstract. Geometric understanding of multi-variable functions will help us think straight when doing word problems and algebraic manipulations.

1.1 Multivariate functions

To understand functions of several variables, start by recalling the ways in which you understand a function f of one variable.

(i) As a rule, e.g., "double and add 1"

(ii) As an equation, e.g., $f(x) = 2x + 1$

(iii) As a table of values, e.g.,

x	0	1	2	5	20	-95	π
$f(x)$	1	3	5	11	41	-189	$2\pi + 1$

(iv) As a graph, e.g.,
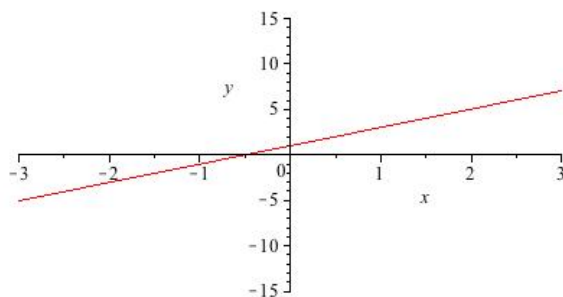
Similarly, a function f of two variables is a way of associating to any pair of values for x and y (two real numbers) a real number $f(x, y)$. The same options apply for understanding f.

(i) We can give the rule if it is easily stated, e.g., "multiply the two inputs."

(ii) We could give an equation, such as $f(x, y) = xy$.

(iii) We could make a table, e.g.,

x	1	1	1	2	2
y	0	1	5	0	π
$f(x, y)$	0	1	5	0	2π

(iv) One might graph f. We next discuss says of doing that.

Exercise 1.1 (+)**.** *What is the simplest verbal rule that corresponds to this definition by cases?*

$$f(x, y) := \begin{cases} x & \text{if } x > y \\ y & \text{if } x \leq y \end{cases}$$

The + indicates that this is a more open-ended exercise—there are lots of ways to describe this rule, and you'll have to decide if you think you've really found the simplest rule.

1.2 Plots: surface, contour, intensity

You can think of a function of two variables as having two real inputs x and y, or as having one input that is a pair (x, y). The second way makes the domain of the function into the xy-plane, or some subset of the plane where the formula or rule makes sense. We can use geometry to understand f via its various visual depictions.

The most common way to make a graph of f is to plot the three-dimensional surface $z = f(x, y)$ as in the following figure.
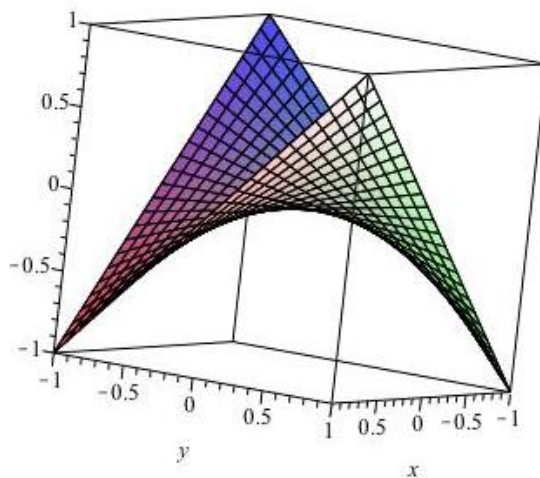

Figure 1: graph of $z = f(x, y)$ where $f(x, y) := xy$

Notice the base, in the x-y plane, of this figure. This tells us which pairs (x, y) are being illustrated. In this case, both the x and y values go between -1 and 1. We write this domain

$$\{(x, y) \mid -1 \leq x \leq 1 \text{ and } -1 \leq y \leq 1\}$$

or, more compactly, $[-1, 1] \times [-1, 1]$.

16

For single variable functions, domains are sets of x values. For a function of two variables, the domain is a set of pairs.

In the figure above, the domain was a square where the x and y values ranged over the same values. In general, that doesn't need to be the case.
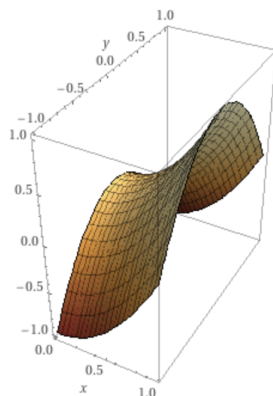
Figure 2: graph of $z = f(x, y)$ where $f(x, y) := x^2 - y^2$

Exercise 1.2. *What is the domain in the x-y plane over which the function xy is plotted in Figure 2?*

What are the values of $f(x, y)$ at the four corners of the domain in this figure? Does the picture match the value you calculated? (You can only see three of the four corners, but you can see enough of the graph to have a good guess what's happening at the fourth corner.)

Another way to help visualize a function of two variables is to plot the level curves.

Definition 1.1. *When $f(x, y)$ is a function of two variables, a* level curve *of $f(x, y)$ is a curve $f(x, y) = c$ where c is a constant.*

For example, with our function $f(x, y) = xy$ above, one of the level curves is $xy = 1$, which we can plot like a normal single variable function $y = 1/x$.

A single level curve doesn't tell us much about $f(x, y)$, but drawing a lot of them together does. The points on the same curve are all at the same "height"—they have the same value of $f(x, y)$. This is shown in Figure 4. The convention when drawing level curves is to pick some fixed increment and draw the level curves corresponding to regular intervals.
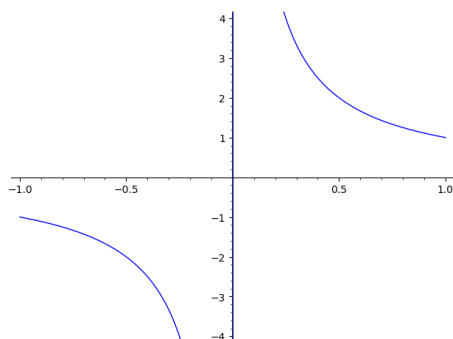
Figure 3: A single level curve for $f(x, y) = xy$.
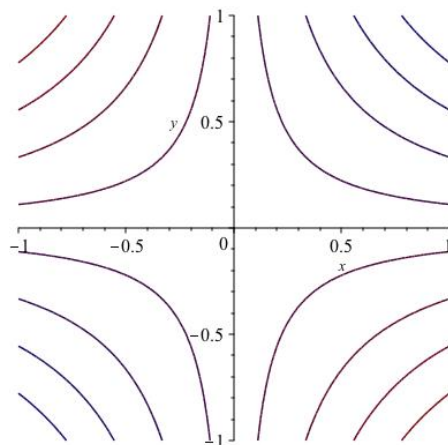
Figure 4: level curves for $f(x, y) = xy$ plotted over the square $-1 \leq x, y \leq 1$.

If we took these curves and drew them on the surface of the 3D picture in Figure 1, these curves would be flat.

Exercise 1.3 (+)**.** *Take the level curve in the upper right of Figure 4 (the little bit that looks almost like a straight line). Estimate the value of c so that this is the curve $xy = c$.*

What, roughly, are the increments between the levels of xy shown in Figure 4?

Level curves are also called **contour plots**. The US Geological Service produces a series of maps drawn this way (see an example on the next page). These are contour plots of $f(x, y)$, where f is the elevation and x and y are distance east and distance north of the center of a

18

quadrangle. Figure 5 shows part of a contour plot of a quadrangle in the Sierra mountains with level curves plotted every 40 feet of elevation.
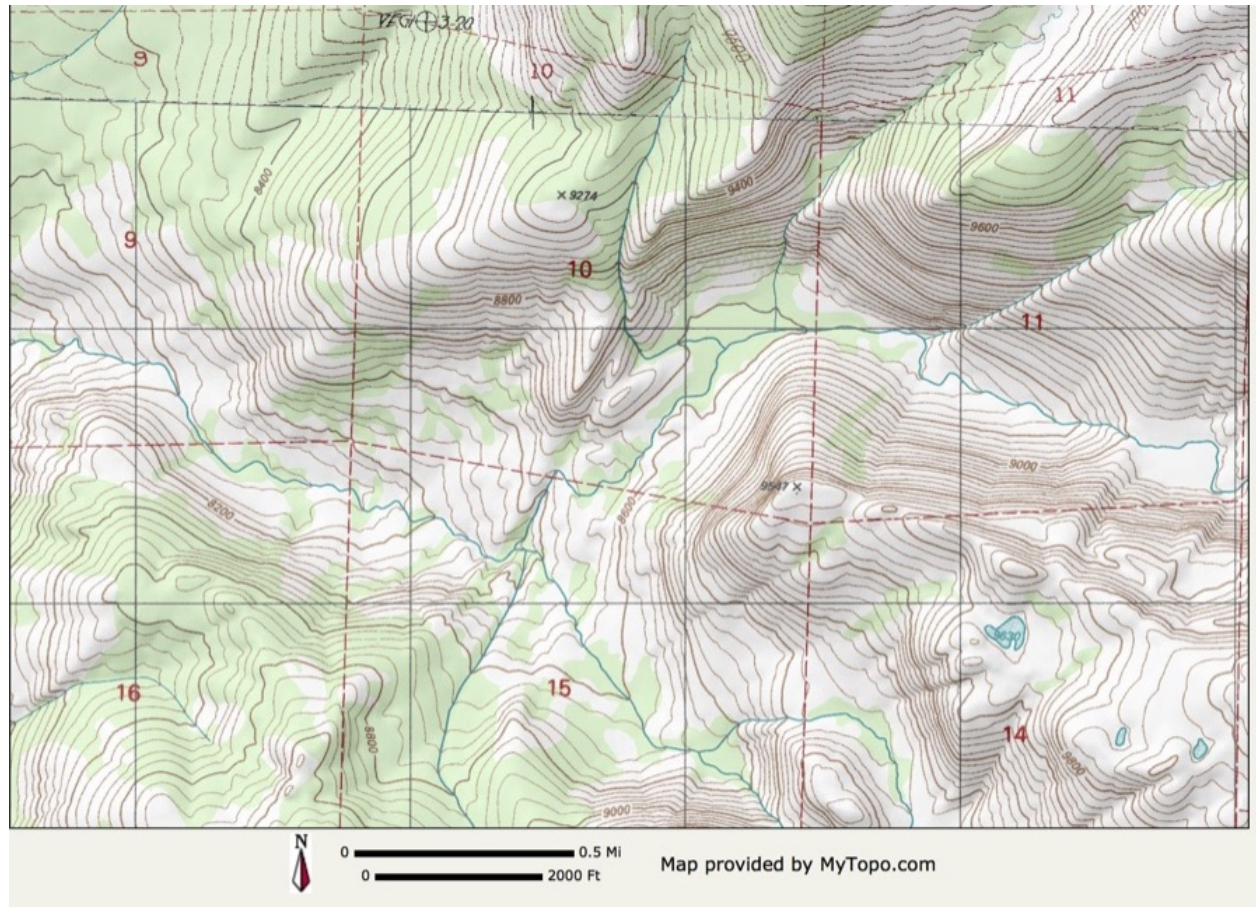
Figure 5: USGS contour plot

The elevation example is very important even if you don't care about hiking. This is because the traditional way to plot f is to plot the surface $z = f(x, y)$, which means that our brains are primed to accept $f(x, y)$ as an elevation at the point with coordinates (x, y). However, this is far from the only use of contour plotting. The most important application of this is when $f(x, y)$ is profit or some other kind of a utility function (e.g., the level of satisfaction when you have x dollars in the bank and a car that costs y dollars). The contour plot of f shows the **indifference curves**. Later we can use this interpretation of contour plots along with some calculus to compute optimal allocations.

Exercise 1.4. *The figure to the right on the next page shows a plot of revenue in millions of dollars as a function of market size in tens of thousands (x-coordinate) and price in hundreds of dollars per unit (y-coordinate).*

(a) *Which of these functions is possibly the function plotted?*

    (i) $f(x, y) = y^2 + y * x^2$

    (ii) $f(x, y) = x^2 + x * (y - 2)^2$

    (iii) $f(x, y) = x + y + x * y$

(b) *Which of these would produce a level curve for this plot?*

    (i) *The intersection of the plot with the surface* $z = f(x, y)$

    (ii) *A curve drawn on plot from the lower left corner to the upper right corner*

    (iii) *The intersection of the plot with a horizontal plane*
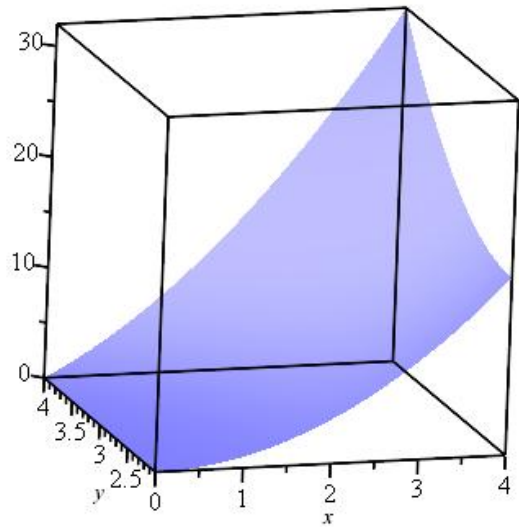
All we are doing in this first section is getting used to functions of more than one variable and their visual depictions. We're almost done, except that we haven't talked about functions of three or more variables. We don't have four dimensions handy, so we can't graph $z = f(x_1, x_2, x_3)$. We can still think of f as a function mapping points in an abstract n-dimensional space to the real numbers, and in the case of exactly three variables, we can make a contour plot which now has contour surfaces in three dimensions. For now, it suffices to practice going back and forth between the equation for a function of two variables and its visual representations.

Exercise 1.5 (+). *Just northeast of the center of the topographical map in Figure 5 is a three-way juncture of a stream. Which direction is the water flowing?*

1.3 Integration over rectangular regions

This section is a bit heftier than the previous one because multiple integrals are really, really important. This is a tricky topic for two reasons. First, students often confuse the *definition* of a double integral with the *computation* of a double integral. We will try to help you keep these straight. Secondly, non-rectangular regions of integration, which are the topic of Unit 2.1, require a solid understanding of free and bound variables; please review this topic in Unit 0 as needed.

Definition and meaning

Let R be a region in the xy-plane and let $f(x, y)$ be a function. The notation $\int_R f(x, y) \, dA$ is read as "the double integral of f over the region R" and defined as follows.

> Divide R into small rectangular regions (ignore for now the fact that these rectangles might not quite cover R, or perhaps extend a litle beyond R). Multiply the area of each rectangle by f evaluated at some point in that rectange, and add up all of these products. The integral is defined to be the limit of this sum of products as the rectangles get small.

What does this compute? In general it computes the total amount of stuff when f is a density of stuff per unit area. For example, suppose the density of iron ore over a patch of ground is a function $f(x, y)$ that varies due to proximity to some pre-historic lava flow. Then $\int_R f(x, y) \, dA$ will be the total amount of iron ore in the region R. Do you see why? You can get the total by adding up the amount in regions small enough that f doesn't vary significantly; then the amount of ore in the region is roughly the area times f evaluated at any point in the region, so we should expect that adding up these products approximates the total; in the limit, it *is* the total.

Exercise 1.6. *Suppose a circular region of radius 5 miles in the Mojave desert has been contaminated with Strontium-90 since a 1953 nuclear test. Contamination is worse near the center but extends throughout the region. What integral represents the total amount of Strontium-90 in the region? Please give interpretations of any functions and variables.*

Here are a bunch of conceptual remarks about integrals over 2-dimensional regions.

1. Notice there is now a quantity dA rather than dx or dy. This means, literally, "the teeny amount of area". Starting now, it will be very important to keep track of the infinitesimals.

2. The units of $\int_R f(x,y)\,dA$ are units of f times units of A. The units of A can be area, but more generally, they are whatever unit x represents times whatever unit y represents.

3. Try to see how this is analogous to integrals in one variable. In each case you break up the (interval / region), then in each small part you evaluate f somewhere, use this as a proxy for f throughout the small part, multiply by the (length / area) of the small part, sum and take the limit.

4. You can integrate in three variables. Just chop a 3-D region into subregions, sum their volumes times the value of $f(x,y,z)$ somewhere in the region, and take a limit. In fact, you can do this in any number of variables even though we can't visualize space in dimensions higher than three. In this course, we'll stick to two variables.

Here are some more meanings for a double integral.

Volume. If $f(x,y)$ is the height of a surface at the point (x,y), then $\int_R f(x,y)\,dA$ gives the volume underneath the surface but above the region R in the xy-plane. That's because the summands (namely the area of a little region times $f(x,y)$ evaluated at a point in the region) is the volume of a tall skinny rectangular shard, many of which together physically approximate the region. Figure 6 shows these shards for a sequence of increasingly fine partition of the region.
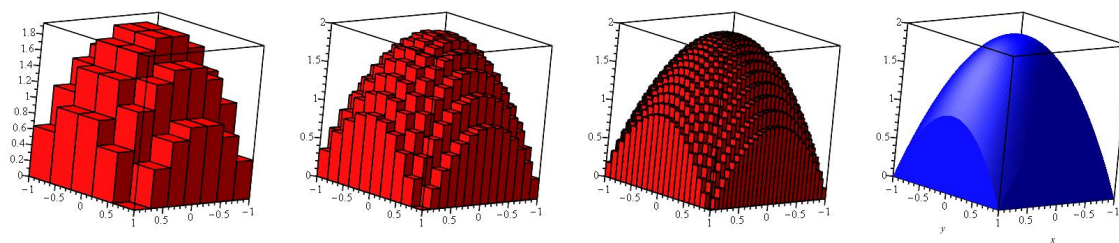

Figure 6: increasingly fine approximation to a volume (far right) by rectangular shards

Notice that the units check out: when talking about volume, we would most often use the same unit of distance for the height, the x units, and the y units. For instance, if all three

distances are in inches, the units of $\int_R f(x,y)\,dA$ are inches times inches^2 (the units of f times the units of A), which gives us inches^3, which is a unit of volume.

Exercise 1.7. *The volume of the blue solid is $\int_R f(x,y)\,dA$ for what region R?*

Remarks:

1. Just as in the case of area, the integral computes **signed volume**, positive above the xy-plane and negative below.

2. Notice that the units work as expected: f is height (units of length) and $\int_R f(x,y)\,dA$ is volume, which does indeed have units of length times area.

Area. A special case is when $f(x,y)$ is the constant function 1. Who would have thought that integrating 1 could be at all important? But it is. If you build a surface of height 1 over a region R, then the volume of each shard is the area at the base of the shard and the integral is just the limiting sum of these, namely the total area. Figure 7 shows a three-dimensional region whose volume computes the area of a heart-shaped figure. Notice again that the units work out correctly: in the example f is unitless, and $\int_R f(x,y)\,dA$ is the area of R, which has units of area.
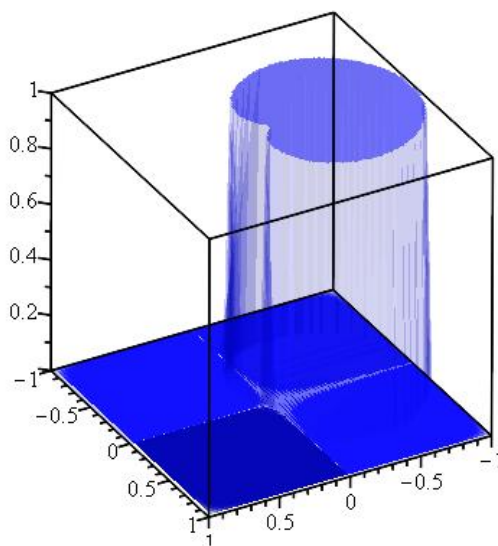

Figure 7: area of heart-shaped region is the volume over it to height 1

Exercise 1.8. *What do you compute if you integrate $\int_R 3\,dA$ instead of $\int_R 1\,dA$?*

Averages. By definition, the average of a varying quantity $f(x,y)$ over a region R is the total of f divided by the area of the region:

$$\text{Average of } f \text{ over } R = \frac{\int_R f(x,y)\, dA}{\text{Area of } R}.$$

Exercise 1.9. *In terms of units of x, units of y and units of f, what are the units of the Average of f over the region R?*

Probability. The integral represents the probability that a random pair with the specificed two-variable density function will be in the region R. This application will get its own treatment in Unit 2.3.

Computing the integral: rectangular regions

Remember how it worked when you learned integration in one variable? You defined an integral as the limit of Riemann sums, which intuitively captures the notion of area under a curve. You can figure out the value of the integral over an interval by computing an anti-derivative and subtracting its values at the two endpoints. Similarly, we have already defined the integral conceptually, now we need to say something about using calculus to compute it. A lucky fact: we don't need anything as difficult as the Fundamental Theorem of Calculus like we did for one variable integrals. That's because everything you already know about computing single variable integrals can be harnessed to compute two-variable integrals.

Let's see how this works, recalling that for now we're sticking to to the case where R is a rectangle. Let R be the rectangle $a \leq x \leq b$ and $c \leq y \leq d$. Some new notation: this rectangle is denoted $[a, b] \times [c, d]$. Chop the domain up into smaller rectangles, m of them in the x-direction and n of them in the y-direction. Let Δx and Δy denote the width and length of little squares, respectively.

Exercise 1.10 (+). *What are Δx and Δy in terms of m, n, a, b, c and d?*

Exercise 1.11. *Consider the function $f(x,y) = xy^2$. Let R be the square region where x and y are both between 1 and 2. Using a calculator or by hand, approximate $\int_R f(x,y)\, dA$ by dividing into four regions of size 1/2x1/2 and computing a Riemann sum. If you use the midpoint of each little region, you should get an answer that's pretty close to the true answer.*

We get a clue for a route to calculating integrals if we add up the little bits in an organized way. First add all the tall skinny rectangles over a given x interval as y varies. In the volume interpretation this gives the volume of the slice of the solid that has width Δx. There is a slice for each x-value in the grid. Figure 8 illustrates this with $m = 5$, showing a Riemann sum with a 5×7 grid of vertical shards broken into five seven-shard strips.
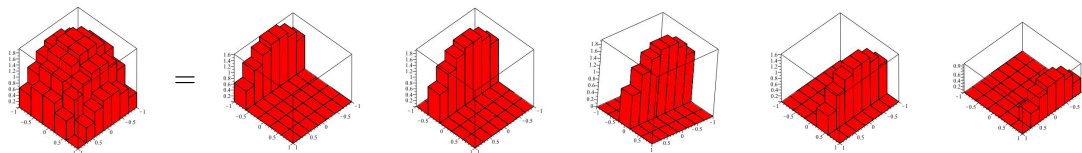
Figure 8: breaking a Riemann sum into 5 slices

Here's the thing. If you fix a value $x = x_j$ in the jth x-interval, then all the shards in the jth slice will be able to use the x-value x_j along with separate y-values y_k chosen in each shard. The volume of the shard will be $(\Delta x)(\Delta y)f(x_j, y_k)$. Factoring out Δx, we have

$$\text{Slice volume for } j^{th} \text{ slice } = \Delta x \left(\sum_{k=1}^{n} f(x_j, y_k)\Delta y \right). \tag{1.1}$$

In pictures, the factored out Δx is the thickness of each of the slices in Figure 8 and the sum is the total cross-sectional area of the slice. The sum should look familiar because **it is a Riemann sum for the integral giving the area under the one-variable cross-section function** $f(x_j, y)$. You know what happens to these Riemann sums. As the number of y intervals gets larger, the quantity inside the parentheses in (1.1) approaches $\int_c^d f(X_j, y)\,dy$. In other words,

$$\lim_{n \to \infty} (\text{Slice area with } n \text{ } y\text{-intervals}) = \int_c^d f(x_j, y)\,dy. \tag{1.2}$$

This integral of course depends on x_j. Call it $g(x_j)$. Multiplying by the width Δx, the sum of all the slice volumes is

$$\sum_{j=1}^{m} g(x_j)\Delta x. \tag{1.3}$$

Here we go again. This is a Riemann sum approaching the limit $\int_a^b g(x)\,dx$. With a little extra work to check that things are continuous and the choices of x_j don't matter, we get this theorem.

Theorem 1.2 (Fubini's Theorem). *Let f be a continuous function on the rectangle $R :=$ $[a, b] \times [c, d]$. With the notation as above, no matter how x_j and y_k are chosen within the x- and y-intervals, the following limit exists and has the value on the right:*

$$\lim_{m,n \to \infty} \text{ Total volume of } m \times n \text{ Riemann sum } = \int_a^b \left(\int_c^d f(x, y) \, dy \right) \, dx. \qquad (1.4)$$

Exercise 1.12. *Suppose $f(x, y) = xy$. What is the inner integral $\int_c^d f(x, y) \, dy$ when $x = 3/2$? What is it as a function of x?*

Pay attention to free and bound variables! In the so-called **inner integral** $\int_c^d f(x, y) \, dy$, the variable y is bound, but x is free. In other words, this integral represents a quantity that depends on x (but not y). That's why we can integrate it against dx in the outer integral, to finally get a number. In the discussion just before equation (1.3) we used the function $g(x)$ to denote the inner integral $\int_c^d f(x, y) \, dy$.

Example 1.3. The function plotted in blue on the right of Figure 6 is the function $f(x, y) := 2 - x^2 - y^2$. The rectangle R in this case is $[-1, 1] \times [-1, 1]$. We compute $\int_R (2 - x^2 - y^2) \, dA$. By (1.2), fixing m and a x-value x_j, the slice areas converge to $\int_c^d 2 - x_j^2 - y^2$ as $n \to \infty$. Denoting this by $g(x_j)$, we have

$$\begin{aligned} g(x_j) &= \int_c^d f(x_j, y) \, dy = \int_c^d (2 - x_j^2) - y^2 \, dy \\ &= \left[(2 - x_j^2)y - \frac{1}{3}y^3 \right] \Big|_c^d \\ &= (2 - x_j^2)(d - c) - \frac{d^3 - c^3}{3}. \end{aligned}$$

When we plug in -1 for c and 1 for d, we get $g(x) = 4 - 2x^2 - \dfrac{2}{3}$. Therefore,

$$
\begin{aligned}
\text{Volume} \quad &= \quad \int_R f(x,y)\,dA \\
&= \quad \int_{-1}^{1} \left(\int_{-1}^{1} f(x,y)\,dA \right) \\
&= \quad \int_{-1}^{1} g(x)\,dx \\
&= \quad \int_{-1}^{1} \left(4 - 2x^2 - \frac{2}{3} \right) dx \\
&= \quad \left. \frac{10x - 2x^2}{3} \right|_{-1}^{1} \\
&= \quad \frac{16}{3}\,.
\end{aligned}
$$

Exercise 1.13. *Use Fubini's theorem to integrate xy^2 over the rectangle $[0,4] \times [2,5]$.*

You may have noticed we made an arbitrary choice to sum over slices with x fixed rather than slices with y fixed. Of course we could have done this the other way around; the math is identical. Fubini's Theorem could therefore have made the stronger statement

$$
\int_a^b \left(\int_c^d f(x,y)\,dy \right) dx = \int_R f(x,y)\,dA = \int_c^d \left(\int_a^b f(x,y)\,dx \right) dy\,. \tag{1.5}
$$

Magic product formula

Suppose your function $f(x,y)$ is of the form $g(x) \cdot h(y)$ and your region of integration is a rectangle $[a,b] \times [c,d]$. Then

$$
\int_R f(x,y)\,dA = \left(\int_a^b g(x)\,dx \right) \times \left(\int_c^d h(y)\,dy \right)\,. \tag{1.6}
$$

Can you see why? It's due to the distributive law. The Riemann sum for the double integral actually factors into the product of two Riemann sums.

Exercise 1.14. *Use the magic product formula to evaluate $\displaystyle\int_R e^x(1 + y + y^2)\,dA$ where $R = [0,1] \times [0,6]$.*

One parting word: circling back to the issue of distinguishing the definition from the computation, the left-hand side of (1.4) refers to the definition – a limit of Riemann sums; the two expressions after the equal signs are single variable integrals, computable as anti-derivatives. The theorem is asserting that they are all equal.

2 Writing and interpreting multivariate integrals

2.1 General regions

The trickiest thing about learning double integration is when R is not a rectangle. Then, when you cut into slices, the limits of integration will change with each slice. That's OK as long as you can write them as a function of the variable you are not integrating and evaluate properly. Before diving into these, we need to review notation for sets of points in the plane.

Writing sets of points in the plane

Recall the notation $\{(x, y) : \text{blah blah blah}\}$ denoting the set of points in the plane satisfying the condition we have called "blah blah blah". For example, $\{(x, y) : x^2 + y^2 \leq 1\}$ is the **unit disk**. You will need to become an expert at writing sets of points in a very specific manner: the set of points where x is in some interval $[a, b]$ and y lies between two functions of x, call them g and h. It looks like

$$\{(x, y) : a \leq x \leq b, \ g(x) \leq y \leq h(x)\}.$$

Figure 9 shows a picture of what this region looks like for $a = 0, b = 1$ and the functions $h(x) := e^x$ and $g(x) := 1/2 + x$.
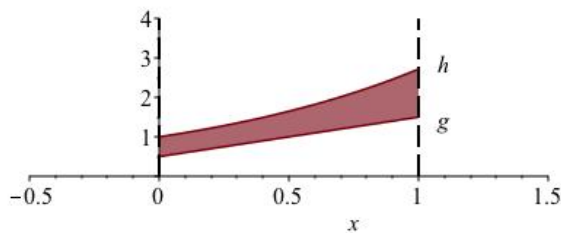

Figure 9: the region $\{(x, y) : a \leq x \leq b, \ g(x) \leq y \leq h(x)\}$

Example 2.1. Let's write the unit disk in this format. For a and b you need the least and greatest x values that appear anywhere in the region. For the unit disk, that's -1 and 1. Then, for each x, you need to figure out the least and greatest y values that can be
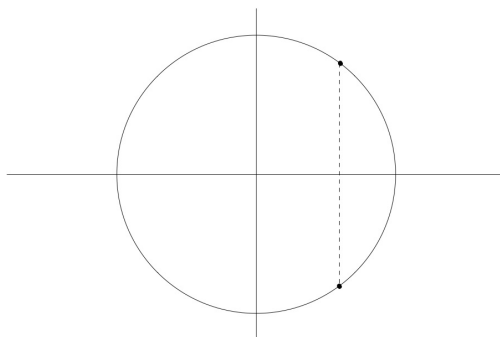
Figure 10: x goes from -1 to 1; for each x, we see y going from $-\sqrt{1-x^2}$ to $+\sqrt{1-x^2}$

associated with that x. For the unit disk, the least value is $-\sqrt{1-x^2}$ and the greatest is $+\sqrt{1-x^2}$. Figure 10 illustrates the least and greatest y values for $x = 3/5$. Summing up, the unit disk $\{(x,y) : x^2 + y^2 \leq 1\}$ can be written in our standard form as

$$\{(x,y) : -1 \leq x \leq 1,\ -\sqrt{1-x^2} \leq y \leq \sqrt{1-x^2}\}. \tag{2.7}$$

This way of writing it naturally breaks the unit disk into vertical strips where x is held constant and y varies from some least to some greatest value depending on x. We should have said this is "*a* standard form" not "*the* standard form" because it is equally useful to break into horizontal strips. These correspond to the format

$$\{(x,y) : c \leq y \leq d,\ g(y) \leq x \leq h(y)\}$$

where for each fixed y, the x values range from some minimum to some maximum value depending on y. You will be practicing a lot with these two formats!

Exercise 2.1. *Write the unit disk in the other standard form (horizontal strips).*

Exercise 2.2. *The region $\{(x,y) \mid 0 \leq x \leq y \leq 1\}$ is one of the triangles in this picture.*
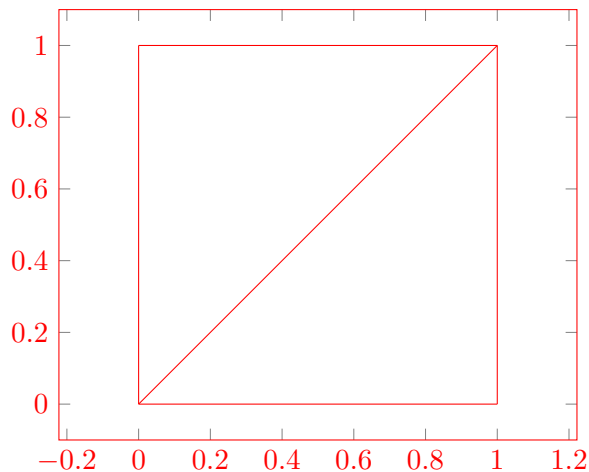
Limits of integration for non-rectangular regions

The reason for writing regions in vertical or horizontal strip format is so you can use Fubini's Theorem to evaluate integrals. When computing $\int_R f(x, y)\, dA$, if you can write R as a region in the form above. We explain first with an example.

Example 2.2. Let's integrate the function 1 over the unit disk. Recall, this should yield the area. Of course, if you are asked to compute the area of a circle, you should do so by recalling that the area is πr^2, so the area should be π. However, in this example we will compute the integral via vertical strips. Having already written the unit disk in vertical strips (2.7), we can make a figure akin to Figure 7 but with a circular base. The slice with x held constant at x_j will not always have the same y-extent. In fact the formula expressing the unit disk in vertical strips tells us y goes from $-\sqrt{1-x^2}$ to $\sqrt{1-x^2}$. Therefore when we use Fubini's Theorem to write $\int_R 1\, dA$ as an iterated integral with vertical strips, we get

$$\text{Area} = \int_R 1\, dA = \int_{-1}^{1} \left(\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} 1\, dy \right) dx\,.$$

Computing the inner integral is easy. The antiderivative of 1 is y, so

$$\int_{-1}^{1} \left(\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} 1\, dy \right) dx = \int_{-1}^{1} y|_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}\, dx = \int_{-1}^{1} \sqrt{1-x^2} - (-\sqrt{1-x^2})\, dx = \int_{-1}^{1} 2\sqrt{1-x^2}\, dx.$$

31

You probably don't immediately know what the antiderivative of $2\sqrt{1-x^2}$ is, but fortunately, we can look it up on a large table of integrals: the antiderivative is $x\sqrt{1-x^2} + \arcsin(x)$. Thus

$$
\begin{aligned}
\int_R 1\,dA &= \int_{-1}^{1} 2\sqrt{1-x^2}\,dx \\
&= 1\sqrt{1-1^2} + \arcsin(1) - ((-1)\sqrt{1-(-1)^2} + \arcsin(-1)) \\
&= \arcsin(1) - \arcsin(-1) \\
&= \frac{\pi}{2} - (-\frac{\pi}{2}) = \pi\,.
\end{aligned}
$$

Exercise 2.3 (+). *When we took the antiderivative of the constant function 1, why did we get y and not x?*

We can turn this into a general rule.

Proposition 2.3. *Let R be a region in the plane written in vertical strips as $\{(x,y) : a \le x \le b, g(x) \le y \le h(x)\}$. Let f be a continuous function on R. Then*

$$
\int_R f(x,y)\,dA = \int_a^b \left(\int_{g(x)}^{h(x)} f(x,y)\,dy \right) dx\,.
$$

Exercise 2.4 (+). *How can you know whether the limits on the inner integral will be functions of x or functions of y?*

Switching the order of integration

You have seen how to take a region R and write it in either standard form: horizontal or vertical strips. Sometimes, in order to make an integral do-able, you will want to switch from horizontal strips to vertical strips or vice versa. Starting with one standard form, you convert to a geometric region R, then write that in the other standard form. This allows you to switch between an iterated integral with x in the inside and one with y on the inside.

Example 2.4. Compute $\int_0^1 \int_y^1 \frac{\sin x}{x}\,dx\,dy$. Unfortunately you can't integrate $\sin x/x$. But wait! The region $\{0 \le y \le 1,\ y \le x \le 1\}$ is triangular.
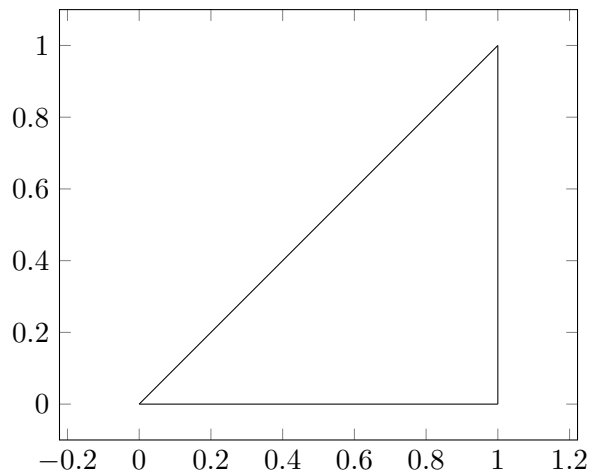
and can also be written in vertical strips: $\{0 \le x \le 1,\ 0 \le y \le x\}$. The integral is therefore equal to $\int_0^1 \int_0^x \frac{\sin x}{x}\, dy\, dx$. Integrating $\sin(x)/x$ is much easier against y than against x! We can now see that this is equal to

$$\int_0^1 \left(y \frac{\sin x}{x} \right) \Big|_0^x dx = \int_0^1 \sin x\, dx = 1 - \cos(1)\,.$$

Exercise 2.5. *Suppose a region R in vertical strips is described as $\{(x,y) : 1 \le x \le 3, 2 \le y \le 1 + x^2\}$. Draw a sketch, then write R in horizontal strips. Then write $\int_1^3 \left(\int_2^{1+x^2} xy\, dy \right) dx$ as an integral of the form $\int_?^? \left(\int_?^? xy\, dx \right) dy$.*

33

2.2 Application: indifference curves

An important type of level curve is the **indifference curve**. Suppose that the independent variables x and y represent quantities of two different things that will rival each other for importance in a single scenario.

Example 1: x is the horsepower of a car and y is its MPG.

Example 2: x is ounces of pizza at a meal and y is pints of FroYo.

An indifference curve is a set of points in the x-y plane corresponding to bundles that the agent (often a consumer) likes equally well. The two examples above are taken from Berheim and Whinston (current textbook for BEPP 250). The indifference curve for horsepower versus fuel economy shown in Figure 11 is taken from actual data. The indifference curve for pizza versus FroYo is a made-up model.
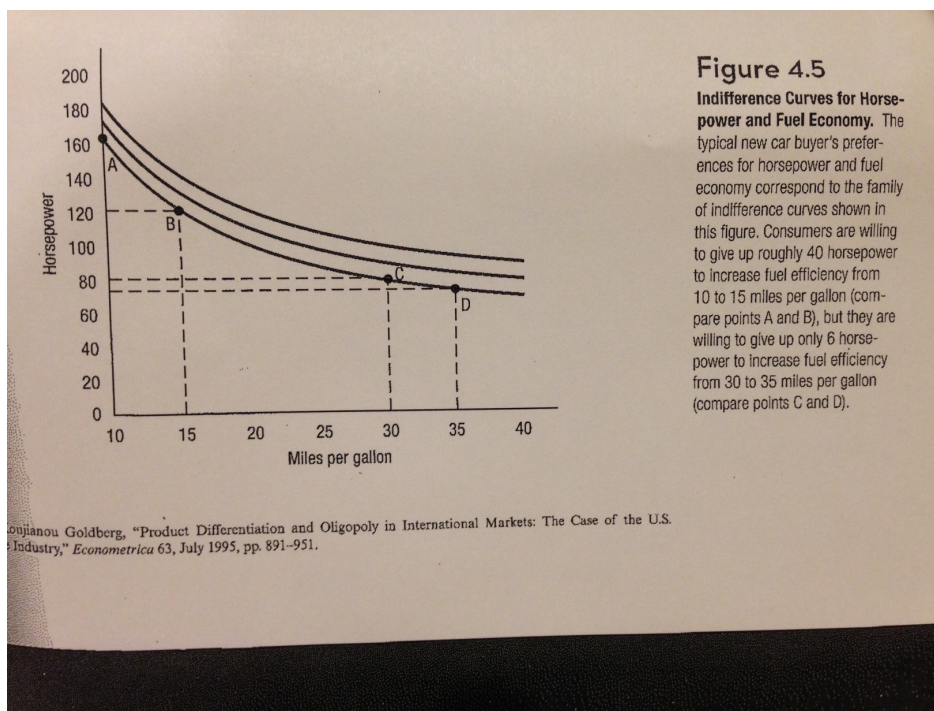

Figure 11: indifference curve for power versus efficiency taken from Berheim and Winston

Often economists assume a utility function $u(x, y)$ that measures how appealing the bundle (x, y) is. The level curves $u(x, y) = c$ represent curves where a consumer would like one point (x_1, y_1) on the curve equally well as any other point (x_2, y_2) on the curve. What is meaningful here is not so much the value of $u(x, y)$, rather the preferences it defines. Sometimes a simplified model assumes a particular utility function u, while other times it is assumed u exists and obeys certain rules, but a particular function is not specified. In the car example, the indifference curves are defined by the data but no utility function is assumed.

Indifference curves are important for several reasons, one of which is that they describe incentives and reactions to changes in the quantities x and y. The **marginal rate of substitution** is the amount of x an agent would be willing to give up in order to increase y by a small amount, as represented as a ratio of amount of x to amount of y. This is not a static quantity, rather it depends on the present levels of x and y. If a group of diners has 10 pints of FroYo and only three ounces of pizza, they will not be willing to give up much pizza for one more pint of FroYo, whereas a group with 60 ounces of pizza and half a pint of FroYo might well give up a lot of pizza for a pint of FroYo.

Two points on the same indifference curve, such as $(60, 1/2)$ and $(20, 3/2)$ in Figure 12 below, determine an equivalence of utility. The slope of the line segment between these two points is a ratio for a trade the agent is willing to make in either direction (see the straight line in the figure). But the point $(20, 3/2)$ is quite far from $(60, 1/2)$ and does not represent the rate of substitution if the consumer is able to make continuous small adjustments. As the point (x, y) on the curve $u(x, y) = 30$ approaches $(60, 1/2)$, the slope of the line segment approaches the slope of the tangent line to the curve $u(x, y) = 30$ at $(60, 1/2)$ (dashed line in the figure).

Exercise 2.6. *What are the units of the marginal rate of substitution of FroYo for pizza?*

Mathematically, the marginal rate of substitution is defined to be the negative of the slope of this tangent line (negative because the slope represents one quantity going down while the other goes up). This slope is just dy/dx, which we know how to compute via implicit differentiation. In the pizza and FroYo example, the level curve is $xy = 30$ and implicit differentiation gives $y + x(dy/dx) = 0$. Thus $dy/dx = -y/x$. At the point $(60, 1/2)$, this gives a marginal rate of substitution of $1/120$ pint of FroYo per ounce of pizza. On the other hand, at the point $(3, 10)$, the marginal rate of substitution is $10/3$ pints of FroYo per ounce of pizza. Whether or not you think xy is a reasonable utility function for this scenario, this model sheds light on consumer behavior and how to model it.
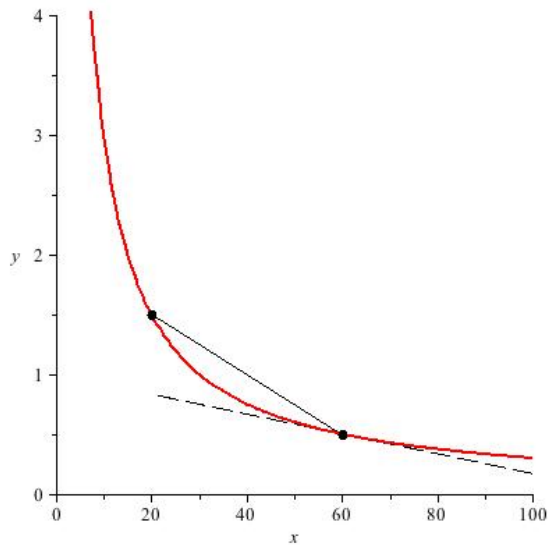
Figure 12: difference quotients and tangents represent rates and marginal rates

Exercise 2.7. *Suppose I change the utility function $u(x, y) = xy$ to some other function, say $xy^2/(1 + x + y^2)$.*

  (i) *Do the units of the marginal rate of substitution of FroYo for pizza change?*

 (ii) *Does the actual marginal rate change?*

2.3   Applications: spatial totals, averages, and joint probabilities

There's no new formulas or techniques in this section, just some applications. Two of them are pretty straightforward: integrals to yield total amounts and integrals to compute averages. The third, probability densities in two variables, will involve a couple of new concepts.

Integrals to compute totals

This is essentially just a reminder that the integral of stuff per unit area over an area yields total stuff.

36

Example 2.5. The population density east of a river running north-south is $f(x, y) = 6000e^{-x^2}$ people per square mile. The county is divided into quadrants as shown in the figure. Roughly how many people are there in the east quadrant?
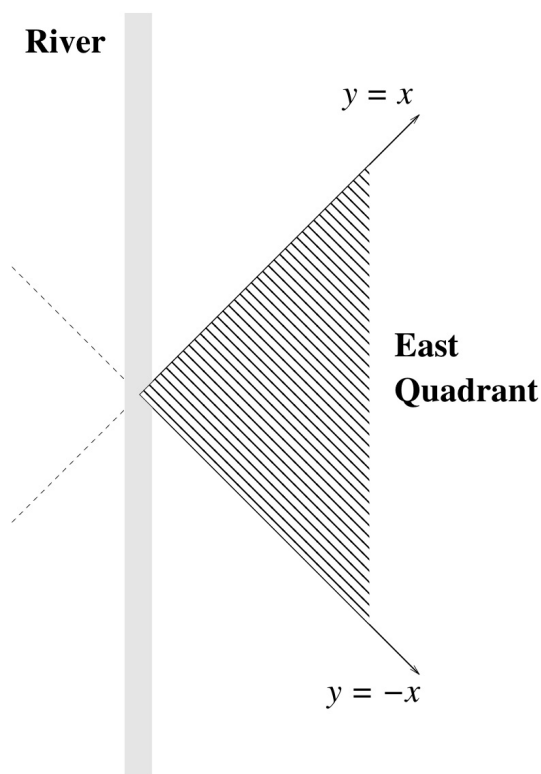
SOLUTION: Make coordinates in which the quadrant is the region represented in standard form by the set

$$\{(x, y) : 0 \le x < \infty, -x \le y \le x\}.$$

We're not going to re-do the theory of improper integrals in two variables, we'll only deal with cases where you can just plug in ∞ and get the right answer. The region is in standard form, so the total population is given by

$$\int_0^\infty \int_{-x}^x 6000e^{-x^2}\, dy\, dx.$$

The inner integral might look tough but it's not (look carefully at which is the variable of integration):

$$\int_{-x}^x 6000\, e^{-x^2}]\, dy = 6000\, ye^{-x^2}\Big|_{-x}^x = 12000\, xe^{-x^2}.$$

The outer integral can then be done by the substitution $u = e^{-x^2}$, leading to

$$\int_0^\infty 12000\, xe^{-x^2}\, dx = -6000\, e^{-x^2}\Big|_0^\infty = 0 - (-6000) = 6000\,.$$

This is a good example of an integral which is not too hard one way but impossible the other. Try to integrate e^{-x^2} against dx rather than dy and you will be stuck at the first step! If you come across an integral like this one, you will always want to switch the order of the integrals because $\int e^{-x^2}\, dx$ cannot be expressed in terms of functions you know!

Averages

The average of a quantity over a region is just the total of the quantity divided by the size of the region.

Example 2.6. What is the average of e^x over the triangular region where $0 \le x \le y \le 1$?
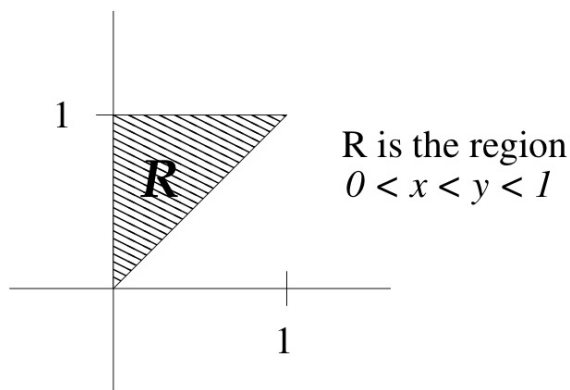
SOLUTION:

$$\int_0^1 \int_0^y e^x\, dx\, dy = \int_0^1 \left(e^x\big|_0^y \right) dy = \int_0^1 (e^y - 1)\, dy = e - 2\,.$$

The average value is the total, $e - 2$, divided by the area. The area is $1/2$, so the average value is $2(e - 2)$.

Exercise 2.8.
1. Compute the average over this region of e^{-x}, which is of course the reciprocal of e^x.
2. Is the average from the previous part greater or less than the reciprocal of $2(e - 2)$?
Bonus question: could you have predicted that?

Example 2.7. The cost of providing fiber optic service to a resident is proportional to the distance to the nearest hub, with constant of proportionality 5 dollars per meter. If a township is a square, two kilometers on a side, and there is a single hub in the center, what is the average of the service cost over this area?

SOLUTION: Make coordinates with the hub in the center. The township is the square $[-1000, 1000] \times [-1000, 1000]$, with x and y representing East-West displacement and North-South displacement in meters. The cost of providing service to the point (x, y) is given by $f(x, y) = 5\sqrt{x^2 + y^2}$. The average is therefore given by

$$\text{Ave.} = \frac{1}{1000^2} \int_R 5\sqrt{x^2 + y^2} \, dA = \frac{1}{1000^2} \int_{-1000}^{1000} \int_{-1000}^{1000} 5\sqrt{x^2 + y^2} \, dy \, dx \, .$$

If you can do this integral, you are much better at integration than we are. We tried it numerically with a 5×5 grid (I used midpoints and we used symmetry to restrict to the quadrant $[0, 1000] \times [0, 1000]$ in order to make my grid squares smaller) and got roughly $3812, which is pretty close to what my computer tells me is the correct numeric value of $3826.

Two-variable probability densities

It is often useful to consider a random pair of real numbers, that is, a random point in the plane. A probability density on the plane[3] is a nonnegative function $f(x, y)$ such that $\int f(x, y) dA = 1$. As before, the mean of the X variable is $\int x f(x, y) \, dA$ and the mean of the Y variable is $\int y f(x, y) \, dA$. Here are a couple of special cases.

Example 2.8 (uniform density on a region). Let R be a finite region and let $f(x, y) = C$ on R and zero elsewhere. For this to be a probaiblity denstiy, the normalizing constant C must be the reciprocal of the area of R (that's because the integral of $1 \, dA$ over R is just the area of R). For example, if R is the unit disk then C would be $1/\pi$. If R is the rectangle $[a, b] \times [c, d]$ then $C = 1/((b - a)(d - c))$.

Exercise 2.9. *Suppose (X, Y) is uniform over the rectangle $[0, 2] \times [0, 5]$. What is the average of $x^2 y$ over this interval? Hint: shorten your work a little with the magic product formula.*

---

[3]The integral, if it is over the whole plane, is technically an improper integral, but we won't worry about that; in all our examples either the density will be nonzero on just a finite region or it will be obvious that there is a limit as the region becomes infinite.

Example 2.9 (planar standard normal distribution). Let $f(x, y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2}$. If you know the density of the standard normal in one variable, then you know this two-variable density integrates to 1 over the whole plane. The easiest way to see this is to use the magic product formula, which works because

$$\frac{1}{2\pi} e^{-(x^2+y^2)/2} = \left(\frac{1}{\sqrt{2\pi}} e^{-x^2/2} \right) \left(\frac{1}{\sqrt{2\pi}} e^{-y^2/2} \right)$$

each factor of which we already know integrates to 1 because it is the one-variable standard normal density. [We slipped in the notion of improper integrals in two variables, but that topic works the way you expect it to, so there's no reason to stop for a big detour now.]

A two-variable probability density corresponds to picking simultaneously two numbers X and Y such that the probability of finding the pair (X, Y) in some region A is equal to the integral of the density over the region A.

Exercise 2.10 (+). *Suppose we ask for the chance that a pair (X, Y) picked from the planar standard normal distribution is in a small region of area 0.01 near the point $(1, 0)$. What would this be, approximately?*

Example 2.10. A probability density on the rectangle $[0, 3] \times [0, 2]$ is given by Ce^{-x}. What is C, and what is the probability of finding the pair (X, Y) in the unit square $[0, 1] \times [0, 1]$?

SOLUTION: We are integrating over a rectangle and Ce^{-x} is a product of $g(x) = Ce^{-x}$ and $h(y) = 1$. By the magic product formula, the integral is

$$\left(\int_0^3 Ce^{-x} \, dx \right) \times \left(\int_0^2 dy \right) = 2C \cdot \left(1 - e^{-3} \right) .$$

Therefore $C = \dfrac{1}{2(1 - e^{-3})}$ which is just a shade over $1/2$. Now using the product formula again to integrate over the unit square gives a probability of

$$\int_{[0,1] \times [0,1]} Ce^{-x} \, dA = C(1 - e^{-1}) = \frac{1 - e^{-1}}{2(1 - e^{-3})} \approx 0.3326.$$

Exercise 2.11 (+). *A probability density on the rectangle $[0, 1] \times [0, 1]$ is given by $x + y$. What is the probability that $x < y$?*

3 Linear algebra 1: vectors and matrices

3.1 Vectors

In order to understand functions of multiple variables, we will need to understand spaces with more than one dimension. This is the subject of *linear algebra.*

A full course in linear algebra occupies a semester (this is the topic of Math 312 and Math 314). In this course, we will devote a few weeks to introducing some of the basic concepts:

- Familiarity with vectors, matrices and the various notations commonly used;

- Understanding vectors and matrices from several points of view: the algebraic rules they follow, their geometric properties, and the concepts they represent;

- How to compute matrix products, the dot product, determinants, inverses and so on;

- Some ways vectors and matrices can be useful: solving equations and inequalities, solving multidimensional max-min problems, encoding geometric concepts such as projections and coordinates;

- An introduction to linear maps on n-dimensional space.

We begin in the plane, also known as $\mathbb{R}^2$, where pictures are easy to draw and understand, then discuss how everything works in higher dimensions. In fact, this is one of the values of linear algebra: it allows you to harness existing intuition in two and three dimensions to think and solve problems in dimensions too high to easily envision.

Two dimensions

Think of a vector as an arrow drawn from one point in the plane (or three dimensional space) to another. The arrow from $(1, 1)$ to $(2, 3)$ is shown in the figure. The only tricky thing about the definition is that we don't care where the arrow is drawn; we only care about its magnitude (length) and direction. So for example, in Figure 13, the dashed arrow represents the same vector, with its **tail** at the point $(5/2, 0)$ instead of $(1, 1)$ and its **head** at $(7/2, 2)$ instead of $(2, 3)$. In other words, the vector represents the *move* from the tail to the head of the arrow, regardless of the absolute location of the tail.
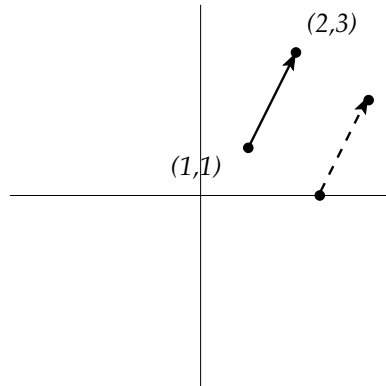
Figure 13: two depictions of the same vector

Exercise 3.1. *Draw the vector in Figure 13, but with its tail at $(-2, -3)$.*

The vector of unit length in the x-direction is called $\hat{\mathbf{i}}$ and the vector of unit length in the y-direction is called $\hat{\mathbf{j}}$. A vector that goes a units in the x-direction and b units in the y-direction is denoted $a\hat{\mathbf{i}} + b\hat{\mathbf{j}}$. It's called that because you can add vectors and multiply them by real numbers (coming shortly in Definition 3.1 below). For example, the vector in Figure 13 can be written as $\hat{\mathbf{i}} + 2\hat{\mathbf{j}}$.

Exercise 3.2. *Draw $\hat{\mathbf{i}}$ with its tail at the origin and $\hat{\mathbf{j}}$ with its tail at the head of $\hat{\mathbf{i}}$.*

Here follow the basic definitions for adding vectors and multipliying vectors by real numbers.

Definition 3.1.

1. *In the context of vectors, a **scalar** is just a real number.*

2. ***Adding two vectors**: First make the move represented by the first vector, then make the move represented by the second. You can do this by sliding one of the arrows (don't rotate it!) so it starts where the other one ends, then following them both. If you add $a\hat{\mathbf{i}} + b\hat{\mathbf{j}}$ to $c\hat{\mathbf{i}} + d\hat{\mathbf{j}}$ you get $(a + c)\hat{\mathbf{i}} + (b + d)\hat{\mathbf{j}}$.*

3. ***Multiplying a vector by a scalar**: Don't change the direction, just multiply the length. As a formula: multiply $a\hat{\mathbf{i}} + b\hat{\mathbf{j}}$ by c you get $ac\hat{\mathbf{i}} + bc\hat{\mathbf{j}}$. This easy formula hides an important fact: if you mutliply both the $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ coefficients by the same real number, the direction doesn't change. That's why the two vectors in the right-hand figure below are on top of each other.*
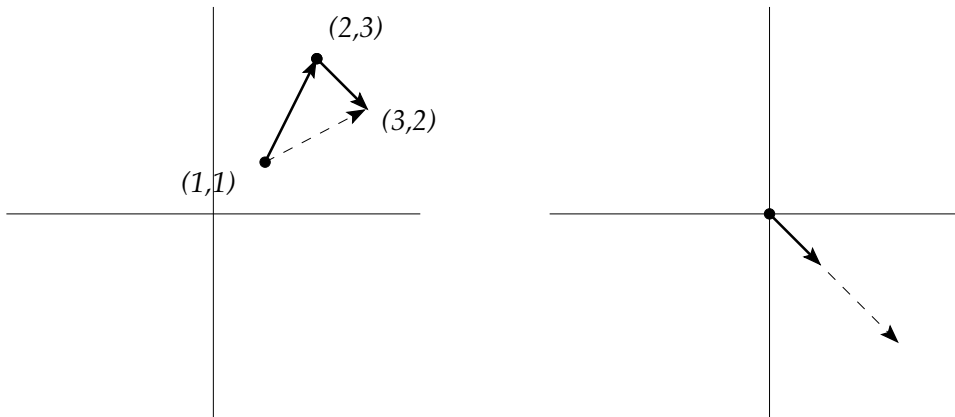
Figure 14: adding vectors and multiplying a vector by a scalar

The left-hand side of Figure 14 shows the vector $\hat{\mathbf{i}}+2\hat{\mathbf{j}}$ being added, tip to tail, to the vector $\hat{\mathbf{i}}-\hat{\mathbf{j}}$. The result is the vector $2\hat{\mathbf{i}}+\hat{\mathbf{j}}$ show by the dotted arrow. On the right-hand side, the vector $\hat{\mathbf{i}}-\hat{\mathbf{j}}$ is multiplied by the real number $\sqrt{6}$ which is a little under $2\frac{1}{2}$.

Exercise 3.3. *Let* $\mathbf{v} := \hat{\mathbf{i}} + 2\hat{\mathbf{j}}$ *and* $\mathbf{w} := \hat{\mathbf{i}} - \hat{\mathbf{j}}$. *Draw a picture illustrating* $3\mathbf{v} + \mathbf{w}$.

The length of a vector can be computed by the Pythagorean Theorem. The length of $a\hat{\mathbf{i}}+b\hat{\mathbf{j}}$ is $\sqrt{a^2 + b^2}$. For example, the vector $\hat{\mathbf{i}}+2\hat{\mathbf{j}}$ which appears in the previous figures has length $\sqrt{5}$. The length of the vector $\mathbf{v}$ is denoted $|\mathbf{v}|$. A *unit vector* is any vector whose length is 1. Often we want to know a unit vector in a given direction: what vector, having the same direction as $\mathbf{v}$, has length 1? Answer: divide $\mathbf{v}$ by $|\mathbf{v}|$. (That is, multiply $\mathbf{v}$ by the reciprocal of its length.)

Exercise 3.4. *What is the unit vector in the direction of* $\mathbf{v} := \hat{\mathbf{i}} + 2\hat{\mathbf{j}}$?

The dot product

Definition 3.2 (dot product)**.** *The* **dot product** *of the vectors* $a\hat{\mathbf{i}}+b\hat{\mathbf{j}}$ *and* $c\hat{\mathbf{i}}+d\hat{\mathbf{j}}$ *is defined to be* $ac+bd$. *Warning: the dot product does not produce a vector, it produces a scalar. For this reason it is sometimes called the* **scalar product***.*

At this point you have no reason to believe the dot product is either natural or important.

We promise: all will be revealed shortly! The other thing you need to know about the dot product is a fact: the dot product of two vectors $\mathbf{v}$ and $\mathbf{w}$ is equal to the product of the lengths times the cosine of the angle θ between them. As an equation:

$$\mathbf{v} \cdot \mathbf{w} = |\mathbf{v}||\mathbf{w}| \cos \theta \,. \tag{3.8}$$

There is something important hidden in the content of this formula. You already know one way of computing the dot product: multiply corresponding components and add them. The formula gives you another way. The first way is algebraic. The second way is completely geometric: you could do it by seeing only the picture. The dot product theorem says that these two computations produce the same result[4]. Take a minute to register this, because it will come up in the problem sets, as well as in applications.

Exercise 3.5. *Let $\mathbf{v}$ denote the vector in the Southeast (down and to the right) diagonal direction with length $\sqrt{18}$. What is $\mathbf{v} \cdot \hat{\mathbf{j}}$?*

Parallel vectors

Vectors in the same direction are called **parallel**. How do you tell whether $\mathbf{v} = a\hat{\mathbf{i}} + b\hat{\mathbf{j}}$ is parallel to $\mathbf{w} = c\hat{\mathbf{i}} + d\hat{\mathbf{j}}$? This is the same as saying you can multiply one vector by real number to get the other. This is the same as asking when the fraction c/a is equal to d/b. To test this you can cross-multiply, arriving at the condition

$$ad - bc = 0 \,. \tag{3.9}$$

Exercise 3.6. *Which one or more of these vectors is parallel to $3\hat{\mathbf{i}} - 4\hat{\mathbf{j}}$?*

(i) $3\hat{\mathbf{i}} + 4\hat{\mathbf{j}}$

(ii) $31\hat{\mathbf{i}} - 41\hat{\mathbf{j}}$

(iii) $-3\hat{\mathbf{i}} + 4\hat{\mathbf{j}}$

(iv) $0.3\hat{\mathbf{i}} - 0.4\hat{\mathbf{j}}$

(v) $4\hat{\mathbf{i}} - 3\hat{\mathbf{j}}$

---

[4]The dot product theorem says that (3.8) is true. The proof is some analytic geometry. We don't give it here because it is an unenlightening combination of analytic geometry and similar triangles.

Three or more dimensions

The situation in three dimensions is pretty much like it is in two dimensions. Now there is a third direction, $\hat{\mathbf{k}}$. Vectors may still be defined by length and direction. Every vector can be written as $a\hat{\mathbf{i}} + b\hat{\mathbf{j}} + c\hat{\mathbf{k}}$. Algebraic rules for adding and multiplying by a scalar are the same. The geometric versions of these rules are the same too: add by putting tip to tail; multiply by increasing or decreasing the length, keeping the same direction. The dot product is analgous too:

$$(a\hat{\mathbf{i}} + b\hat{\mathbf{j}} + c\hat{\mathbf{k}}) \cdot (d\hat{\mathbf{i}} + e\hat{\mathbf{j}} + f\hat{\mathbf{k}}) = ad + be + cf \,.$$

After three dimensions, we start to run out of intuition (can you picture tip to tail adding in four dimensions?), and will eventually run out of reserved letters as well $(\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}, \ldots)$.

Definition 3.3 (n-dimensional vectors)**.** *A column vector* $\mathbf{v}$ *in n-dimensional real space* $\mathbb{R}^n$ *is a column of n real numbers* $v_1, \ldots, v_n$ *written as*

$$\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}.$$

This represents the arrow going from the origin to the point $(v_1, \ldots, v_n)$, *even if we can't really visualize points in* $\mathbb{R}^n$.

Notation: v_j will always stand for the jth coordinate of $\mathbf{v}$, and similarly for other boldface letters, $\mathbf{x}, \mathbf{y}, \mathbf{u}$, etc.

This gives us a new way to write our $\mathbb{R}^2$ and $\mathbb{R}^3$ vectors, as well: instead of writing $2\hat{\mathbf{i}} + 3\hat{\mathbf{j}}$, we could also write $\begin{bmatrix} 2 \\ 3 \end{bmatrix}$. There is a third way we will write vectors (because vertical columns look odd in the middle of horizontal text): we will sometimes write them as a horizontal list using parentheses, like $(2, 3)$. (When we write like this, we have to be careful not to confuse vectors with points.) Which notation we use will be a matter of convenience, and you should be prepared to translate between them.

A few vectors are important enough to get their own names. We write $\mathbf{0}$ for the vector all of whose values are 0. We usually let the reader figure out the dimension from context—$\mathbf{0}$

might mean $\begin{bmatrix} 0 \end{bmatrix}$ or $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ or $\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$, and we should write carefully enough that it's easy to figure out which of these we mean.

Definition 3.4 (Standard Unit Vectors). *The **standard unit vectors** $\mathbf{e}_1, \mathbf{e}_2, \ldots \mathbf{e}_n$ in $\mathbb{R}^n$ are the vectors*

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \ , \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \ , \quad \cdots \ , \quad \mathbf{e}_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \ .$$

That is, $\mathbf{e}_j$ is the vector which has a 1 in the j-th row and 0's everywhere else.

These vectors generalize $\hat{\mathbf{i}}, \hat{\mathbf{j}}$ and $\hat{\mathbf{k}}$ in $\mathbb{R}^3$.

Again, when we write $\mathbf{e}_j$, we have to be able to figure out the dimension from context. In $\mathbb{R}^3$, the standard unit vectors are $\mathbf{e}_1 = \hat{\mathbf{i}} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, $\mathbf{e}_2 = \hat{\mathbf{j}} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$, and $\mathbf{e}_3 = \hat{\mathbf{k}} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$.

In $\mathbb{R}^5$, the standard unit vectors are

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \mathbf{e}_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \text{ and } \mathbf{e}_5 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

The standard unit vectors have an important property: we can write every vector as a sum using the standard unit vectors:

$$\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \sum_{j=1}^{n} v_j \mathbf{e}_j.$$

Exercise 3.7. *Let* **v** *be the 4-dimensional vector whose tail is at the point* $(1, 2, 3, 4)$ *and whose head is at the point* $(7, 3, -1, 12)$. *Write* **v** *as a column vector, then write it also as a sum of standard unit vectors.*

This property is important enough that it has a name: we say $\mathbf{e}_1, \ldots, \mathbf{e}_n$ is a *basis* for $\mathbb{R}^n$, and we sometimes call the vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ the *standard basis*. We will not have much opportunity to investigate other bases in this course, but they are a crucial idea if one continues to study linear algebra beyond this course.

The next definitions are natural extensions of the analogous definitions in the plane.

Definition 3.5.

1. *The length of a vector* **v** *with coordinates* $v_1, \ldots, v_n$ *is the quantity* $\sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$.

2. *The sum of two vectors in* $\mathbb{R}^n$ *is defined by*
$$\mathbf{v} + \mathbf{w} := \sum_{j=1}^{n} (v_j + w_j)\mathbf{e}_j \, .$$

3. *The product of the scalar* λ *with the vector* $\mathbf{v} \in \mathbb{R}^n$ *is defined by*
$$\lambda \mathbf{v} := \sum_{j=1}^{n} \lambda v_j \mathbf{e}_j \, .$$

4. *The dot product in* $\mathbb{R}^n$ *is defined by*
$$\mathbf{v} \cdot \mathbf{w} := \sum_{j=1}^{n} v_j w_j \, .$$

5. *Informally, because it is easier to write on paper, we sometimes say, "let* $\mathbf{x} := (x_1, \ldots, x_n)$,*" instead of, "let* $\mathbf{x} := x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n$ *be the vector from the origin to* $(x_1, \ldots, x_n)$.*"*

Example 3.6. Let $\mathbf{v} := (0, 0, 3)$ and $\mathbf{w} := (1, 1, 0)$, and describe the set of all $(x, y, z) \in \mathbb{R}^3$ that can be written as $\mathbf{v} + \lambda \mathbf{w}$ for some real λ. What is the length of $\mathbf{v} + \lambda \mathbf{w}$?

We could also have written $\mathbf{v} = 3\hat{\mathbf{k}}$ and $\mathbf{w} = \hat{\mathbf{i}} + \hat{\mathbf{j}}$. The set $\{\mathbf{v} + \lambda \mathbf{w}\}$ may be written as the set
$$\left\{ \begin{bmatrix} \lambda \\ \lambda \\ 3 \end{bmatrix} \; : \; \lambda \in \mathbb{R} \right\} \, .$$

Consider what these vectors look like: if we put all the tails at the origins, the heads of the vectors will be at the coordinates $(\lambda, \lambda, 3)$; these points are a line in $\mathbb{R}^3$. Because its coordinates are $(\lambda, \lambda, 3)$, the length of $\mathbf{v} + \lambda\mathbf{w}$ is given by $|\mathbf{v} + \lambda\mathbf{w}| = \sqrt{2\lambda^2 + 9}$.

3.2   Matrices

Once we have written vectors, it is only a small step to writing *matrices.* A $n \times m$ matrix is an array of numbers $\{a_{ij} : 1 \le i \le n, 1 \le j \le m\}$ with n rows. It is written in rectangular form with i counting the rows $1, \dots, n$ and j counting the columns $1, \dots, m$. For example, if $m = n = 2$ and a 2×2 matrix M has a_{11}, a_{12}, a_{21} and a_{22} equal to 2,3,5 and 7 respectively, then the matrix would be denoted

$$M = \begin{bmatrix} 2 & 3 \\ 5 & 7 \end{bmatrix}.$$

When we have big matrices with variables, we sometimes write them like

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix}.$$

When we write a_{ij}, the first coordinate, i, denotes which *row* we're in, and the second coordinate, j, denotes which *column* we're in.

Sometimes matrices can be multiplied, but only when their coordinates match in the right way.

Rule for when matrices can be multiplied. When P is a $p \times q$ matrix, and M is a matrix, we may only multiply PM when M is a $q \times n$ matrix. When this happens, PM will be a $p \times n$ matrix.

Unlike the multiplication we are used to, matrix multiplication is *non-commutative*—the order we multiply in matters. When PM is a matrix, MP may not even be an allowed multiplication; even when MP is defined, it may not have the same dimensions as PM, and even if it does have the same dimensions, it may be different from PM.

When PM is a $p \times n$ matrix, we need to figure out what the value is at each coordinate.

Rule for calculating the coordinates in matrix multiplication When we multiply PM, we get a matrix R. The value r_{ij}—the value at the i-th row and j-th column of R—is the dot product of the *i-th row of P* with the *j-th column of M*.

Let's see what that looks like. We'll let $p = 3$ and $q = n = 2$, and take the matrices

$$P = \begin{bmatrix} 4 & -3 \\ 0 & 1 \\ 2 & 1 \end{bmatrix} \text{ and } M = \begin{bmatrix} 2 & 3 \\ 5 & 7 \end{bmatrix}.$$

Then PM will be a 3×2 matrix as well. We will calculate the coordinates of PM one at a time:

$$PM = \begin{bmatrix} \boxed{-7} & - \\ - & - \\ - & - \end{bmatrix} \text{ from } P = \begin{bmatrix} \boxed{4} & \boxed{-3} \\ 0 & 1 \\ 2 & 1 \end{bmatrix} \text{ and } M = \begin{bmatrix} \boxed{2} & 3 \\ \boxed{5} & 7 \end{bmatrix} \text{ since } -7 = 4{\cdot}2 + (-3){\cdot}5.$$

$$PM = \begin{bmatrix} -7 & \boxed{-9} \\ - & - \\ - & - \end{bmatrix} \text{ from } P = \begin{bmatrix} \boxed{4} & \boxed{-3} \\ 0 & 1 \\ 2 & 1 \end{bmatrix} \text{ and } M = \begin{bmatrix} 2 & \boxed{3} \\ 5 & \boxed{7} \end{bmatrix} \text{ since } -9 = 4{\cdot}3 + (-3){\cdot}7.$$

$$PM = \begin{bmatrix} -7 & -9 \\ \boxed{5} & - \\ - & - \end{bmatrix} \text{ from } P = \begin{bmatrix} 4 & -3 \\ \boxed{0} & \boxed{1} \\ 2 & 1 \end{bmatrix} \text{ and } M = \begin{bmatrix} \boxed{2} & 3 \\ \boxed{5} & 7 \end{bmatrix} \text{ since } 5 = 0 \cdot 2 + 1 \cdot 5.$$

$$PM = \begin{bmatrix} -7 & -9 \\ 5 & \boxed{7} \\ - & - \end{bmatrix} \text{ from } P = \begin{bmatrix} 4 & -3 \\ \boxed{0} & \boxed{1} \\ 2 & 1 \end{bmatrix} \text{ and } M = \begin{bmatrix} 2 & \boxed{3} \\ 5 & \boxed{7} \end{bmatrix} \text{ since } 7 = 0 \cdot 3 + 1 \cdot 7.$$

$$PM = \begin{bmatrix} -7 & -9 \\ 5 & 7 \\ \boxed{9} & - \end{bmatrix} \text{ from } P = \begin{bmatrix} 4 & -3 \\ 0 & 1 \\ \boxed{2} & \boxed{1} \end{bmatrix} \text{ and } M = \begin{bmatrix} \boxed{2} & 3 \\ \boxed{5} & 7 \end{bmatrix} \text{ since } 9 = 2 \cdot 2 + 1 \cdot 5.$$

$$PM = \begin{bmatrix} -7 & -9 \\ 5 & 7 \\ 9 & \boxed{13} \end{bmatrix} \text{ from } P = \begin{bmatrix} 4 & -3 \\ 0 & 1 \\ \boxed{2} & \boxed{1} \end{bmatrix} \text{ and } M = \begin{bmatrix} 2 & \boxed{3} \\ 5 & \boxed{7} \end{bmatrix} \text{ since } 13 = 2 \cdot 3 + 1 \cdot 7.$$

Σ notation lets us write this very compactly.

Definition 3.7 (matrix multiplication). *Suppose $P = (p_{ij})$ is a $p \times q$ matrix and $M = (m_{ij})$ is a $q \times n$ matrix. Then PM is defined to be the $p \times n$ matrix R whose (i, j)-element is defined by*

$$r_{ij} := \sum_{k=1}^{n} p_{ik} m_{kj}. \tag{3.10}$$

50

$$PM, MP, PS, SP, MS, SM.$$

Which of these six potential products are defined, and when defined, what are their dimensions?

Definition 3.8 (matrix-vector multiplication). *A column vector $\mathbf{v} \in \mathbb{R}^m$ can be viewed an a $m \times 1$ matrix. Let $A = (a_{ij})$ be a $n \times m$ matrix. By the definition of matrix multiplication, $A\mathbf{v} = \mathbf{y} \in R^n$ where $\mathbf{y}_i = \sum_{j=1}^{m} a_{ij} v_j$.*

Example 3.9.

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} a \\ 0 \end{bmatrix}.$$

In other words, if A is the matrix on the left, then multiplying $\mathbf{v}$ by A zeroes out the second entry of $\mathbf{v}$ but leaves the first untouched.

Theorem 3.10 (matrix multiplication is associative). *Suppose matrices A, B and C are respectively $m \times n, n \times p$ and $p \times q$. Then $(AB)C = A(BC)$. We may write ABC for this product.*

One last piece of business is to take care of a couple of pretty trivial notations.

Definition 3.11 (transpose). *The **transpose** of a $n \times k$ matrix A with entries a_{ij} is the $k \times n$ matrix, denoted A^T, defined by $a_{ij}^T := a_{ji}$.*

That is, the transpose is what we get if we flip the rows and columns:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}.$$

The transpose gives us an alternative method of writing the dot product, which will be useful at times. Let $\mathbf{v}$ and $\mathbf{w}$ be two vectors in $\mathbb{R}^n$. Then

$$\mathbf{v} \cdot \mathbf{w} = \mathbf{v}^T \mathbf{w} \tag{3.11}$$

where the left side is a dot product between two $n \times 1$ column vectors and the right side is the matrix product of a $1 \times n$ vector with a $n \times 1$ vector.

Exercise 3.9. *What are the dimensions of the result of the matrix product $\mathbf{v}^T \mathbf{w}$ in the equation 3.11 above?*

Multivariable Functions

You are used to functions $f : \mathbb{R} \to \mathbb{R}$, meaning functions that map reals to reals, and have started to work with functions $f : \mathbb{R}^2 \to \mathbb{R}$ mapping the plane to the reals. We are now ready to start thinking about more general functions from $\mathbb{R}^m$ to $\mathbb{R}^n$.

First, let us think about functions from $\mathbb{R}^m$ to $\mathbb{R}$. When $m = 1$, this might be a function $f(x)$, and when $m = 2$, we often write $f(x, y)$. When $m = 3$, we might write $f(x, y, z)$, but using new letters for each variable quickly gets unwieldy. Another solution is to write the variables with subscripts: a rather complicated example of a function from $\mathbb{R}^7$ to $\mathbb{R}$ could be

$$f(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = \frac{x_1 + x_2 x_3}{x_4 + x_5 e^{x_6}} + x_7.$$

Now that we have seen vectors, however, we realize we have another option: we could write the input to f as a vector, like

$$f(\mathbf{x}) = \frac{x_1 + x_2 x_3}{x_4 + x_5 e^{x_6}} + x_7.$$

When we think about functions whose input is from $\mathbb{R}^2$, we have a bunch of options: we can write $f(\mathbf{x})$, or $f(x_1, x_2)$, or $f(x, y)$, and we'll want to be comfortable switching between these as convenient.

Now that we have functions with vector *inputs*, we can also think about functions with vector *outputs*: functions from $\mathbb{R}^m$ to $\mathbb{R}^n$ where n can be bigger than 1.

We often write the output as a horizontal vector, like

$$f(r, \theta) = (r \cos \theta, r \sin \theta).$$

This is a function from $\mathbb{R}^2$ to $\mathbb{R}^2$: the output is the 2-dimensional vector $(r \cos \theta, r \sin \theta)$.

Exercise 3.10. *$f(x, y, z) = (x^3, y - z)$ is a function from $\mathbb{R}^m$ to $\mathbb{R}^n$ for which values of m and n?*

Exercise 3.11. *Write an example of a function from $\mathbb{R}^2$ to $\mathbb{R}^4$.*

3.3   Linear Maps

Functions of the form $f(\mathbf{v}) = A\mathbf{v}$ will have an important role in multivariable calculus. When A is an $n \times m$ matrix and $\mathbf{v}$ is a vector, this makes sense when $\mathbf{v}$ is an $m \times 1$ vector and $f(\mathbf{v}) = A\mathbf{v}$ is an $n \times 1$ vector, so $f : \mathbb{R}^m \to \mathbb{R}^n$—that is, f is a function whose *input* is an $m \times 1$ vector which then *outputs* an $n \times 1$ vector.

Later we will consider more general functions from $\mathbb{R}^m$ to $\mathbb{R}^n$. Just like we use tangent lines to simplify and understand the beahvior of functions with complex graphs, the functions which can be represented by matrix multiplication—functions of the form $f(\mathbf{v}) = A\mathbf{v}$—represent a special class of functions which will help us understand multivariable functions.

We want to notice how these functions are special: how they have properties that other functions might not.

Definition 3.12. *A function $f : \mathbb{R}^m \to R^n$ is said to be a* linear map *if these two conditions hold for all* $\mathbf{x}$ *and* $\mathbf{y}$, *and all real* λ.

1. $f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$

2. $f(\lambda \mathbf{x}) = \lambda f(\mathbf{x})$.

This definition works for ordinary single variable functions $f : \mathbb{R} \to \mathbb{R}$, but it is also perfectly suited for vector functions $f : \mathbb{R}^m \to \mathbb{R}^n$ because both sides call for only addition of two vectors or multiplication of a vector by a scalar.

Theorem 3.13. *When A is an $n \times m$ matrix, the function $f : \mathbb{R}^n \to \mathbb{R}^m$ given by $f(\mathbf{v}) = A\mathbf{v}$ is a linear map.*

As we will discuss below, with some work we can discover that these are the *only* linear maps.

When seeing multivariable functions for the first time, we should remember to rewrite things in coordinate notation for small values of m and n to help us see what is going on. For instance, when $f : \mathbb{R}^2 \to \mathbb{R}$, these conditions say

$$f(x_1 + y_1, x_2 + y_2) = f(x_1, x_2) + f(y_1, y_2), \quad \text{and } f(\lambda x_1, \lambda x_2) = \lambda f(x_1, x_2).$$

Exercise 3.12. *If an ordinary function $f : \mathbb{R} \to \mathbb{R}$ is linear, does it follows that $f(0) = 0$?*

Exercise 3.13. *Show that the function $f(x, y)$ given by $f(x, y) = xy$ is not linear by choosing an input $\mathbf{x} = (x, y)$ and a λ so that $f(\lambda \mathbf{x}) \neq \lambda f(\mathbf{x})$.*

Example 3.14. Suppose $f : \mathbb{R} \to \mathbb{R}$ is linear and $f(5) = 1$. Do we have enough information to determine $f(12)$? The answer is yes. Because any nonzero real number is a multiple of any other, we can write $12 = 5\lambda$, where in this case $\lambda = 12/5$. By property 2 of linear functions, $f(12) = f(5\lambda) = \lambda f(5) = (12/5)1 = 12/5$.

As you saw from this example, there are not very many linear functions on $\mathbb{R}^1$. Once you know the value of $f(x)$ for any nonzero x, it determines the entire function. Similarly, for functions $f : \mathbb{R}^1 \to \mathbb{R}^n$, if you know $f(x) = \mathbf{v}$ then for any input y, you know $f(y) = (y/x)\mathbf{v}$, so you know the whole function.

What about functions $f : \mathbb{R}^m \to \mathbb{R}^n$ when m and n are greater than 1? Let's look at the example $m = n = 2$, always a good place to obtain intuition. If you know $f(\hat{\mathbf{i}})$, you can figure out $f(t\hat{\mathbf{i}})$ for any t, but those are the only values you can determine: you don't know the whole function. On the other hand, if you know $f(\hat{\mathbf{i}}) = \mathbf{v}$ and $f(\hat{\mathbf{j}}) = \mathbf{w}$ then you know $f(a\hat{\mathbf{i}} + b\hat{\mathbf{j}}) = f(a\hat{\mathbf{i}}) + f(b\hat{\mathbf{i}}) = af(\hat{\mathbf{i}}) + bf(\hat{\mathbf{j}})$, so you know the value of f at every vector in $\mathbb{R}^2$.

We will reason through this in class. Having done so, we will have proved the following theorem.

Theorem 3.15. *Suppose the map $f : \mathbb{R}^m \to \mathbb{R}^n$ is linear. If we know the values of f at the standard basis vectors $\mathbf{e}_1, \ldots, \mathbf{e}_m$, say they are $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(m)}$, then the function f is completely determined.*

PROOF: Given $\mathbf{v} = \sum_{j=1}^{n} v_j \mathbf{e}_j \in \mathbb{R}^m$, Definition 3.12 implies that

$$f(\mathbf{v}) = f(v_1 \mathbf{e}_1 + v_2 \mathbf{e}_2 + \cdots + v_m \mathbf{e}_m) = v_i f(\mathbf{e}_1) + v_2 f(\mathbf{e}_2) + \cdots + v_m f(\mathbf{e}_m) = v_1 \mathbf{y}^{(1)} + v_2 \mathbf{y}^{(2)} + \cdots + v_m \mathbf{y}^{(m)}.$$

$\square$

Example 3.16. Suppose $f : \mathbb{R}^3 \to \mathbb{R}^3$ is linear and we know $f(\mathbf{e}_1) = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}$, $f(\mathbf{e}_2) = \begin{bmatrix} -1 \\ 3 \\ 1 \end{bmatrix}$, and $f(\mathbf{e}_3) = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$.

To find $f(\begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix})$, we remember that $\begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} = 2\mathbf{e}_1 + 4\mathbf{e}_2 + 6\mathbf{e}_3$, so

$$
\begin{aligned}
f(\begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}) &= f(2\mathbf{e}_1 + 4\mathbf{e}_2 + 6\mathbf{e}_3) \\
&= f(2\mathbf{e}_1) + f(4\mathbf{e}_2) + f(6\mathbf{e}_3) \\
&= 2f(\mathbf{e}_1) + 4f(\mathbf{e}_2) + 6f(\mathbf{e}_3) \\
&= 2\begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} + 4\begin{bmatrix} -1 \\ 3 \\ 1 \end{bmatrix} + 6\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 6 \\ 20 \\ 8 \end{bmatrix}.
\end{aligned}
$$

Representing linear maps via matrices

Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be linear, specified by $f(\mathbf{e}_j) = \mathbf{y}^{(j)}$ for $1 \le j \le m$, where $\mathbf{y}^{(j)}$ are column vectors in $\mathbb{R}^n$. Place the vectors $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(m)}$ into a matrix Y, from left to right. In other words, let $Y_{ij} = \mathbf{y}_i^{(j)}$:

$$
Y := \begin{bmatrix} \begin{pmatrix} \mathbf{y}_1^{(1)} \\ \vdots \\ \mathbf{y}_n^{(1)} \end{pmatrix} \cdots \begin{pmatrix} \mathbf{y}_1^{(m)} \\ \vdots \\ \mathbf{y}_n^{(m)} \end{pmatrix} \end{bmatrix}
$$

The notation with multiple indices can get confusing. Keep in mind the key idea: the *columns* of Y are the vectors $f(\mathbf{e}_j)$.

Matrix multiplication allows us to multiply the $n \times m$ matrix Y by any $m \times 1$ column vector $\mathbf{v}$. What do we get? Multiplication by the matrix Y computes the function f. To see this, we just need to compute it on a basis vector $\mathbf{e}_j$.

Exercise 3.14. *Check that the matrix multiplication $Y\mathbf{e}_j$ picks out column j of Y.*

Recalling that the value of G on standard basis vectors determines all of G, we see that $Y\mathbf{v} = G(\mathbf{v})$ for all $\mathbf{v}$.

Interpreting Linear Maps Geometrically

Since we can interpret vectors geometrically, we can also interpret linear maps and matrices geometrically by how they transform vectors.

Example 3.17. The matrix $A = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ transforms vectors like this:
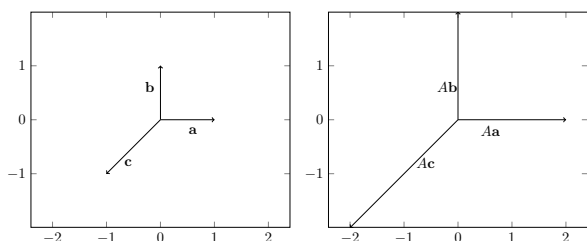
We can see that A *stretches* the vectors: it doubles their length while keeping the direction the same.

Example 3.18. The matrix $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ transforms vectors like this:
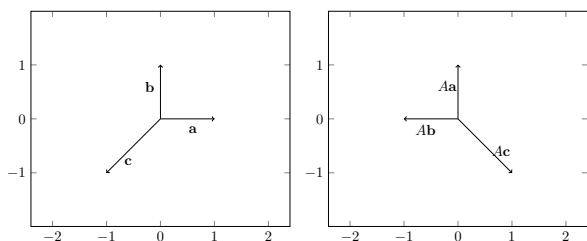
We can see that A *rotates* the vectors counterclockwise by $90°$.

Composing linear maps

Theorem 3.19. *Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^p \to \mathbb{R}^n$ are linear maps represented by the $m \times n$ matrix A and the $n \times p$ matrix B, respectively. Then the composition $f \circ g$ is a linear map from $\mathbb{R}^p$ to $\mathbb{R}^m$ represented by the matrix $m \times p$ matrix AB.*
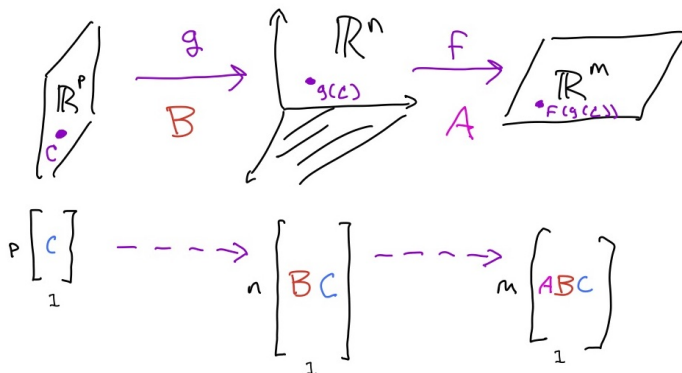
Figure 15: On top, the maps between spaces. On the bottom, what happens to a single column vector starting in the first space.

Exercise 3.15. *The matrix $A := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ rotates vectors $90°$ counterclockwise. What geometric map does A^2 represent?*

Some interesting examples and properties

The last self-check exercise is about as much work as we're comfortable foisting off on you during the intuition-building phase of this unit. Here are some more examples of linear maps $f : \mathbb{R}^2 \to \mathbb{R}^2$ and their properties.

Example 3.20. Let $f(\mathbf{v}) := A\mathbf{v}$ is a linear map from $\mathbb{R}^2$ to $\mathbb{R}^2$. For which A is this map invertible? After some experimenting, we might notice that that the matrix $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ is not, because it is not one to one on its image: $A \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = A \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Indeed, it squishes the whole plane to the diagonal line $y = x$. Formally, $A(b\mathbf{e}_1 + c\mathbf{e}_2) = (b + c)\mathbf{e}_1 + (b + c)\mathbf{e}_2$.

How can we diagnose this problem (non-invertibly squishing the plane to a line) in other maps? Letting $A := \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, we see that A maps $\mathbf{e}_1$ to $\begin{bmatrix} a \\ c \end{bmatrix}$, and $\mathbf{e}_2$ to $\begin{bmatrix} b \\ d \end{bmatrix}$. If these

57

are parallel, or if one of them is the zero vector, then everything gets squished to a single line $\left\{ \lambda \begin{bmatrix} b \\ d \end{bmatrix} : \lambda \in \mathbb{R} \right\}$ as illustrated in Figure 16.
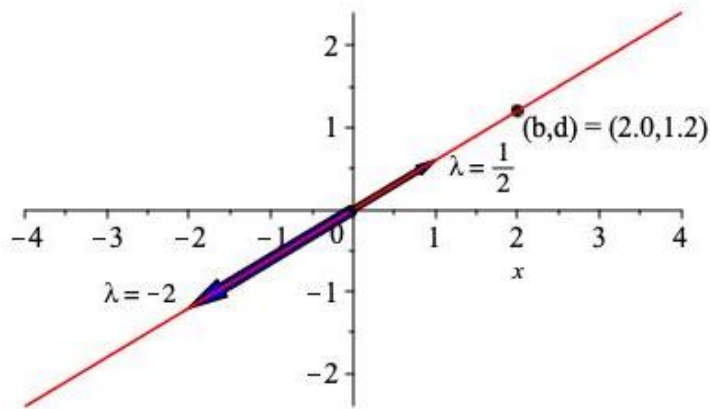
Figure 16: the set $\{\lambda(b\mathbf{e}_1 + d\mathbf{e}_2) : \lambda \in \mathbb{R}\}$

Recall we can test whether vectors (a, c) and (b, d) are parallel by checking whether $ad - bc = 0$. We conclude that when $ad - bc = 0$, the range of f is a line and f is not invertible. If $ad - bc$ is nonzero, then $\mathbf{e}_1$ and $\mathbf{e}_2$ get mapped to vectors that are not parallel. When we discuss inverse matrices, we will see that the converse is also true: $ad - bc \neq 0$ implies that f is invertible. We will also discuss criteria in dimensions great than 2, leading to the notion of a **determinant**.

Example 3.21 (computer vision and rendering). Suppose a 3-D image is to be rendered on a screen for virtual reality. This is accomplished via a function $f : \mathbb{R}^3 \to \mathbb{R}^2$ such that $f(x, y, z)$ is the intersection point with a plane (the screen or your VR glasses) with a ray from your eye to the point (x, y, z). In coordinates where your eye is at the origin, this is a linear map.

4 Derivatives in more than one variable

4.1 Partial Derivatives

One thing we would like to point out is that you've been taking partial derivatives all your calculus-life. When you compute df/dt for $f(t) = Ce^{-kt}$, you get $-Cke^{-kt}$ because C and k are constants. The notation df/dt tells you that t is the variable and everything else you see is a constant. If we use the notation f' instead, then we are relying on your knowing which is the independent variable. It's usually called something like "t" or "x", not "C" or "k", but every now and then we end up computing df/dk or df/dC, so watch out! The only rule is: everyone should understand which is the independent variable.

So now, studying partial derivatives, the only difference is that the other variables aren't constants—they vary—but you treat them as constants anyway. It's not a big difference because really, what is a constant? It's always possible to imagine some quantity changing. Mathematically we just need to be precise about what is holding steady and what is changing. In this section, only one variable at a time will change. Then in the next section (chain rule), we'll change more than one independent variable at a time and keep track of the total effect on the independent variable.

Definition 4.1. *The notation* $\dfrac{\partial f}{\partial x}$ *denotes the rate of change of f with respect to x when other variables are held fixed. This is also sometimes denoted by f_x. Formally,*

$$\left.\frac{\partial f}{\partial x}\right|_{(x_0,y_0)} = \lim_{h \to 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h} .$$

Example 4.2. Let $f(x, y, z) = ze^{x/y}$. What are $\dfrac{\partial f}{\partial x}, \dfrac{\partial f}{\partial y}$, and $\dfrac{\partial f}{\partial z}$?

When we find the partial derivative, we treat all variables other than the independent variable like constants. So $\dfrac{\partial}{\partial x} ze^{x/y} = \dfrac{z}{y} e^{x/y}$. You might be more familiar with doing this with numbers: $\dfrac{d}{dx} 7e^{x/3} = \dfrac{7}{3} e^{x/3}$, but doing it with letters y and z representing constants is no different.

Similarly, we can find $\dfrac{\partial}{\partial y} ze^{x/y}$ using the chain rule: it is $ze^{x/y} \cdot \frac{\partial}{\partial y}(x/y) = \frac{-xz}{y^2} e^{x/y}$.

Finally, $\dfrac{\partial}{\partial z} ze^{x/y} = e^{x/y}$: the whole term $e^{x/y}$ acts like a constant.

Exercise 4.1. *Let* $f(x, t, q) = \dfrac{e^q - 1}{1 + xtq}$. *What is* $\dfrac{\partial f}{\partial t}$ *at the point* $(3, 1, 1)$, *and what does this quantity signify?*

Remarks.

1. This notation only makes sense when the arguments of f are given names. For example, if f is the function $(x, y) \mapsto xy^2$ then f is also the function $(t, x) \mapsto tx^2$; all these variables are bound and their names don't matter. But in the first instance, $\partial f/\partial x = y^2$ and in the second instance $\partial f/\partial x = 2tx$.

2. To be unambiguous, it might be best always to write expressions such as $\dfrac{\partial f}{\partial x}(x, y, z)\bigg|_{(x_0, y_0, z_0)}$

so that you know what variables are implied as arguments of f and where the partial derivative is being evaluated. That is so formal that no one does it.

3. There are some real problems with the notation. We will go over some cases where ambiguities arise. You don't have to worry about it as long as you remember it can be important to specify which variables depend on which others.

Application: instantaneous rates

Suppose the cost of a proposed building is a function $f(A, q, \ell)$ where A is the area of usable space in square feet, q is an index of the quality (thickness of walls, gauge of wiring, level of insulation, quantity of lighting, etc.) and ℓ is a location parameter measuring, for example, the desirability of the location. The average cost per square foot for a given proposed building is, by definition, $f(A, q, \ell)/A$. However, this statistic is far less useful than the *marginal* cost per square foot, that is, $\partial f/\partial A$. That's because most decisions are about whether to put a few extra dollars into one of these categories or to trim a few bucks from another category. Therefore, it is most useful to look at the *instantaneous rate*: how many dollars more you will spend or save with each square foot, rather than a measure of the cost that is already in all the proposals being compared.

Example 4.3. The total number P of people exposed to an recurring ad is a function of its market share, M, and the length of time, t, that the ad stays in rotation[5]. The marginal increase in exposure per time run is $\partial f/\partial t$. The right time to yank the ad is when $v \cdot \partial f/\partial t$ drops below the cost per time to run the ad, where v is the value in dollars per unit of exposure. Note that the units match: v has units of dollars per exposure, $\partial f/\partial t$ has units

---

[5]It is not just the product of these because the longer it runs, the more redundancy there is in people seeing it multiple times.

of exposure per time and the cost to run the ad is priced in dollars per time: (\$/exp) (exp/t) = \$/t.

Note: the notion of marginal rates should already be familiar from univariate calculus. There isn't much added here, except to say that it makes sense to compute marginal rates when there are many quantities that could vary, by varying only one.

Exercise 4.2. *The monthly payment on a mortgage loan is a function $M(P, r, N)$ of the principal borrowed, P, the interest rate, r, and the number of years, N, over which the loan will be paid back. Suppose you are offered a \$250,000 loan at a 3.5% annual interest rate for 30 years. You are considering going to great lengths to increase your credit score, which will reduce your interest rate.*

Use partial derivative notation to write an expression for the amount you reduce your monthly payment for each percentage point of interest rate reduction. (You don't need to look up any formulas from outside this chapter or evaluate anything to do this. Just use partial derivative notation to write down which partial derivative you want to consider, and at what values.)

Linearization and the Increment Theorem

By now I'm sure you remember the linearization in one-variable. The value of $f(x)$ near the point $x = a$ is well approximated by $L(x) = f(a) + f'(a) \cdot (x - a)$. Suppose we now want to approximate $f(x, y)$ near a point (a, b) where we know the value. Suppose, in fact that we change only x but not y. Then we might as well treat y as a constant and write

$$f(x + \Delta x, y) \approx f(x, y) + (\Delta x) \cdot \frac{\partial f}{\partial x}(x, y).$$

It's a partial derivative, not a total derivative, because there is another variable y which is being held fixed. Similarly, if we moved only y we would have

$$f(x, y + \Delta y) \approx f(x, y) + (\Delta y) \cdot \frac{\partial f}{\partial y}(x, y).$$

I hope it doesn't seem like too much of a leap to say that if you move both x and y you'll get both of these effects:

$$f(x + \Delta x, y + \Delta y) \approx f(x, y) + (\Delta x) \cdot \frac{\partial f}{\partial x}(x, y) + (\Delta y) \cdot \frac{\partial f}{\partial y}(x, y). \tag{4.12}$$

61

Exercise 4.3. *What would this estimate look like for a function $f(x, y, z)$ of three variables?*

Linear algebra gives us a convenient notation for writing this compactly even when there are more variables. In order to do this, we need to take the assorted partial derivatives $\partial f / \partial x_i$ and gather them up into a single vector. We call this vector the *gradient*, ∇f. Take note of this definition, as it will have a number of other uses!

The gradient is a new kind of object: a **vector field**, namely a vector-valued function on $\mathbb{R}^n$—that is, a function from $\mathbb{R}^n$ to $\mathbb{R}^n$. In other words, at every $\mathbf{x}$ in $\mathbb{R}^n$ there is a vector, $\nabla f(\mathbf{x})$. You don't need to worry too much about vector fields in this course; we mention this mainly so that you are not confused about whether ∇f is a single vector versus a different vector at every point (it is the latter).

Definition 4.4 (gradient). *Let $f : \mathbb{R}^n \to \mathbb{R}$ have continuous partial derivatives. The* **gradient** *of f, denoted ∇f is the vector of partial derivatives. It can be written as a summation $\sum_{j=1}^{n} \frac{\partial f}{\partial x_j} \mathbf{e}_j$, or more formally, $\sum_{j=1}^{n} \frac{\partial f}{\partial x_j} \mathbf{e}_j(\mathbf{x})$, or in vector notation:*

$$\nabla f(\mathbf{x}) := \begin{bmatrix} \partial f / \partial x_1 \\ \vdots \\ \partial f / \partial x_n \end{bmatrix} \quad \text{evaluated at the point } \mathbf{x}.$$

All we're doing here is taking the n partial derivatives of f and, instead of writing them as n separate values, we've combined them into a single vector.

Example 4.5. If $f(x, y, z) = ze^{x/y}$, what is $\nabla f(x, y, z)$?

We already found the partial derivatives $\frac{\partial}{\partial x} ze^{x/y} = \frac{z}{y} x^{x/y}$, $\frac{\partial}{\partial y} ze^{x/y} = \frac{-xz}{y^2} e^{x/y}$, and $\frac{\partial}{\partial z} ze^{x/y} = e^{x/y}$, so

$$\nabla f(x, y, z) = \begin{bmatrix} \frac{z}{y} x^{x/y} \\ \frac{-xz}{y^2} e^{x/y} \\ e^{x/y} \end{bmatrix}.$$

Theorem 4.6 (Increment theorem). *Suppose a function $f : \mathbb{R}^n \to \mathbb{R}$ has continuous partial derivatives in a neighborhood of a point $\mathbf{x}$. Then*

$$f(\mathbf{x} + \Delta \mathbf{x}) \approx f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot \Delta \mathbf{x}. \tag{4.13}$$

Note that this is almost exactly analogous to the single variable increment theorem:

$$f(x + \Delta x) \approx f(x) + f'(x) \Delta x,$$

except that $\mathbf{x}$ is a vector, not a single real value, and we have the gradient ∇x as the appropriate analog of the derivative $f'(x)$.

As often happens, the vector notation is good for writing things compactly, but can obscure what's happening if you're not comfortable with it. Let's think about what this says when $n = 3$, so f has three inputs. Then we can write $\mathbf{x} = (x, y, z)$, $\Delta\mathbf{x} = (\Delta x, \Delta y, \Delta z)$, and

$$\nabla f = \begin{bmatrix} \partial f/\partial x \\ \partial f/\partial y \\ \partial f/\partial z \end{bmatrix}, \text{ so the Increment Theorem says}$$

$$f(x + \Delta x, y + \Delta y, z + \Delta z) \approx f(x, y, z) + \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y + \frac{\partial f}{\partial z}\Delta z.$$

It is often convenient to write this with both $f(\mathbf{x} + \Delta\mathbf{x})$ and $f(\mathbf{x})$ on the same side: remembering that $f(\mathbf{x} + \Delta\mathbf{x}) - f(\mathbf{x}) = \Delta f$, we have

Corollary 4.7.

$$\Delta f \approx \nabla f(\mathbf{x}) \cdot \Delta\mathbf{x}.$$

Example 4.8 (estimating via the increment theorem)**.** Roughly how much does $x^2/(1+y)$ change if (x, y) changes from $(10, 4)$ to $(11, 3)$? Here $\Delta x = 1$ and $\Delta y = -1$.

Remembering that the increment theorem is stated in vector notation, we can rewrite it for two values:

$$\Delta f \approx \frac{\partial f}{\partial x}(x, y)\Delta x + \frac{\partial f}{\partial y}(x, y)\Delta y.$$

We compute $\frac{\partial f}{\partial x} = 2x/(1+y)$ and $\frac{\partial f}{\partial y} = -x^2/(1+y)^2$ so $\frac{\partial f}{\partial x}(10, 4) = 4$ and $\frac{\partial f}{\partial y}(10, 4) = -4$. Thus,

$$\Delta f \approx \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y = 4(1) + (-4)(-1) = 8\,.$$

In fact, f changes from 20 to 30.25, so the 8 was kind of a crude estimate, but that's because Δx and Δy were pretty big. If we choose 0.1 and -0.1 instead, we get a linear estimate of $\Delta f = 0.8$, which is very close to the actual $0.818\ldots$.

The Increment Theorem can be applied to compute a messy function based on known values at a nearby point, but there are many other applications. One is relative error estimation, as in the following example, and absolute error estimation, which is a straightforward application of the increment theorem as in self-check exercise 4.4.

Example 4.9. An important related example is relative error propagation under multiplication. Suppose we expect relative errors of no more than $E_x = \Delta x/x$ and $E_y = \Delta y/y$. What kind of relative error might we expect on the product xy? Using the Increment Theorem for $f(x, y) := xy$ and plugging in relative errors,

$$
\begin{aligned}
(x + \Delta x)(y + \Delta y) &= f(x + \Delta x, y + \Delta y) \\
&\approx xy + \Delta x \frac{\partial f}{\partial x} + \Delta y \frac{\partial f}{\partial y} \\
&= xy + (\Delta x)(y) + (\Delta y)(x)
\end{aligned}
$$

and therefore the relative error E_{xy} in xy is given by

$$
\begin{aligned}
E_{xy} &= \frac{(x + \Delta x)(y + \Delta y) - xy}{xy} \\
&= \frac{f(x + \Delta x, y + \Delta y) - xy}{xy} \\
&\approx \frac{xy + (\Delta x)(y) + (\Delta y)(x) - xy}{xy} \\
&= \frac{\Delta x}{x} + \frac{\Delta y}{y} \\
&= E_x + E_y \, .
\end{aligned}
$$

The take home point is that relative errors add under multiplication. The function xy is simple enough that you can arrive at the same result by algebra. However, in complicated expressions such as $x^{3.2} y^2 / z$, algebra is less useful but calculus still gives the correct relative total error.

Exercise 4.4 (+). *The cost of fuel for a year's worth of trucking will be xy where x the average cost of fuel per gallon over the coming year and y is the number of miles trucked over the coming year. The company estimates these based on last year's numbers: \$2.86 per gallon and 12,000,000 miles trucked. If their price might fluctuate up to 10 cents per gallon and their miles trucked could be off by $\pm 1,000,000$, what error could there be in the fuel budget?*

With the concept of gradient and the increment theorem in hand, we can go back to the concept of linearization. The increment theorem defines a multivariate linearization, $L_{\mathbf{a}}$, analogous to the tangent line approximation L_a from the first section. Let $f : \mathbb{R}^n \to \mathbb{R}$ have continuous partial derivatives in a neighborhood of a point $\mathbf{a}$ in $\mathbb{R}^n$.

Definition 4.10. *The linearization $L := L_{\mathbf{a}}$ of f near $\mathbf{a}$ is defined by*

$$L_{\mathbf{a}}(\mathbf{x}) := f(\mathbf{a}) + \nabla f(\mathbf{a}) \cdot (\mathbf{x} - \mathbf{a}).$$

Exercise 4.5. *Write the linearization explicitly in coordinate notation when $f : \mathbb{R}^2 \to \mathbb{R}$, so it should say $L_{a,b}(x, y) = \cdots$.*

With this notation, we can restate the Increment Theorem very compactly.

Theorem 4.11 (Increment theorem, restated). *Suppose a function $f : \mathbb{R}^n \to \mathbb{R}$ has continuous partial derivatives in a neighborhood of a point $\mathbf{a}$. Then*

$$f(\mathbf{a} + \mathbf{x}) \approx L_{\mathbf{a}}(\mathbf{x}). \tag{4.14}$$

Mapping to $\mathbb{R}^k$

We have talked about functions from $\mathbb{R}^n$ to $\mathbb{R}^1$, mostly with $n = 2$ or 3, but not much about functions $f : \mathbb{R}^n \to \mathbb{R}^k$ with $k > 1$, other than to discuss linear mappings. It turns out we have already done the hardest work. Making the range multidimensional is much easier than making the domain multidimensional. A function $f : \mathbb{R}^n \to \mathbb{R}^k$ can be thought of as a k-tuple of functions $(f_1, \ldots, f_k)$. You could just do calculus on these independently, then put the results together. With matrix notation, in fact this is already done for you, as we now discuss.

Let $f = (f_1, \ldots, f_k) : \mathbb{R}^n \to \mathbb{R}^k$ be a function with continuous partial derivatives, mapping R^n to $\mathbb{R}^k$. To examine the linearizations of such a map at the point $\mathbf{a} \in \mathbb{R}^n$, we can write down k linearizations

$$f_j(\mathbf{x}) \approx f_j(\mathbf{a}) + \nabla f_j(\mathbf{x})^T (\mathbf{x} - \mathbf{a}), \text{ for } 1 \le j \le k. \tag{4.15}$$

We have written the gradient in its transposed form, as a row vector, so we have room to make a column of them. The matrix product of a $k \times n$ matrix M with a $n \times 1$ column vector is a $k \times 1$ column vector whose entries are the products of the successive rows of M. Assembling the k row vectors $\nabla f(\mathbf{a})^T$ into a $k \times n$ matrix, we see that the k linearizations in (4.15) can be assembled to yield

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \begin{bmatrix} \nabla f_1(\mathbf{a})^T \\ \nabla f_2(\mathbf{a})^T \\ \vdots \\ \nabla f_k(\mathbf{a})^T \end{bmatrix} (\mathbf{x} - \mathbf{a}).$$

The matrix in this formula is used often enough to give it a name and a notation, after which we can state a version of the increment theorem with range in $\mathbb{R}^k$ in place of $\mathbb{R}^1$.

Definition 4.12. *Suppose $f : \mathbb{R}^n \to \mathbb{R}^k$ is differentiable. The derivative Df is the $k \times n$ matrix whose (i,j)-entry is $\partial f_i / \partial x_j$.*

This means that the i-th row of Df is precisely the vector $\nabla f_i(\mathbf{x})$.

The function f has nk partial derivatives: for each i between 1 and k and each j between 1 and n, we have a derivative $\partial f_i / \partial x_j$. The matrix Df just takes all these partial derivatives and combines them into a single matrix.

Example 4.13. Suppose $f(x, y, z) = (ze^{x/y}, xyz)$. Then $Df = \begin{bmatrix} \frac{z}{y} x^{x/y} & \frac{-xz}{y^2} e^{x/y} & e^{x/y} \\ yz & xz & yz \end{bmatrix}$.

Note that the top row is precisely $\nabla(ze^{x/y})$ and the second row is precisely $\nabla(xyz)$.

Exercise 4.6. *Suppose $f(r, \theta) := (r \cos \theta, r \sin \theta)$. Compute the matrix $Df(r, \theta)$.*

Example 4.14 (special cases of the derivative)**.**

- If $k = 1$ and $n = 1$ then Df is a 1×1 matrix, in other words a scalar. This is the ordinary derivative: $Df(x) = f'(x)$.

- If $k = 1$ and $n > 1$ then Df is the gradient, as a row vector. Thus $(Df(\mathbf{x}))(\mathbf{v}) = (\nabla f(\mathbf{x})^T)\mathbf{v} = \nabla f(\mathbf{x}) \cdot \mathbf{v}$.

- If $n = 1$ and $k > 1$ then $f : \mathbb{R} \to \mathbb{R}^k$ is a **parametrized curve** and $Df(t)$ is the **velocity vector** at time t.

Definition 4.15. *The linearization $\mathbf{L_a}(\mathbf{x})$ of $f : \mathbb{R}^n \to \mathbb{R}^k$ near $\mathbf{a}$ is defined by*

$$\mathbf{L_a}(\mathbf{x}) := f(\mathbf{a}) + [Df(\mathbf{a})](\mathbf{x} - \mathbf{a}).$$

The linearization $\mathbf{L_a}(\mathbf{x})$ is an estimate of the function $f(\mathbf{x})$, and the error is $|\mathbf{L_a}(\mathbf{x}) - f(\mathbf{x})|$. Generalizing the increment theorem, while also making an honest statement out of (4.15), we have the following proosition.

Proposition 4.16. *The linearization* $\mathbf{L_a}(\mathbf{x})$ *is a closer than linear approximation:*

$$|\mathbf{L_a}(\mathbf{x}) - f(\mathbf{x})| \ll |\mathbf{x} - \mathbf{a}|$$

as $\mathbf{x} \to \mathbf{a}$.

When the range of $f : \mathbb{R}^n \to \mathbb{R}$—that is, when the outputs are a single value—this is precisely the increment theorem.

Example 4.17. Consider for example the map $f : \mathbb{R}^2 \to \mathbb{R}^3$ defined by $f(x, y) = (\cos(y), \sin(y), x^3)$. The y-axis in $\mathbb{R}^2$ maps to a curve going around and around the unit circle in the xy-plane of $\mathbb{R}^3$, while the x-axis in $\mathbb{R}^2$ maps to a line $(1, 0, x^3)$ parallel to the z-axis. The point $\mathbf{a} := (1, 0)$ maps to $f(\mathbf{a}) = (1, 0, 1)$. Computing partial derivatives, we see that

$$Df = \begin{bmatrix} 0 & -\sin y \\ 0 & \cos y \\ 3x^2 & 0 \end{bmatrix}$$

and evaluating at $(1, 0)$ gives

$$Df(1,0) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 3 & 0 \end{bmatrix}.$$

So the linearization is given by

$$\mathbf{L}_{(1,0)}(x, y) = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 3 & 0 \end{bmatrix} \begin{bmatrix} x - 1 \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ y \\ 1 + 3(x - 1) \end{bmatrix}.$$

Exercise 4.7 (+)**.** *Explain why* $f(x, y) = f(x, y + 2\pi)$. *Then compute* $\mathbf{L}_{1,2\pi}$ *and explain any differences between this and* $L_{1,0}$.

4.2 Chain rule

Branch diagrams

In applications, computing partial derivatives is often easier than knowing what partial derivatives to compute. With all these variables flying around, we need a way of writing down what depends on what. We do this by writing a branch diagram. Some common ones are shown in Figure 17.

The branch diagram for the ordinary chain rule.

z is a function of x and y, both of which are functions of a single variable t.

z depends on x and y but y is really a function of x

w is a function of x, y and z, but z is really a function of the other two.
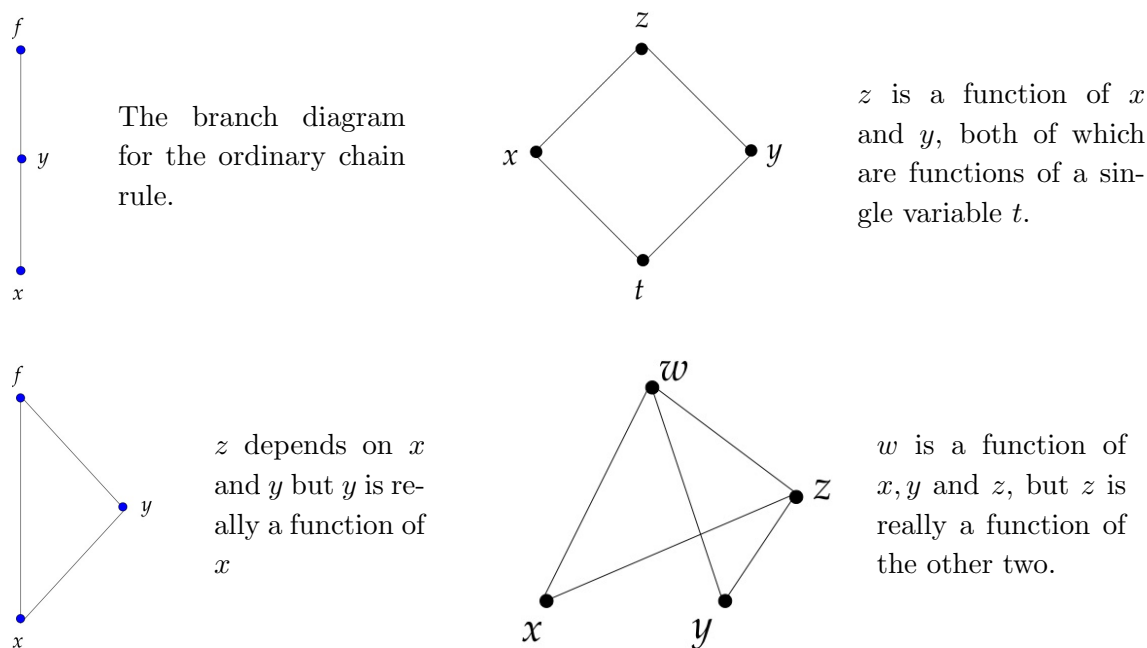
Figure 17: some common branch diagrams

Any variable at the top is an dependent variable. Any variable at the bottom is an independent variable; these drive the other variables and are the only ones we tweak directly. The variables in the middle are called *intermediate* variables. The independent variables drive them and they drive the dependent variables.

Exercise 4.8. *Draw a brach diagram for the situation where a variable u depends on variables x and y, while the variable v depends only on y.*

68

Summing over chains

Think about the ordinary chain rule, which is what the first branch diagram represents. A useful metaphor is that it is like a gear assembly: f depends on y, which in turn depends on x. Each unit increase of x increases y by $y'(x)$ many units. Each unit increase of y inceases f by $f'(y)$ units. Therefore each unit increase in x produces $y'(x) \cdot f'(y)$ units increase in f.

In the second diagram, there is a single independent variable t, which we think of as a gear driving both x and y, while both x and y drive z. We are going to try now to explain why

$$\frac{dz}{dt} = \frac{\partial z}{\partial y}\frac{dy}{dt} + \frac{\partial z}{\partial x}\frac{dx}{dt} \, . \tag{4.16}$$

Exercise 4.9. *Apply this with $z = x + xy$, $x = e^t$ and $y = 1 + t^2$. After you do this, check the computed value of dz/dt by plugging in to write z as the explicit function $e^t + e^t(1 + t^2)$ and differentiating.*

When t increases by Δt, both x and y increase. The increases are roughly $(\Delta t)(dx/dt)$ and $(\Delta t)(dy/dt)$ respectively. Each increase in x produces an increase in z that is $\partial z/\partial x$ times as great. So the increase in x of $\Delta t \dfrac{dx}{dt}$ gives gives an increase in z of roughly $\Delta t \dfrac{dx}{dt}\dfrac{\partial z}{\partial x}$. Simultaneously, the increase in t has produced an increase in y which produces another increase in z of roughly $\Delta t \dfrac{dy}{dt}\dfrac{\partial z}{\partial y}$. Thus the total increase in z is roughly

$$\Delta t \left[\frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt} \right] \, .$$

This means that the rate of change of z per change in t is given by equation (4.16). Note that we use partial derivative notation for derivatives of z with respect to x and y, as both x and y vary, but we use total derivative notation for derivatives of x and y with respect to t because each is a function of only the one variable; we also use total derivative notation dz/dt rather than $\partial z/\partial t$. Do you see why? Partial derivative notation would mean that t was changing while something else was being held fixed, which is not the case. Rather, all variables are functions of the single variable t. Sometimes this distinction can be useful; other times it can be ambiguous which is even going on. Ambiguities when they arise are not your fault - they're built into the poor notation.

That's the basic story. There are lots of variations, depending on how many independent variables there are (up till now there has been only one, all the others ultimately being functions of the one), how many intermediate variables and how they are related.

Theorem 4.18 (multivariate chain rule). *Let z and t be variables with z lying somewhere above t in a branch diagram. Then the derivative of z with respect to t is obtained by multiplying derivatives along every chain going from t up to z, then summing all such products.*

Example 4.19. The fourth diagram in Figure 17 has two chains from x to w—the direct link from x to w, and the chain that goes from x to z to w. So $\frac{dw}{dx}$ will be a sum of two terms. The direct link between x and w means one of those terms will be simply $\frac{\partial w}{\partial x}$. The chain from x to z to w should give a product of two terms, $\frac{\partial w}{\partial z}\frac{\partial z}{\partial x}$. Adding this gives

$$\frac{dw}{dx} = \frac{\partial w}{\partial x} + \frac{\partial w}{\partial z}\frac{\partial z}{\partial x}.$$

Exercise 4.10. *Draw a branch diagram that would lead to the following chain rule:*

$$\frac{dz}{dx} = \frac{\partial z}{\partial u}\frac{\partial u}{\partial t}\frac{\partial t}{\partial x} + \frac{\partial z}{\partial v}\frac{\partial v}{\partial t}\frac{dt}{dx} + \frac{\partial z}{\partial v}\frac{\partial v}{\partial s}\frac{ds}{dx}.$$

Where to evaluate?

The one thing you need to be careful about is evaluating all derivatives in the right place. It's just like the ordinary chain rule. For example, in (4.16), the derivatives dx/dt and dy/dt are evaluated at some time t_0. The partial derivative $\partial z/\partial x$ is evaluated at $x(t_0)$ and the partial derivative $\partial z/\partial y$ is evaluated at $y(t_0)$.

Example 4.20 (Chain rule for $f(x, y)$ when y is a function of x). The heading says it all: we want to know how $f(x, y)$ changes when x and y change but there is really only one independent variable, say x, and y is a function of x. This is captured by the third of the four branch diagrams on the previous page. Applying the chain rule gives

$$\frac{df}{dx} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} \cdot y'. \tag{4.17}$$

The notation really makes a difference here. Both df/dx and $\partial f/\partial x$ appear in the equation and they are not the same thing!

70

Gradients and directional derivatives

Example 4.21 (Derivative along an explicitly parametrized curve). One common application of the multivariate chain rule is when a point varies along a curve or surface and you need to figure the rate of change of some function of the moving point. The classical economics application is that price and quantity are moving together along the *demand curve* and we want to figure out how revenue changes along this curve (and in particular, we want to find where the revenue is maximized). In this section we solve the problem when the curve is known explicitly, saving the case of implicitly defined curves until we have discussed implicit differentiation.

Suppose a point varies along a curve in $\mathbb{R}^2$ as a function of time, and its coordinates are explicitly known: the coordinates at time t are $(x(t), y(t))$. This is illustrated in Figure 18.
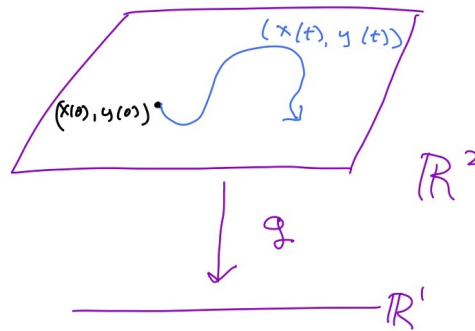

Figure 18: a function g evaluated along a parametrized curve in $\mathbb{R}^2$

Assuming all functions to be differentiable, the rate of change of the function $g(x, y)$ with respect to *time* along the curve is given by the formula we just computed: x and y are functions of t and g is a function of x and y, so

$$\frac{dg}{dt} = \frac{\partial g}{\partial x}\frac{dx}{dt} + \frac{\partial g}{\partial y}\frac{dy}{dt}. \tag{4.18}$$

Exercise 4.11. *This says the same thing as equation (4.16), provided we interpret g, x and y here as which symbols in (4.16)?*

In general, suppose $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))$ is a differentiable parametrized curve in $\mathbb{R}^n$

71

and $g : \mathbb{R}^n \to \mathbb{R}$ is a real differentiable function. Then

$$\frac{dg(\mathbf{x}(t))}{dt} = \sum_{j=1}^{n} \frac{\partial g}{\partial x_j} \frac{dx_j}{dt}$$

which we can write as $\nabla g \cdot \mathbf{x}'(t)$. This brings us to the idea of directional derivatives. Let $\mathbf{u}$ be a unit vector in $\mathbb{R}^n$ and g be a differentiable function from $\mathbb{R}^n$ to $\mathbb{R}$. If $\mathbf{x}(t)$ is a parametrized curve in $\mathbb{R}^n$ and $\mathbf{x}'(0) = \mathbf{u}$, then we have just seen that $dg(\mathbf{x}(t))/dt$ at $t = 0$ is given by $\nabla g \cdot \mathbf{u}$.

Definition 4.22 (directional derivative)**.** *The directional derivative of a continuously differentiable function $g : \mathbb{R}^n \to \mathbb{R}$ in the direction of a unit vector $\mathbf{u}$ is the quantity $D_{\mathbf{u}}g(\mathbf{x}) := \nabla g(\mathbf{x}) \cdot \mathbf{u}$. It is the derivative of g along any parametrized curve whose value at $t = 0$ is $\mathbf{x}$ and whose derivative at zero is $\mathbf{u}$.*

The gradient lets you evaluate all directional derivatives at the same time. Once you know ∇g, you know $D_{\mathbf{u}}g = \nabla g \cdot \mathbf{u}$ for all unit vectors $\mathbf{u}$ simultaneously. Using the geometric definition of the dot product, and noting that $|\mathbf{u}| = 1$ for all unit vectors, we see that

$$D_{\mathbf{u}}g = |\nabla g| \cos \theta$$

where θ is the angle between u and the gradient vector. This has some easy but useful consequences.

- The directional derivative is zero in the direction perpendicular to the gradient.

- The directional derivative is the greatest going parallel to the gradient and the maximum directional derivative is the magnitude of the gradient vector.

- The least directional derivative is in the opposite direction and is the negative of the greatest directional derivative.

Exercise 4.12. *Sketch the level curve of $F(x, y) = x + y^2$ through the point $(3, 1)$ in the first quadrant and draw the gradient of F at the point $(3, 1)$ as a vector whose tail is at the point $(3, 1)$.*

Nice clean chain rule for composition of two maps
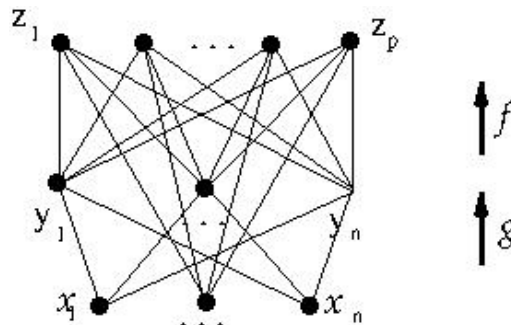
Suppose we have continuously differentiable functions $g : \mathbb{R}^m \to \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}^p$. The composition $f \circ g$ maps $\mathbf{x} \in \mathbb{R}^m$ to $\mathbb{R}^p$ as in the diagram.
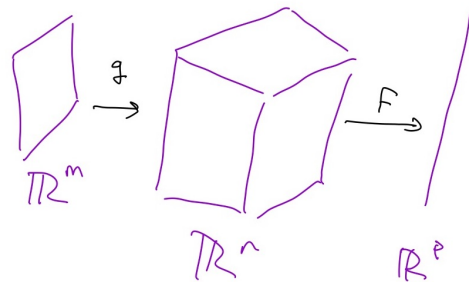
What is the derivative $D(f \circ g)$ of the composition $f \circ g$?

In terms of coordinates $x_1, \ldots, x_m$ in $\mathbb{R}^m$ and likewise $(y_1, \ldots y_n)$ in $\mathbb{R}^n$ and $(z_1, \ldots, z_p)$ in $\mathbb{R}^p$, the branch diagram has three levels, with arrows connecting each point in a level to each in the adjacent level.

Recalling that Dg is the linearization of g and Df is the linearization of f, the linearization of $f \circ g$ is $Df \circ Dg$. For linear maps, composition is matrix multiplication; hence $D(f \circ g) = [Df][Dg]$ as matrix multiplication.

5 Multivariate graphing and optimization

5.1 Graphing in three dimensions

Review of curves in two dimensions

Before talking about curves in three dimensions, let us recall what we know about graphing curves in two dimensions. You are probably most familiar with the situation where you have a function $y = f(x)$.

Figure 19: Graph of $y = x^2$

This may be so familiar that you've forgotten the formal definition it comes from.

Definition 5.1. *The graph of $f(x)$ is the set of points (x, y) satisfying the equation $y = f(x)$.*

That is, we obtained the graph by placing a dot at precisely those pairs (x, y) which satisfied the equation, and left all others empty. (And since we placed infinitely many dots, it ends up being the nice smooth curve we're familiar with instead of discrete dots.) We can plot points on the curve of a function by trying out some specific values of x: for instance, we can quickly check that the graph of $y = x^2$ must contain the points $(0, 0)$, $(1, 1)$, $(-1, 1)$, and $(1/2, 1/4)$.

There are two other ways of describing curves in two dimensions that you might be familiar with. An **implicitly defined curve** is the set of pairs (x, y) such that $g(x, y) = 0$ for some function $g : \mathbb{R}^2 \to \mathbb{R}$. This generalizes the graph of a function, because the graph $y = f(x)$ can always be written as the curve implicitly defined by $g(x, y) = y - f(x)$. But we can use implicitly defined curves to represent curves which are not functions, like the unit circle.

Figure 20: Curve defined by $x^2 + y^2 = 1$, which we can rewrite $x^2 + y^2 - 1 = 0$

It is not hard to check if a point belongs to an implicitly defined curve: if we want to confirm that $(\sqrt{2}/2, -\sqrt{2}/2)$ belongs to the curve defined by $x^2 + y^2 = 1$, we need only confirm that $(\sqrt{2}/2)^2 + (-\sqrt{2}/2)^2 = 1/2 + 1/2 = 1$. Plotting points can be tricker; we can quickly tell that the curve defined by $x^2 + y^2 = 1$ contains the points $(1,0)$, $(-1,0)$, $(0,1)$, and $(0,-1)$, but finding any other points requires some algebra, and for a more complicated formula, finding particular points can be difficult.

A third way is to define a curve *parametrically*: we pick two functions $x(t)$ and $y(t)$, and we take the graph to be all points of the form $(x(t), y(t))$ as t varies over some interval. This also generalizes the graph of a function $y = f(x)$, since we can always take the parameterization $x(t) = t, y(t) = f(t)$. For instance, the graph of $f(x) = x^2$ is the same as the graph of $x(t) = t$, $y(t) = t^2$. But some curves are much easier to represent using parametric descriptions.

Figure 21: Deltoid curve defined by $x(t) = 2\cos t + \cos 2t, y(t) = 2\sin t - \sin 2t$ where $0 \le t \le 2\pi$

Plotting points on an implicitly defined curve is not so difficult: we plug in values of t. For instance, in the deltoid curve in the picture above, when $t = 0$ we get the point

$$(x(0), y(0)) = (2 + 1, 0 - 0) = (3, 0),$$

while when $t = \pi/2$ we get the point

$$(x(\pi/2), y(\pi/2)) = (0 + (-1), 2 - 0) = (-1, 2).$$

Exercise 5.1. *Which point would you rather plug in to the functions defining the deltoid curve, $t = 1$ or $t = \pi$? Why?*

**Exercise 5.2.** *Which point on the deltoid curve corresponds to $t = \pi/4$?*

We have seen that one of the x-intercepts for the deltoid is when $t = 0$ and $x = 3$. There is a second x-intercept where x is negative. What are the t and x values at this point?

Curves in three and more dimensions

We can take the idea of parametric descriptions of curves into higher dimensions.

Definition 5.2. *When $\mathbf{x} : \mathbb{R} \to \mathbb{R}^n$ is a function, the* curve defined by $\mathbf{x}$ *is the set of points of the form $\mathbf{x}(t)$ for some value of t.*

For instance, if $\mathbf{x}(t) = (t, t^2, t^3)$, we get a graph which contains points like $\mathbf{x}(0) = (0, 0, 0)$, $\mathbf{x}(1) = (1, 1, 1)$, and also $\mathbf{x}(2) = (2, 4, 8)$. We could plot many more points and draw a smooth curve connecting them to get a picture like the one below.
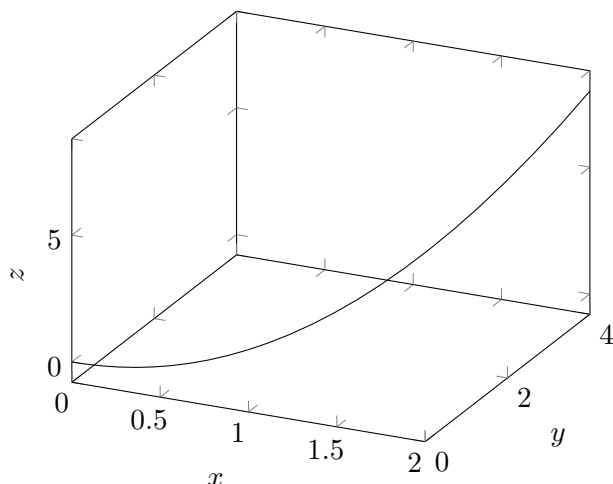
Exercise 5.3. *When $y = 2$ in the graph above, what is the z coordinate of the curve?*

Exercise 5.4. *Sketch the parametrized curve $\{(\cos\theta, \sin\theta, \theta) : 0 \leq \theta \leq 6\pi\}$ in $\mathbb{R}^3$, maintaining approximately the right ratios of the three coordinates.*

As the exercise illustrates, we don't always have to call the parametric variable t.

We should mention one specific kind of curve: lines. Lines in $\mathbb{R}^3$ have particularly nice parametric equations. If we want to talk about the line going through the point (x_0, y_0, z_0) in the direction (a, b, c), the equation is

$$\mathbf{x}(t) = (x_0, y_0, z_0) + (a, b, c)t = (x_0 + at, y_0 + bt, z_0 + ct).$$

Although we will not try to sketch them, we can give parametric descriptions of curves in more than three dimensions as well, like the curve $\mathbf{x}(t) = (t, t^2, \cos t, \sin t, \tan t)$ in $\mathbb{R}^5$.

Surfaces in three dimensions

What if, instead, we look at the set of points in $\mathbb{R}^3$ satisfying an equation like $z = f(x, y)$?

Exercise 5.5. *Sketch, as best you can, the set of points $z = x + y$. (Try plugging in about nine values of (x, y) and then fill in the sketch with your best guess of what you're getting.)*

When we graph a function, there's a value of z for every single choice of x and y (at least as long as f is defined). So we're graphing a *surface*. It's a pain to do by hand, so we'll mostly use computers when we want to actually look at the graphs. What's important is for us to remember what the graph means: the z value above the coordinate (x, y) is the value of $f(x, y)$ at that point.
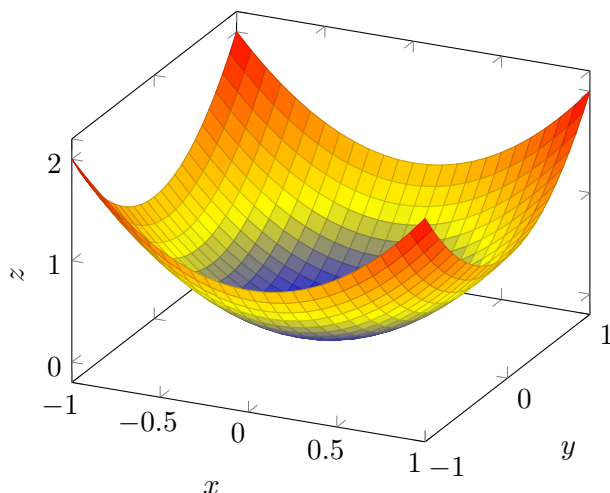
We can also define surfaces implicitly, with a function $g(x, y, z) = 0$. For instance, the unit sphere is the graph of $x^2 + y^2 + z^2 - 1 = 0$.

What is dimension?

We've used the word "dimension" a lot, and its time to be at least a little more precise about what it means.

Our prototypical example where we know what the dimension is $\mathbb{R}^n$, which we even refer to as "n-dimensional space". For instance, $\mathbb{R}^2$, is two dimensional. One important fact

about $\mathbb{R}^2$ is that it takes two numbers to describe a point: we need an x coordinate and a y coordinate. You might know that there are other schemes for describing points in $\mathbb{R}^2$; for instance, you might have encountered polar coordinates, where a point is described by a distance r from the origin and an angle θ. But, importantly, however we describe the points in $\mathbb{R}^2$, it always takes two numbers to do it.[6]

Similarly, describing a point in $\mathbb{R}^3$ requires three coordinates, like an x, y, and z coordinates. If we wanted to extend the idea of polar coordinates to $\mathbb{R}^3$, we would similarly need a third coordinate. (Indeed, there are actually two common ways of doing this, *spherical coordinates* and *cylindrical coordinates*, which are an interesting topic in multivariable calculus if you like geometry, but not something we cover in this course. If you're curious, you can get a decent idea of how these coordinate systems work with a quick web search.)

We will take this to be our definition of what dimension means.

"Definition" 5.1. *The* dimension *of a curve/surface/space/etc is the smallest number of variables we need to describe a point in it.*

This isn't really a precise definition, since "describe a point" is a pretty vague idea. Indeed, making it completely precise turns out to be very difficult. Fortunately, this informal intuition is good enough for us.

Using this definition, we can see that the surface of the earth is two dimensional. A point on the surface of the earth can be described using two coordinates, longitude and latitude which, by our definition, means the surface of the earth is two dimensional. (The earth itself—that is, the solid thing inside the surface—is three dimensional; for instance, you could describe a point using the longitude, the latitude, and how far beneath the surface it is. Or you could put the center at the origin, draw a z-axis from the north pole to the south pole, and pick some points on the equator to intersect the x- and y-axes.)

This illustrates an important distinction: the surface of the earth is located in three dimensional space, but it's a two dimensional thing: we distinguish between the dimension of the surface itself, and the dimension of the three dimensional universe it sits inside of. (The fancy term we mostly won't use is that the "ambient space" of the surface of the earth is three dimensional.)

Exercise 5.6 (+)**.** *What is the dimension of Walnut street, and why? (There are a couple of defensible answers here.)*

---

[6]This isn't completely true, but it's true as long as we stick to reasonable ways of describing points.

Curves are always one dimensional. Indeed, that's more or less the definition of a curve. For instance, the curve $y = x^2$ is a one dimensional thing that we draw inside of $\mathbb{R}^2$. That's one reason we like to describe curves parametrically: no matter which $\mathbb{R}^n$ we're working in, the parametric variable t is the single coordinate we need to describe a point along the curve.

When we look at the set of points satisfying an equation, we typically reduce the dimension by one. For instance, the set of points in $\mathbb{R}^2$ satisfying an equation like $y = f(x)$ is one dimensional—all we need to know is the x value to find a point satisfying the equation. And the set of points in $\mathbb{R}^3$ satisfying $z = f(x, y)$ is two dimensional, since we need to pick the x and y coordinates to specify a point satisfying the equation. We call a two dimensional thing a *surface*.

Working in $\mathbb{R}^4$, the set of points satisfying an equation like $g(x, y, z, w) = 0$ will be three dimensional. We eventually need some new terms here. A *hypersurface* is a set of points with dimension one less than the ambient space. That is, a hypersurface in $\mathbb{R}^n$ is a set of points with dimension $n - 1$, and we can represent it as the set of solutions to a single equation.

In the plane, when $n = 2$, there is no difference between 1 and $n - 1$. Parametrized curves and the set of solutions to an equation are both one-dimensional. Given a curve γ, we have a choice whether to represent it as a parametrized curve or a level set $g(x) = c$. In three dimensions, level surfaces are two dimensional. For example the unit sphere $\{\mathbf{x} : x^2 + y^2 + z^2 = 1\}$ (not the interior, just the surface) is a level surface for the distance to the origin function. Parametrized curves, such as the spiral $\{(\cos t, \sin t, t) : 0 \le t \le 6\pi\}$, are one dimensional.

In four and higher dimensions, not only are there hypersurfaces $\{\mathbf{x} : F(\mathbf{x}) = 0\}$ and curves $\{\mathbf{x}(t) : a \le t \le b\}$, but there are sets of every dimension in between. These can be described from the top down, by multiple constraints, or from the bottom up, via multiple parameters. This course won't go into how one does calculus on these in-between sets. We give one example so you can see what we are talking about.

Example 5.3 (just for fun)**.** The set $\{\mathbf{x} \in \mathbb{R}^4 : x_1^2 + x_2^2 = 1 = x_3^2 + x_4^2\}$ is a 2-surface in $\mathbb{R}^4$. We have described it via two relations: it is where the two functions $x_1^2 + x_2^2 - 1$ and $x_3^2 + x_4^2 - 1$ both vanish. We could also parametrize it with two parameters: it is the set $\{(\cos \alpha, \sin \alpha, \cos \beta, \sin \beta) : 0 \le \alpha, \beta < 2\pi\}$.

It turns out that this set is the same as the surface of a donut, but it's very hard to visualize because it lives inside four-dimensional space instead of the usual three.

Exercise 5.7. *Describe the infinite cylinder of radius 1, centered around the z-axis in $\mathbb{R}^3$ as a hypersurface by finding a function $f(x, y, z)$ for which the cylinder is the set $\{(x, y, z) : f(x, y, z) = 0\}$.*

*(Hint: $f(x, y, z)$ is one of the functions $z^2 - 1$, $x^2 + y^2 - 1$, or $x^2 + y^2 + z^2 - 1$.)*

Exercise 5.8 (+). *Suppose you are working with 7 variables and have two equations $f(x_1, x_2, \ldots, x_7) = 0$ and $g(x_1, x_2, \ldots, x_7) = 0$. How many dimensions would you expect the graph to have? (In this case the graph is the set of all the points in $\mathbb{R}^7$ which satisfy both equations at once.)*

Regions and Boundaries

In addition to describing sets of points with equalities, we often use *inequalities* to describe sets of points. You already saw this in single variable calculus, where you might have talked about things like "the maximum of $f(x)$ on the interval $[a, b]$". The interval $[a, b]$ just means $\{x : a \leq x \leq b\}$.

Similarly, in two dimensions, we've seen things like the unit disk, $\{(x, y) : x^2 + y^2 \leq 1\}$, the "filled in" version of the unit circle. We can do similar things in more dimensions. For instance, in $\mathbb{R}^3$, the region defined by $x \geq 0, y \geq 0, z \geq 0, x + y + z \leq 1$ is the solid tetrahedron:
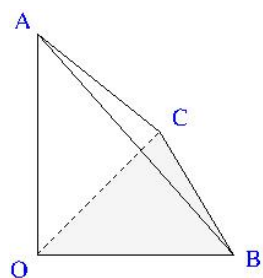

A region defined only by inequalities generally has the same dimension as the space it's cut out of: the interval $[a, b]$ is one dimensional, the unit disk $\{(x, y) : x^2 + y^2 \leq 1\}$ is two dimensional, a solid tetrahedron is three dimensional.

Regions defined by inequalities usually have a *boundary*, the points that separate the inside of the region from the outside. Most of the time, we can find the boundary by turning an inequality into an equality. For instance, the boundary of the unit disk $\{(x,y) : x^2 + y^2 \leq 1\}$ is the unit circle $\{(x,y) : x^2 + y^2 = 1\}$.

Often the boundary of a region naturally breaks down into several parts. For instance, the upper half of the unit disk is $\{(x,y) : x^2 + y^2 \leq 1, y \geq 0\}$.
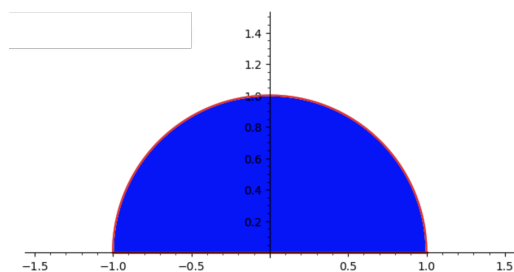
It boundary can be described as consisting of two curves: the line $y = 0$ on the interval $-1 \leq x \leq 1$ as well as the curve $y = \sqrt{x^2 + 1}$, also on the interval $-1 \leq x \leq 1$.

You're familiar with the boundary of a region, at least in some special cases. In one dimension, the boundary of an interval is called the end points: the boundary of the interval $[a, b]$ is just the two points $x = a$ and $x = b$. In two dimensions, you're probably used to talking about the "sides" of a shape: that's what we sometimes call the boundary when it's a bunch of line segments. Similarly, the "faces" of a three dimensional shape are what we call the boundary when it's a bunch of pieces of planes. For instance, the boundary of the tetrahedron above has four parts, one for each face:

- the triangle OCB given by $z = 0$, $x \geq 0$, $y \geq 0$, $x + y \leq 1$,

- the triangle OAB given by $y = 0$, $x \geq 0$, $z \geq 0$, $x + z \leq 1$,

- the triangle OAC given by $x = 0$, $y \geq 0$, $z \geq 0$, $y + z \leq 1$, and

- the triangle ABC given by $x + y + z = 1$, $x \geq 0$, $y \geq 0$, $z \geq 0$.

Notice that each of these faces comes from turning one of the inequalities into an equality.

So far all these examples of regions include the boundary. We can also have regions which exclude the boundary, which correspond to strict inequalities. For instance, $\{(x,y) : x^2 +$

$y^2 < 1\}$ is *just* the interior of the unit disk, excluding the boundary. Later on, we will often want to take a region which includes its boundary, given by $\leq$ inequalities, and divide it into two pieces—the *boundary* given by an equality and the *interior* given by the strict inequality.

You might have already noticed that boundaries almost always have dimension one less than the region. The unit disk is two dimensional, and its boundary is the one dimensional unit circle. The solid tetrahedron is three dimensional, and its boundary is four two dimensional triangles. An interval $[a, b]$ is one dimensional, and its boundary is the two points a and b—individual points are zero dimensional.

Planes and Hyperplanes

One particularly important kind of surface in $\mathbb{R}^3$ is a plane. The analog in $\mathbb{R}^2$ is a line, and you're probably most used to describing lines with equations like

$$y = mx + b$$

or

$$y - y_0 = m(x - x_0).$$

The latter form, which we call *point slope*, tends to be more useful in calculus: we describe a line by knowing one point on the line, and what its slope is. For instance, the tangent line to $y = \ln(x)$ at $(1, 0)$ has the equation $y - 1 = (1)x$ beacuse the derivative when $x = 1$ is $1/1 = 1$.

You can probably come up with the equations of some specific planes, like $z = 0$ or $z = 4$. You might also notice that an equation which would have given you a line in $\mathbb{R}^2$ gives you a vertical plane in $\mathbb{R}^3$.

But we will need to be able to talk about the equation of a plane $\mathbb{R}^3$ in general. It turns out that the simplest way to describe a plane in $\mathbb{R}^3$ is to talk about what it's perpendicular to. For example, $x + y + z = 0$ is the equation of a plane. It contains the point $(0, 0, 0)$ (check this), and it has the following property:

if (x_0, y_0, z_0) and (x_1, y_1, z_1) are two points on the plane $x + y + z = 0$ then the vector $\begin{bmatrix} x_1 - x_0 \\ y_1 - y_0 \\ z_1 - z_0 \end{bmatrix}$ is perpendicular to the vector $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

Exercise 5.9 (+). *Check the claim in the previous paragraph. (Hint: How can you check if two vectors are perpendicular? How does knowing the points are on the plane help you simplify that equation?)*

This idea motivates a general way of writing equations for planes. Suppose we want the equation for a plane which contains a point (x_0, y_0, z_0) and is perpendicular to a vector $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$. (We call this vector the **normal vector**, since "normal" is another word for perpendicular.) Then we are asking for every point (x, y, z) on the plane to satisfy the equation

$$\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix} - \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} \right) \cdot \begin{bmatrix} a \\ b \\ c \end{bmatrix} = 0.$$

Expanding and simplifying the dot product, we get the following:

Definition 5.4. *If (x_0, y_0, z_0) is a point and $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ then the equation for the plane through (x_0, y_0, z_0) normal to $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ is*

$$a(x - x_0) + b(y - y_0) + c(z - z_0) = 0.$$

As usual, we can generalize from $\mathbb{R}^3$ to $\mathbb{R}^n$. A *hyperplane* is a hypersurface (that is, it has dimension one less than n) which is the graph of a linear equation

$$\mathbf{a} \cdot (\mathbf{x} - \mathbf{x_0}) = 0.$$

Of course, a hyperplane in $\mathbb{R}^2$ is just a line, so to complete the circle of ideas, we should check that this is really what lines look like. If we have a line $y - y_0 = m(x - x_0)$, we can rearrange this equation to $m(x - x_0) - (y - y_0) = 0$, which is

$$\begin{bmatrix} m \\ -1 \end{bmatrix} \cdot \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix} = 0.$$

So our new form for an equation of a hyperplane is really a minor variant of point slope form.

The Tangent Hyperplane

A hyperplane in $\mathbb{R}^2$ is just a line, and the lines we're most interested in come from lineariza-tions. The linearization of a function $f(x)$ at x_0 is the function

$$y = L_{x_0}(x) = f(x_0) + f'(x_0)(x - x_0).$$

In order to rearrange this into the matrix form like above, we can get everything on the same side, writing

$$f'(x_0)(x - x_0) - (y - f(x_0)) = 0,$$

which we can then rewrite as the matrix

$$\begin{bmatrix} f'(x_0) \\ -1 \end{bmatrix} \cdot \begin{bmatrix} x - x_0 \\ y - f(x_0) \end{bmatrix} = 0.$$

This is the equation of a line, which you probably recall that we call the *tangent line* (to the function f at the point x_0).

The same idea works for any differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, even when $n > 1$. When $f : \mathbb{R}^n \to \mathbb{R}$ is a function which is differentiable at $\mathbf{a}$, we have seen the linearization

$$L_{\mathbf{a}}(\mathbf{x}) = f(\mathbf{a}) + \nabla f(\mathbf{a}) \cdot (\mathbf{x} - \mathbf{a}).$$

When we rewrite this as a matrix using the same steps, we get

$$\begin{bmatrix} \nabla f(\mathbf{a}) \\ -1 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{x} - \mathbf{a} \\ y - f(\mathbf{a}) \end{bmatrix} = 0.$$

Note that these are $(n + 1)$-dimensional vectors; we're writing $\begin{bmatrix} \nabla f_{\mathbf{a}} \\ -1 \end{bmatrix}$ to mean the vector whose first n terms are ∇f and whose bottom term is -1.

Naturally, we call this the *tangent hyperplane* to f at $\mathbf{a}$. (When $n = 2$, a hyperplane is just a plane, so we call it the tangent plane, too.)

Example 5.5. Find an equation for the tangent plane to $z = xy$ at the point $(2, -3)$.

Since $\nabla z = \begin{bmatrix} y \\ x \end{bmatrix}$, $\nabla z(2,-3) = \begin{bmatrix} -3 \\ 2 \end{bmatrix}$,

$$L_{2,-3} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} + \begin{bmatrix} -3 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} x - 2 \\ y + 3 \end{bmatrix},$$

so an equation for the tangent plane is

$$\begin{bmatrix} -3 \\ 2 \\ -1 \end{bmatrix} \cdot \begin{bmatrix} x - 2 \\ y + 3 \\ z + y \end{bmatrix} = 0.$$

Exercise 5.10. *Give equations for the tangent plane to the surface $z = 25 - x^2 - y^2$ at the point $(3, 4, 0)$.*

There is one last specific case we should think about explicitly, in preparation for the next section. What happens if $\nabla f(\mathbf{a}) = \mathbf{0}$? In this case, we get the hyperplane

$$\begin{bmatrix} 0 \\ \vdots \\ 0 \\ -1 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{x} - \mathbf{a} \\ z - f(\mathbf{a}) \end{bmatrix} = 0.$$

This is exactly the hyperplane $z = f(\mathbf{a})$. So when $\nabla f(\mathbf{a}) = \mathbf{0}$, the tangent hyperplane is just the horizontal hyperplane at the right height. The tangent hyperplane is an approximation of f, so when $\mathbf{x}$ is close to $\mathbf{a}$, we have $f(\mathbf{x}) \approx L_{\mathbf{a}}(\mathbf{x}) = f(\mathbf{a})$: the function f isn't changing very much near $\mathbf{a}$.

5.2 Critical points

Unit 6 discusses constrained maxima and minima, where the problem is restricted to a curve or surface in a higher dimensional space. This involves new concepts, some of which are mind-bending, at least if you are among the majority who do not easily visualize in more than two dimensions (and certainly not in more than three). By contrast, this unit considers unconstrained maxima and minima, where analogies with the one-variable case make the concepts much more manageable. We begin the discussion in two variables, where functions and their graphs can be sketched.

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a function. We seek the maxima and minima of f. We will focus on functions which are "nice enough": where f is *continuously differentiable*, which means that not only does the derivative exist where f is defined, but the derivative is continuous at every point. Suppose f attains a maximum or a minimum at $\mathbf{a}$. We are going to use what we know about differentiable functions of one variable: wherever they have a maximum or minimum, the derivative vanishes. Looking at values on the line $\mathbf{a} + t\mathbf{e}_1$, we see that if the directional derivative $D_{\mathbf{e}_1} f(\mathbf{a}) = \partial f/\partial x_1$ does not vanish, then $\mathbf{a}$ cannot be a maximum or minimum because the one-variable function $t \mapsto f(\mathbf{a} + t\mathbf{e}_1)$ has nonzero derivative there. The same thing holds for $\mathbf{e}_2$ of course: the directional derivative $D_{\mathbf{e}_2} f = \partial f/\partial x_2$ must vanish.

Now we know two necessary conditions for a maximum or minimum at $\mathbf{a}$: both partial derivatives must vanish. There are a lot of other directions, so we might expect many more conditions. Due to the magic of directional derivatives, however, $D_{\mathbf{u}} f = \nabla f \cdot \mathbf{u}$, so if both partial derivatives vanish, the gradient vanishes and all directional derivatives vanish. This generalizes easily to n dimensions.

Definition 5.6. *Let $\mathbf{a}$ be an interior point in the domain of definition of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. We say that f has a **critical point at a** if $\nabla f(\mathbf{a}) = \mathbf{0}$.*

What we have argued, in logical terms, is that in order for $\mathbf{a}$ to be an extreme (either a max or a min), it is a *necessary condition* that $\mathbf{a}$ be a critical point. As we know from one-variable functions, it is not *sufficient*. For example, the function x^3 has a critical point at zero but neither a maximum nor a minimum. Also, if $\mathbf{a}$ is not in the interior of the region of definition, we might find an extreme where there is no critical point. This is analogous to finding the minimum value of f on $[a, b]$ at one of the endpoints: endpoints can be a maximum or a minimum even if $f' \neq 0$ there.

Exercise 5.11 (+). *Suppose $\mathbf{a}$ is a critical point for f and $\mathbf{x}(t)$ is a curve in $\mathbb{R}^n$ with $\mathbf{x}(0) = \mathbf{a}$. Can you conclude anything about $\dfrac{d}{dt}(f(\mathbf{x}(t)))$?*

To pave the way for what is to come, let's assemble a pictorial taxonomy of behaviors of functions near critical points in one and two variables. We exclude the trivial case where the function is constant. In one variable we have really only three possibilities: a maximum, a minimum, or a point of inflection something looking like x^3. We exhibit simple examples of each of these in Figure 22.

Now let's look at some of the possibilities in two variables. The first two pictures in Figure 23
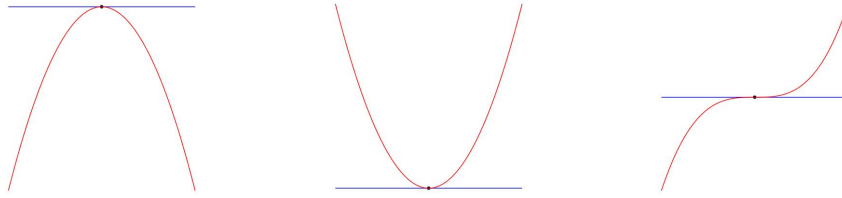
Figure 22: a maximum, a minimum and an inflection point

show a maximum and a minimum. So as not to muddy the picture, we didn't draw the tangent plane in blue, but it is horizontal. The third picture shows a **saddle point**. It is not a maximum, nor a minimum. It is not an inflection point, just (to anthropomorhpize) a very ambivalent function: the second derivative is positive along the x-axis but negative along the y-axis.
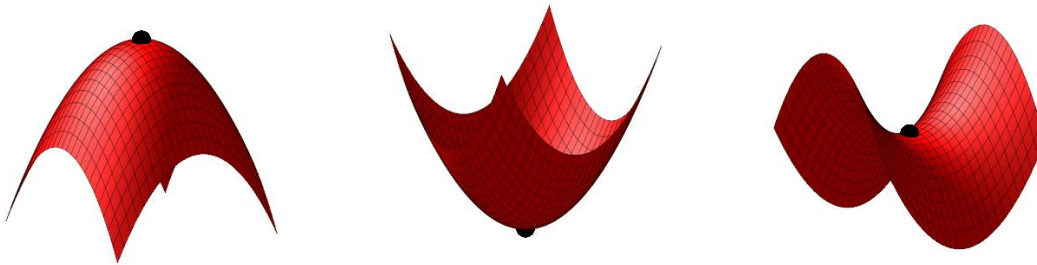

Figure 23: a maximum, a minimum and a saddle

In two variables, there is room for many other behaviors. Figure 24 shows two more examples. The term **inflection point** is not well defined in two variables, but these two examples have at least one second derivative equal to 0.

Exercise 5.12. *Match the five pictures in Figures 23 and 24 to these five graphs of functions. (Two are pretty hard to distinguish, so just give your best guess and move on.)*

(a) $z = xy$

(b) $z = 3x^2 + 3y^2$
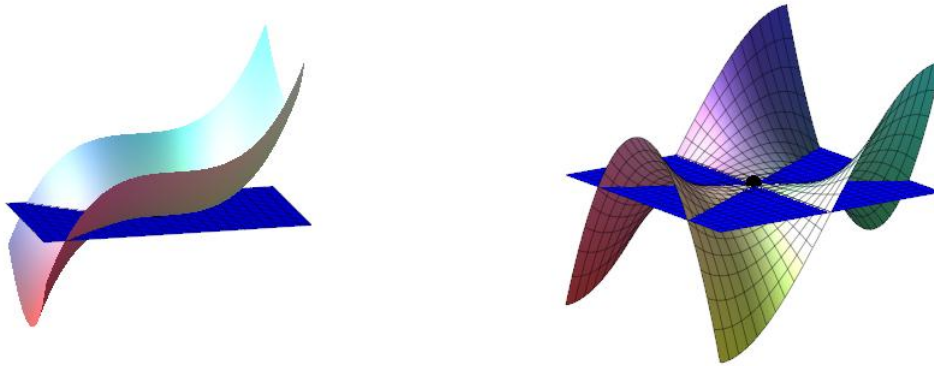
(c) $z = x^3 + 3xy^2$

87

Figure 24: two possible critical point geometries

(d) $x = -3x^2 - 3y^2$

(e) $z = 3x^2 + y^3$

Maxima and minima

When you are looking for a maximum or minimum, you might not need any fancy tests to tell you whether a critical point **a** is the point you're looking for. You might be able to tell "by inspection".

Example 5.7. Suppose the value of a harvested mushroom increases with each mile you go north or east from the center of the valley: in coordinates it is proportional to $1 + x + 2y$. Suppose also the density of the mushroom decreases away from the valley center, inversely proportional to $e^{x^2+y^2}$. If we can only search for mushrooms in one location, where can we maximize the value of our harvest?

Recall that this means the value has the form $\frac{C(1+x+2y)}{e^{x^2+y^2}} = C(1 + x + 2y)e^{-x^2-y^2}$ for some constant C. To maximize the value $C(1 + x + 2y)e^{-x^2-y^2}$ over all of $\mathbb{R}^2$, we find where the gradient vanishes. The gradient is

$$\nabla C(1+x+2y)e^{-x^2-y^2} = \begin{bmatrix} (1 + x + 2y)(-2xe^{-x^2-y^2}) + e^{-x^2-y^2} \\ (1 + x + 2y)(-2ye^{-x^2-y^2}) + 2e^{-x^2-y^2} \end{bmatrix} = \begin{bmatrix} (-2x - 2x^2 - 4xy + 1)e^{-x^2-y^2} \\ (-2y - 2xy - 4y^2 + 2)e^{-x^2-y^2} \end{bmatrix}$$

88

Since $e^{-x^2-y^2}$ can never be 0, the gradient is 0 when both

$$-2x + 2x^2 - 4xy + 1 = 0 \text{ and } -2y - 2xy - 4y^2 + 2 = 0.$$

Mutiplying the first equation by -1 and the second by $-1/2$ to simplify a bit, and then factoring, we get

$$\begin{aligned} 2x^2 + (4y+2)x - 1 &= 0\,; \\ 2y^2 + (x+1)y - 1 &= 0\,. \end{aligned}$$

After some tricky algebra, $x = (\sqrt{11} - 1)/10 \approx 0.23$ is the root of $10x^2 + 2x - 1$ and $y = 2x \approx 0.46$. Is this a local max, local min or neither? This is the only critical point in the positive orthant. We know there's no point going west or south, decreasing both the price and the density. So it's either this point, or there is no maximum. The function goes to zero as we go to infinity in any direction. Therefore this point, with harvest value roughly $1.65C$, is a global maximum.

Exercise 5.13 (+). *How do we know the value goes to zero as we go to infinity in any direction? Give a quick informal reason.*

You might remember that in single variable calculus, you often used the second derivative test to determine if a critical point is a local minimum or local maximum. There is a version of the second derivative test for multivariable functions, but it's a bit more complicated, and we'll be in a better position to make sense of it in a couple weeks.

Boundaries

Suppose we want to find the maximum or minimum of f over a region R that has a boundary. For example, perhaps R is the unit disk, whose boundary ∂R is the unit circle, or perhaps R is the positive quadrant in the plane, whose boundary ∂R is the union of the positive x- and y-axes.

The maximum could occur in the interior of R, or on the boundary of R or not at all. To understand which, we need to check all the interior critical points, compare them to the maximum on the boundary to see which is greater, then also check that $f(\mathbf{x})$ cannot exceed this as $\mathbf{x} \to \infty$ in any directions in which R is not bounded. In Unit 6 we will see how to find maxima on the boundary. Until then, we will examine cases where maxima or minima occur in the interior.

Exercise 5.14 (+). *Let R be a nice big blobby region in the plane. Let $f(x,y) = 3x - 5y$. Do you expect the maximum of f on R to occur in the interior or on the boundary?*

6   Calculus on curves and surfaces

6.1   Two variables

We will consider two sorts of questions on curves in $\mathbb{R}^2$. Later, we will ask similar questions in more than two variables.

Q1 How are the rates of change of the x and y variables related if the pair (x, y) is constrained to stay on the curve γ?

Q2 Where are the extrema of a function $g : \mathbb{R}^2 \to \mathbb{R}$ on the curve γ?

Example 6.1. Due to variation in demand, the profit from making x sedans and y SUVs (in thousands of dollars) is $g(x, y) = 2x - 0.3x^2 + 3y - 0.6y^2 - 0.1xy$. Operating at its maximum capacity making only sedans, a factory can produce 1000 sedans per day; by switching over to SUV production, the factory can make 2 SUVs for every 3 fewer sedans it produces.

The total production therefore satisfies the equation $x + 1.5y = 1000$, which is the constraint curve γ. We will see below how to calculate $\frac{dx}{dy}$ or $\frac{dy}{dx}$ assuming the factory remains at maximum capacity, as well as how to find the production schedule that maximizes profit.

Recall that there are several ways to understand a curve in $\mathbb{R}^2$.

(i) The **graph of a function** f is the set of pairs (x, y) such that $y = f(x)$ and x varies over some interval (possibly infinite).

(ii) A **parametrized curve** is the set of pairs $(x(t), y(t))$ as t varies over some interval. This is a generalization of the first way because the graph $y = f(x)$ can always be parametrized by $x(t) = t, y(t) = f(t)$.

(iii) An **implicitly defined curve** is the set of pairs (x, y) such that $g(x, y) = 0$ for some function $f : \mathbb{R}^2 \to \mathbb{R}$. This generalizes the first way also becase the graph $y = F(x)$ can always be written as the set where the function $g(x, y) := y - F(x)$ is zero..

Exercise 6.1. *Match these descriptions of the quarter-circle to the three methods above.*

- $\{(\cos(\theta), \sin(\theta)) : 0 \le \theta \le \pi/2\}$

- $\{(x, y) : x^2 + y^2 = 1, x \geq 0, y \geq 0\}$
- $\{(x, y) : y = \sqrt{1 - x^2}, 0 \leq x \leq 1\}$

Once we have a curve in one of these three formats, we will want to understand $\frac{dy}{dx}$—the relative rate of change of the two variables.

(i) The relation of change in y to change in x is easy if one travels along a graph $y = f(x)$, because of course dy/dx is just the derivative $f'(x)$. This ratio can be seen visually as the slope of the tangent line to the curve at the point $(x, f(x))$.

(ii) For a parametrized curve, the slope of the tangent line is the slope of the velocity vector, defined in Unit 4, (4.17). That is, $dy/dx = y'(t)/x'(t)$.

Exercise 6.2. *Suppose x, y, and t all of different units ("units of x", "units of y", and "units of t"). Check that the units in (4.17) are the same on both sides.*

(iii) Perhaps the most interesting is to compute this slope in the third case, where the curve γ is specified as a zero set or level curve. This brings us to a technique known as **implicit differentiation**.

Implicit differentiation

Suppose a curve is defined by $F(x, y) = c$, for some constant c. What is the slope of its tangent line? Classic implicit differentiation goes like this. Differentiate both sides with respect to x, giving

$$F_x + F_y \frac{dy}{dx} = 0 \,. \tag{6.19}$$

You then solve for dy/dx to get

$$\frac{dy}{dx} = -\frac{F_x}{F_y} \,. \tag{6.20}$$

We're not sure whether this sets off your bullshit detector, but the writers of this text find the lack of explanation unsatisfying. The mysterious part is how we know that the left side differentiates to $F_x + F_y \frac{dy}{dx}$. Why not just F_x, or $F_x + F_y$, or $\nabla F \cdot \mathbf{v}$, or any of a number of other possibilities?

The key is to identify which variables vary freely and which depend on which others; that is, we need to know the branch diagram. The correct diagram in this case is the third diagram from Figure 17, shown at the right. As x changes, and y changes along with x to stay on the curve, this causes F to maintain its value of c. So $dF/dx = 0$, but also the multivariate chain rule says $d/dx = F_x + F_y \, dy/dx$, giving (6.19).
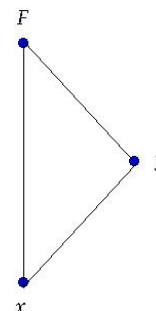
Solving for dy/dx gives $y' = -F_x/F_y$ as above.

Exercise 6.3. *Suppose x, y and F all have different units ("units of x", "units of y" and "units of F"). Do a unit check on equation (6.20).*
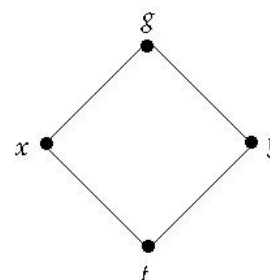
Derivative along a parametrized curve

Having disposed of the first question, **Q1**, we turn to **Q2**, the question of maximizing (or minimizing) along a curve. The first step is to figure out how a function $g : \mathbb{R}^2 \to \mathbb{R}$ varies when (x, y) is constrained to lie on a curve, γ.

If γ is parametrized, say by $\{(x(t), y(t)) : a \le t \le b\}$, it makes sense to ask what is the rate of change of g per change in the parameter t.

In this case it is the second diagram in Figure 17, shown to the right, and the multivariate chain rule gives

$$\frac{dg}{dt} = \frac{\partial g}{\partial x}\frac{dx}{dt} + \frac{\partial g}{\partial y}\frac{dy}{dt} \, . \qquad (6.21)$$

One can also arrive at this by writing g as $g(x(t), y(t))$ and using the ordinary chain rule.

Exercise 6.4. *Suppose $g(x, y) = x + y^2/x$. What is dg/dt along the path $x = \cos t, y = \sin t$?*

Exercise 6.5 (+). *The graph of a function $y = f(x)$ is supposed to be a special case of a parametrized curve, $x = t$, $y(t) = f(t)$. Compute the rate of change of g in this case, and say what it means.*

We say $(x(t), y(t))$ is a critical point for the parametrized path γ if $dg/dt = 0$. By the logic of one variable calculus, the maximum and minimum for g on γ can only occur at a

critical point or at the endpoints. In preparation for higher dimensions, we write (6.21) as $g' = \nabla g \cdot \mathbf{v}$ where the prime denotes the "total derivative with respect to t," and $\mathbf{v}$ denotes the velocity vector $x'(t)\hat{\mathbf{i}} + y'(t)\hat{\mathbf{j}}$.

Derivative along an implicitly defined curve

Suppose instead γ is implicitly defined, for instance as the level curve $F(x, y) = 0$. There are several ways we could proceed. We could find a parametrization and check if the rate of change of g is zero for that parametrization. We could compute the directional derivative in the direction tangent to the curve and see if that is zero. These two methods yield the same result because for any parametrization $\mathbf{x}(t)$, the velocity vector $\mathbf{v}(t)$ is tangent to γ, hence a multiple of the unit tangent vector $\mathbf{u}$. Therefore $g'(t) = \nabla g \cdot \mathbf{x}(t)$ is a multiple of $D_{\mathbf{u}}g = \nabla g \cdot \mathbf{u}$. Either both vanish or neither does.

Here is a simple way to do this for a level curve γ defined by $F(x, y) = 0$. We know (6.19) that the slope of the curve at (x, y) is $-F_x/F_y$. Therefore, if $F_y \neq 0$ then $\hat{\mathbf{i}} - (F_x/F_y)\hat{\mathbf{j}}$ is a tangent vector to the curve. Clearing denominators, we see that $F_y\hat{\mathbf{i}} - F_x\hat{\mathbf{j}}$ is tangent to the curve as long as both F_x and F_y are not simultaneously zero. Therefore, to check whether (x, y) is a critical point for g on γ, we need only check whether the derivative of g in the direction of $F_y\hat{\mathbf{i}} - F_x\hat{\mathbf{j}}$ vanishes.
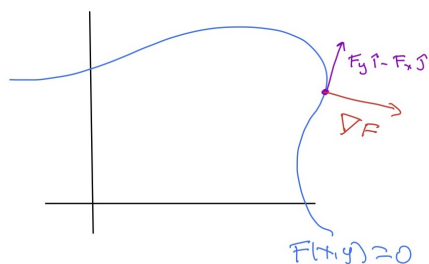

Figure 25: rotating the gradient of F by $90°$ gives a tangent vector

In other words we need $\nabla g \cdot (F_y\hat{\mathbf{i}} - F_x\hat{\mathbf{j}}) = 0$: ∇g must be perpendicular to $F_y\hat{\mathbf{i}} - F_x\hat{\mathbf{j}}$. That is the same as being parallel to the vector $F_y\hat{\mathbf{i}} - F_x\hat{\mathbf{j}}$ rotated by 90 degrees, namely $\nabla F = F_x\hat{\mathbf{i}} + F_y\hat{\mathbf{j}}$. This is illustrated in Figure 25. We conclude the following proposition.

Proposition 6.2. *Let $F : \mathbb{R}^2 \to \mathbb{R}$ be differentiable and let γ be the level curve $\{(x, y) : F(x, y) = 0\}$. The point $(x, y) \in \gamma$ is a critical point for g on γ if and only if ∇g is parallel*

to ∇F at the point (x, y). □
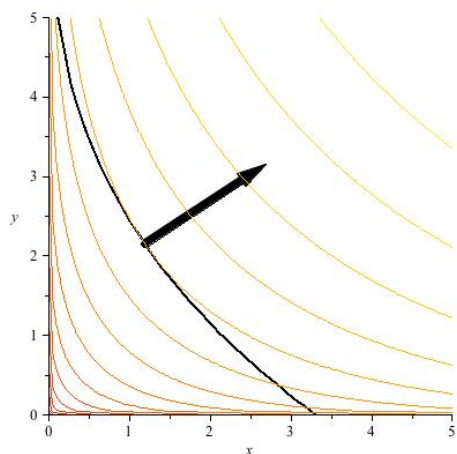
Figure 26

Example 6.3 (Pictorial interpretation). A picture clarifies that ∇g being parallel to ∇F is the condition for a critical point for g along the curve $F = 0$. Figure 26 shows a black constraint curve, $F(x, y) = 0$, along with contours on an orange-yellow scale for another function $g(x, y)$ (yellower means greater). The maximum of g along the curve $F(x, y) = 0$ is the place where the level curves of g moving from higher values to lower, first hit the black curve. At this point, the curves are tangent and the gradients are parallel. The black arrow represents the directions of both gradients.

6.2 Three dimensions and higher

Parametrized curves in n dimensions

Parametrized curves $\mathbf{x}(t)$ in three or more variables behave just like parametrized curves in two variables. They are defined by functions $\mathbf{x} : \mathbb{R} \to \mathbb{R}^n$, which represents an n-tuple $(x_1(t), \ldots, x_n(t))$. Assuming $\mathbf{x}$ is differentiable, the velocity vector $\mathbf{v}(t) = D\mathbf{x} = \mathbf{x}'(t) = (x_1'(t), \ldots, x_n'(t))$ is a vector tangent to the curve γ parametrized by $\mathbf{x}$. If $g : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then a particle moving along the curve parametrized by $\mathbf{x}$ experiences a rate of change of g described by $g'(t) = \nabla g \cdot \mathbf{v}(t)$.

Exercise 6.6 (+). *The gradient is a "vector field"—a function giving a vector at each point—not a single vector. Make an honest statement out of $\nabla g \cdot \mathbf{v}(t)$ by saying at what point the vector field ∇g should be evaluated.*

Sometimes it is easiest to define a curve by a vector function, because these can be added and multiplied by scalars.

Example 6.4. The function $\mathbf{x}(t) := t \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + (6t - 16t^2) \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ defines a function whose value at any time is a vector in $\mathbb{R}^3$, which we think of as an arrow from the origin to the position of a particle at time t. This is a formula representing a particle with initial velocity $\hat{\mathbf{i}} + \hat{\mathbf{j}} + 6\hat{\mathbf{k}}$ in ft/s acting under the force of the earth's gravity (the 16 ft/sec^2 is an approximation to one half the true gravitational constant at the surface of the earth). We can write the function more familiarly as $\mathbf{x}(t) = (t, t, 16t - t^2)$.

Exercise 6.7. *Write the parametrized curve $\mathbf{x}(t) := \cos t [1, 1, 1]^T + \sin t [1, 0, -1]^T$ as a vector function $(x(t), y(t), z(t))$. Bonus: Can you tell what simple shape this curve describes?*

Hypersurfaces

So now we turn to the fun part of the unit – hypersurfaces! Mostly we will show pictures of hypersurfaces in $\mathbb{R}^3$. Every now and then we will try to stretch your mind with discussions in higher dimensions. exercises in this subsection that deals with hypersurfaces in complete generality, leaving these ideas for class discussion.

Suppose $f : \mathbb{R}^n \to \mathbb{R}^1$ is a differentiable function. The equation $f(\mathbf{x}) = 0$ defines an $(n-1)$-dimensional surface. Is this a nice smooth surface or does it have features that will make it tricky to analyze, such as corners, holes, discontinuities and so forth?

Let $\mathbf{x}$ be a point on the surface. Let $\mathbf{v} = \nabla f(\mathbf{x})$ be the gradient of f at the point $\mathbf{x}$. If we move in the direction $\mathbf{u}$ from $\mathbf{x}$, then $D_{\mathbf{u}} f(\mathbf{x}) = \mathbf{v} \cdot \mathbf{u}$. When $\mathbf{v} \cdot \mathbf{u} > 0$, the function f will grow and we can't stay on the surface $f = 0$. Similarly if $\mathbf{v} \cdot \mathbf{u} < 0$ we can't stay on the surface $f = 0$. The only possible directions in which we can stay on the surface are therefore the directions $\{\mathbf{u} : \mathbf{v} \cdot \mathbf{u} = 0\}$. Recall that in $\mathbb{R}^3$, this set of directions is a plane perpendicular to $\mathbf{v}$. The plane on the left of Figure 27 is the plane perpendicular to the gradient, shown by the arrow.
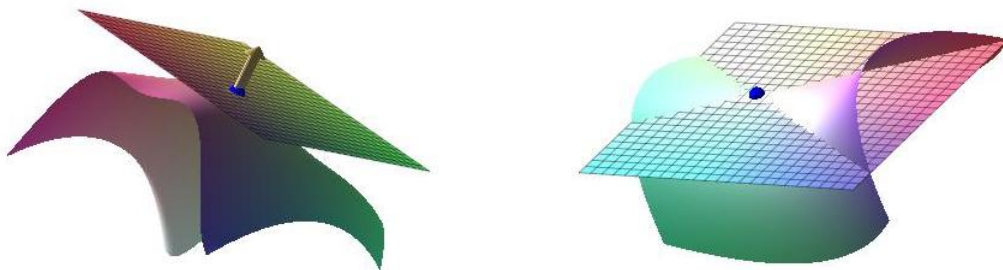

Figure 27: surfaces and tangent hyperplanes

In higher dimensions we don't yet have the vocabulary, but it's not too surprising that the set of vectors perpendicular to a given vector $\mathbf{v}$ is called a **hyperplane**. This hyperplane $\mathcal{H}$ is the **tangent hyperplane** to the surface at $\mathbf{x}$, meaning that if $\gamma(t)$ is any smooth curve in the surface with $\gamma(0) = \mathbf{x}$, then the velocity vector $\gamma'(0)$ is in $\mathcal{H}$. For a drawing the best we can do is to illustrate in $\mathbb{R}^3$. Figure 27 shows examples of a surface, imagined to be the set $f = 0$ in $\mathbb{R}^n$, and its tangent hyperplane at $\mathbf{x}$. The left figure shows a tangent hyperplane that stays on one side of the surface. The right figure shows the situation, common in high dimensions, of a tangent hyperplane that cuts through the surface.

We gave the intuition for this but not a mathmatical proof, which would indeed be too advanced for this course. In fact the above picture does not always work. One clue to this effect is that we made an assumption: we assumed that $\nabla f(\mathbf{x})$ is not the zero vector. If it is, then there is no well defined hyperplane perpendicular to it. The function $x^2 - y^2 - z^2$ has zero gradient at the origin. Figure 28 shows what it looks like for there to be no tangent plane at the origin. We summarize the discussion by stating but not proving a theorem.
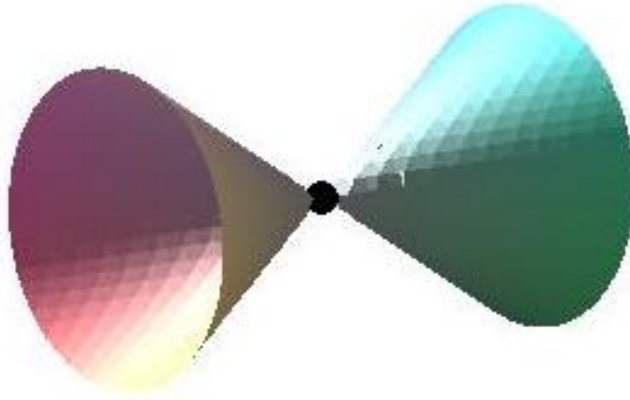
Figure 28: where the gradient vanishes, the surface may not be smooth

Theorem 6.5. *If $f(\mathbf{a}) = 0$ and $\nabla f(\mathbf{a}) \neq \mathbf{0}$ then the surface $\{\mathbf{x} : f(\mathbf{x}) = 0\}$ in $\mathbb{R}^n$ is smooth in a neighborhood of $\mathbf{a}$ and the hyperplane through $\mathbf{a}$ perpendicular to $\nabla f(\mathbf{a})$ is tangent to the surface at $\mathbf{a}$.*

This means that the tangent plane at $\mathbf{a}$ is precisely those $\mathbf{x}$ such that

$$(\mathbf{x} - \mathbf{a}) \cdot \nabla f(\mathbf{a}) = 0.$$

Exercise 6.8 (+)**.** *Consider the plane in $\mathbb{R}^3$ given by $x + 2y - 3z = 0$. What is its tangent plane?*

Computing on hypersurfaces

Going back to three variables, surfaces in $\mathbb{R}^3$ defined by an equation are just like curves defined by an equation in $\mathbb{R}^2$, only now there are three variables. Any equation $F(x, y, z) = 0$ defines a surface. If two variables vary freely, the third changes as a function of the other two (as long as $\nabla F \neq 0$). When this happens, we can ask for the rate of change of one with respect to another. What should $\partial z / \partial x$ mean in this context? It means: consider z as a function of x and y, then find out the rate of change in z when x varies, y is held constant, and z changes in order still to satisfy the equation $F(x, y, z) = 0$.

Exercise 6.9. *We are going to differentiate the relation $F(x, y, z) = 0$ implicitly. Take a moment to try to figure out the branch diagram for this before reading ahead. Record your answer here: did you anticipate the correct branch diagram or not?*

Computationally, the way we compute the implicit partial derivative $\partial z / \partial x$ when $F(x, y, z) = 0$ is to keep in mind the following branch diagram.
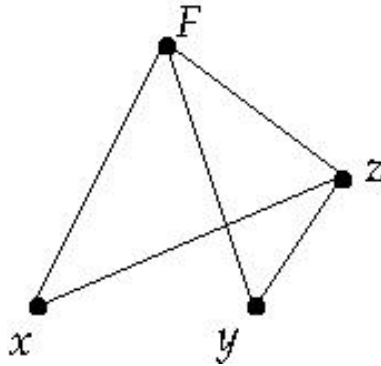
The variables vary in such a way that F remains at zero. Taking the partial derivative with respect to x of the equation $F(x, y, z) = 0$ gives

$$0 = \frac{\partial F}{\partial x}(x, y, z) = \frac{\partial F}{\partial x} + \frac{\partial F}{\partial z} \cdot \frac{\partial z}{\partial x}.$$

The variable y is not in any chain from x to F, so y appears nowhere in the multivariate chain rule[7] expression for $\partial F / \partial x$. Solving for $\partial z / \partial x$ we see that

$$\frac{\partial z}{\partial x} = \frac{-F_x}{F_z}.$$

This looks exactly the same as for two variables, x and z only; compare to equation (6.20). This is not a coincidence. If z is a function of x and y and we hold y constant, we really do have a two variable problem where z is a function of x, and the function happens to have a constant in it.

Example 6.6. Find $\partial z / \partial x$ when the equation $F(x, y, z) = x + y + \ln z - yz = 0$ defines z as a function of x and y. We compute $F_x = 1$ and $F_z = \frac{1}{z} - y$ therefore

$$\frac{\partial z}{\partial x} = \frac{-1}{1/z - y} = \frac{z}{yz - 1}.$$

[7]Did you notice the notational dilemma? We want the total derivative of F with respect to x only, not y. So we used the partial derivative symbol on the left. But this means something different from the partial derivative on the right, which holds z constant as well as y. Don't say we didn't warn you!

6.3 Optimizing over regions with boundary

Many optimization problems are over regions with a boundary—a company might want to maximize profit while borrowing at most the amount of its credit line, or while emitting no more than the legally permitted amount of polution; we want to build a house that maximizes utility while using no more than the available amount of size; we want to minimize the cost per unit of producing a widget while keeping the quality above a certain level. Often the boundary of the region is made up of several separate constraints—for instance, the number of widgets we can make might be bounded by the amount of money we can invest in production, but also by the logical constraint that we can't make fewer than 0 widets.

In all these cases, the boundaries are given by inequalities $\leq$ or $\geq$: we can go up to some bound, but aren't required to. We can divide our region into a *boundary*, where we exactly satisfy one of the constraints, and an *interior* given by replacing all the inequalities with strict inequalities $<$ or $>$. The methods for finding potential extrema on a boundary are different from those for finding them in the interior, so we have to check the boundary and interior separately.

A couple of examples will help to clarify this.

Example 6.7. What is the maximum of the function $g(x, y) := x + 2y$ on the region R where the closed unit disk intersects the first quadrant?

SOLUTION: We begin by describing the region and its boundary mathematically. The region R is illustrated in Figure 29. It is a quarter-disk: a 90° piece of pie. Its boundary consists of three arcs, $\gamma_1, \gamma_2, \gamma_3$. The second and third of these are straight lines. The first, γ_1 will need to be described by one of the three methods from the previous section. In addition there are three vertices, marked A, B and C, at which g will need to be evaluated in case its maximum of over the arcs occurs at an endpoint.

To look for interior maxima we compute the gradient of g and look for critical points on the interior. The gradient of g is the constant vector $\hat{\mathbf{i}} + 2\hat{\mathbf{j}}$, which never vanishes. Therefore R has no interior critical points and neither the absolute maximum nor the absolute minimum can occur in the interior of R.

Turning attention to the boundary, we compute critical points on each of the three bounding arcs. The maximum of g must occur at one of these, or at an endpoint of one of the arcs. The value of g at each endpoint A, B and C is respectively 1, 2 and 0. The arcs γ_2 and γ_3
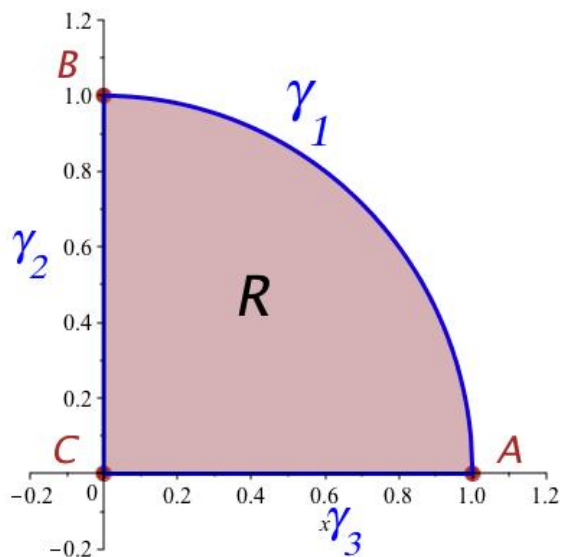
Figure 29: intersection of the unit disk with the postive quadrant

are easy to deal with because they coincide with portions of the coordinate axes and we can simply set one of the coordinates to zero. Along γ_3 (where $y = 0$), the function $x + 2y$ is just x and has no critical points along the arc; the maximum is reached at the right endpoint, A. Along γ_2 (where $x = 0$), the function is just $2y$, again reaching its maximum of 2 at the endpoint B.

Finally we turn our attention to γ_1. By now you know many ways to describe this arc. For instance, we could parameterize it as $(\cos t, \sin t)$. Then, on this arc, $g(x, y) = g(\cos t, \sin t) = \cos t + 2\sin t$. This is a single variable function: $\frac{dg}{dt} = 2\cos t - \sin t$. If the maximum is on this curve, it must be when $0 = \frac{dg}{dt} = 2\cos t - \sin t$, so when $2\cos t = \sin t$, which is when $2x = y$. Plugging in $x^2 + (2x)^2 = 1$, we find that $x = 1/\sqrt{5}$ and $y = 2/\sqrt{5}$. There,

$$g(x, y) = x + 2y = \frac{1}{\sqrt{5}} + \frac{4}{\sqrt{5}} = \frac{5}{\sqrt{5}} = \sqrt{5} \approx 2.23606 \,.$$

This beats all the other maxima, therefore the global maximum of $x + 2y$ over the unit disk in the first quadrant is $\sqrt{5}$ and is achieved at the point $(1/\sqrt{5}, 2/\sqrt{5})$.

At this point, you should stop and work through a self-check that is slightly long but very much follows the example just completed.

Exercise 6.10 (+). *We seek to maximize the function $g(xy) := x/(1 + 2y)$ on the region R trapped between the curve $y = x^2$ and the line $y = x$.*

(a) Draw the region R.

(b) Describe the bounding arcs in any of the ways (i)–(iii) from the beginning of Unit 6.

(c) List all vertices, that is, endpoints of the bounding arcs.

(d) Compute the gradient of g and determine from this whether the maximum of g occurs in the interior.

(e) Compute critical points of g on the bounding arcs according to your description of the bounding arcs in step (ii).

(f) Compare the value(s) of g at the critical point(s) to the values at the endpoints and any interior critical points to determine the overall maximum.

Example 6.8. Where is the maximum of $\dfrac{x}{1 + x^2 + y^2}$ on the disk of radius 2 around the origin?

SOLUTION: Let $f(x, y) := \dfrac{x}{1 + x^2 + y^2}$. The region R is easily describe as a disk, whose boundary is a circle – no vertices. Figure 30 shows a plot of f over R (spoiler alert!).
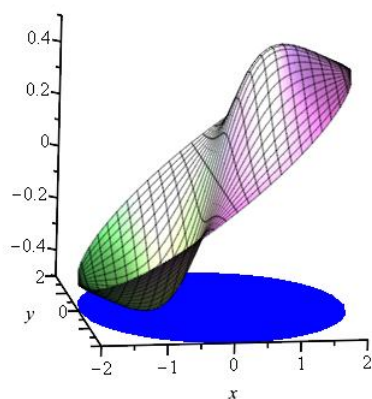

Figure 30: an interior maximum

Step 1: The critical points on the interior are where both partial derivatives of f vanish.

The partial derivatives are as follows.

$$\frac{\partial f}{\partial x} = \frac{1 + y^2 - x^2}{(1 + x^2 + y^2)^2}$$
$$\frac{\partial f}{\partial y} = \frac{-2xy}{(1 + x^2 + y^2)^2}$$

For these both to vanish, the numerators must both vanish, meaning either x or y is zero and $x^2 = 1 + y^2$. The only possibilities are $(x, y) = (\pm 1, 0)$. The corresponding values of f are $\pm 1/2$.

Step 2: The boundary is described by $x^2 + y^2 = 4$. This is nice because when $x^2 + y^2$ is known, the function f simplifies somewhat from $x/(1 + x^2 + y^2)$ to $x/(1 + 4) = x/5$. This provides a bit of a short cut because the maximum is obviously when x is maximized, which occurs at $(2, 0)$ and gives a value of $2/5$.

Step 3: comparing the maximum of $2/5$ on the boundary to $1/2$ on the interior, we find the maximum on R to be the value $1/2$, occurring at the point $(1, 0)$, corroborated in Figure 30.

Exercise 6.11. *What is the absolute minimum of $x/(1 + x^2 + y^2)$ on R, and where is it achieved?*

An example in three variables with lots of boundary

Suppose we want to maximize the expected profit $g(x, y, z)$ when x, y and z represent spending in three different categories, say labor, materials and advertising. The main point of this example is to deal with constraints. We will then try to solve the problem with various objective functions such as $g(x, y, z) = x^2 + y^2 + 2z^2$, $g(x, y, z) = \ln(3 + x^2 + y^2 + 2z^2)$, $g(x, y, z) = xyz - x/2 - y/3 - z/5$, and so on.

The typical constraint is that we only have a finite amount of money to spend on these, thus $x + y + z \leq c$, and that the individual amounts of spending cannot be negative: $x, y, z \geq 0$. Let R be the region

$$\{(x, y, z) \in \mathbb{R}^3 : x + y + z \leq c \text{ and } x, y, z \geq 0\}.$$

To optimize a smooth function over R, we need to understand R and mathematize it. You might already have a glimmer of what R looks like. It is a pyramid (more precisely, a
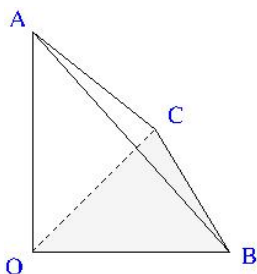
103

Figure 31: the region R, its faces, edges and vertices

tetrahedron) sitting in the positive orthant, with its apex tucked into the origin and three faces coinciding with coordinate planes; see Figure 31.

The region R has an interior, four triangular faces, six line segment edges, and four vertices, which we label for convenience O, A, B, C. With no further information, the maximum of f could occur in the interior or on a two-dimensional face, or an edge or a vertex. The computation is different for each of these 15 pieces. In general there is no way around this, although for some functions things might be simpler, as in this self-check exercise.

Exercise 6.12 (+). *If you know that $f(x, y, z) = x^2 + y^2 + 2z^2$ is increasing in each coordinate, which of the 15 pieces mentioned about is a possible location of the maximum? For example, the vertex O is not. What about the line segement OA (not counting its endpoints, which get counted separately)?*

In general, optimizing on each piece is a different problem. If f is a complicated function, you might have to optimize on each piece separately and compare to find the global maximum on R. To begin this process, we need mathematical descriptions of each piece. For example, the line segment OA is the parametrized set $\{(0, 0, t) : 0 \le t \le c\}$. What about the faces? One of them is characterized by $x + y + z = c$, together with some boundaries. For reasons you will see shortly, we won't worry about the boundaries until doing the optimization.

Exercise 6.13. *What (deceptively simple) equations characterize the other three faces?*

104

7 Linear algebra 2: equations and determinants

The central topic of this chapter is solving systems of linear equations. The solution set is always a **linear** or **affine** space (defined below) so we will spend a little time first talking about those. We will then discuss some algorithms for solving systems of linear equations and introduce matrix algebra. The last section is on **determinants**, which arise in many ways in the study of both linear equations and multivariate calculus.

Throughout the chapter we will approach these topics in several ways: theoretical understanding, computational understanding and geometric understanding. We aim to give you an understanding of how to work with any number of variables (geometrically, to work in real space of any dimension) but will use mainly examples in dimension at most three, so that we can visualize them. The following exercise is just there to check that you understand how simultaneous solutions to more than one equation show up when the equations are graphed.

Exercise 7.1. *Suppose in the plane you graph the zero set of two curves; if you want to be concrete, think of graphing the circle $x^2 + y^2 = 5$ in blue and the cubic $y = x^3 - x + 1$ in red. The solution to the two simultaneous equations corresponds to what in the picture?*
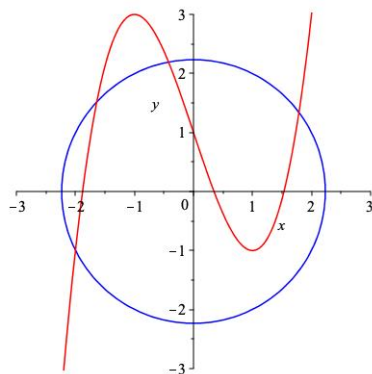
This chapter is a bit of an anomaly because all the methods of computing solutions to linear systems and determinants and inverses of matrices are laborious. After presenting the method of row-reduction, which computes determinants and inverses as well as solving systems of equations, we will present several other computational methods. We leave it to the individual instructor to clarify what methods you must know, and what computing platforms you may be allowed to use.

7.1 Linear and affine spaces in $\mathbb{R}^n$

In three dimensions, the set of points satisfying one linear equation is a plane, that is, hyperplanes in $\mathbb{R}^3$ are just regular planes. The set of points satisfying several linear equations is an intersection of planes. There are only a few geometric objects that an intersection of planes can be: a plane, a line, a single point, or the empty set. Intuitively, these are all the "flat" spaces: spaces that don't curve. Generalizing from $\mathbb{R}^3$ to $\mathbb{R}^n$, one might imagine (and it is true) there there are flat spaces of every dimension, from points (dimension 0), lnes (dimension 1) and planes (dimension 2) to those of any dimension k, the highest-dimenional being hyperplanes (dimension $n-1$), unless you count the whole of $\mathbb{R}^n$ as an n-dimensional flat subspace of itself.

By analogy with other constrained surfaces, we can characterize these subspaces by the equations they satisfy (collections of linear equations in this case) or by parametrizing them. A k-dimensional flat space is easily parametrized by k real variables, using a fixed set of k vectors, called a **basis** (definition upcoming). There is yet another way to characterize which spaces are flat, namely by checking that sums of elements of the space remain in the space. It will be useful to understand all three ways and the connections between them: via equations, geometrically, and via parametrizations. We start by selecting a special category of flat spaces, namely linear spaces, which are just flat spaces containing the origin. You will see how this simplifies things. We will then go back and discuss more general flat spaces, which are called "affine spaces".

Linear spaces

A linear space in $\mathbb{R}^n$ is a subset of $\mathbb{R}^n$ that is flat and passes through the origin. There are several equivalent ways to define this notion precisely: you can take any of them to be the definition and then prove that the others define exactly the same property. We won't prove the equivalences but any linear algebra course would. In the following definition we identify each point in $\mathbb{R}^n$ with the corresponding vector from the origin to the point.

Definition 7.1 (linear space in $\mathbb{R}^n$)**.** *A linear space $L \subseteq \mathbb{R}^n$ is a collection of vectors that satisfies any one (and hence all) of these equivalent properties.*

   (i) (algebraic) A space $L \subseteq \mathbb{R}^n$ is linear if and only if there is a matrix A representing L in the sense that L is the solution set to the equation $A\mathbf{v} = \mathbf{0}$.

(ii) *(geometric) A space $L \subseteq \mathbb{R}^n$ is linear if and only if it is an intersection of hyperplanes through the origin.*

(iii) *(parametrized) A space $L \subseteq \mathbb{R}^n$ is linear if and only if it there is some positive integer k and some vectors $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k)}$ such that L is the set of possible values of $a_1\mathbf{v}^{(1)} + \cdots + a_k\mathbf{v}^{(k)}$ as $a_1, \ldots, a_k$ range over all k-tuples of real numbers.*

(iv) *(formal)*

    (a) *If $\mathbf{v} \in L$ and λ is any real number then $\lambda\mathbf{v} \in L$;*

    (b) *If $\mathbf{v}$ and $\mathbf{w}$ are in L then so is $\mathbf{v} + \mathbf{w}$.*

Definition 7.1 should remind you of the two linearity properties of the map $\mathbf{v} \mapsto A\mathbf{v}$ established in Theorem 3.13, which indicates why the word linear was chosen to represent this property of subspaces.

Example 7.2. We check that the set S of all points (t, t, t) with identical coordinates in $\mathbb{R}^3$ is a linear space; remember, we are identifying this with the set of all vectors $t\mathbf{e}_1 + t\mathbf{e}_2 + t\mathbf{e}_3$; another way of writing S is the set of all column vectors $\begin{bmatrix} t \\ t \\ t \end{bmatrix}$.

Checking the geometric definition (ii) first, we see that $S = L_1 \cap L_2$ where L_1 is the hyperplane $x - y = 0$ through the origin and L_2 is the hyperplane $x - z = 0$ also through the origin. Checking the algebraic definition (i) is similar: S is the solution set of

$$\begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Parametrizing a linear space can be hard but this one was described in parametric terms so it's easy: we can take $k = 1$ and $v^{(1)} = (1, 1, 1)$. Clearly S is the set of $a\mathbf{v}^{(1)}$ as a varies over all real numbers so definition (iii) is satisfied with $k = 1$; thus S is a linear space. We will see below that it is one-dimensional. Finally, for definition (iv), if $\mathbf{v} = (t, t, t) \in S$ then $\lambda\mathbf{v} = (\lambda t, \lambda t, \lambda t)$ which is also in S and if $\mathbf{w} = (s, s, s)$ then $\mathbf{v} + \mathbf{w} = (t + s, t + s, t + s)$ and that too is in S. So we have checked (a) and (b) of (iv).

We might have added a fifth equivalent definition: a linear space is the solution set to simultaneous **homogeneous** linear equations, where "homogeneous" means the constant

107

term is zero. This is easily seen from the algebraic formulation. A vector $\mathbf{v} \in \mathbb{R}^n$ with its tail at the origin corresponds to a point with coordinates $(v_1, \ldots, v_n)$. If A is a $k \times n$ matrix (it has to have n columns if $A\mathbf{v}$ is to make sense) then $A\mathbf{v} = \mathbf{0}$ if every row of A dotted with $\mathbf{v}$ produces zero. For the first row, that means $A_{11}v_1 + A_{12}v_2 + \cdots + A_{1n}v_n = 0$. In general for every $j \leq k$ it means that $A_{j1}v_1 + \cdots + A_{jn}v_n = 0$. In other words, there are j equations that the n-tuple $(v_1, \ldots, v_n)$ must satisfy and they are all linear and homogeneous. The fact in this paragraph may have been obvious to some of you, but it is worth pointing out so we remember what the equation $A\mathbf{v} = \mathbf{0}$ has to do with solving systems of homogeneous linear equations.

Exercise 7.2. *(+) Why should (i) imply (ii), that is, why should the solution set L of $A\mathbf{v} = \mathbf{0}$ be an intersection of hyperplanes? (The significance of intersection is clarified in Exercise 7.1.) Why, more specifically, are they hyperplanes through the origin? A one sentence answer to each of these two questions should be sufficient.*

Affine spaces

Building on the definition we already have for linear spaces, an affine space is define as a linear space moved over (so it no longer has to pass through the origin). Formally, the definition looks like this. The mathematical term for "moved over" is "translated".

Definition 7.3 (affine space). *A space $L' \subseteq \mathbb{R}^n$ is an affine space if there is some vector $\mathbf{v}_*$ and some linear space $L \subseteq \mathbb{R}^n$ such that L' is L translated by $\mathbf{v}_*$. Formally: $L' = \{\mathbf{x} + \mathbf{v}_* : \mathbf{x} \in L\}$.*

There is a matrix version of this: affine spaces are solution sets of any system of linear equations (not just homogeneous linear equations). Solutions of $A\mathbf{v} = \mathbf{b}$ form an affine space because these are just any particular solution $\mathbf{v}_*$ plus a solution to $A\mathbf{v} = \mathbf{0}$.

Exercise 7.3. *Suppose L is a linear space, the solution set to $A\mathbf{v} = \mathbf{0}$. The affine space L' which is L translated by $\mathbf{v}_*$ is supposed to be the solution space to some matrix equation $M\mathbf{v} = \mathbf{b}$. You can always take $M = A$, but what is $\mathbf{b}$?*

In class we will discuss the geometric formulations of linear and affine spaces as much as you want. The parametrized version is trickier and involves notions that are usually given only in a full course on linear algebra. For us it will suffice to state the parametrized version of affine spaces and to give an example.

Theorem 7.4 (parametrized affine spaces). *A space $L' \subseteq \mathbb{R}^n$ is affine if and only if there is some positive integer k, a vector $\mathbf{v}_*$, and some vectors $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k)}$, such that L' is the set of possible values of $\mathbf{v}_* + a_1 \mathbf{v}^{(1)} + \cdots + a_k \mathbf{v}^{(k)}$ as $a_1, \ldots, a_k$ range over all k-tuples of real numbers.*

Example 7.5. Suppose L' is the solution set to these equations.

$$
\begin{aligned}
x - z &= 2 \\
y - z &= 2
\end{aligned}
$$

In matrix form,

$$
\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}.
$$

In geometric terms, the two planes intersect in the line ℓ where $x = y$ and z is 2 less than either this common value; see Figure 32. The line ℓ is parametrized as $(t, t, t - 2)$, for all real numbers t. This can be written as $(0, 0, -2) + t(1, 1, 1)$, where $(0, 0, -2)$ is taking the role of $\mathbf{v}_*$ and only the one vector $(1, 1, 1)$ is needed to parametrize the line $\{(t, t, t) : t \text{ real}\}$ along with a single initial point on the line, say $(0, 0, -2)$.
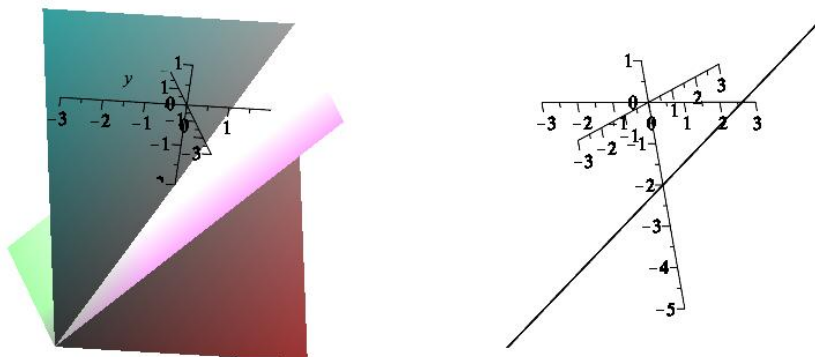

Figure 32: planes $x - z = 2$ and $y - z = 2$ (right) intersect in the line $x = y = z + 2$ (left)

Dimensions and spans

Suppose L is the solution space to k homogeneous equations in n variables. What do you think the dimension of the space will be? Remember, we are intersecting k hyperplanes in $\mathbb{R}^n$. Each one of them has dimension $n-1$. Intersect two and you get a linear space of dimension $n-2$ (well, unless the two were exactly the same hyperplane, but that would be pretty silly). Now intersect with a third: the dimension is now $n-3$ unless the third hyperplane contained the $(n-2)$-dimensional linear space formed by the intersection of the first two; that's not quite so silly. In general, the intersection of k hyperplanes in $\mathbb{R}^n$ will be at least $n-k$ but could be more. In fact it is $n-k'$ where k' is the number of times, when intersecting all the hyperplanes sequentially, that you intersect with a hyperplane that does NOT already contain the intersection of the previous ones. This number k' had better turn out to be the same no matter in which order you do the intersecting! It does. In the next section on row reduction, you'll see one straightforward way to compute dimension.

Definition 7.6 (span)**.** *Let* $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(m)}$ *be vectors in* $\mathbb{R}^n$*. Their* **span** *is defined as the set of all linear combinations of these m vectors. Formally it is the set of possible values of* $\sum_{j=1}^{m} \lambda_j \mathbf{v}^{(j)}$ *as* $\boldsymbol{\lambda}$ *varies over all m-tuples of real numbers.*

Example 7.7. In $\mathbb{R}^3$, the span of the set $\{\mathbf{e}_2, \mathbf{e}_3\}$ is the yz-plane. The span of the set $\{\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_2 + \mathbf{e}_3\}$ is also the yz-plane.

Definition 7.8. *A* **basis** *for a linear space L is a set of vectors that spans L such that no subset spans L. Equivalently it is a set of m vectors that span L, where m is the dimension of L. Equivalently, it is a set of vectors that span L such that none is a linear combination of the others. The* **dimension** *of L is the number of vectors in any basis for L (it will always be the same number).*

Example 7.9. The most obvious basis for the yz-plane in $\mathbb{R}^3$ is $\{e_2, \mathbf{e}_3\}$, as mentioned in Example 7.7. But there are others, for example $\{\mathbf{e}_2, \mathbf{e}_2 + \mathbf{e}_3\}$. In fact any two nonzero vectors in the yz-plane that are not multiples of each other form a basis the yz-plane. Note that the three vectors $\{\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_2 + \mathbf{e}_3\}$ are not a basis. These are redundant: you only need two vectors to span the yz-plane in $\mathbb{R}^3$.

Exercise 7.4. *Let L be the linear space in $\mathbb{R}^4$ defined by the two equations $x+y+z+w=0$ and $x+y-z-w=0$. Find the dimension and a basis for L.*

7.2 Solving linear equations

Let's get straight on what it means to solve a set of equations. It doesn't mean they've been nicely pre-arranged, like a crossword puzzle, to have one and only one solution. Rather it means to find all the solutions there are, and to be sure you have them all; if there are none, you should be sure of that too.

The single most important thing about systems of linear equations is that there is a straight-forward algorithm for solving them (in fact there are many algorithms). Even though linear algebra is taught as an advanced math major course, in some sense we are studying a dead subject, with everything known that there is to know[8].

The rest of this section parallels a video posted on Canvas. You should either read Section 7.2 or watch the video but you don't need to do both unless you want to.

One's first intuition is that if you have equal numbers of equations and variables, then you expect a solution exists and is unique; more variables than equations should correspond to an underdetermined system with many solutions; more equations than variables should correspond to an overdetermnined system with no solutions.

It doesn't always break down this neatly. For example, solving two equations in two variables boils down geoemetrically to finding the intersection of two lines. There can be precisely one solution, but it's also possible there is no solution (two parallel lines, such as $x + y = 1$ and $x + y = 2$) or there are infinitely many solutions (the two equations have the same line as a solution set, for example $x + y = 1$ and $3x + 3y = 3$). A major part of what we want to do in this section is understand correctly which of the three cases we're in. The other major part of course is to compute the solutions when they exist [9].

Simultaneous linear equations are tedious to some and satisfying to others. How would you

---

[8]By the way, the same is true of simultaneous *polynomial* equations. The equations

$$y^4 + xy + x^2 = 3;$$
$$y^3 - x^2 y + x^3 = 13;$$

can be solved just as algorithmically as linear equations can be solved, and you can be just as certain this is true by a glance to identify these as simultaneous polynomial equations. That course, however, is more advanced than the math major linear algebra course, and is taught surprisingly seldom here and elsewhere.

[9]For homogeneous systems, $A\mathbf{v} = \mathbf{0}$, there is always at least one solution, namely all zeros.

solve these equations if you had to get an answer in the next 5 minutes?

$$
\begin{aligned}
3x - 2y + z &= 4 \\
-x + 4y + \frac{7}{3}z &= 3 \\
9x - 20y - 2z &= 0
\end{aligned}
\tag{7.1}
$$

There are a number of ways to do this and we're pretty sure most of you could do it one way or another. In order to study it better, we will begin by agreeing on a precise algorithm. The algorithm is based on the "add multiples of one equation to another" approach, rather than the more popular "solve for one variable in terms of the others" approach, because this is what is easiest to relate to our study of vectors and matrices. We begin with an example.

Example 7.10. Starting with (7.1), we use multiples of the first equation to kill off the x terms in the other equations. The equations transform like this.

$$
\begin{array}{lllll}
3x - 2y + z = 4 & \quad & 3x - 0 + (13/5)z = 33/5 & \quad & 3x - 0 + 0 = 4 \\
0 + (10/3)y + (8/3)z = 13/3 & \quad & 0 + (10/3)y + (8/3)z = 13/3 & \quad & 0 + (10/3)y + 0 = 5/3 \\
0 - 14y - 5z = -12 & \quad & 0 - 0 + (31/5)z = 31/5 & \quad & 0 - 0 + (31/5)z = 31/5
\end{array}
$$

To get from (7.1) to the first of these three sets of equations, add $1/3$ times the first equation to the second equation and -3 times the first equation to the third equation. To get to the next set of equations, use the second equation to kill off y terms in the other two. We add $3/5$ of the second equation to the first equation. The multiple $3/5$ is chosen because $(3/5)(10/3) = 2$. Then we add $42/10$ of the second equation to the third equation. Finally, we use the third equation to kill off z terms in the other two. We need to subtract $13/31$ times the third equation from the first, and $40/93$ times the third equation from the second, yielding the set of equations on the right. This set of equations is in **diagonal** form. Each equation tells us the value of a particular variable. The answer can be read off quickly.

Exercise 7.5. *Read off the answer from the final set of equations.*

What do vectors and matrices have to do with solving linear equations? You can write the three equations all at once like this.

$$
\begin{bmatrix} 3 & -2 & 1 \\ -1 & 4 & \frac{7}{3} \\ 9 & -20 & -2 \end{bmatrix}
\begin{bmatrix} x \\ y \\ z \end{bmatrix}
=
\begin{bmatrix} 4 \\ 3 \\ 0 \end{bmatrix}.
\tag{7.2}
$$

This writes the problem in the form "Solve $A\mathbf{v} = \mathbf{b}$ for $\mathbf{v}$," where

$$A := \begin{bmatrix} 3 & -2 & 1 \\ -1 & 4 & \frac{7}{3} \\ 9 & -20 & -2 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 4 \\ 3 \\ 0 \end{bmatrix} \text{ and } \mathbf{v} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}.$$

In the next section we will discuss how to solve $A\mathbf{v} = \mathbf{b}$ for $\mathbf{v}$ by "dividing by A", or more accurately, multiplying on the left by the inverse of A.

For now, let's see how the matrix A and vector $\mathbf{b}$ transformed when we added multiples of one row to another to eliminate variables. To save space, we write A and $\mathbf{b}$ together, separated by a vertical line. We call this an **augmented matrix** (because we have "augmented" the matrix A with the special column $\mathbf{b}$).

$$\left[\begin{array}{ccc|c} 3 & -2 & 1 & 4 \\ 0 & \frac{10}{3} & \frac{8}{3} & \frac{13}{3} \\ 0 & -14 & -5 & -12 \end{array}\right] \quad \left[\begin{array}{ccc|c} 3 & 0 & \frac{13}{5} & \frac{33}{5} \\ 0 & \frac{10}{3} & \frac{8}{3} & \frac{13}{3} \\ 0 & 0 & \frac{31}{5} & \frac{31}{5} \end{array}\right] \quad \left[\begin{array}{ccc|c} 3 & 0 & 0 & 4 \\ 0 & \frac{10}{3} & 0 & \frac{5}{3} \\ 0 & 0 & \frac{31}{5} & \frac{31}{5} \end{array}\right]$$

Whether manipulating equations or matrices and vectors, the procedure is called **row reduction** (also called **Gaussian elimination**) because the steps involve changing the rows of $[A|\mathbf{b}]$ in ways that don't change the solutions.

Before writing a precise algorithm, let's see what might go wrong. Suppose we tried to solve a nearly identical set of equations, but instead of the matrix A being as in (7.2), the very first matrix entry was 0, not 3. We could not then add multiples of the first row to kill x coefficients in the other rows. This indicates a bad choice of which row is used for what. A way to work around that would be to use the second or third row to clear x coefficients. We could make that happen by interchanging rows 1 and 2 at the start (or equally well, rows 1 and 3). This works, as long as we take this into account not only in A but also in $\mathbf{b}$. We would go from

$$\begin{bmatrix} 0 & -2 & 1 \\ -1 & 4 & \frac{7}{3} \\ 9 & -20 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \\ 0 \end{bmatrix}.$$

to

$$\begin{bmatrix} -1 & 4 & \frac{7}{3} \\ 0 & -2 & 1 \\ 9 & -20 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 0 \end{bmatrix}.$$

113

To keep things neat, we should also allow multiplication or division of any row of $[A|\mathbf{b}]$ by any nonzero constant. For example, we might change the first row from $[-1\ 4\ 7/3|3]$ to $[-3\ 12\ 7|9]$, meaning that the equation $-x+4y+\frac{7}{3}z = 3$ has been replaced by the equivalent equation $-3x + 12y + 7z = 9$.

What we have identified are some special ways of modifying an augmented matrix which have the important property that they don't change the solutions to the system of equations.

Definition 7.11. *When we have an $n \times (n+1)$ augmented matrix $[A|\mathbf{b}]$, the* **row opera-tions** *are the following ways of changing the matrix:*

- *adding a multiple of one row to another row,*

- *multiplying a row by a non-zero number,*

- *swapping the order of two rows.*

The idea that these operations do not change the solutions to the system of equations means that if $[A|\mathbf{b}]$ can be transformed to $[A'|\mathbf{b}']$ using the row operations, for any vector $\mathbf{v}$, we have $A'\mathbf{v} = \mathbf{b}'$ exactly when $A\mathbf{v} = \mathbf{b}$.

Typically our goal will be to use the row operations to turn A into the diagonal matrix where all diagonal entries are 1s and all off-diagonal entries are 0s.

Example 7.12. Consider the system from above:

$$
\left[
\begin{array}{ccc|c}
0 & -2 & 1 & 4 \\
-1 & 4 & \frac{7}{3} & 3 \\
9 & -20 & -2 & 0
\end{array}
\right].
$$

Since we want to reach a diagonal matrix, we first need to get something that isn't 0 in the upper left position, which we can do by swapping rows:

$$
\left[
\begin{array}{ccc|c}
-1 & 4 & \frac{7}{3} & 3 \\
0 & -2 & 1 & 4 \\
9 & -20 & -2 & 0
\end{array}
\right]. \qquad\qquad swap(R1, R2)
$$

114

Now we need to add multiples of the top row to the other rows so the rest of the first column is all 0's. We do that by adding 9 times the top row to the third row:

$$
\left[\begin{array}{ccc|c}
-1 & 4 & \frac{7}{3} & 3 \\
0 & -2 & 1 & 4 \\
0 & 16 & 19 & 27
\end{array}\right]. \qquad add(9 \times R1\ to\ R3)
$$

The second diagonal position has a -2—in particular, it's not 0—so we can use it to eliminate the rest of the second column. We add 2 times the second row to the first row:

$$
\left[\begin{array}{ccc|c}
-1 & 0 & \frac{13}{3} & 11 \\
0 & -2 & 1 & 4 \\
0 & 16 & 19 & 27
\end{array}\right] \qquad add(2 \times R2\ to\ R1)
$$

and then 8 times the second row to the third row:

$$
\left[\begin{array}{ccc|c}
-1 & 0 & \frac{13}{3} & 11 \\
0 & -2 & 1 & 4 \\
0 & 0 & 27 & 59
\end{array}\right]. \qquad add(8 \times R2\ to\ R3)
$$

Finally, we use the third row to eliminate the rest of the third column: we add $-13/81$ times the third row to the first:

$$
\left[\begin{array}{ccc|c}
-1 & 0 & 0 & \frac{124}{81} \\
0 & -2 & 1 & 4 \\
0 & 0 & 27 & 59
\end{array}\right]. \qquad add(-13/81 \times R3\ to\ R1)
$$

(We chose $-13/81$ because $(13/3) + 27 * (-13/81) = 0$. There's some annoying arithmetic to verify that $11 - 59\frac{13}{81} = \frac{124}{81}$.) And the last step is adding $-1/27$ times the third row to the second:

$$
\left[\begin{array}{ccc|c}
-1 & 0 & 0 & \frac{124}{81} \\
0 & -2 & 0 & \frac{49}{27} \\
0 & 0 & 27 & 59
\end{array}\right]. \qquad add(-1/27 \times R3\ to\ R2)
$$

Finally, we can divide each row by the diagonal entry so they're all 1's:

$$
\left[\begin{array}{ccc|c}
1 & 0 & 0 & -\frac{124}{81} \\
0 & 1 & 0 & -\frac{49}{54} \\
0 & 0 & 1 & \frac{59}{27}
\end{array}\right]. \qquad
\begin{array}{l}
mulitply(-1 \times R1) \\
multiply(-1/2 \times R2) \\
multiply(1/27 \times R3)
\end{array}
$$

Then the right column is the solution: $x = -124/81$, $y = -49/54$, $z = 59/27$.

Note that in this example we used a kind of shorthand to keep track of the three kinds of row operations (add, multiply and swap), which rows were involved, and what the mulitplication factor was. These are not standard and you don't have to know them for exams. On the other hand, we do require clear writing and the graders will understand these if you choose to use them.

If the method fails to produce a unique solution, that's because the original system of equations was either underdetermined or overdetermined (many solutions or none at all).

Example 7.13 (an underdetermined system). Start with

$$
\begin{aligned}
2z &= 10 \\
15x - 25y + 5z &= 40 \\
3x - 5y - 4z &= -17
\end{aligned}
$$

This becomes the augmented matrix

$$
\left[\begin{array}{ccc|c}
0 & 0 & 2 & 10 \\
15 & -25 & 5 & 40 \\
3 & -5 & -4 & -17
\end{array}\right].
$$

We start out as in the previous example. Because the upper left is 0, we swap the first two rows:

$$
\left[\begin{array}{ccc|c}
15 & -25 & 5 & 40 \\
0 & 0 & 2 & 10 \\
3 & -5 & -4 & -17
\end{array}\right]. \qquad\qquad swap(R1, R2)
$$

Now we eliminate the first column by adding $-1/5$ of the first row to the third:

$$
\left[\begin{array}{ccc|c}
15 & -25 & 5 & 40 \\
0 & 0 & 2 & 10 \\
0 & 0 & -5 & -25
\end{array}\right]. \qquad\qquad add(-1/5 \times R1 \ to \ R3)
$$

Notice the problem: the only place the second column is non-zero is in the first row, which is now also the only place where the first column is non-zero.

So there's nothing left to do with the second column. We can go on to the third: add $5/2$ times the second column to the third to get

$$
\left[\begin{array}{ccc|c}
15 & -25 & 5 & 40 \\
0 & 0 & 2 & 10 \\
0 & 0 & 0 & 0
\end{array}\right]. \qquad\qquad add(5/2 \times R1 \ to \ R3)
$$

116

We can also subtract $5/2$ times the second column from the first:

$$\left[\begin{array}{ccc|c} 15 & -25 & 0 & 15 \\ 0 & 0 & 2 & 10 \\ 0 & 0 & 0 & 0 \end{array}\right]. \qquad add(-5/2 \times R2 \ to \ R1)$$

We might notice that everything in the top row is divisible by 5, so we can simplify to:

$$\left[\begin{array}{ccc|c} 3 & -5 & 0 & 3 \\ 0 & 0 & 2 & 10 \\ 0 & 0 & 0 & 0 \end{array}\right]. \qquad multiply(1/5 \times R1)$$

Again, we can divide through to get 1s in the key entries:

$$\left[\begin{array}{ccc|c} 1 & -5/3 & 0 & 1 \\ 0 & 0 & 1 & 5 \\ 0 & 0 & 0 & 0 \end{array}\right]. \qquad \begin{array}{l} multiply(1/3 \times R1) \\ multiply(1/2 \times R2) \end{array}$$

We can read the solution easily enough: the first row says $x - (5/3)y = 1$, the second row says $z = 5$, and the third row says $0 = 0$ (which doesn't tell us anything interesting).

In the original equation, if -17 had been -16, then all steps would have been exactly the same, except the final equation would have been $0 = 1$ instead of $0 = 0$. Instead of ignoring the last equation as a tautology, we would conclude that there is no solution.

Exercise 7.6. *Apply the row reduction algorithm to solve this system of two linear equations. Please write the new matrix at every step - there aren't too many steps. Begin by writing down $[A|\mathbf{b}]$.*

$$\begin{aligned} x - 3y &= 11 \\ 2x + 5y &= 77 \end{aligned}$$

7.3 Matrix algebra

Recalling the definition of matrix multiplication on page 49, we are going to try to use matrix algebra to solve systems of linear equations. The advantage over row reduction is that

when it works, matrix algebra is neater, with a lot less to keep track of. The disadvantage is that row reduction always works, whereas solving via matrix algebra only works for square matrices (the number of equations is the same as the number of variables) without redundancies. Later we will say what non-redundancy means using the **determinant**; for now, a clumsy but accurate formulation is the geometric one: each time you intersect with a new hyperplane, it neither contains the existing affine space nor is disjoint from it. Hence the dimension of the solution space reduces by exactly 1 with each new equation and the solution to the entire system is a single point.

We need to be careful of one thing: matrix multiplication is not commutative. In general, $AB \neq BA$. In fact when A and B are not square matrices, but the dimensions match up so that AB is defined, it can happen that BA is not defined. Matrix multiplication is, however, associative: $(AB)C = A(BC)$ and if one side is defined, so is the other.

The matrix algebra method boils down to this. Suppose we want to solve $A\mathbf{v} = \mathbf{b}$ for $\mathbf{v}$, where A is an $n \times n$ matrix and $\mathbf{b} \in \mathbb{R}^n$. The first thing to observe when multiplying $n \times n$ matrices is that one matrix is special: multiplying by this matrix, on the left or the right, does not change any matrix. It is usually called I, the **identity** matrix. Thus, $IM = MI = M$ for all $n \times n$ matrices, M. In fact $IC = C$ for all $n \times anything$ matrices and $CI = I$ for all $anything \times n$ matrices. So for example $I\mathbf{v} = \mathbf{v}$ when $\mathbf{v} \in \mathbb{R}^n$ is viewed as a column vector.

Exercise 7.7. *(+) Figure out the $n \times n$ identity matrix I. You can do this by thinking of the map I defines on $\mathbb{R}^n$: where does it send each $\mathbf{e}_j$? Once you figure it out, remember it, as you will need it later.*

The next thing is that some matrices have **inverses**. The $n \times n$ matrix A is said to have an inverse, if there is some matrix A^{-1} such that $A^{-1}A = AA^{-1} = I$. When A has an inverse, we can indeed solve $A\mathbf{v} = \mathbf{b}$ by multiplying both sides by A^{-1}. Being careful to multiply on the left, we get $A^{-1}A\mathbf{v} = A^{-1}\mathbf{b}$. The left side of this is $I\mathbf{v} = \mathbf{v}$, so we have found the value of $\mathbf{v}$: it is $A^{-1}\mathbf{b}$.

Exercise 7.8. *(+) Somewhere there, we used associativity of matrix multiplication. Can you find where?*

About this inverse matrix, when does it exist and how can we compute it? One correct answer is: you can find A^{-1} when A can be row-reduced to a diagonal matrix. When it can't, you get either an underdetermined system with an equation $0 = 0$ or an overdetermined

equation $0 = 1$. In the next section we will see that there is a number you can compute – the determinant of A – which will be nonzero if A is invertible and zero if A is not invertible. In the remainder of this section, we pull out the row-reduction apparatus to show you a way to compute the inverse of a matrix. We don't expect you to do this operation more than a couple of times; there is online software to do this, to which we will refer you. But you should understand how you can, in principle, do it yourself. In the next section, once you know about determinants, we mention some other methods of inverting matrices and computing determinants that work particularly well for 2×2 and 3×3 matrices[10].

Algorithm 7.14 (computing the inverse of a square matrix via row operations)**.**

1. *Start with the matrix $[A|I]$ where A is any $n \times n$ matrix and I is the $n \times n$ identity matrix. Never mind that it no longer has dimensions $n \times (n+1)$.*

2. *Use row reduction to transform A to the diagonal matrix with 1s along the diagonal. If this can't be done then A has no inverse.*

3. *When A is transformed into the diagonal matrix with 1s, the matrix on the right side of the vertical bar is now A^{-1}.*

Example 7.15. To invert the matrix A in (7.1), begin with $[A|I]$ and row reduce. Without narrating each step, here is the result.

$$\left[\begin{array}{ccc|ccc} 3 & -2 & 1 & 1 & 0 & 0 \\ -1 & 4 & \frac{7}{3} & 0 & 1 & 0 \\ 9 & -20 & -2 & 0 & 0 & 1 \end{array} \right] \qquad \left[\begin{array}{ccc|ccc} 1 & -2/3 & 1/3 & 1/3 & 0 & 0 \\ 0 & 10/3 & 8/3 & 1/3 & 1 & 0 \\ 0 & -14 & -5 & -3 & 0 & 1 \end{array} \right]$$

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 13/15 & 2/5 & 1/5 & 0 \\ 0 & 1 & 4/5 & 1/10 & 3/10 & 0 \\ 0 & 0 & 31/5 & -8/5 & 21/5 & 1 \end{array} \right] \qquad \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 58/93 & -12/31 & -13/93 \\ 0 & 1 & 0 & 19/62 & -15/62 & -8/31 \\ 0 & 0 & 1 & -8/31 & 21/31 & 5/31 \end{array} \right]$$

The LCM of $\{31,62,93\}$ is 186. Clearing fractions we see $A^{-1} = \dfrac{1}{186} \left[\begin{array}{ccc} 116 & -72 & -26 \\ 57 & -45 & -48 \\ -48 & 126 & 30 \end{array} \right].$

Exercise 7.9. *Use Algorithm 7.14 to invert* $\left[\begin{array}{cc} 2 & 5 \\ 3 & 8 \end{array} \right].$

---

[10]The authors of this text routinely use these methods for 4×4 matrices in their research, and sometimes larger sparse matrices (lots of zero entries); we won't be asking you to do this.

7.4 Determinants

It turns out that testing for invertability and computing inverses involve something called the **determinant** of a square matrix. The determinant is important enough that we will spend the better part of a week on it. For example, a square matrix is invertible if and only if it has a nonzero determinant. There are a number of different ways to compute the determinant. Usually one begins with one of these computational methods as a definition, then proves that the other methods produce the same thing. We don't have the time to prove most of these things. Rather, we will define the determinant as a geometric concept, then state various algorithms to compute it. The ingredients of the definition are as follows. You can read them quickly and casually. We won't be working with these definitions, they are just to help your intuition.

- We think of an $n \times n$ matrix A as an ordered list of n column vectors, $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}$.

- The **parallelopiped** spanned by n vectors $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}$ in $\mathbb{R}^n$ is the set $\{a_1 \mathbf{v}^{(1)} + \cdots + a_n \mathbf{v}^{(n)} : 0 \le a_1, \ldots, a_n \le 1\}$. In two dimensions, this is a paralellogram. Figure 33 shows the parallelopiped in two dimensions spanned by the two vectors $\mathbf{v}$ and $\mathbf{w}$.

- The **volume** of a body in n-dimensional space is the usual notion of volume: count the number of little n-dimensional cubes of side ε sitting inside the body, multiply by ε^n and take the limit as $\varepsilon \to 0$.

- The **signed volume** of the parallelopiped is the same but with a negative sign if the linear map A mapping each $\mathbf{e}_j$ to $\mathbf{v}^{(j)}$ is *orientation reversing* (see posted hadout).
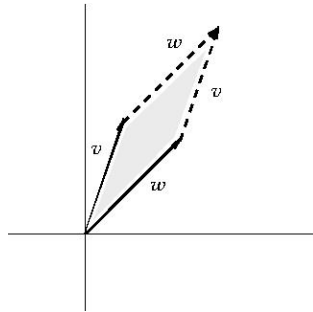

Figure 33: a parallelopiped in two dimensions (a paralellogram)

Definition 7.16 (determinant)**.** *Let $A = [\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}]$ be a $n \times n$ matrix. The determinant of A is the signed volume of the parallelopiped spanned by $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}$.*

We're going to tell you a lot of facts about determinants, and hope to give you enough of the connection between these to make these facts seem plausible and easy to remember. We are not going to prove most of them. That's probably no shocker for you, but it is for us. Mathematicians are generally allergic to facts without proof. We always want our students to know why things work. However, in this case, it's not an option. Proving these things is the subject of a true linear algebra class (which you are highly encouraged to take!).

Theorem 7.17 (properties of the determinant)**.**

  (i) $\det(I) = 1$ *where I is the identity matrix.*

  (ii) If you switch two rows of A, you flip the sign of $\det(A)$.

  (iii) If A has two identical rows, then $\det(A) = 0$.

  (iv) $\det(AB) = \det(A)\det(B)$.

  (v) A is invertible if and only if $\det(A) \neq 0$.

  (vi) $\det(A) = \det(A^T)$, *where A^T is the* **transpose** *of A, defined by $A^T_{ij} := A_{ji}$.*

Exercise 7.10. *One of the first five statements immediately implies the next one. Which one, and why?*

Let's discuss the properties for a bit. It should be easy to convince yourself of property (i): the parallelopiped is the n-dimensional unit cube, which has volume 1, and the map is the identity map, which should not be mirror reversing. Property (ii) follows from the rules for mirror reversing, which we have never stated, but we hope do not seem too implausible. Property (v) says that a matrix is invertible if its map does not squish the n-cube down to volume zero, which we have seen happen, for example, in Figure 16. You may be able to convince yourself of property (iv) by composing the maps $v \mapsto B\mathbf{v} \mapsto AB\mathbf{v}$. If Q is the unit cube in $\mathbb{R}^n$ then BQ is a parallelopiped of volume $\det(B)$. Under A, each unit cube gets mapped to a cube of volume $\det(A)$, therefore $A(BQ)$ should have volume $\det(A)\det(B)$. The determinant can be viewed as the **expansion factor** of the linear mapping.

There are some more properties of determinants, summed up in the following theorem, that relate to row reduction.

Theorem 7.18 (determinants and row reduction)**.**

(i) If you multiply all elements of one row by a constant λ, the determinant is multiplied by λ.

(ii) If A has an all-zero row, then det(A) = 0.

(iii) If you add a multiple of one row to another row, the determinant is unchanged

SKETCH OF ARGUMENT: Figure 33 can be extended to geometric arguments for (i) and (iii). The left side of Figure 34 shows the result in two dimensions of multiplying one vector by λ and leaving the other unchanged. The altitude is unchanged, and the area (base times altitude) is multiplied by λ. The right side of the illustration shows what happens when a multiple of $\mathbf{w}$ is added to $\mathbf{v}$. The endpoint of $\mathbf{v}$ slides along the opposite side of the parallelogram, to become $\mathbf{v}'$; the side of the parallelogram opposite $\mathbf{w}$ slides along itself from its previous starting point at the tip of $\mathbf{v}$ to the new starting point at the start of $\mathbf{v}'$, resulting in no change to the altitude or base, and therefore, to the area.
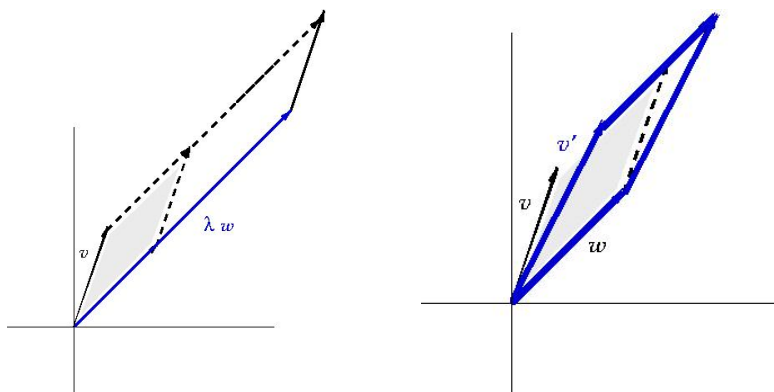

Figure 34: change in area under row operations

Putting together several results from above, we know exactly how the row operations affect the determinant:

- adding a multiple of one row to another row does not change the determinant,
- muliplying a row by λ multiplies the determinant by λ,
- swapping the order of two rows multiplies the determinant by −1.

Using row reduction, we can take a matrix A and reduce it either to I or to a matrix with a row which is all 0's. If we reduce A to the identity matrix I, we know that $\det(I) = 1$, so if we keep track of when we swapped rows and multiplied by λ, we can also figure out the determinant of A. If we reduce A to a matrix with a row which is all 0's, we know that the determinant of A must have been 0.

Example 7.19. What is $\det \left(\begin{bmatrix} 0 & -2 & 1 \\ -1 & 4 & \frac{7}{3} \\ 9 & -20 & -2 \end{bmatrix} \right)$?

In Example 7.12, we reduced this matrix to the identity using the following steps (you may want to compare to the description above, which writes out the intermediate matrices):

1. swapped the top two rows, $\hspace{3cm} swap(R1, R2)$
2. added 9 times the first row to the third row, $\hspace{1cm} add(9 \times R1 \text{ to } R3)$
3. added 2 times the second row to the first row, $\hspace{1cm} add(2 \times R2 \text{ to } R1)$
4. added 8 times the second row to the third row, $\hspace{1cm} add(8 \times R2 \text{ to } R3)$
5. adding $-13/81$ times the third row to the first row, $\hspace{0.3cm} add(-13/81 \times R3 \text{ to } R1)$
6. adding $-1/27$ times the third row to the second row, $\hspace{0.3cm} add(-1/27 \times R3 \text{ to } R2)$
7. multiplying the first row by -1, $\hspace{2cm} multiply(-1 \times R1)$
8. multiplying the second row by $-1/2$, $\hspace{1.5cm} multiply(-1/2 \times R2)$
9. multiplying the third row by $1/27$. $\hspace{2cm} multiply(1/27 \times R3)$

The only steps that affect the determinant are the first step and the last three:

$$(-1) \cdot (-1) \cdot (-1/2) \cdot (1/27) \det \left(\begin{bmatrix} 0 & -2 & 1 \\ -1 & 4 & \frac{7}{3} \\ 9 & -20 & -2 \end{bmatrix} \right) = \det(I) = 1 \,,$$

so

$$\det \left(\begin{bmatrix} 0 & -2 & 1 \\ -1 & 4 & \frac{7}{3} \\ 9 & -20 & -2 \end{bmatrix} \right) = -54 \,.$$

Exercise 7.11. *Find the determinant of* $\begin{bmatrix} 2 & 5 \\ 3 & 8 \end{bmatrix}$.

More ways to compute determinants

This section gives two ways to compute the determinant of a square matrix that are completely different from the row reduction method. For most people they are faster for matrices of sizes 2 and 3, about the same for 4×4 matrices, and slower once the size is 5 or greater. The first algorithm uses permutations.

Algorithm 7.20 (computing determinants using permutations). *There are $n!$ ways of choosing n entries of A with precisely one from each row and column. These are sometimes called* **permutations** *though it is not important here. For each of these, multiply all the entries together, then add together these $n!$ products, flipping the signs of half of them as follows. Count how many row swaps (switching one row with another) it takes to get all the chosen entries on the diagonal; if the number is odd you need to flip the sign of the product.*

Example 7.21. For 2×2 matrices there are only two ways of choosing. In the matrix $A := \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ you can choose either a and d or b and c. For a and d, getting the chosen elements to the diagonal requires zero row swaps; for b and c it requires one row swap. Therefore the product bc gets a minus sign, the product ad does not. This leads to a formula you will likely recognize:

$$\det \left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} \right) = ad - bc \,.$$

For a 3×3 matrix, there are only six possibilities. Three are odd (require 1 row swap) and the other three are even (require 0 or 2 row swaps). Explicitly, if $A := \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$, then the products ceg, afh and bdi have their sign flipped, while the products aei, bfg and cdh do not change sign. This leads to the formula

$$\det \left(\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \right) = aei + bfg + cdh - ceg - afh - bdi \,. \qquad (7.3)$$

Example 7.22. Let $A := \begin{bmatrix} 3 & -2 & 1 \\ -1 & 4 & \frac{7}{3} \\ 9 & -20 & -2 \end{bmatrix}$. The determinant will be the sum of six

124

products. The first three appear without a sign flip. The first is $a_{11}a_{22}a_{33} = (3)(4)(-2) = -24$. The second is $a_{12}a_{23}a_{31} = (-2)(7/3)(9) = -42$. The third is $a_{13}a_{21}a_{32} = (1)(-1)(-20) = 20$. So far these sum to -46.

The rest are odd and need to be sign-flipped. Thus we subtract $a_{11}a_{23}a_{32} = (3)(7/3)(-20) = -140$, $a_{12}a_{21}a_{32} = (-2)(-1)(-2) = -4$ and $a_{13}a_{22}a_{31} = (1)(4)(9) = 36$, which sum to -108. We end up with $-46 - (-108) = \mathbf{62}$.

Exercise 7.12. *Compute* $\det \begin{bmatrix} 1 & 0 & 2 \\ 4 & 3 & 1 \\ 1 & 0 & 5 \end{bmatrix}$ *via permutations.*

Yet another way to compute determinants is via cofactor expansions. This turns out to be a good method when most of the entries of the matrix are zero (this is called a sparse matrix). You don't have to read this, but cofactors do also come in handy also when computing inverse matrices: the authors of this text usually use cofactors when inverting 3×3 matrices (which we have to do surprisingly often!).

Definition 7.23. *Let A be an $n \times n$ matrix. The (i,j)-cofactor of A is the determinant of the $(n-1) \times (n-1)$ matrix you get when you delete the row i and column j from A. Denote this by $\operatorname{cof}(A, i, j)$.*

Theorem 7.24. *Let A be any $n \times n$ matrix. Pick a row, say row i. Then $\det(A)$ is computed by the following sum.*

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} A_{ij} \operatorname{cof}(A, i, j).$$

We remark that because $\det(A) = \det(A^T)$ you can also use cofactor expansions along columns. The following example shows why cofactor exapansion is easy in the 3×3 case: the cofactors are 2×2 determinants which can by computed on the fly because the formula $ad = bc$ is easy to remember.

Example 7.25. We compute

$$\det \left(\begin{bmatrix} 1 & 0 & 2 \\ 4 & 3 & 1 \\ 1 & 0 & 5 \end{bmatrix} \right)$$

via cofactors. The best row column to use is the second column because there is only one nonzero entry. The cofactor expansion for column $j = 2$ boils down to a single term:

$$
\begin{aligned}
\det(A) &= \sum_{i=1}^{n}(-1)^{i+2}A_{i2}\operatorname{cof}(A,i,2) \\
&= (-1)^{2+2}A_{22}\operatorname{cof}(A,2,2) \\
&= 3\det\left(\begin{bmatrix} 1 & 2 \\ 1 & 5 \end{bmatrix}\right) \\
&= 3\cdot 3 = 9.
\end{aligned}
$$

We don't want to twist your arm to learn cofactors (really!) but thought we'd mention one more thing they are good for, namely computing inverse matrices. The following theorem works quite well for sparse matrices and 3×3 matrices.

Theorem 7.26 (inverting matrices via cofactors). *Let A be a $n \times n$ matrix with nonzero determinant. Then the (i,j)-entry of A^{-1} is $\dfrac{(-1)^{i+j}}{\det(A)}\operatorname{cof}(A,j,i)$. Note the order j then i in the last expression!*

Example 7.27. We use cofactors to invert the matrix in Example 7.25. We already know the determinant is 9, so A^{-1} will be $\frac{1}{9}$ times the array of cofactors with alternating signs. Computing the (i,j)-cofactors for $1 \le i,j \le 3$ and putting them in an array gives

$$
\begin{bmatrix}
15 & 19 & -3 \\
0 & 3 & 0 \\
-6 & -7 & 3
\end{bmatrix}.
$$

Putting in the alternating signs, dividing by the determinant 9, and transposing because the (j,i) cofactor gives the (i,j) entry, gives

$$
A^{-1} = \frac{1}{9}\begin{bmatrix}
15 & 0 & -6 \\
-19 & 3 & 7 \\
-3 & 0 & 3
\end{bmatrix}.
$$

A final way is via to invert matrices is "Dodgson condensation" invented by Charles Lutwidge Dodgson (a.k.a. Lewis Carroll, author of *Alice in Wonderland*). You can look this up in Wikipedia if you're interested. It is faster than the permutation or cofactor method for large non-sparse matrices but has trouble when it runs into an entry of zero.

Determinants and derivatives of maps

Recall from Unit 3 that a linear map $f : \mathbb{R}^2 \to \mathbb{R}^2$ can be represented by a 2×2 matrix $A := \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. In Example 3.20, we saw that $f : \mathbb{R}^2 \to \mathbb{R}^2$ fails to be invertible if it maps the two basis vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ to parallel vectors. The map f is invertible if and only if the matrix A is invertible, which means it should have nonzero determinant.

When this map does have a nonzero determinant, it has a uniform effect on areas: the area of the image of any region is the area of the original region multiplied by $\det(A)$. For example, the unit square gets mapped to the parallelopiped spanned by the column vectors of A, which, by definition, is the determinant of A. The same is true in any dimension: a linear map has uniform volume expansion. You will prove this in this week's worksheets.

For nonlinear maps f, expansion of area (or in general, of n-dimensional volume) is not uniform over space, but can be defined infintesimally at each point. The details of how this can be defined as a limit are hairy, so we won't do it. What is more straightforward is to see what this limit must be. Recall from Definition 4.12 on page 66 that the derivative of $f : \mathbb{R}^n \to \mathbb{R}^n$ is the $n \times n$ matrix Df defined by

$$Df_{ij}(\mathbf{a}) = \frac{\partial f_i(\mathbf{a})}{\partial x_j(\mathbf{a})} .$$

The matrix $Df(\mathbf{a})$ is also called the **Jacobian** and its determinant $J(\mathbf{a})$ is called the **Jacobian determinant** of f. Figure 35 illustrates the significance of J. The dimension shown is $n = 2$ but it could be any dimension.
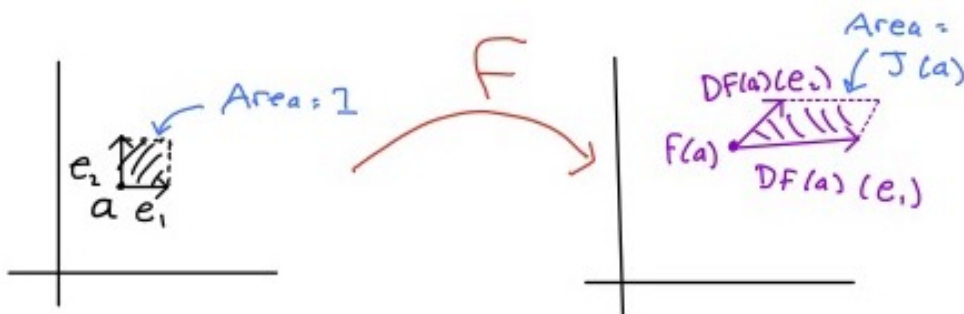

Figure 35: the Jacobian determinant is the factor by which volume is increased near a point

The figure shows where f maps a unit square whose lower left corner is at $\mathbf{a}$. The multivariate increment theorem, Proposition 4.16, says that this square gets mapped to approximately a square spanned by the column vectors of $Df(\mathbf{a})$, drawn starting at the point $f(\mathbf{a})$. Proposition 4.16 is not exact, but gets closer if the unit square is replaced by a square with side length ε and we scale each of the two column vectors by ε. For these parallelopipeds, the ratio of area after mapping to before is $J(\mathbf{a})$. Thus the Jacobian determinant tells us the area magnification factor if f, in the limit for small regions near $\mathbf{a}$.

Exercise 7.13. *Define $f : \mathbb{R}^2 \to \mathbb{R}^2$ by $f(r,\theta) := (r\cos\theta, r\sin\theta)$. Determine $Df(\mathbf{a})$ and $J(\mathbf{a})$ at an arbitrary point $\mathbf{a} := (r_0, \theta_0)$.*

8 Constrained optimization

8.1 Lagrange multipliers

So far we have focused on optimizing boundaries by parameterizing them. In this section, we will encounter a method for optimizing a differentiable function $g : \mathbb{R}^n \to \mathbb{R}$ over a hypersurface R defined implicitly by an equation $R = \{\mathbf{x} : F(\mathbf{x}) = 0\}$. One might go about it this way (though we won't). Find the directions $\mathbf{u}$ tangent to the surface; find critical points $\mathbf{x}$ where $D_{\mathbf{u}}g(\mathbf{x}) = 0$ for all $\mathbf{u}$ tangent to R at $\mathbf{x}$; evaluate at these points and compare to boundary values (if any) and values at infinity (if any). The part we are going to do differently from the above is to deal with the tangent directions $\mathbf{u}$ separately. Instead, we recall Proposition 6.2 which said that a critical point on a curve occurs when the gradients line up. The same thing works in any number of variables.

Lemma 8.1 (gradients should line up). *Let $F, g : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and have nonzero gradients at the point $\mathbf{x}$. Then $D_{\mathbf{u}}g(\mathbf{x}) = 0$ for all vectors $\mathbf{u}$ tangent to the surface $R := \{\mathbf{x} : F(\mathbf{x}) = 0\}$ at $\mathbf{x}$ if and only if $\nabla g(\mathbf{x})$ is parallel to $\nabla F(\mathbf{x})$.*

PROOF: Recall from the discussion preceding Figure 27 in Chapter 6 that the tangent hyperplane to R at $\mathbf{x}$ is the hyperplane of vectors $\mathbf{u}$ such that $\mathbf{u} \cdot \nabla F = 0$. Recall also that $D_{\mathbf{u}}g(\mathbf{x}) = \nabla g(\mathbf{x}) \cdot \mathbf{u}$. Therefore, $D_{\mathbf{u}}g(\mathbf{x}) = 0$ for all tangent vectors $\mathbf{u}$ to R at $\mathbf{x}$ if and only if $\nabla g(\mathbf{x}) \cdot \mathbf{u} = 0$ for all tangent vectors $\mathbf{u}$ to R at $\mathbf{x}$. Therefore $\nabla g(\mathbf{x}) \cdot \mathbf{u} = 0$ for all $\mathbf{u}$ such that $\nabla F(\mathbf{x}) \cdot \mathbf{u} = 0$; this is equivalent to $\nabla F(\mathbf{x})$ being parallel to $\nabla g(\mathbf{x})$. □

According to this result, finding critical points for g on R is equivalent to finding $\mathbf{x} \in R$ such that $\nabla g(\mathbf{x})$ is parallel to $\nabla F(\mathbf{x})$. The computational question remains as to how we find these. Two vectors $\mathbf{u}$ and $\mathbf{v}$ are parallel when $\mathbf{u} = \lambda \mathbf{v}$ for some nonzero real λ. The equations for $\nabla F(\mathbf{x}) = \lambda \nabla g(\mathbf{x})$ are that $\partial F / \partial x_j$ must equal $\partial g / \partial x_j$ for each j. That gives n equations but now there are $n + 1$ unknowns, namely $x_1, \ldots, x_n$ and λ. The last equation is that you are on R, in other words $F(\mathbf{x}) = 0$. We have proved the following theorem.

Theorem 8.2 (method of Lagrange multipliers). *Let $F, g : \mathbb{R}^n \to \mathbb{R}$ be differentiable. The any maxima or minima of g in the interior of $R := \{\mathbf{x} : F(\mathbf{x}) = 0\}$ satisfy these $n + 1$*

equations in $n + 1$ variables.

$$\begin{aligned}
\frac{\partial F}{\partial x_1} &= \lambda \frac{\partial g}{\partial x_1} \\
\frac{\partial F}{\partial x_2} &= \lambda \frac{\partial g}{\partial x_2} \\
\vdots \quad &\vdots \quad \vdots \\
\frac{\partial F}{\partial x_n} &= \lambda \frac{\partial g}{\partial x_n} \\
F(\mathbf{x}) &= 0 \, .
\end{aligned} \tag{8.1}$$

$\square$

Exercise 8.1. *Write the Lagrange multiplier equations to maximize the utility function $u(x, y) := xy$ from the Pizza and Froyo example in Section 2.2, if the constraint is that it costs 10 cents per ounce to produce pizza and 1 dollar per pint to produce FroYo, and production costs for the meal must be kept to m dollars.*

The name **Lagrange multiplier** refers to the extra variable λ that you introduce and solve for. Let's see some examples.

Example 8.3. PROBLEM: Maximize $x + 2y + 3z$ on the unit sphere.
SOLUTION: Letting $g(x, y, z) := x + 2y + 3x$ and $F(x, y, z) = x^2 + y^2 + z^2$, we have a problem in the form to which Theorem 8.2 applies. Computing $\nabla g = \hat{\mathbf{i}} + 2\hat{\mathbf{j}} + 3\hat{\mathbf{k}}$ and $\nabla F = 2x\hat{\mathbf{i}} + 2y\hat{\mathbf{j}} + 2z\hat{\mathbf{k}}$, we need to solve the equations

$$\begin{aligned}
2x &= \lambda \\
2y &= 2\lambda \\
2z &= 3\lambda \\
x^2 + y^2 + z^2 &= 1.
\end{aligned}$$

Plugging the first three equations into the fourth yields $(\lambda/2)^2 + \lambda^2 + (3\lambda/2)^2 = 1$, hence $\lambda = 2/\sqrt{14}$ and $(x, y, z) = (1/\sqrt{14}, 2/\sqrt{14}, 3/\sqrt{14})$. This gives the location of the maximum. The value there is $x + 2y + 3z = 14/\sqrt{14} = \sqrt{14}$.

Exercise 8.2. *How many equations will there be if you use Lagrange multipliers to maximize $g(x, y, z, w) := xyzw - 2x(y + z)$ over the unit sphere in four dimesion? Write the first and last one. (Don't try to solve them!)*

The method of Lagrange multipliers gives you a method of computing equations that specify all critical points on a hypersurface. Unfortunately, there is no general method for solving such equations. When the equations are linear, you do know how to solve them. Here's an example with linear equations.

Example 8.4.
PROBLEM: Maximize $g(x, y, z) := xy + 2xz + 3yz$ subject to $x + y + z \leq 1$ and $x, y, z \geq 0$.
SOLUTION: You have seen these constraints before: the tetrahedron in Figure 31. The objective function g is increasing in each coordinate, hence we only need to consider the upper face $F(x, y, z) := x + y + z = 1$ and its boundary. The interesting computation is the interior point of the upper face. We find it with Lagrange multipliers. Computing $\nabla g = (y + 2z)\hat{\mathbf{i}} + (x + 3z)\hat{\mathbf{j}} + (2x + 3y)\hat{\mathbf{k}}$ as well as $\nabla F = \hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}}$, we have the equations

$$
\begin{aligned}
y + 2z &= \lambda \\
x + 3z &= \lambda \\
2x + 3y &= \lambda \\
x + y + z &= 1 \, .
\end{aligned}
$$

We write these in the form

$$
\begin{bmatrix}
0 & 1 & 2 & -1 \\
1 & 0 & 3 & -1 \\
2 & 3 & 0 & -1 \\
1 & 1 & 1 & 0
\end{bmatrix}
\begin{bmatrix}
x \\ y \\ z \\ \lambda
\end{bmatrix}
=
\begin{bmatrix}
0 \\ 0 \\ 0 \\ 1
\end{bmatrix} .
$$

Letting M denote this 4×4 matrix and inverting gives $M^{-1} = \begin{bmatrix} -\frac{3}{4} & \frac{1}{2} & \frac{1}{4} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & 0 & -\frac{1}{4} & \frac{1}{2} \\ 0 & -\frac{1}{2} & -\frac{1}{2} & \frac{3}{2} \end{bmatrix}$.

(We could also have row-reduced.)

Multiplying both sides by M^{-1} gives $(x, y, z, \lambda)^T = M^{-1}\mathbf{e}_4$, which is the last column of M^{-1}, that is, $x = 0, y = 1/2, z = 1/2, \lambda = 3/2$. We need to check whether the critical point we found is actually on the upper triangular face, where $x + y + z = 1$ in the positive octant. In fact it's not. It is on the boundary of this face, on a line segment parametrized by $\{(0, t, 1 - t) : 0 \leq t \leq 1$. On this segment, $g(x, y, z) = g(0, t, 1 - t) = 3t(1 - t)$. The maximum value of $3/4$ occurs at $t = 1/2$, and we quickly verify this beats anything on the other two bounding edges, hence g has a global maximum on R of $3/4$, occurring at $(0, 1/2, 1/2)$.

Finally, let's try a messier but more realistic example.

Example 8.5. The profit from a line of sneakers is revenue minus cost. It does not seem too controversial to model this by $g(d, p, c, a) = d(p-c) - a$ where g is the profit in dollars, d is the demand (number that will be sold) in units of pairs of sneakers, p is the price of a pair of sneakers, c is the manufacturing cost of a pair of sneakers (not including the fixed costs that don't change with the number of sneakers, so don't affect where the maximum lies), and a is the advertising budget in dollars. The demand for sneakers d is not a constant. It depends on the price and the advertising budget. The dependence is probably quite complicated and must be worked out by the Marketing Deparment from past sales figures, similiarly to the way the indifference curves were plotted in Figure 11.

To get a feel for things, let's try a reasonably simple dependence of demand on price and advertising that captures the flavor of typical data. Say, for example, that the dependence of demand on price and advertising is captured by

$$F(d, p, a) := \log_{10}\left(\frac{a + 10^6}{a}d\right) + k(p - c) - 5 = 0$$

where c is the fixed cost of production per sneaker and the constant k measures the responsivity of demand to cost. The model includes a diminishing return from advertising once the budget is in the hundreds of thousands, an exponential decay in demand once the price is above cost, and a maximum demand (at cost, given sufficient advertising) of $100,000$ pairs of sneakers.

Exercise 8.3. *What is $\lim_{a \to \infty} d$ when $p = c$?*

Let us assume that the cost c is fixed at \$20 and attempt to find critical points. We set $\nabla g = \lambda \nabla F$ to obtain

$$\begin{bmatrix} p - 20 \\ d \\ -1 \end{bmatrix} = \lambda \begin{bmatrix} \frac{1}{d \ln 10} \\ \frac{1}{20} \\ -\frac{10^5}{a(a+10^5)\ln 10} \end{bmatrix}$$

as well as $F = 0$. This is too complicated to solve, but easy to approximate numerically. We tried it with $c = \$20$ and $k = (1/20)\$^{-1}$, obtaining $a = 78755, d = 16208, p := 28.69$ and $\lambda = 324,160$. Thus, in this model it was better to keep the price around 40% over cost (that would be \$28) and spend a little under \$80,000 on advertising (about \$5 per sneaker pair, as it turns out). The demand will be a little over $16,000$ pairs, and the profit a little under \$4 per pair, from a mark up of \$8.68 per pair and an advertising cost of a little under \$5 per pair.

8.2 Multiple Constraints

Another nice feature of the Langrange multiplier technique is that it generalizes without much change to the situation where we have more than one constraint, which is more representative of what real problems tend to look like.

Let's start with an example. Three dimensions is the smallest number of dimensions in which this situation really makes sense, so suppose we're trying to maximize $g(x, y, z) = x - y/2 + z$ on this purple curve:
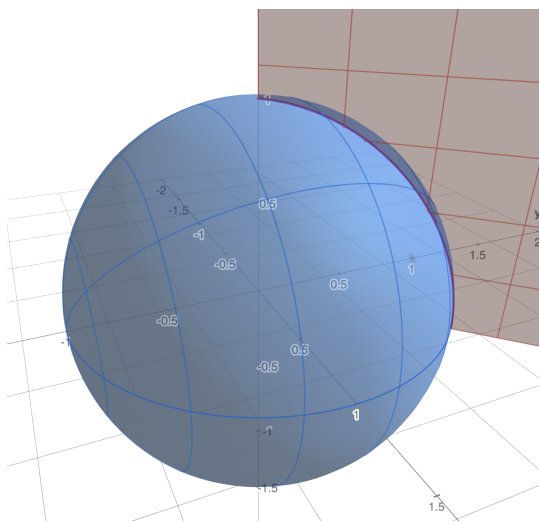

That is, the arc is on the unit sphere, but also satisfies $x = y$. We can express this with two constaints, $x^2 + y^2 + z^2 - 1 = 0$ and $x - y = 0$, as well as the inequalities $x \geq 0$, $y \geq 0$, $z \geq 0$. (In this case we could also substitute $x = y$ and reduce the number of variables, but let's stick to Lagrange multipliers so we can illustrate the method.) The two surfaces given by the individual constaints are represented in blue and red, and their intersection is the arc we're restricted to. We should give the constraints names: $F(x, y, z) = x^2 + y^2 + z^2$ and $H(x, y, z) = x - y$.

Suppose we have a maximum at some point on the constraint curve. ∇g at this point would tell us the direction to move in in order to increase g. Since we're at a maximum, moving in the direction of ∇g, or even a direction which is a little bit like ∇g, must take us off the curve. That is, the direction of the curve must be perpendicular to ∇g.

The catch is that there are now a lot of directions perpendicular to ∇g. At each point on

the curve, the tangent line to the curve gives the direction of the curve; the perpendicular directions are exactly the *plane which is normal* to the tangent line:
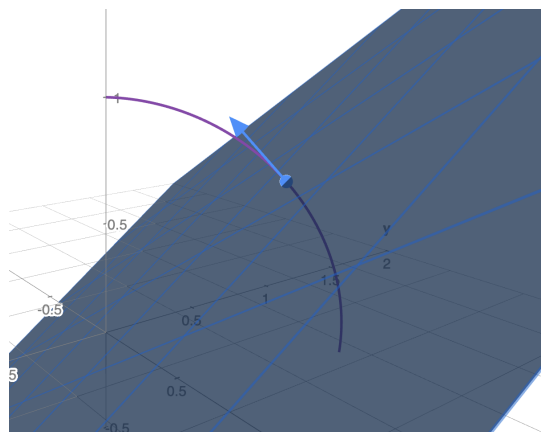
With one constraint, we knew exactly what direction ∇g had to be, and it turned out to the direction parallel to ∇F. Here, there's a whole plane of directions ∇g can be in. We do, however, know something about this plane: it contains both ∇F and ∇H. Two non-parallel vectors are enough to determine a plane[11] : we know that every vector perpendicular to the curve, and especially ∇g, is a linear combination $\lambda \nabla F + \mu \nabla H$.

In our case, that means that in addition to the equations $x^2 + y^2 + z^2 - 1 = 0 = 0$ and $x - y = 0$ from the constraints, we should have an extra vector equation

$$(1, -1/2, 1) = \nabla g = \lambda(2x, 2y, 2z) + \mu(1, -1, 0).$$

Breaking these out into individual equations, we have

$$1 = 2\lambda x + \mu \quad -1/2 = 2\lambda y - \mu \quad 1 = 2\lambda z \quad x^2 + y^2 + z^2 - 1 = 0 \quad x - y = 0.$$

This is five equations in five variables, so we expect it to have a small number of solutions. The last equation tells us that $x = y$. Adding the first two equations gives $1/2 = 4\lambda x$. Since $x = y$, we have $z^2 = 1 - 2x^2$. Recall that we are looking at the region $z \geq 0$, so we actually

---

[11] Technically, there's an issue here: it could be that ∇F and ∇H happen to be parallel themselves, in which case they don't determine a plane. In that case our constraints are actually redundant—we don't really have two different constraints, at least at that point. Like the case where ∇F is actually 0, this happens rarely enough that we won't worry about it.

have $z = \sqrt{1 - 2x^2}$. We also have $\lambda = 1/2z$, so $\lambda = \frac{1}{2\sqrt{1-2x^2}}$. Substituting this in gives

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

This looks ugly, but actually it leads to a quadratic we can solve: multiply both sides by the denominator and square, so we have $\frac{1-2x^2}{4} = 4x^2$, so $x^2 = 1/18$; again, $x \geq 0$, so $x = \frac{1}{2\sqrt{3}}$. Now we can solve for the other values: $y = x = \frac{1}{2\sqrt{3}}$, and $z = \sqrt{1 - \frac{1}{9}} = \frac{2\sqrt{2}}{3}$.

As usual, we aren't guaranteed that the points we get are actually extrema, but this method works in general for finding the candidates

Theorem 8.6 (Lagrange multipliers with two constraints). *Let $g, F, H : \mathbb{R}^n \to \mathbb{R}$ be differentiable. Then any maxima or minima of g in the interior of $R := \{\mathbf{x} \mid F(\mathbf{x}) = 0 \text{ and } H(\mathbf{x}) = 0\}$ satisfies these $n + 2$ equations in $n + 2$ variables:*

$$
\begin{aligned}
\frac{\partial g}{\partial x_1} &= \lambda \frac{\partial F}{\partial x_1} + \mu \frac{\partial H}{\partial x_1} \\
\frac{\partial g}{\partial x_2} &= \lambda \frac{\partial F}{\partial x_2} + \mu \frac{\partial H}{\partial x_2} \\
\vdots \quad \vdots \quad \vdots & \\
\frac{\partial g}{\partial x_n} &= \lambda \frac{\partial F}{\partial x_n} + \mu \frac{\partial H}{\partial x_n} \\
F(\mathbf{x}) &= 0 \\
H(\mathbf{x}) &= 0 .
\end{aligned}
\tag{8.2}
$$

$\square$

Exercise 8.4. *Suppose we wish to optimize $g(x, y, z, w) = x + y + z + w$ subject to $xyzw = 0$ and $e^{x+y} = z + w$. How many equations does the method of Lagrange multipliers give us, and what are they?*

There's nothing special about two constraints here. Although we won't use it much, we might as well state a version with as many constraints as we want.

Theorem 8.7 (Lagrange multipliers with k constraints). *Let $g, F_1, \ldots, F_k : \mathbb{R}^n \to \mathbb{R}$ be differentiable. Then any maxima or minima of g in the interior of $R := \{\mathbf{x} \mid F_i(\mathbf{x}) =$*

0 *for every i}* *satisfies these* $n + k$ *equations in* $n + k$ *variables:*

$$
\begin{aligned}
\frac{\partial g}{\partial x_1} &= \sum_{i \leq k} \lambda_i \frac{\partial F_i}{\partial x_1} \\
\frac{\partial g}{\partial x_2} &= \sum_{i \leq k} \lambda_i \frac{\partial F_i}{\partial x_2} \\
\vdots \quad \vdots \quad \vdots & \qquad\qquad\qquad\qquad (8.3) \\
\frac{\partial g}{\partial x_n} &= \sum_{i \leq k} \lambda_n \frac{\partial F_i}{\partial x_n} \\
F_1(\mathbf{x}) &= 0 \\
\vdots \quad \vdots \quad \vdots & \qquad\qquad\qquad\qquad (8.4) \\
F_k(\mathbf{x}) &= 0 \,.
\end{aligned}
$$

$\square$

9   The Multivariable Second Derivative Test

We still owe you a version of the second derivative test for multivariable functions. Remember that for single variable functions, the second derivative test says that when a is a critical point of f, we can often tell what kind of critical point it is by looking at $f''(a)$: when $f''(a) > 0$, a is a local minimum, and when $f''(a) < 0$, a is a local maximum.

At a local minimum, the first derivative goes from negative to positive, so the second derivative is positive (if it's defined and non-zero): $\smile$

At a local maximum, the first derivative goes from negative to positive, so the second derivative is positive (if it's defined and non-zero): $\frown$

In order to talk about a second derivative test for multivariable functions, we need to say what the second derivative is.

Exercise 9.1. *Before going on, what do you think the analog of the second derivative will be for a function $f : \mathbb{R}^n \to \mathbb{R}$? (Hint: What is the analog of the first derivative?)*

Second Order Partial Derivatives

One answer (but not, as we'll eventually see, the only one) is to simply take two partial derivatives, possibly with respect to different values. We haven't talked about second order partial derivatives yet, but you might be able to guess how they work.

Definition 9.1. *When $f : \mathbb{R}^n \to \mathbb{R}$ is a function, we write $\frac{\partial^2 f}{\partial x_i \partial x_j}$ for the derivative $\frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right)$.*

Example 9.2. If $f(x, y) = x^2 y$ then

- $\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}(2xy) = 2y$,

- $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}(x^2) = 2x$,

- $\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}(2xy) = 2x$,

- $\frac{\partial^2 f}{\partial y^2} = \frac{\partial}{\partial y}(x^2) = 0$.

Exercise 9.2. *If $g(x, y, z) = xyz$, what is $\frac{\partial^2 g}{\partial y \partial z}$?*

You might notice that when we calculated $\frac{\partial^2 f}{\partial x \partial y}$ and $\frac{\partial^2 f}{\partial y \partial x}$, even though we took the partial derivatives in a different order, we got the same answer. This isn't a coincidence.

Theorem 9.3 (Clairaut's Theorem). *If $f : \mathbb{R}^n \to \mathbb{R}$ is a function with all second partial derivatives defined and continuous near (a, b) then $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$.*

In words, as long as f is nice enough, the order we take the partial derivatives in doesn't matter. In this class, we won't worry about the sort of functions where this goes wrong, so we'll be able to assume that we can always take multiple partial derivatives in whatever order we want.

Sometimes this will save us effort, because we want to know both $\frac{\partial^2 f}{\partial x \partial y}$ and $\frac{\partial^2 f}{\partial y \partial x}$, and we can calculate just one of them. Sometimes choosing the right order will make the calculation easier.

Exercise 9.3. *Let $f(x, y) = e^{y^2 \sin y} + x^2 y$. Find $\frac{\partial^2 f}{\partial x \partial y}$.*

Because these multiple partial derivatives get messy to write out, there's a common shorthand: we write $f_{xx} := \partial^2 f / \partial x \partial x$, $f_{xy} := \partial^2 f / \partial x \partial y$ and $f_{yy} := \partial^2 f / \partial y \partial y$. Then Clairaut's Theorem says that, for the sorts of functions we'll consider this semester, $f_{xy} = f_{yx}$.

Although they don't come up that much in this class, nothing stops us from continuing this out to higher derivatives: if we have a function $f(x, y, z, w)$, we can write either $\frac{\partial^4 f}{\partial x^2 \partial y \partial w}$ or f_{xxyw} for the result of taking the partial derivative of f four times—twice with respect to x, once with respect to y, and once with respect to w. As long as f is a reasonable function, Clairaut's Theorem still applies, and it doesn't matter which order with take the partial derivatives in.

Exercise 9.4. *Let $f(x, y) = e^{y^2 \sin y} + x^2 y$. Find f_{xxy}.*

Deriving the Second Derivative Test

Suppose we have a function $f : \mathbb{R}^2 \to \mathbb{R}$ with a critical point at $\mathbf{a}$. Imagine starting on the surface of $z = f(\mathbf{v})$, starting at $(\mathbf{a}, f(\mathbf{a})$ and traveling along the surface so that the shadow in $\mathbb{R}^2$ moves in a straight line in the direction of a unit vector $\mathbf{v}$. Our position at time t is

$(\mathbf{a} + t\mathbf{v}, f(\mathbf{a} + t\mathbf{v}))$. We denote this by the position function $g_\mathbf{v}(t) := f(\mathbf{a} + t\mathbf{v})$. The curve parametrized by $g_\mathbf{v} : [0, \infty) \to \mathbb{R}^3$ lies on on the surface given by $z = f(x, y)$.

If $\mathbf{a}$ is a local minimum of f then, no matter which direction $\mathbf{v}$ is pointing, 0 must be a local minimum of $g_\mathbf{v}$. The picture below shows a graph with a local minimum and curves $g_\mathbf{v}$ for three different vectors $\mathbf{v}$ going through the minimum, so we can see that each curve $g_\mathbf{v}$ has a local minimum.
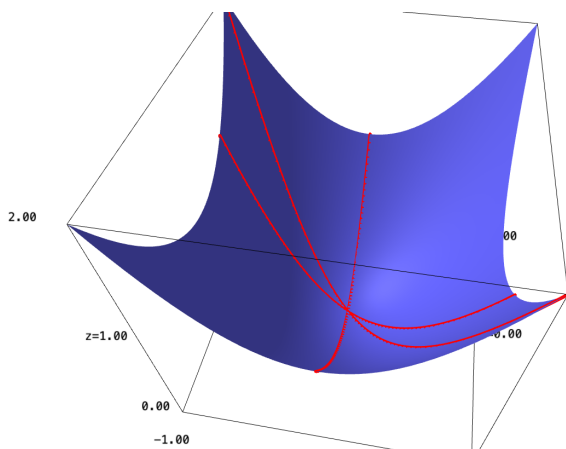

Similarly, if $\mathbf{a}$ is actually a local maximum of f then 0 must be a local maximum of $g_\mathbf{v}$, no matter which direction $\mathbf{v}$ is pointing.
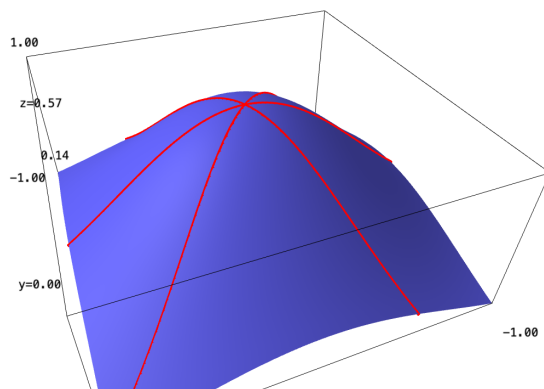
The new phenomenon in two or more dimensions, a saddle point, is what happens when $g_\mathbf{v}$ is sometimes a local maximum and sometimes a local minimum, depending on which direction we look in. In the picture below, we see that along one axis, $g_\mathbf{v}$ has a local minimum at the

critical point, along the other axis, $g_{\mathbf{v}}$ has a local maximum, and in a third direction $\mathbf{v}$, $g_{\mathbf{v}}$ is actually flat (though it's hard to tell from just the picture that it's exactly flat).
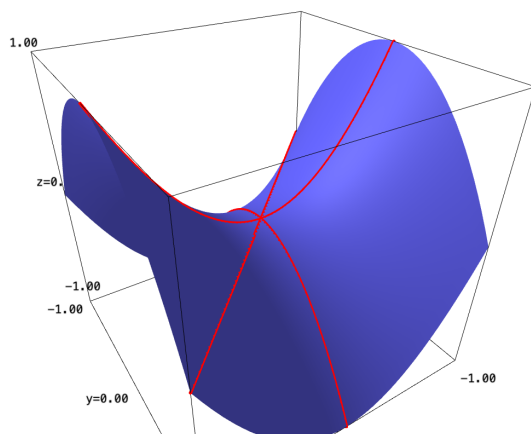
Exercise 9.5. *Consider the function $f(x, y) = x^2 + y^2$. If $\mathbf{v}$ is the unit vector in the direction $\pi/3$ radians counter-clockwise from the x-axis, what is the function $g_{\mathbf{v}}$ as above? Using the first and second derivative tests for single variable functions, confirm that 0 really is a local minimum of $g_{\mathbf{v}}$.*

The function $g_{\mathbf{v}}$ is a single variable function, so we could hope that the second derivative test works for it.

Theorem 9.4. *If f is twice continuously differentiable in a neighborhood of the point $\mathbf{a}$ and $\mathbf{a}$ is a critical point of f, then:*

- *if $g_{\mathbf{v}}''(\mathbf{a}) > 0$ for all unit vectors $\mathbf{v}$ then $\mathbf{a}$ is a local minimum of f, and*

- *if $g_{\mathbf{v}}''(\mathbf{a}) < 0$ for all unit vectors $\mathbf{v}$ then $\mathbf{a}$ is a local maximum of f,*

- *if there is a unit vector $\mathbf{v}$ with $g_{\mathbf{v}}'(\mathbf{a}) > 0$ and a different unit vector $\mathbf{w}$ with $g_{\mathbf{w}}''(\mathbf{a}) > 0$ then $\mathbf{a}$ is a saddle point of f.*

This theorem doesn't cover all possible cases—for instance, if $g_{\mathbf{v}}''(\mathbf{a}) \geq 0$ for every value of $\mathbf{v}$, but there is at least one direction $\mathbf{v}$ where $g_{\mathbf{v}}''(\mathbf{a}) = 0$, then none of the cases apply and the second derivative test is inconclusive. For example, on the "half-pipe" $z = x^2$, every point on the line $x = 0$ is a critical point, and the second derivative is positive in most

directions. But if we stay perfectly on the line $x = 0$, the surface is flat—for instance, $g''_{(0,1)}(0,0) = 0$—so this isn't covered by the theorem.
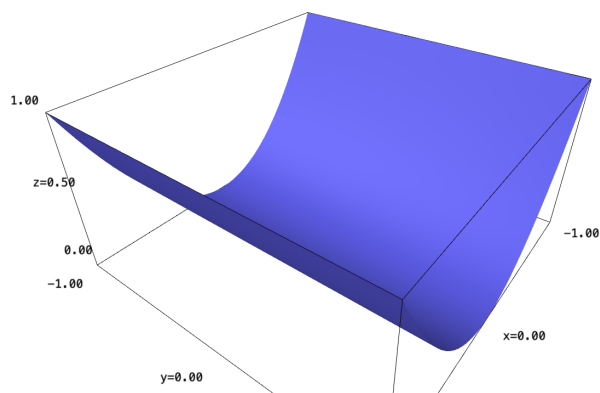
The origin is a local minimum, but it's a wide flat local minimum because the second (and third!) derivatives are all 0, and the second derivative test doesn't help us tell what's going on. But these sorts of examples are unusual, so the second derivative test will usually apply.

This theorem also isn't very useful in practice, since to check what kind of critical point $\mathbf{a}$ is, we would have to check every possible direction $\mathbf{v}$. Fortunately, there's a way to gather all this information up in one place.

Definition 9.5. *The **Hessian matrix**, is the $n \times n$ matrix*

$$H := \left[\frac{\partial^2 f}{\partial x_i \partial x_j} \right]_{i,j \leq n}$$

is the matrix of second partial derivatives.

Note that the Hessian matrix is really a matrix of functions, since each of the second partial derivatives is really a function. However we will always be concerned with the values of these second partial derivatives at a critical point.

Exercise 9.6. *Let $f(x,y) = x^2 + y^2$. What is H? What is the value of H at the unique critical point of this function?*

Really, we could think of the Hessian as the more complete analog of the second derivative for multivariable functions, in the same way the gradient was the more complete analog of the first derivative.

141

We are not going to discuss multivariable Taylor series too much in this course. However they are useful in understanding how multivariable functions behave near critical points, so we will quote without proof what quadtratic Taylor polynomials look like. Recall that in one variable, the quadratic case of Taylor's theorem says that $P_2(x) := f(a) + f'(a)(x - a) + f''(a)(x - a)^2/2$ approximates f with a remainder that is $\ll |x - a|^2$. See if you can guess the multivariable generalization before reading on.

Theorem 9.6 (quadratic approximation in more than one variable). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable in a neighborhood of the point $\mathbf{a} \in \mathbb{R}^n$ where $\mathbf{a}$ is a critical point of f. Let H be the Hessian matrix at $\mathbf{a}$. Define*

$$P_2(\mathbf{x}) := f(\mathbf{a}) + \nabla f(\mathbf{a})^T (\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^T H (\mathbf{x} - \mathbf{a}).$$

Then $|P_2(\mathbf{x}) - f(\mathbf{x})| \ll |\mathbf{x} - \mathbf{a}|^2$.

Example 9.7. Let $f(x, y) = 3 + x + xy + x^2 + x^3 + xy^2$. $(0, 0)$ is a critical point. Find $P_2(\mathbf{x})$.

The partial derivatives are $f_x = 1 + y + 2x + 3x^2 + y^2$, $f_y = x + 2xy$, $f_{xx} = 2 + 6x$, $f_{xy} = f_{yx} = 1 + 2y$, $f_{yy} = 2x$. $(\nabla f(0))^T = [1\ 0]$. The Hessian matrix at $(0, 0)$ is $\begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix}$.

Therefore

$$P_2(\mathbf{x}) = f(0, 0) + [1\ 0] \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \frac{1}{2} [x\ y] \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

$$= 3 + x + \frac{1}{2}(2x^2 + xy + xy)$$

$$= 3 + x + x^2 + xy.$$

In this case, f was actually a polynomial, and $P_2(\mathbf{x})$ is exactly the quadratic part. (The xy term counts as quadratic.)

The error term $|P_2(\mathbf{x}) - f(\mathbf{x})|$ is $|x^3 + xy^2|$. When $|\mathbf{x}|$ is very small—that is, when our point is close to the origin—$|x^3 + xy^2|$ will be much smaller than $|\mathbf{x}|^2 = x^2 + y^2$.

This means that for most functions, if we can tell whether $\frac{1}{2}(\mathbf{x} - \mathbf{a})^T H (\mathbf{x} - \mathbf{a})$ has a local minimum or local maximum, we can also figure out what happens near $f(\mathbf{a})$. (As with the second derivative test from single variable calculus, it's possible to have critical points

where both the first and second derivatives are 0, in which case the second derivative test is inconclusive.)

The following theorem answers that question for us, though it might look a little mysterious at the moment.

Theorem 9.8 (multivariable second derivative test). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable in a neighborhood of the point $\mathbf{a} \in \mathbb{R}^n$ where $\mathbf{a}$ is a critical point of f. Then:*

- *if, for every λ such that $H - \lambda I$ is not invertible, $\lambda > 0$, then f has a local minimum at $\mathbf{a}$,*

- *if, for every λ such that $H - \lambda I$ is not invertible, $\lambda < 0$, then f has a local maximum at $\mathbf{a}$,*

- *if there is a $\lambda < 0$ so that $H - \lambda I$ is not invertible and also a $\mu > 0$ so that $H - \mu I$ is invertible, then f has a saddle point at $\mathbf{a}$.*

Remember that I is the identity matrix, so λI is the matrix with λ's along the diagonal and 0's everywhere else.

Example 9.9. When $f(x, y, z) = x^2 + xy + 2y^2 + xz + 2z^2$, what are the critical points and are they local minima, local maxima, or saddle points?

The partial derivatives are $\nabla f = \begin{bmatrix} 2x + y + z \\ x + 4y \\ x + 4z \end{bmatrix}$. If $\nabla f = \mathbf{0}$ then the equations $x + 4y = 0$ and $x + 4z = 0$ give us $y = -x/4$, $z = -x/4$, so plugging these into $2x + y + z = 0$ gives us

$$2x - \frac{x}{4} - \frac{x}{4} = 0.$$

The only way this can happen is when $x = 0$, so also $y = z = 0$, so the only critical point is at the origin.

The Hessian is $H = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 4 & 0 \\ 1 & 0 & 4 \end{bmatrix}$. To find the determinant of $H - \lambda I = \begin{bmatrix} 1 - \lambda & 1 & 1 \\ 1 & 4 - \lambda & 0 \\ 1 & 0 & 4 - \lambda \end{bmatrix}$, we could use row reduction, but for a 3×3 matrix it's easier to use the permutation method or simply the resulting formula (7.3). This results in

$$\det(H - \lambda I) = (4 - \lambda)\left(\lambda^2 - 5\lambda + 2\right).$$

143

This is 0 when either $4 - \lambda = 0$ or when $\lambda^2 - 5\lambda + 2 = 0$. We can't factor $\lambda^2 - 5\lambda + 2$, but we can use the quadratic formula: this is 0 when $\lambda = \frac{5 \pm \sqrt{25-8}}{2} = \frac{5 \pm \sqrt{17}}{2}$. We can tell that both these roots are positive (we know that $\sqrt{17}$ is a little bit more than 4, and definitely less than 5, since $16 < 17 < 25$).

In a 3×3 matrix, we saw that $\det(H - \lambda I)$, which is called the **characteristic polynomial**, was a cubic. We found all three solutions. Our conclusion is that the determinant of $H - \lambda I$ is 0 when λ is either 4 or $\frac{5 \pm \sqrt{17}}{2}$. Since all three of these values are positive, the critical point is a local minimium.

What happens when we do this for a general $n \times n$ Hessian matrix? Using the permutation method, you can see that the characteristic polynomial will be degree n: there is only one permutation that gives you a λ^n term, namely the one that consists of all the diagonal elements. Not all degree-n polynomials have n real roots, but the Hessian is a symmetric matrix and characteristic polynomials of symmetric matrices have real roots (this linear algebra fact is not so easy to prove so we won't try to argue it here).

Where did λ come from, and what does $H - \lambda I$ being invertible have to do with anything? As it turns out, there's an optimization problem involved. We'll discuss where this theorem comes from in class. (In fact, the λ in the theorem will turn out to be same λ as in a Lagrange multiplier problem, though as far as I know, it's a coincidence that they're both traditionally written using the same Greek letter.)

For now, we'll just note that this is a theorem we can use, because we know how to tell when $H - \lambda I$ is not invertible: when $\det(H - \lambda I) = 0$. Since the determinant is a polynomial, this gives us a systematic way to figuring out what kind of critical point we're dealing with: write out $\det(H - \lambda I)$ as a polynomial with the variable λ, find the roots of the polynomial, and check if they're all positive or all negative.

Exercise 9.7. *Use this method to determine what kind of critical point $f(x, y) = x^2 - y^2$ has at $\mathbf{0}$.*

Further facts, FYI

These values of λ so that $H - \lambda I$ fails to be invertible turn out to be one of the fundamental ideas in linear algebra. They're called the **eigenvalues** (which is German for, roughly, "characteristic values"), and they'll play a central role if you decide to take a full linear

algebra course. A square matrix that fails to be invertible maps some nonzero vector to zero. A vector $\mathbf{v}$ such that $(H - \lambda I)\mathbf{v} = 0$ satisfies $H\mathbf{v} = \lambda\mathbf{v}$ and is called an **eigenvector**. We will look a little more closely at these in class as well.

Unless there are doubled roots, there will be n real eigenvalues and the corresponding eigenvectors will be a basis for $\mathbb{R}^n$. Even if there are doubled roots, $\mathbb{R}^n$ will have a basis of eigenvectors. You are not responsible on an exam for knowing the facts in this optional two paragraphs, but they might come in handy on the worksheets or homework.

<u>Please do learn the terminology though.</u> The terms characteristic polynomial, eigenvalue and eigenvector appear on the vocabulary for this week. The term "characteristic polynomial" is used for $\det(M - \lambda I)$ for any matrix M, not just for the Hessian matrix of a function.

10 Introduction to differential equations

10.1 Modeling with differential equations

A differential equation is an equation, involving derivatives, in which the quantity that you must solve for in order to make the equation true is an unknown *function*. A classical example is Malthusian population growth. This simple population growth model (due to Malthus) postulates that the (instantaneous) rate of growth of a population is proportional to its present size. If we let $A(t)$ denote the size of the population at time t, then the equation representing this is

$$\frac{dA}{dt} = kA(t).\tag{10.1}$$

Here k is a constant of proportionality. It is important that you understand its units! To make the equation work out, k must have units of inverse time. The value $k = 0.03\text{years}^{-1}$ for example, would mean that if you measure time in years, and the present population is one million, then the instantaneous growth rate would be 30,000 people per year.

A *solution* to a differential equation such as (10.1) is any function $A(t)$ that makes the equation true. Typically there will be more than one such equation. For example, the general solution to $A'(t) = 0.03A(t)$ is $Ce^{0.03t}$, where C is any real number.

Exercise 10.1. *If $A(t) = Ce^{0.03t}$, what is $A'(t)$? What is $kA(t)$ when $k = 0.03$?*

We are going to postpone until Unit 11 the business finding nice solutions such as this one. In Unit 10 we will concentrate instead on understanding pretty much everything else: how to set up a differential equation, what it means, and what the solution will look like qualitatively.

Verifying that you have found a solution is a lot easier than finding a solution. To check that Ce^{kt} solves (10.1), just evaluate both sides when $A(t) = Ce^{kt}$. The left side is the derivative of $A(t)$ which is Cke^{kt}. The right side is k times $A(t)$ which is Cke^{kt}. They match – whoopee!

The reason you might expect there to be many solutions to an equation such as (10.1) is that it is an equation of evolution. Once you know where you start, everything else should be deterimined, but there is nothing in the equation that tells you where you start. A differential equation together with a value at a certain time is called an *initial value problem*. For example, $A'(t) = 0.03A(t);\ A(0) = 1,000,000$ is an initial value problem.

146

Exercise 10.2. *Solve the initial value problem $A'(t) = 0.03A(t)$, $A(0) = 1,000,000$. That is, write down an explicit function of t satisfying both of those equations.*

Standard form for first-order differential equations

A differential equation could be arbitrarily complicated. The equation

$$f(x)\sqrt[3]{1 + \left(\frac{df}{dx}\right)^2} - \ln f(x + a) + \exp\left(\frac{d^2 f}{dx^2}\right) = \arctan(x + f(x))$$

is a differential equation, but way too complicated for us to have any hope of figuring out what functions f satisfy it. Note the appearance of a second derivative, the square of the first derivative, a big messy cube root and the appearance of the unknown function f as the argument of the arctangent. We will stick to much simpler classes of differential equations, starting with **first order differential equations in standard form**. This is the form

$$\frac{dy}{dx} = F(x, y). \tag{10.2}$$

Be sure you understand what this means. The unknown function in this case is the function $y(x)$. We call y the "dependent variable" and x the "independent variable". The function F is an abstraction representing that the right-hand side is some function of x and y. Here are some examples:

$$\frac{dy}{dx} = x^{-2}$$

$$\frac{dy}{dx} = ky$$

$$\frac{dy}{dx} = x - y$$

$$\frac{dy}{dx} = \sqrt[3]{y + e^x}$$

Even though you don't yet know much about differential equations, there is a lot you can say looking at these examples. (i) It is possible that $F(x, y)$ will be a function of just x, as in the first equation. This means that $y(x)$ is just the integral of $F(x) = x^{-2}$. So you can already solve this one: $y(x) = -1/x + C$. (ii) It is possible that $F(x, y)$ will be a function

147

of just y, as is the case in the second equation. In that case it's not so obvious how to solve it, but you actually already know the solution to this particular equation because it is just (10.1). (*iii*) In general, a first order equation in standard form can be simple, like the third one, or complicated, like the fourth. The simple ones are usually exactly solvable (we will solve the third one in Unit 11.3), and the more complicated ones often are not. The fourth equation, while not exactly solvable, will still yield plenty of information; this is what this unit is mostly about.

A point of notation: should we use y' or dy/dx? Both mean the same thing, but dy/dx is clearer because it tells you which is the dependent variable. If you wrote $y' = -ce^{tx}$ it would be unclear whether t or x was the independent variable (or c could be too, but we never choose c for a variable name because it sounds too much like it should be a constant). We will use both notations, as both are common in real life. One more point: when we want to emphasize that the unknown variable is a function, we sometimes use a name like f or g instead of y. For example, $f' = -xf$ is a differential equation. (It is understood that the indpendent variable is x). The most common independent variable names are x and t because t usually represents time.

Integral equations

Certain equations with integrals in them can be made into differential equations by differentiating both sides. (This uses the Fundamental Theorem of Caculus.) For example the integral equation

$$f(t) = 12 - \int_5^t 3f(s)\,ds$$

can be differentiated with respect to t to obtain

$$f'(t) = -3f(t)\,. \tag{10.3}$$

Exercise 10.3. *What happened to the* 5 *when we differentiated? Why?*

The integral equation has only one solution but (10.3) has many. This means that there was initial value data in the integral equation that we forgot to include in the differential equation. Can you spot it? Really we should have translated the integral equation into the initial value problem:

$$f'(t) = -3f(t) \ ; \ \ f(5) = 12\,.$$

148

Exercise 10.4 (+). *How did we use the integral equation to figure out that $f(5) = 12$?*

Exercise 10.5. *Convert the integral equation $f(t) = 3 + \int_2^t f(s)^2 \, ds$ into an initial value problem by differentiating both sides.*

Exercise 10.6. *Convert the initial value problem $f'(t) = f(t)$, $f(2) = 3$ into an integral equation by integrating both sides.*

Time constants

We will be studying many differential equations, of which the one in Equation (10.1) is the most basic and deserves a few more words. The constant k is a constant of proportionality. It is important that you understand its units! The units of A are population and the units of A' are population per time, so to give both sides the same units, k must have units of inverse time. Such constants are called **time constants**. Checking the solution $A(t) = Ce^{kt}$, we notice that kt is unitless, which is as it should be when we exponentiate it.

What is the interpretation of a time constant? The most literal interpretation is that, if the value is say $k = 0.03\text{years}^{-1}$, this would mean that if you measure time in years, and the present population is one million, then the instantaneous growth rate would be 30,000 people per year. Another interpretation is to take the reciprocal of a time constant, which has units of time. For example, if $k = 0.04\text{years}^{-1}$, then $1/k = 25\,\text{years}$. This length of time has an interpretation: it is the length of time for the population to change by a factor of e (because setting $t = 1/k$ means that $e^{kt} = e^1 = e$). Instead of $t = 1/k$, we could take $t = (\ln 2)/k$, in which case $e^{kt} = 2$ and t would be the doubling time.

When $k > 0$, the solution $A(t) = Ce^{kt}$ 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, $f' = kf$.

When $k = -\ell < 0$, we have 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. 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.

To explain further why temperature *differences* occur on this list, consider an object that is at a different temperature from its surrounding environment. The function $f(t) := C + Ae^{-\ell t}$ models a quantity that is decaying exponentially fast not to zero but to C.

Exercise 10.7. *Let C be the ambient temparature. If the rate of change of temperature is supposed to be proportionals to the* difference *between the object's temperature and the ambient temperature, explain what differential equation you would expect f to solve. Then check that $f(t) := C + Ae^{-\ell t}$ solves this equation.*

Let's look at some other examples of time constants.

Exercise 10.8. *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?*

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 10.9. *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.*
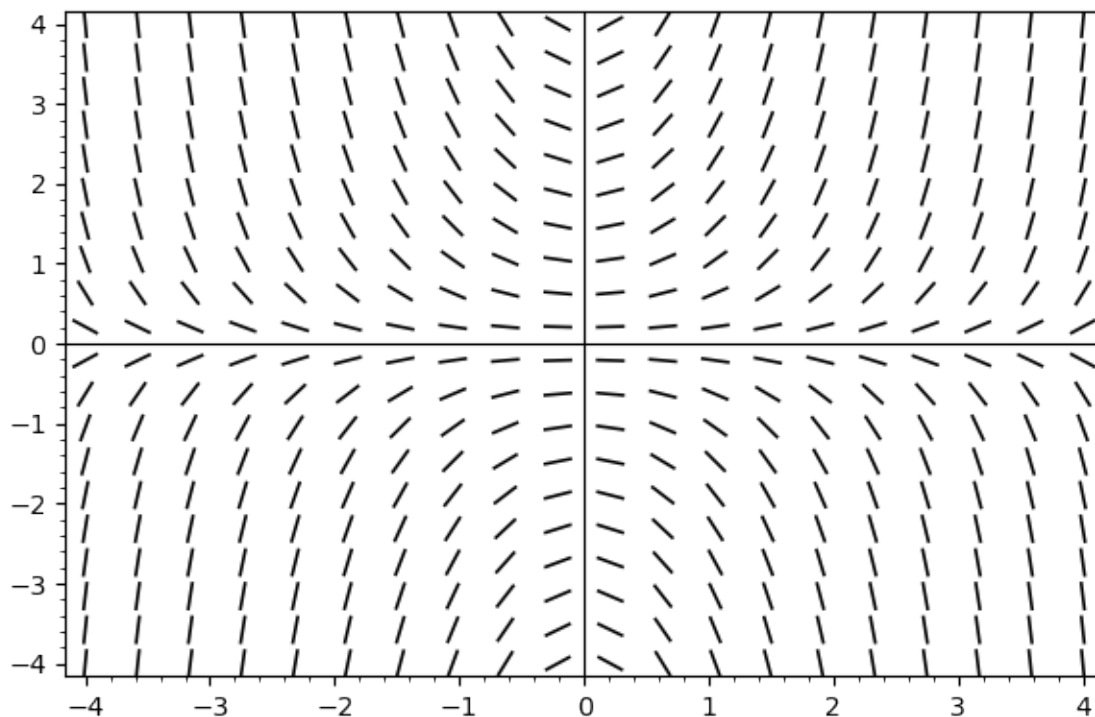
10.2 Slope fields

Slope field drawings enable you to sketch solutions to differential equations. The equation $dy/dx = f(x, y)$ tells you what the slope of the graph of y should be at any point (x, y), if indeed that point is on the graph. We make a grid of points and through every (x, y) in the grid we put a little line segment of slope $f(x, y)$.

For example, here is a slope field of the differential equation $y' = y$.
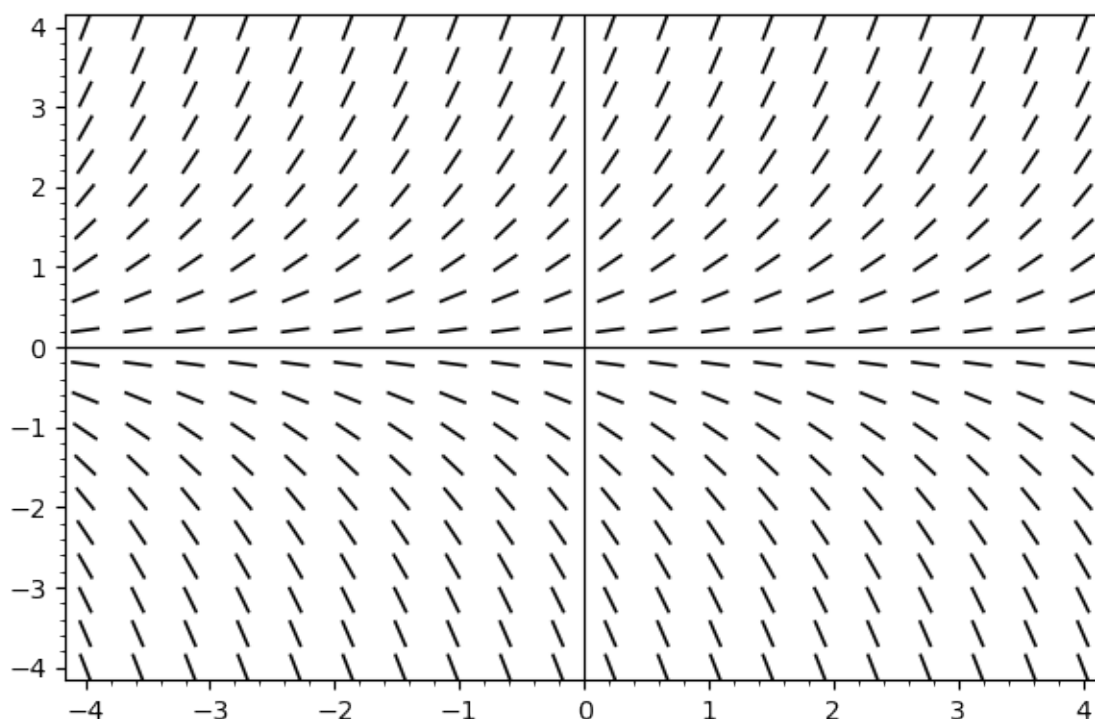
Each little line segment represents a solution to the differential equation at that point: the line segment has the slope determined by the equation, so in this slope field, the slope y' is always equal to the y value. Along the x axis, these lines are all horizontal, because the slope is $y' = y = 0$. When $y = 1$, the line segments all have slope $y' = y = 1$, and so on. Because y' only depends on y, the slopes only depend on the y value as well. A more typical slope field might depend on both. Here's a slope field for $y' = xy$.

For instance, at the point $(2, 2)$, the function is already getting rather steep: $y' = xy = 4$, so the line segment has slope 4. On the left side, where the x value is -4, most of the lines are so steep we can't really see them clearly.

Exercise 10.10. *Plot a slope field for the differential equation $y' = y + x$ by hand. (A 4 by 4 grid—16 total line segments—is enough to get a decent picture, and won't take too long.)*

The general solution to $y' = xy$ is $y = Ce^{x^2/2}$. (We'll see how to figure this out next week, but right now we can at least check that it's right: if $y = Ce^{x^2/2}$ then the chain rule tells us that y' is indeed $Cxe^{x^2/2}$, which is xy.) Let's take a look at how some of these solutions compare to the slope field.
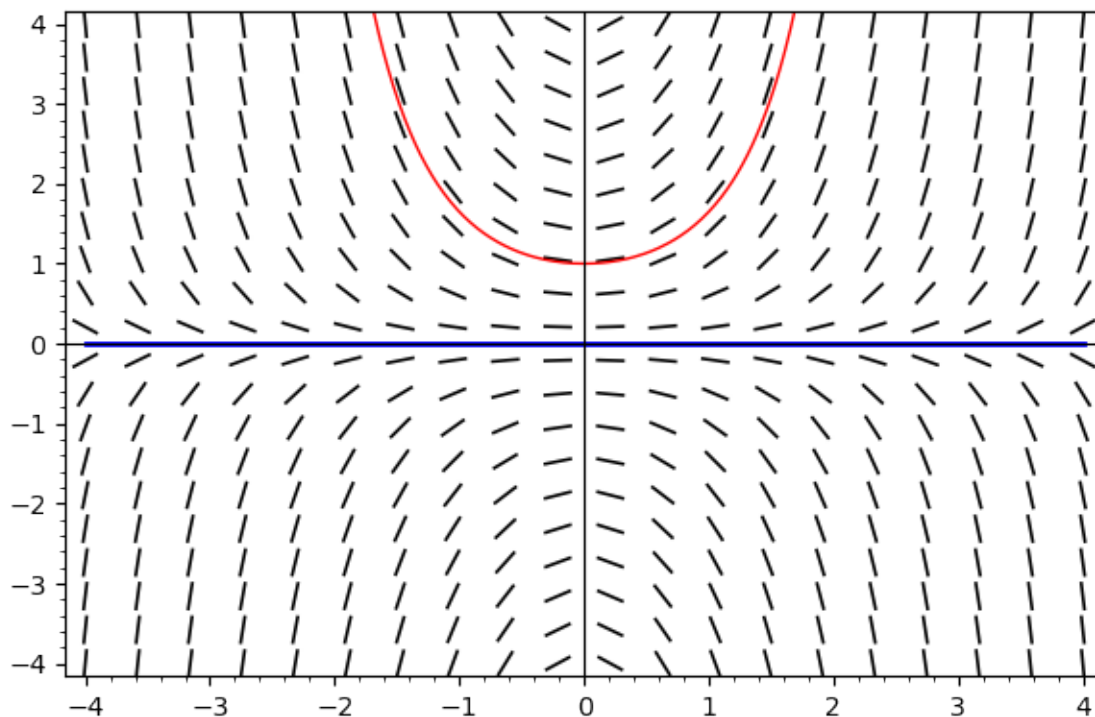
The blue line is the solution where $C = 0$—the constant function $y = 0$—and the red curve is the solution where $C = 1$, so $y = e^{x^2/2}$. The slope field is always the tangent line to our solution at that point, so we can sketch an approximation of what our solution looks like by "following the slope lines".

Slope fields are a *qualitative* approach to understanding the solution to a differential equation, meaning that you get information about the nature of the solution even when you can't find the exact solutions. For instance, we can often tell how solutions behave as x gets large are small.

What can we hope to know about a differential equation?

Our last order of business in this section is some philosophy. Sometimes we can find exact solutions to a differential equation, and next week we'll study some cases where this happens. Often, though, finding an exact solution to a differential equation is very difficult, or even impossible.

So often, instead of looking for an exact formula for a solution, we ask for *qualitative infor-mation* about a solution. Qualitative information takes many forms, but for our purposes, we'll focus on two questions: given an initial value problem $\frac{dy}{dt} = F(t, y)$, $y(0) = c$, we can ask:

- what happens for a short time - for instance, when t is small, is y going to be increasing or decreasing?

- what happens after a very long time - does y approach infinity, or negative infinity, or some other value?

Slope fields are very good at answering both these questions.

**Example 10.1.** For instance, recall the slope field for $y' = xy$ which we saw above:
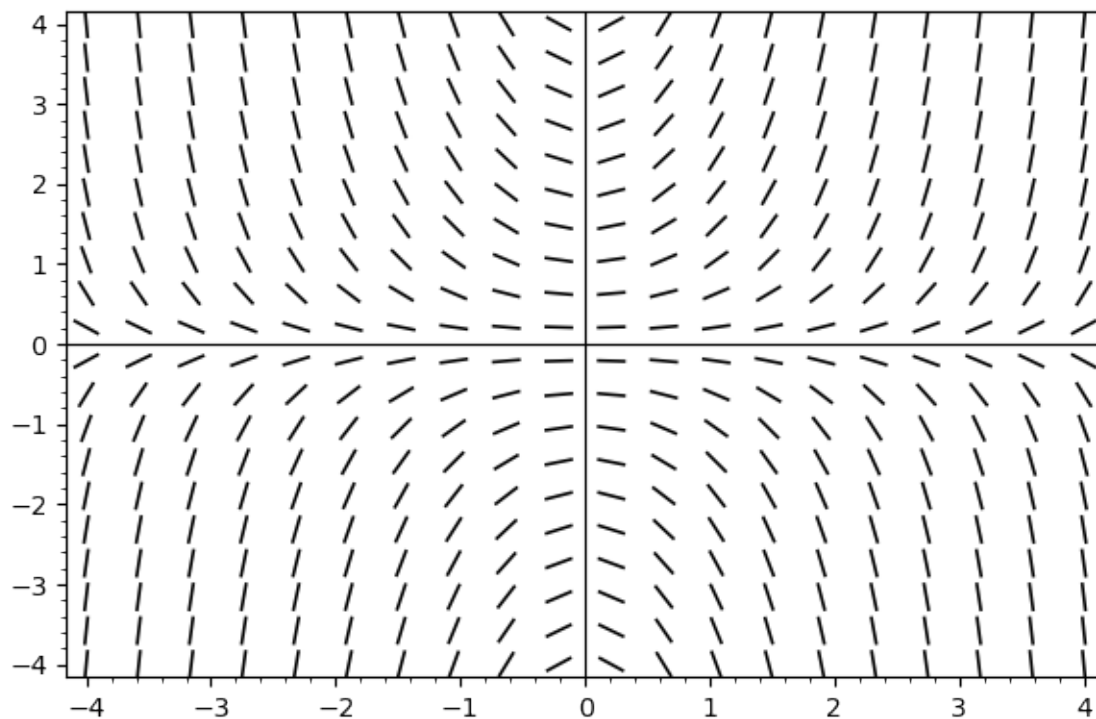
We can see that, in the short term (and, in this case, forever), if $y(0)$ is positive then y is increasing, while if $y(0)$ is negative then y is decreasing. When $y(0)$ is exactly 0, we have an *equilibrium*: $y = 0$ is a constant solution. As we'll discuss in a couple weeks, this

equilibrium is *unstable*—if there's even a small change to the value of y, the function starts diverging further away from 0.

In the long run, we can see that (except for the equilibrium solution), y is going to diverge to either infintiy or negative infinity.

Example 10.2. In a solution to the differential equation $y' = 2 - e^{y/2}$, what happens as x approaches infinity? Take a look at the slope field.
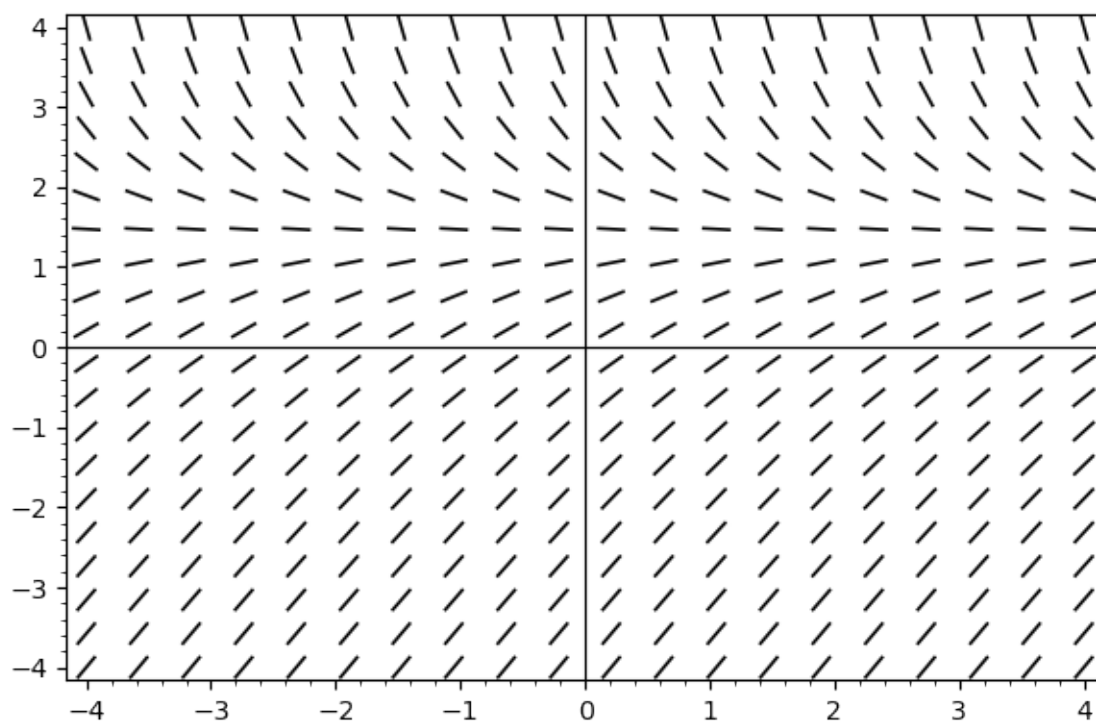
We notice that there seems to be a horizontal line which looks like a constant solution somewhere between $y = 1$ and $y = 2$, and above that line, the slope field points down towards it, while below that line the slope field points up towards it.

We can identify the line: it's where $y' = 0$, so it should be where $2 - e^{y/2} = 0$; solving for y, we get $y = 2\ln 2$. When $y(x) < 2\ln 2$, $2 - e^{y/2}$ is positive, so the function increases. When $y(x) > 2\ln 2$, then $2 - e^{y/2}$ is negative and the function therefore decreases. It seems clear from this that if the function begins below $2\ln 2$ it will steadily increase but at a lesser and lesser rate and never reach $2\ln 2$, while if the function begins above $2\ln 2$ then it will

steadily decrease but at a lesser and lesser rate and never reach $2 \ln 2$. We therefore have a very good idea what this function looks like without ever solving the equation.

Exercise 10.11. *Give a similar analysis of the slope field for $y' = y$ at the start of Section 10.2: if $y(0) = c$, which values of c will lead to y increasing or decreasing? Which values of c will lead to a solution approaching positive infinity?*

10.3 Euler iteration

Euler's method, or **Euler iteration** is a way of finding a numerical approximation for the solution of an initial value problem at some later time. In other words, for the equation $y' = F(t, y); y(t_0) = y_0$, you can compute an approximation to $y(t_1)$ when t_1 is any time greater than t_0.

The idea behind Euler iteration is that you follow the slope field for a small amount of time Δt, which is fixed at some value such as 0.5 or 0.1. Let $t_1 = t_0 + \Delta t$ be the new time and let y_1 be the approximation you get by followingthe slope field fo time Δt. In other words, $y_1 = y_0 + (\Delta t)F(t_0, y_0)$. The slope at the point (t_1, y_1) will in general be different. Follow that slope for time Δt, and repeat.

To be more formal about it, what you need in order to apply Euler's method is this:

1. A first order differential equation in stadard form $y = F(t, y)$;

2. An initial value $y(t_0) = y_0$;

3. A step size Δt;

4. A final value t_1 of the independent variable t, at which to compute $y(t_1)$.

Exercise 10.12. *Suppose $dy/dx = y$. We know this to have solution $y = Ce^x$, but forget this for now.*

  (i) *With the intial condition $y(0) = 1$, compute $y(1)$ using Euler iteration with 1 step*

 (ii) *Now do it with 2 steps*

(iii) *Now dow it with 3 steps.*

156

We will not be talking too much about existence of solutions of differential equations. Suffice it to say some conditions are needed on f so that $y' = F(t, y)$ will have a unique solution, valid for all times $t > t_0$. One way to prove it is to show that Euler's method converges as $\Delta t \to 0^+$ to the unique solution. This will not always be true and requires further assumptions on F, as we will discuss later.

Different notions of solution

Our last order of business in this section is some philosophy. You need to understand what is meant by a solution to a differential equation. The simplest differential equation is of the form $y' = f(x)$, in other words, the right-hand side does not depend on y. You already know how to solve this: $y(x) = \int f(x)\,dx$. But wait, what if it's something you can't integrate? An example of this would be $dy/dx = e^{x^2}$. We could write a solution like $y(x) = y(0) + \int_0^x e^{t^2}\,dt$, but is this really a solution? The answer is yes. Here's why.

Euler's method allows you to approximate values of the dependent variable. For example, given $y' = F(x, y)$ and $y(0) = 5$ we could use Euler's method to evaluate $y(2)$. What you need to understand is that yes we can do it but it's tedious and not all that accurate unless you use a miniscule step size. By contrast, using Riemann sums to estimate $\int_0^2 e^{t^2}\,dt$ is a piece of cake. Keep in mind the relative difference in difficulty between Riemann sums and Euler's method as we discuss three levels of possible solution to a differential equation.

1. If you can find a solution $y = f(x)$ where f has an explicit formula then that is obviously the best. Your calculator or computer (or maybe even your phone) can evaluate this, and typically you have other information associated with f such as how fast it grows, whether it has asymptotes, and so forth.

2. Next best is if you can write a formula for y that involves functions without nice names, defined as integrals of other functions. You already realize that many simple looking functions such as e^{x^2} and $\ln(x)/(1+x)$ have no simple anti-derivative. The differential equation $y' = e^{x^2}$ is trivial from a differential equations point of view (it is in the form $y' = f(x)$ which we discussed above) but still we can do no better than to write the solution as $y = \int e^{x^2}\,dx$. This is perfectly acceptable and counts as solving the equation.

3. Lastly, for the majority of equations, we can't write a solution even if allowed to use integrals of functions. In this case the best we can do is to numerically approximate particular values and to give limiting information or orders of growth for y. For example, if $y' = 2 - e^y$ then $\lim_{x \to \infty} y(x) = \ln 2$.

Exercise 10.13. *Would you use Euler's method or some kind of Riemann sums to estimate the value of $y(2)$, where y solves the initial value problem $y' = t/\sqrt{1 + t^3}$ with $y(1) = 4$?*

One last thing that Euler iteration does for us is to convince us that an initial value problem should have a solution. After all, if you look at an equation with functions and derivatives, there is no reason to believe that there is a function satisfying the equation. But Euler iteration shows you that there has to be. Just do Euler iteration and make the steps smaller and smaller; in the limit it will produce a function satisfying the differential equation. This is the basis for a theorem. The theorem is not officially part of this course but you might be interested to know what it says.

> **Theorem:** Let $F(x, y)$ be a continuous function. Then the initial value problem $y' = f(x, y), \quad y(x_0) = y_0$ has a unique solution, at least for a small amount of time (after that it might become discontinuous). This solution can be obtained by taking the limit of what you get from Euler iteration as the step sizes go to zero.

11 Exact solutions to differential equations

Exactly solving differential equations is like finding tricky integrals. You have to recognize the equation as a type for which you know a trick, then apply the trick. You will learn precisely two tricks. The first works for a class of equations called **separable equations**. The trick involves getting all the x variables on one side of the equation and the y variables on the other (hence the name "separable"). The second class is the class of **linear first order equations**. The trick there will be to find a so-called integrating factor. Before learning either of these tricks, we will get familiar with the easiest but single most important differential equation. This one is both separable and linear.

11.1 $f' = kf$ and exponential trajectories

The single most important differential equation is, as luck would have it, very easy to solve:

$$\frac{dy}{dx} = ky \tag{11.1}$$

where k is a constant. You can solve it by guessing the answer, but let's solve it a way that will generalize.

Step 1: Separate. To get the dependent variables on the left and the independent variables on the right, we divide both side by y and multiply both sides by dx:

$$\frac{dy}{y} = k\,dx\,. \tag{11.2}$$

If you are worried about whether it's OK to multiply dy/dx by dx, you're right to be concerned, because a dy without a $/dx$ is meaningless, but it works anyway and we'll show you why later.

Step 2: Integrate. The integral of dy/y is $\ln|y|+$C. What's the integral of k? You might think it's $k^2/2$ but it's not. Pay attention to the variable of integration, which is dx. The integral of $k\,dx$ is just $kx + C$. So now (combining the constants) we have

$$\ln|y| = kx + C\,. \tag{11.3}$$

Step 3: exponentiate both sides to get

$$|y| = e^{kx+C} = e^C e^{kx}\,. \tag{11.4}$$

159

Step 4: The constant e^C can be any *positive* real number. If the absolute value of y can be any positive multiple of e^{kx}—that's the same as saying y can be any multiple of e^{kx}, positive or negative. Call this multiple C_1. We write the solution in its final form:

$$y = C_1 e^{kx} \tag{11.5}$$

where C_1 can be any real number.[12]

Note: A lot of people like to call the new constant C_1 the same name as the old constant and write $y(x) = Ce^{kx}$, where C is any real number. This is a correct solution, but we don't want you changing the value of C midstream if it leads to writing incorrect equations such as $e^C e^{kx} = Ce^{kx}$.

In an application, the independent variable will be expressed in some natural unit, often time, and the function variable will have another unit such as money, volume, total quantity of something, etc. The units of dy/dx are y-units divided by x-units, so in the equation $dy/dx = ky$, the units of the constant k must be in units of "reciprocal x". For example, if x is in seconds then k is in $(\sec)^{-1}$: the name for this unit is Hertz, abbreviated Hz. In the solution $y = Ce^{kx}$, notice that the exponent is unitless (as I have previous claimed must be true of exponents). It is good to be aware of your units; when units don't make sense, you know to check your work for a mistake.

The meaning of these equations in applications is that the rate of change is proportional to the present quantity. Therefore we should think of k as a **relative rate of change**, that is a rate expressed as a fraction of the whole. One example is an interest rate. Interest, even though it produces dollars, is measured in units of inverse time because the dollars cancel: each dollar begets a similar number of future dollars. The number of dollars produced in a year by a single dollar is equal to the number of pennies produced in a year by a single penny or the number of gigabucks produced in a year by a single gigabuck.

---

[12]We explained where the solutions where C_1 is positive or negative came from, but can C_1 really be 0? We can check that $y = 0$ is a solution to the original equation by plugging it in to the differential equation and confirming that when $y = 0$, $y' = 0 = ky$. In (11.2), we divided by y, which implicitly meant we were assuming that y was not 0. Formally, should deal with that case separately—if $y = 0$, we get the constant solution $y = 0$, and if $y \neq 0$, the derivation we gave works. Combining those two cases gives the general solution (11.5).

Applications

We will look at a number of applications. There is a fundamental difference in the behavior depending on whether k is positive or negative. When k is positive, e^{kx} grows (in fact very rapidly). When k is negative, e^{kx} shrinks.

One quantity that tends to grow exponentially is wealth. Wealth can be negative if it's a debt, but both debts and assets tend to grow rather than shrink. Another is population. A characteristic of applications in which $f' = kf$ with $k > 0$ is that each unit of whatever quantity is growing contributes to the growth independently of each other unit. So for example, if you put two chunks of money in two accounts at the same interest rate, it's just like putting in one chunk that's the sum of the original two. This is reflected in the fact that the growth rate is a proportion per time, not an absolute amount per time. Likewise, a population of five times as many rabbits will produce five times as many baby rabbits.

Let's look at $y' = -ky$ more closely. It approaches zero. We can come up with a closely related equation where the limiting value is something other than zero. For example, suppose an object at temperature A is placed in a large bath at temperature B. The temperature of the object approaches the temperature of the bath at a rate proportional to the difference in tempearatures. Mathematically, if the temperature as a function of time is denoted by $y(t)$, we have

$$\frac{dy}{dt} = k(B - y).$$

Let's see two ways of solving this. One is using the same method as before. We get the chain of equations:

$$\frac{dy}{B - y} = k\,dt$$

$$-\ln|B - y| = kt + C$$

$$|B - y| = e^{-C-kt}$$

$$B - y = \pm e^{-C} e^{-kt} = C_1 e^{-kt}$$

$$y = B - C_1 e^{-kt}$$

where C_1 is any real number.

What is this saying qualitatively? If the dependent variable is approaching a target, B, at a rate proportional to the distance from the target, then the value at time t will be

the equilibrium value B plus an offset that decreases exponentially. Many physical systems behave this way (thermal equilibria, radioactive decay, resistor-capacitor networks) but also systems in social science where there is negative feedback (population approaching a natural limit, price corrections after an economic shock, etc.).

The second way to solve this is to move upward by B: algebraically, replace y by $y - B$. We will discuss this method in class.

Continuous versus annualized rates

Interest is fundamentally a continuous time phenomenon, especially when the amounts involved are so large that the interest is substantial even in a minute or a few seconds (think National Debt). Many consumers, however, are barely able to handle simple interest and haven't a clue about continous time interest. This has led to regulation where interest rates must be quoted in Annualized Percentage Yield (APY) as well as a simple growth rate.

To see how this works, suppose an asset grows at the rate of 6% per year. If $A(t)$ is value at time t, this means that $A'(t) = 0.06A(t)$ when t is measured in years; the units of the 0.06 are inverse years. After one year, an amount A_0 will grow to $A_0 e^{0.06}$. That means the gain was $A_0(e^{0.06} - 1)$. Because $e^{0.06} \approx 1.061837$, this means that the percentage growth of the asset in one year was roughly 6.1837%. In other words:

Continuous interest rate of 6% leads to annualized interest rate of 6.1837%.

We can do this for any rate. Let r be the continuous rate (in the above example $r = 0.06$) and let A be the annualized rate (in the above example 0.061837). Then r and A are related by the equations:

$$A = e^r - 1 \ ; \ r = \ln(1 + A).$$

If you write these as percentages, you have to remember to multiply and divide by 100 at the apropriate places:

$$A = 100(e^{r/100} - 1) \ ; \ r = 100\ln(1 + A/100).$$

11.2   Separable equations

The method we used to solve the exponential equation works for any differential equation where we can separate the x and y variables like a product.

Definition 11.1. *A differential equation $y' = f(x, y)$ is separable if it can be factored into the form $y' = \frac{g(x)}{h(y)}$.*

For instance, $y' = x^2 y^3 = \frac{x^2}{y^{-3}}$ is separable, as is $y' = e^{x+y}$ (because $e^{x+y} = e^x e^y = \frac{e^x}{e^{-y}}$). But $y' = x + y$ is not separable, because the right hand side cannot be written as a product of a function of x and a function of y.

Exercise 11.1. *Write down an example of a separable differential equation and an example of a non-separable differential equation. (See if you can write down examples where it's not so easy to tell which it is.)*

When we try to solve a separable equation, it's usually convenient to write the derivative using $\frac{dy}{dx}$ notation instead of y', because then we can describe the steps like cross-multiplication: once we realize we have an equation $\frac{dy}{dx} = \frac{g(x)}{h(y)}$, we "cross-multiply" to get

$$h(y)\, dy = g(x)\, dx,$$

and then we can integrate both sides to get

$$\int h(y)\, dy = \int g(x)\, dx.$$

Assuming we can find the antiderivatives, this becomes

$$H(y) = G(x) + C.$$

Note that we only need one constant. If we had $H(y) + D = G(x) + E$, we could subtract D from both sides to get $H(y) = G(x) + (E - D)$. But we definitely do need a constant after we integrate—we always expect to get a family of solutions to a differential equation, not just one.

Ideally, we can solve for y to get some equation $y(x) = H^{-1}(G(x) + C)$. Sometimes we can't solve, and the best we can do is get an implicit description of our solution.

Example 11.2. Find all solutions to $y' = x + xy$.

You can write this differential equation as $y' = x(1 + y)$ so it is of the right form with $g(x) = x$ and $h(y) = 1/(1 + y)$. Continuing, we write

$$\frac{dy}{1 + y} = x\,dx.$$

(We're assuming here that $y \neq -1$; we'll come back to this case later.)

Next we integrate both sides leading to

$$\ln|1 + y| = \frac{x^2}{2} + C\,.$$

In other words, $H(y) = \ln(1 + y)$ and $G(x) = x^2/2$. The calculus is now done. Because the natural logarithm has a known inverse function, we won't have to settle for having an H^{-1} in the final formula and instead can use the known inverse which is the exponential function. To isolate y, exponentiate both sides and subtract one. Here is the sequence of equations.

$$
\begin{aligned}
\ln|1 + y| &= \frac{x^2}{2} + C \\
|1 + y| &= ce^{x^2/2} \text{ where } c = e^C \text{ is any positive constant} \\
1 + y &= ce^{x^2/2} \text{ where } c = \pm e^C \text{ is any non-zero constant} \\
y &= -1 + ce^{x^2/2} \text{ where } c = \pm e^C \text{ is any non-zero constant.}
\end{aligned}
$$

When $y = -1$, the equation $y' = x + xy$ becomes $y' = 0$, so $y = -1$ is a constant solution; conveniently this is just the missing case of our solution above, so the general solution is

$$y = -1 + ce^{x^2/2}$$

where c can be any constant.

Notice that we originally had a constant C, and in the process of simplifying, we ended up replacing it with a different constant related to the original one; this often happens when we simplify our solutions.

Exercise 11.2. *Find all solutions to the differential equation $y' = x/y$.*

If you're a stickler for details (and most mathematicians are!), it might bother you that $h(y) \, dy = g(x) \, dx$ is not a real equation. So here's a slightly more honest explanation. What we should really do is work with the equation

$$h(y)\frac{dy}{dx} = g(x).$$

This is a perfectly good equation, and it's valid to integrate both sides with respect to x:

$$\int h(y)\frac{dy}{dx} \, dx = \int g(x) \, dx.$$

The integral of the right side is $G(x) + C$, just like we want.

But what about the left hand side? $\int h(y)\frac{dy}{dx} \, dx$ is exactly the setup for a u-substitution: we have $u = y(x)$, and then $\int h(y)\frac{dy}{dx} \, dx = \int h(u) \, du = H(u) + D = H(y) + D$. So we could say that what we're really doing is a substitution on the lefthand side, and our cross-multiplication notation is a convenient and suggestive shorthand.

11.3 Integrating factors and first order linear equations

The other special form of differential equation we'll learn how to solve is

$$y' = Q(x) - P(x)y \, .$$

This first order differential equation (only the first derivative appears) is called first-order **linear** because the expression for y' is a linear function of the dependent variable y; in this instance we don't care whether P and Q are linear functions of x, it's strictly the y variable we care about, and we use the broad sense of the word linear to include affine functions, allowing not only the $P(x) \, y$ term but also the $Q(x)$ term.

We'll first do an example where the motivation for the steps might seem a little mysterious and then explain the general method.

Example 11.3. Find the solution to $y' = 1 - y/x$.

We first gather the terms involving y on one side: $y' + \frac{y}{x} = 1$. Now we're going to multiply both sides by x to get

$$xy' + y = x.$$

Then we notice that the left side is actually the result of the product rule: $xy' + y = (xy)'$. So we can rewrite this as

$$(xy)' = x.$$

But now the left side is itself a derivative, so the Fundamental Theorem of Calculus tells us we can integrate both sides to get rid of the derivative:

$$xy = \int x \, dx = \frac{x^2}{2} + C.$$

Finally, we can solve for y:

$$y = \frac{x}{2} + \frac{C}{x}.$$

Now we need to figure out how to generalize these steps for any $P(x)$ and $Q(x)$. We'll always start by rearranging our linear equation into the form

$$y' + P(x)y = Q(x).$$

The equation is what we multiply by so that we can repeat the trick where the left-hand side looks like the product rule. We need to pick the right function $I(x)$ to multiply by. We know that $(I(x)y)' = I(x)y' + I'(x)y$. So if we could change $y' + P(x)y$ to look like $I(x)y' + I'(x)y$, we would know that the antiderivative is $I(x)y$. So we need to multiply by a function $I(x)$ so that $P(x)I(x) = I'(x)$.

Here's where we get lucky: we know what this $I(x)$ is, it's exactly $e^{\int P(x)\,dx}$. We can check that: if $I(x) = e^{\int P(x)\,dx}$ then, using first the chain rule and then the fundamental theorem of calculus,

$$I'(x) = \left(e^{\int P(x)\,dx}\right)' = e^{\int P(x)\,dx}\left(\int P(x)\,dx\right)' = P(x)e^{\int P(x)\,dx}.$$

We call $I(x) = e^{\int P(x)\,dx}$ the *integrating factor*. (Because we multiply by it in order to make the equation possible to integrate.) We multiply both sides of the equation by it to get

$$e^{\int P(x)\,dx}y' + P(x)e^{\int P(x)\,dx}y = Q(x)e^{\int P(x)\,dx}.$$

Then we integrate both sides with respect to x. We know what the integral of the left hand side is—it's $ye^{\int P(x)\,dx}$—and the right hand side is a function of x we hope we can integrate

$$ye^{\int P(x)\,dx} = \int Q(x)e^{\int P(x)\,dx}\,dx.$$

Example 11.4. Find the solution to $y' = x + y$.

We rewrite this differential equation as $y' + (-1)y = x$. Then $P(x) = -1$, the constant function. We need our integrating factor $I(x) = e^{\int -1\, dx} = e^{-x}$. We multiply both sides by this to get

$$y'e^{-x} - e^{-x}y = xe^{-x}.$$

We integrate both sides. Since the left side is $(ye^{-x})'$, its antiderivative is ye^{-x}. We could calculate this using the product rule to see that, but we don't need to unless we're checking out work—the point is that we picked $I(x)$ exactly to make this happen.

The right hand side is a classic integration by parts you've seen before: $\int xe^{-x}\, dx = -e^{-x}(x+1) + C$. So

$$ye^{-x} = -e^{-x}(x+1) + C,$$

so

$$y = -(x+1) + Ce^x.$$

Notice that when we found the integrating factor, we *didn't* have a constant, even though we took an integral. We could also have taken $I(x) = e^{-x+D}$. Why didn't we? The reason is that *any* of the functions e^{-x+D} would have worked as our integrating factor, so we chose the simplest one—we just needed to multiply by *a* function that makes the left hand side easy to integrate. (If we had taken $I(x) = e^{-x+D}$, the D's would have canceled out in the end—it doesn't matter which one we choose because they all lead to the same solution.)

3. $y' = 2\cos^2(x)\sin(x) - \sec(x) - \tan(x)y$

Exercise 11.5. *Solve one of the equations from the previous exercise.*

12   Equilibria of Differential Equations

12.1   The Logistic Equation

The logistic equation is a model for describing the growth of a population of some species. Let us use $P(t)$ for the population at time t. It modifies the familiar exponential model, $\frac{dP}{dt} = rP$, which says that the growth of the population is proportionate to how big the population already is. r is the constant of proportionality; a higher r means a faster rate of growth—r takes into account things like how many children each member of the population produces, how long it takes new creatures to reach reproductive age, and so on.

As you know, the solution to $\frac{dP}{dt} = rP$ is Ce^{rP}. In particular, the model predicts that population grows exponentially forever, which is of course absurd. The popultion of rabbits might grow approximately exponentially for a while, but what happens when there isn't enough food for them all? The logistic equation corrects for this by taking into account that the environment has only limited resources, so once the population gets big enough, the population grows more slowly or shrinks because of competition for these limited resources.

Formally, we add a new constant, K, the *carrying capacity*—the maximum population the environment can sustainably support. The logistic equation is

$$\frac{dP}{dt} = rP(1 - \frac{P}{K}).$$

When P is much smaller than K, $\frac{dP}{dt}$ is close to rP—the population grows almost as quickly as in the exponential model. When P gets close to K, though, $1 - \frac{P}{K}$ gets small, and the rate of growth slows. And if P bigger than K, which means the population is overcrowded, $\frac{dP}{dt}$ becomes negative as the population shrinks back towards K.

For simplicity, let's take the (totally unrealistic) values $r = K = 1$ and look at the slope field:
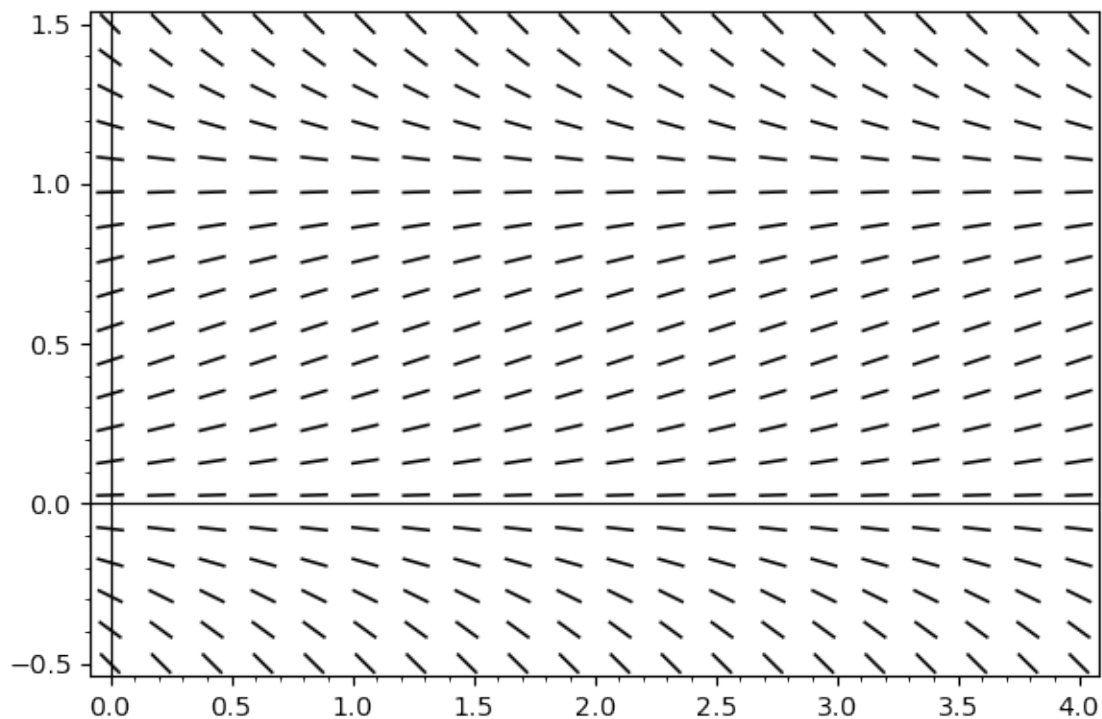
Looking at the slope field, you might notice that there are two horizontal lines which stand out. These are the constant solutions—solutions of the form $P(t) = c$. We call a constant solultion to a differential equation an *equilibrium solution*, because it's a situation where the solution "remains in equilibrium": because $\frac{dP}{dt} = 0$, the value of P does not change.

In this case, one of the equilibria is $P(t) = K$ and the other is $P(t) = 0$. If the population is right at the carrying capacity, the number of new creatures being born is exactly canceled by the number dying due to limited resources. And if there are no creatures, of course we do not expect any to spontaneously generate.

We notice that the equilibria help us understand the other solutions as well.
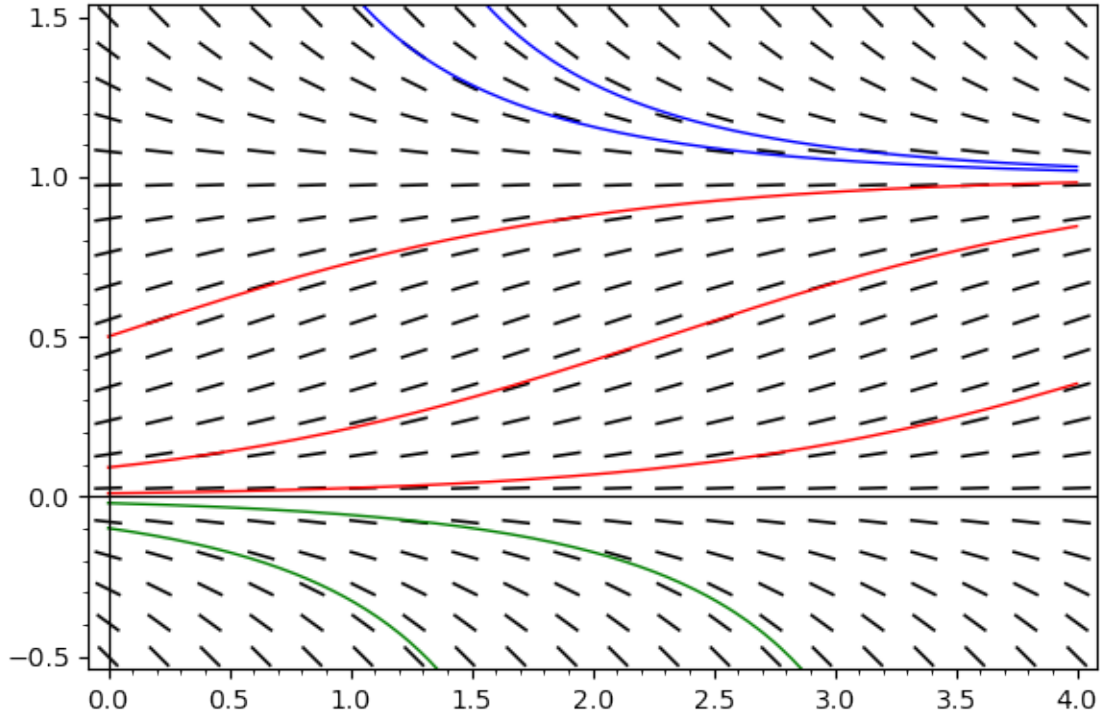
There are only really three kinds of solutions to this equation besides the equilibria: solutions which start below $P = 0$ and decrease away from it, solutions which start between the equilibria and increase towards $P = 1$ without ever reaching it, and solutions which start above $P = 1$ and decrease towards it without ever reaching it.

To put this more formally: if $P(t)$ is a solution to the equation $P' = P(1 - P)$ then

$$\lim_{t \to \infty} P(t)$$

can only be:

- 0 if $P(0) = 0$,

- 1 if $P(0) > 0$,

- $-\infty$ if $P(0) < 0$.

More generally, for the same reasons, if $P(t)$ is a solution to $P' = rP(1 - \frac{P}{K})$ then

$$\lim_{t \to \infty} P(t)$$

can only be:

- 0 if $P(0) = 0$,

- K if $P(0) > 0$,

- $-\infty$ if $P(0) < 0$.

Exercise 12.1. *Which of the initial conditions $P(0) = 0$, $P(0) > 0$, and $P(0) < 0$ would make sense when modeling a population?*

12.2 Autonomous Equations

This approach to analyzing the solutions of a differential equation works for an important class of differential equations.

Notice that the variable t does not appear on the right hand side of the logistic equation $\frac{dP}{dt} = rP(1 - \frac{P}{K})$. This makes sense: the formula on the right reflects a (very simple) model of some underlying biological phenomenon. How much the population grows depends on how big the population is—the variable P—and some constants representing facts about the situation—r, representing how quickly the species reproduces, and K, representing the resources available to support the population—but it doesn't depend on the date on the calendar, which is what t represents.

This situation comes up a lot, especially when the independent variable is time. For instance, many differential equations which come up in physics have this property: the differential equation represents some physical law, which is going to be the same at all times, so the equation doesn't depend on t. So we give these equations a name.

Definition 12.1. *An autonomous differential equation is an equation $y' = F(y)$ where the independent variable does not appear on the right hand side.*

Autonomous equations are always separable, but the integrals needed to solve them quickly become unwieldy. We're instead going to be interested in the qualitative features of solutions, especially what happens to our solutions as t goes to infinity.

One thing you might notice is that slope fields of autonomous equations are very simple: because y' does not depend on the t value, we see the same pattern repeated across each

row. (Go take another look at the slope field for the differential equation and notice how everything in the same row has the same angle.)

This means we can always take a solution to an autonomous equation and shift it horizontally to get other solutions.

We already saw that there are (at least) two kinds of equilibria: in the logistic equation, when P was close to K but not exactly K, the solution approached K, while when P was close to 0 but not exactly 0, the solution diverged away from 0. We say that K is a *stable equilibrium* of the logistic equation while 0 is an *unstable equilibrium.*

Definition 12.2. *When* $y' = F(y)$ *is an autonomous differential equation, we say an equilibrium solution* $y = c$ *is* stable *if every solution whose initial condition is near* c *coverges to* c *as* $t \to \infty$.

We say $y = c$ *is* unstable *if no solution whose initial condition is near (but not equal) to* c *converges to* c *as* $t \to \infty$.

If we take a stable equilibrium and "perturb" it a little, the solution will start returning to the equilibrium. But if we take an unstable equilibrium and "perturb" it, the solution moves *away* from the equilibrium. This difference matters a lot when we use differential equations to model actual problems: the real world is never as simple as our mathematical model, so our equilibria are basically always being shifted a little by small errors in measurements and by details we didn't incorporate into our model.

The good news about autonomous differential equations is that it's usually easy to tell if an equilibrium is stable or unstable. It's common to represent the information we need using a diagram. We start by marking the equilibria on a vertical line; for instance, if we are looking at the logisitic equation $\frac{dP}{dt} = rP(1 - \frac{P}{K})$, our equlibria are 0 and K:

$- - - P = K$

$- - - P = 0$
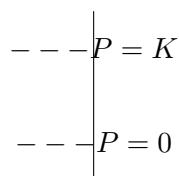
This divides the values into several intervals. Then in each interval, we ask whether $\frac{dy}{dt} = F(y)$ is positive or negative. As long as F is continuous (which it usually is in practice), the intermediate value theorem tells us that, on each interval, F is either positive on the whole

interval or negative on the whole interval. (Because F can only transition from positive to negative by going through 0, and we have marked all the 0's already.) That means that we can check whether F is positive or negative by testing out a single value from that interval.

If $F(y)$ is positive on an interval, we draw an arrow pointing up, representing the fact that $\frac{dy}{dt} = F(y)$ is positive, so y is increasing on that interval. If $F(y)$ is negative, we drawn an arrow pointing down, because y is decreasing on that interval.

For instance, with the logistic equation, when $P > K$, $rP(1 - \frac{P}{K})$ is negative. When $0 < P < K$, both rP and $1 - \frac{P}{K}$ are positive, so $rP(1 - \frac{P}{K})$ is positive as well. And when $P < 0$, rP is negative while $1 - \frac{P}{K}$ is positive, so $rP(1 - \frac{P}{K})$ is negative. Our diagram is:
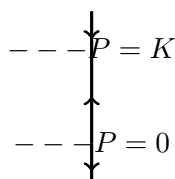

When the arrows both point towards the equilibrium, it's a stable equilibrium: solutions nearby move towards the equilibrium. And when the arrows point away from the equilibrium, it's an unstable equilibrium: nearby solutions move away from the equilibrium.

Exercise 12.2. *Find the equilibria of the differential equation $y' = -3(y - 5)(y - 10)$ and determine which are stable and which are unstable.*

You might have noticed that our discussion of stable and unstable equilibria left out a case: what if *one* arrow points towards the equilibrium while the other points away. This is less common, but it is possible; we call it a *semistable* equilibrium.

Exercise 12.3 (+). *Write down an autonomous differential equation $y' = F(y)$ which gives rise to this diagram:*
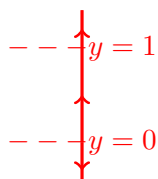
(Think about what has to be true about F near 0 and near 1 for something like this to happen.)

12.3 Systems of Equations

In keeping with what we have done already this semester, we can think about what happens to our differential equations when y is a function from $\mathbb{R}^m$ to $\mathbb{R}^n$ for values m, n which might be bigger than m.

The case where m is bigger than 1 leads to *partial differential equations*, since there are multiple input variables, leading us to need partial derivatives to talk about them. Partial differential equations are quite important, but also quite difficult, and we will not discuss them here. (Traditionally one first takes an entire course on *ordinary* differential equations—the case we have been discussing where there is just one independent variable—before studying partial differential equations.)

The case where $n > 1$, on the other hand, will not be so different from what we have already seen. As usual, when we have a function $\mathbf{y} : \mathbb{R} \to \mathbb{R}^n$, we can think of it as being n separate functions $y_1, \ldots, y_n$ from $\mathbb{R}$ to $\mathbb{R}$. A *system of differential equations* is a list of n equations, one for the derivative of each of these. Or, in vector notation,

$$\mathbf{y}' = \mathbf{F}(\mathbf{y}, x)$$

where $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$.

Often we write these equations out separately. For instance, when $n = 2$, we might call the two dependent functions y and z, and a typical system might be

$$y' = y + z + x$$
$$z' = y + z - x.$$

Example 12.3. Consider the system of differential equations $y' = z - x$, $z' = 1 + \frac{2y}{x^2}$. Check that $y = 3x^2$, $z = 7x$ is one of the solutions.

We have $y' = 6x$, $z' = 7$. Then $6x = 7x - x$ and $7 = 1 + \frac{2 \cdot 3x^2}{x^2}$ are both true.

Exercise 12.4. *Write a system of differential equations so that $y = \sin x$ and $z = \cos x$ is one of the solutions.*

There's another way we could try to generalize what we've done with differential equations: we could start looking at equations with more than one derivative, like $y'' = y' + y$. But we don't need a whole new subject for that: "higher order" differential equations are just systems of differential equations in disguise.

Example 12.4. Let us rewrite $y'' = y' + y$ as a system of differential equations. We give y' its own name, say z, and one of our equations is just $y' = z$. Then we rewrite $y'' = y' + y$ using the fact that $y'' = z'$, so it becomes $z' = z + y$. So

$$z' = z + y, y' = z$$

is a system of equations representing the same problem as solving $y'' = y' + y$.

Exercise 12.5. *Write a system of three differential equations which is equivalent to $y''' = y \cdot y' \cdot y''$. (You'll need to introduce two new variables, one to represent y' and one to represent y''.)*

We won't worry too much about finding exact solutions to systems of differential equations—we're going to focus on qualitative features of their solutions. But there's one thing you should know about the solutions to systems of differential equations, and that's how many constants they have: if you have a system of n differential equations, the general solution should have n constants. That means that an initial value problem for a system of differential equations should specify n initial values—most commonly, the values of all the equations at one x value.

Exercise 12.6. *Consider the system of differential equations*

$$y' = \frac{8}{3}y + \frac{2}{3}z, \quad z' = \frac{1}{3}y + \frac{7}{3}z.$$

Check that, for all values of C and D, $y = -Ce^{2x} + 2De^{3x}$, $z = Ce^{2x} + De^{3x}$ is a solution.

Exercise 12.7. *Find the solution to*

$$y' = \frac{8}{3}y + \frac{2}{3}z, \quad z' = \frac{1}{3}y + \frac{7}{3}z.$$

such that $y(0) = 1$ and $z(0) = -1$.

12.4   Vector Fields

We can't easily draw slope fields for systems of differential equations the way we could for individual equations. When our system of equations has two variables, we could imagine a 3D version of a slope field, but drawn on 2D paper, the result isn't very readable.

However if the system is *autonomous*, we can make a useful 2D drawing. Remember that in a slope field for a single autonomous differential equation, $y' = F(y)$, the slope doesn't depend on the x coordinate, so we just draw the same slope all the way across each horizontal line.

That suggests that if we have an autonomous system of two equations,

$$\frac{dx}{dt} = F(x, y), \ \frac{dy}{dt} = G(x, y),$$

we could make a 2D drawing using only the x and y coordinates and leave out the independent variable t entirely. At each coordinate (x, y), we draw the vector (x', y'). We call this a *vector field*, because we have a vector for each point. (More precisely, we usually draw a vector *proportional to* (x', y')—often the actual vector (x', y') would be too long, so we draw something like $\frac{1}{2}(x', y')$ or $\frac{1}{10}(x', y')$.)

For example, consider the system of differential equations $x' = y$, $y' = -x$.
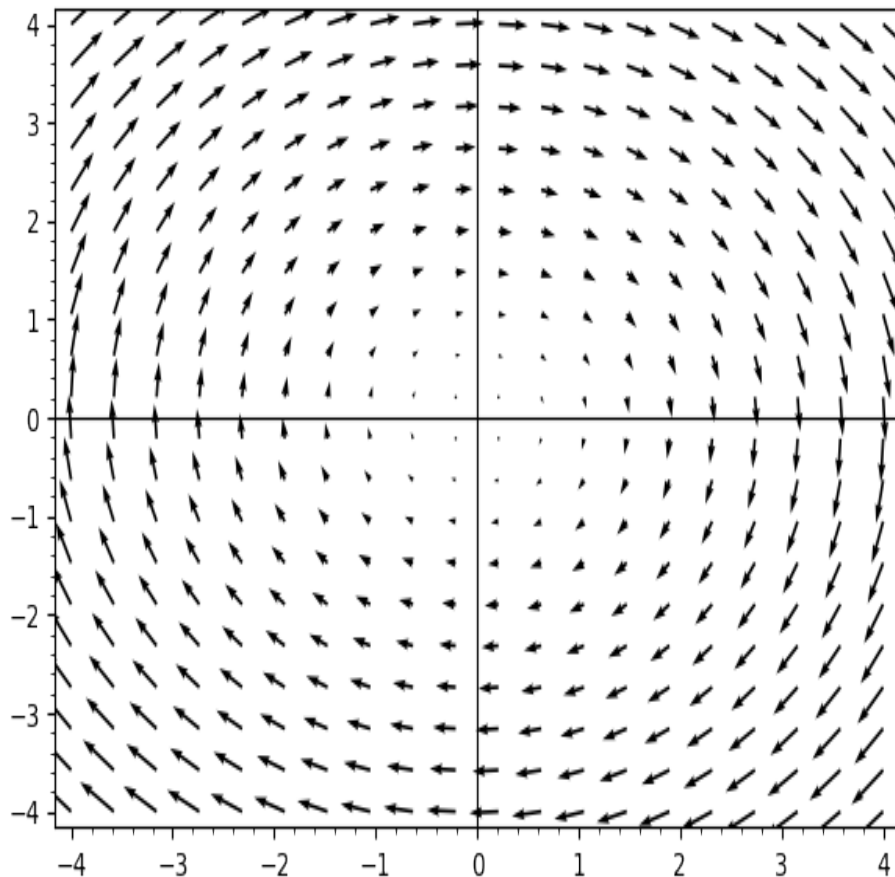
In a slope field, we draw *line segments*, without arrows. That's because the x-axis is the independent variable: the direction of movement from left to right is built into the graph. With a vector field, we have to draw arrows to indicate which direction the vectors point.

The vector field tells us that if at some time t we happen to be at the point (x, y), then the arrow at that point tells us what direction to move in, and the length of the arrow tells us how fast we'll be moving. And if that's what happens at some time t when we're at the point (x, y), the same thing happens at *any* time t that we're at (x, y), because in an autonomous system of equations, x' and y' don't depend on t.

For instance, if our solution is at the point $(1, 0)$ at time t, then at time t the solution will be moving straight down—the y value will be decreasing and the x value will be staying the same. If instead our solution is at the point $(.1, 0)$ at time t, the solution will still be

178

moving straight down, but at a much slower speed.

So this picture is telling us that our solutions go around in clockwise circles centered on the origin. Especially after seeing the vector field, you might be able to guess that one of the solutions to this differential equation is $x = \cos(t)$, $y = -\sin(t)$, which we can confirm by calculating the derivatives.
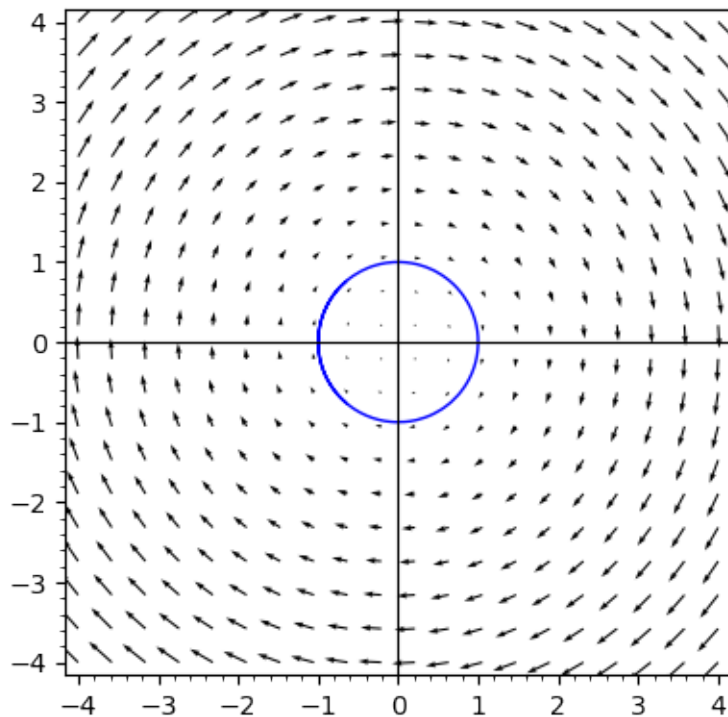
Exercise 12.8. *Sketch a field for the system of differential equations $x' = -y$, $y' = -x$. (5x5 should be enough to get the flavor.)*

13   Autonomous Systems of Differential Equations

13.1   Qualitative Behavior for Autonomous Systems

An autonomous system of equations is a system like

$$\frac{dx}{dt} = f(x, y) \quad \frac{dy}{dt} = g(x, y),$$

where the functions on the right do not depend on t. If $\mathbf{y} = \begin{bmatrix} x \\ y \end{bmatrix}$, we can write this very compactly as $\mathbf{y}' = F(\mathbf{y})$. Just like individual autonomous differential equations, autonomous systems of differential equations often have equilibria: the equilibria of $\mathbf{y}' = F(\mathbf{y})$ are precisely where $F(\mathbf{y}) = 0$.

Exercise 13.1. *Find all equilibrium solutions of the system $\frac{dx}{dt} = x^2 + y$, $\frac{dy}{dt} = x + y$.*

Because real applications always involve some amount of measurement and modeling error, knowing where the equilibria are doesn't help us much unless we also understand what happens *near* the equilibria. We'll continue focusing on the case where our system has two equations so we can nicely visualize the different kinds of behavior.

The most important behavior we are worried about is whether solutions near the equilibrium tend move towards the equilibrium or away from it.

Definition 13.1. *When $\mathbf{y}' = F(\mathbf{y})$ is an autonomous system of differential equations, we say an equilibrium solution $\mathbf{y} = \mathbf{c}$ is a* sink, *or* stable, *if every solution whose initial condition is near $\mathbf{c}$ coverges to $\mathbf{c}$ as $t \to \infty$.*

We say $\mathbf{y} = \mathbf{c}$ is a source, *or* unstable, *if every solution whose initial condition is near $\mathbf{c}$ diverges away from $\mathbf{c}$ as $t \to \infty$.*

These definitions are easier to understand from a picture, since all sinks and all sources look roughly the same. A sink looks roughly like this: all the arrows are pointing in towards the equilibrium.
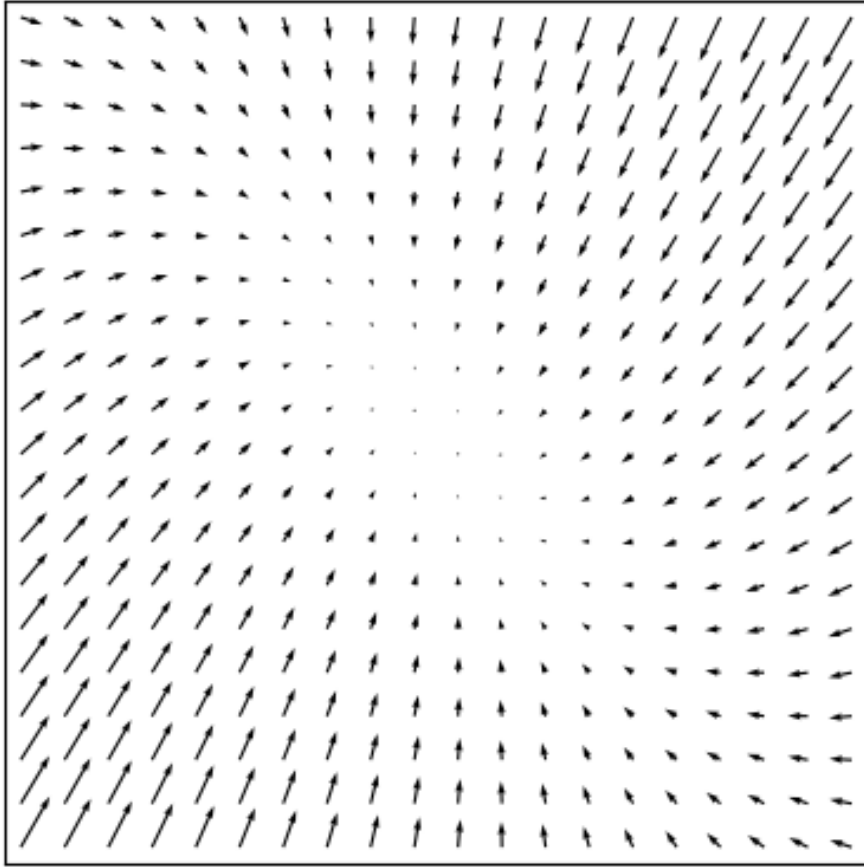
A source looks like the opposite: all the arrows are pointing away from it.
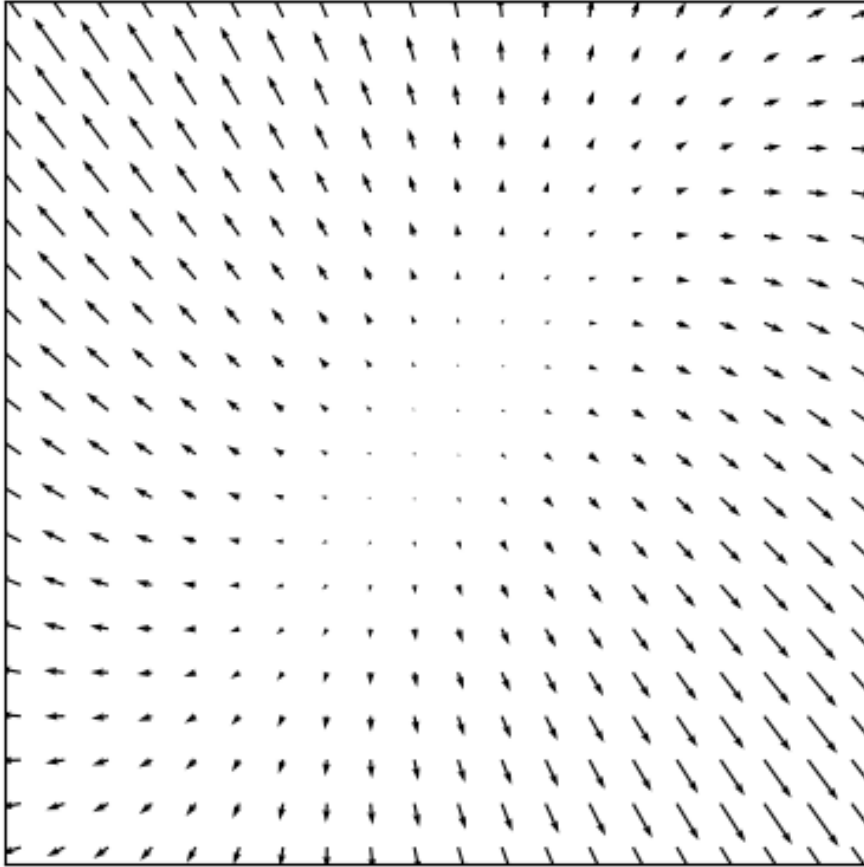
An equilibrium might be neither. One way to be neither is if solutions neither converge towards the equilibrium nor diverge away from it, but instead circle around it forever, like the example we saw in the previous section:
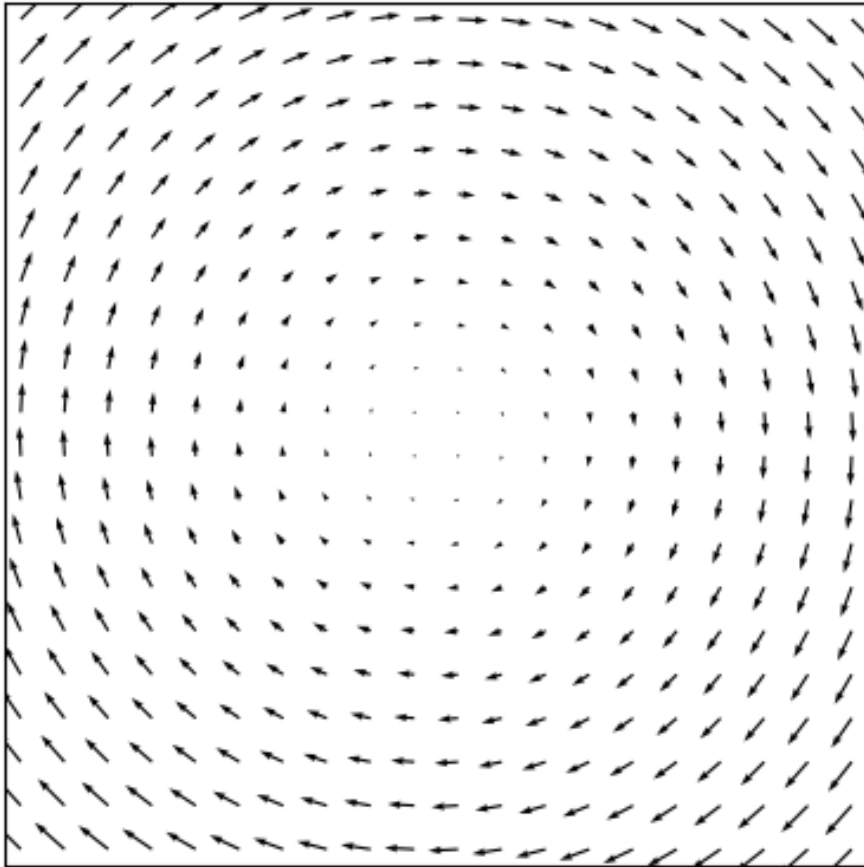
We call this kind of equilibrium a *center point*.

Another way an equilibrium could be neither is if some solutions converge towards the equilibrium while others diverge away.
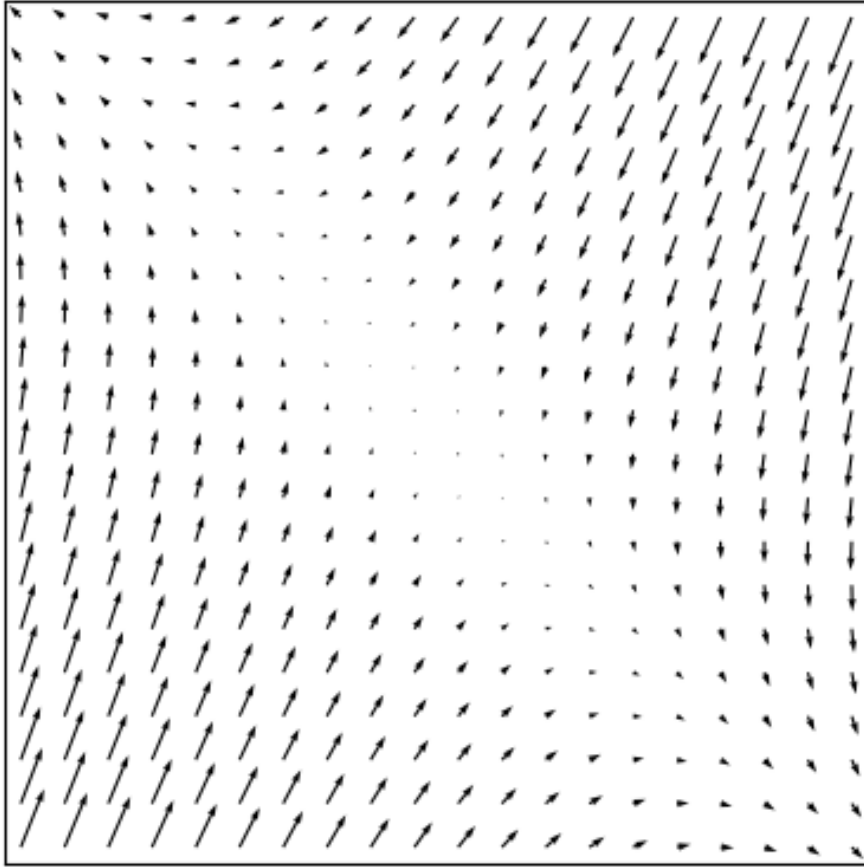
We call this a *saddle point*, and it has a lot in common with the saddle points we saw when optimizing. The behavior here is typical: if we started at, say, $(-2, -3)$, our solution would initially approach the equilibrium, but then, just as it got close, start to twist away from it. If we happened to start in exactly the right place (in this picture, it's anywhere on the line $y = 2x$), our solution would approach the equilibrium without ever slipping away to a divergent solution, but if our initial position is just a tiny bit off that line, it will eventually start diverging away from the equilibrium.
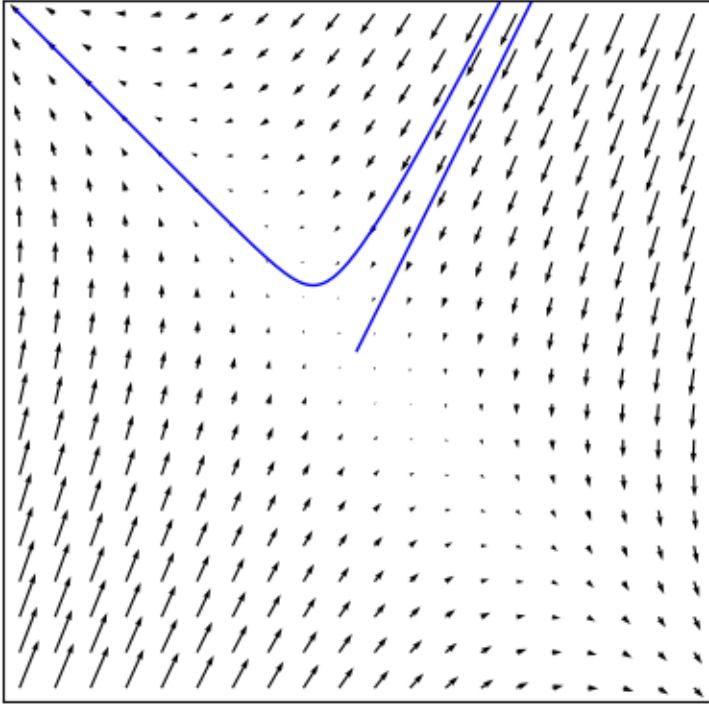
There's a second behavior we care about: when our equilibrium is a source or a sink, we can separately ask whether the solutions spiral around the equilibrium, like in these two examples.
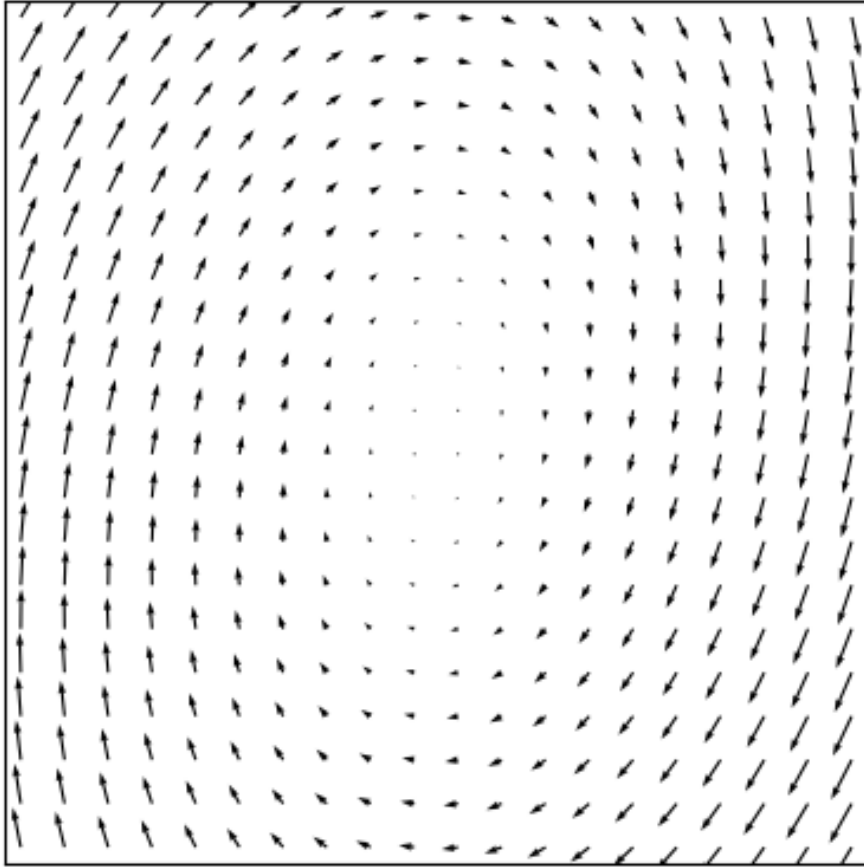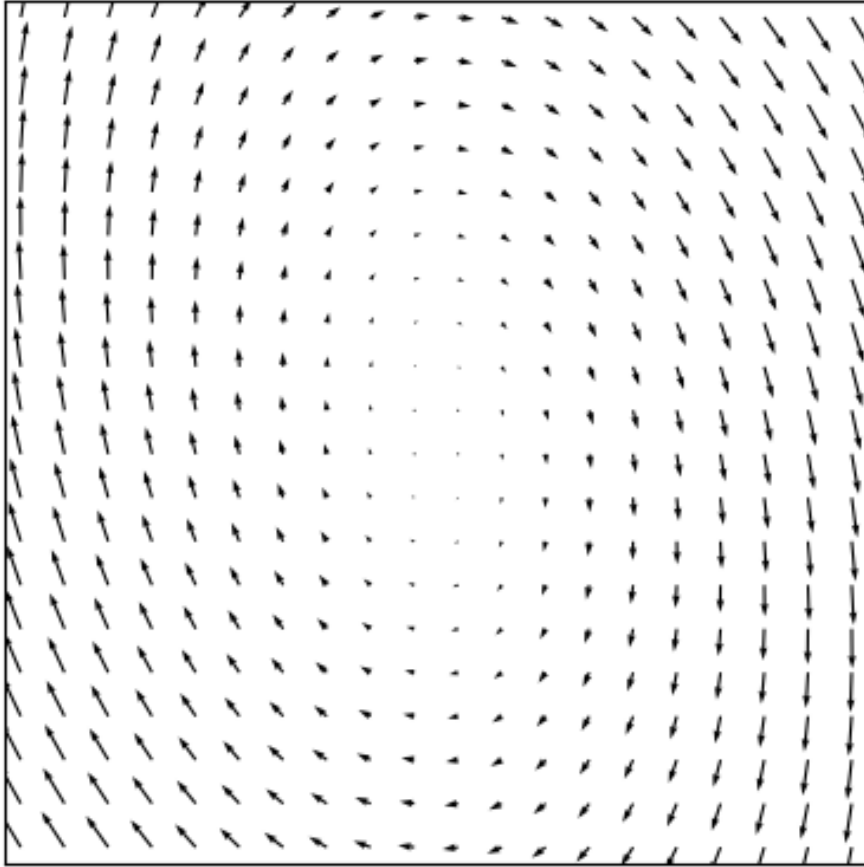
In both of these examples, the solutions are spiraling around the equilibrium. But, unlike a center point, the solutions don't just spiral forever: in the first one, they're slowly spiraling inwards towards the equilibrium. We call this a *spiral sink*. In the second, the solutions are slowly spiraling away from the equilibrium. We call this a *spiral source*. We could think of this spiral behavior as being somewhere between an ordinary sink or source and a circle point.

Exercise 13.2. *The system of differential equations $x' = x^3$, $y' = y^3$ has an equilibrium at the origin. Sketch a vector field and use it to identify what kind of equlibirum the origin is.*

Do the same with the system $x' = -y$, $y' = x - y$.

13.2   Determining the Behavior of Linear Autonomous Systems

When, in addition to being autonomous, our system of differential equations is *linear and homogeneous*—that is, it has the form $\mathbf{y}' = A \cdot \mathbf{y}$ for some matrix A—then there's always an equilibrium at $\mathbf{y} = \mathbf{0}$, and there's a systematic way to tell what kind of equilibrium we're dealing with.

Exercise 13.3. *Why does an equation $\mathbf{y}' = A \cdot \mathbf{y}$ always have an equilibrium at $\mathbf{y} = \mathbf{0}$?*

Let's start by looking at what happens with an ordinary (not spiraling) sink. All the solutions point towards the equilibrium, but there are two solutions which are straight lines pointing directly towards the equilibrium.
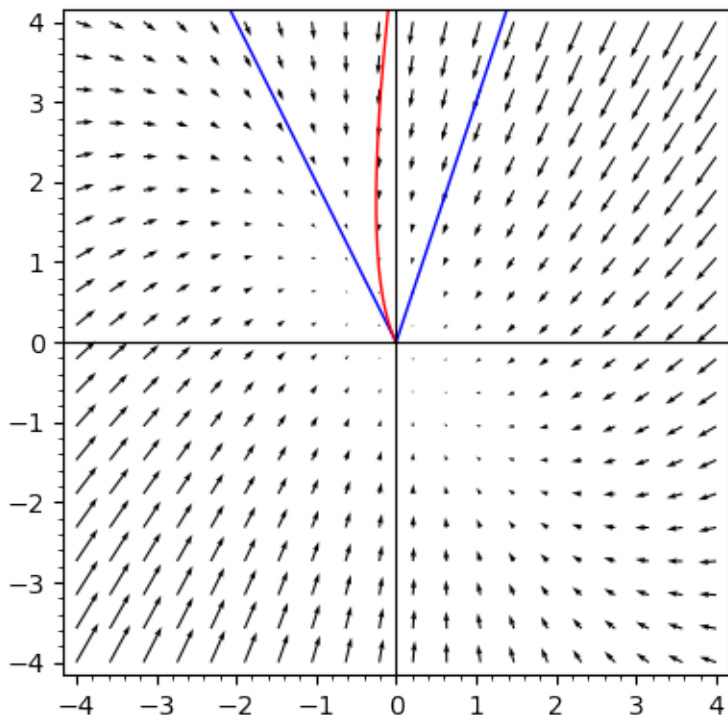
All the other solutions, like the one in red, have a curved behavior. (In fact, they always approach one of the purely linear solutions as they get near the equilibrium.)

Say our initial value is $x(0) = a$, $y(0) = b$. What has to happen for this initial condition

to give us one of the straight line solutions? We would need $(x', y') = A \begin{bmatrix} a \\ b \end{bmatrix}$ to point

directly towards the origin. That is, we would need $A \begin{bmatrix} a \\ b \end{bmatrix}$ to point in the same direction

as $- \begin{bmatrix} a \\ b \end{bmatrix}$. If that happened then, when we started at $\begin{bmatrix} a \\ b \end{bmatrix}$, the differential equation would tell us to start walking straight towards the origin. And then, after we'd moved a

little, we'd be at $0.99 \begin{bmatrix} a \\ b \end{bmatrix}$, and the linearity of matrix multiplication would tell us that

$A \cdot 0.99 \begin{bmatrix} a \\ b \end{bmatrix} = 0.99 A \begin{bmatrix} a \\ b \end{bmatrix}$, so we'd still be pointing in the same direction, namely right

towards the origin. Since this would keep happening forever, we'd just stay on the same

line $\begin{bmatrix} a \\ b \end{bmatrix}$.

What does it mean for $A \begin{bmatrix} a \\ b \end{bmatrix}$ to point in the direction $- \begin{bmatrix} a \\ b \end{bmatrix}$? It means $A \begin{bmatrix} a \\ b \end{bmatrix} = \lambda \begin{bmatrix} a \\ b \end{bmatrix}$ where λ is negative.

Similarly, if $A \begin{bmatrix} a \\ b \end{bmatrix} = \lambda \begin{bmatrix} a \\ b \end{bmatrix}$ and $\lambda > 0$ then we would have a linear solution pointing

directly away from the origin.

So if we're given the differential equation $\mathbf{y}' = A \cdot \mathbf{y}$ and we want to know what the solutions which are lines are, we should try to find vectors $\mathbf{v}$ so that $A\mathbf{v} = \lambda\mathbf{v}$. We've seen this before: that's the same as saying $A\mathbf{v} = \lambda I \cdot \mathbf{v}$, which is the same as $A\mathbf{v} - I \cdot \lambda \mathbf{v} = 0$, which is the same as $(A - \lambda I)\mathbf{v} = 0$. We recall that this has a solution exactly when $A - \lambda I$ is not invertible. You might remember that these values of λ have a name—they're the eigenvalues of the matrix A.

Theorem 13.2. *If there is a real number $\lambda < 0$ so that $A - \lambda I$ is not invertible, then $\mathbf{y}' = A \cdot \mathbf{y}$ has a solution which is a line pointing towards the origin.*

If there is a real number $\lambda > 0$ so that $A - \lambda I$ is not invertible, then $\mathbf{y}' = A \cdot \mathbf{y}$ has a solution which is a line pointing away from the origin.

When A is a 2×2 matrix, we already know how to find these values of λ, because we found them when we were looking for minima and maxima: we set $\det(A - \lambda I) = 0$, write out the polynomial, and solve it.

Theorem 13.3.
- *If all values λ so that $A - \lambda I$ is not invertible are negative real numbers, then the origin is a sink of $\mathbf{y}' = A \cdot \mathbf{y}$.*

- *If all values λ so that $A - \lambda I$ is not invertible are positive real numbers, then the origin is a source of $\mathbf{y}' = A \cdot \mathbf{y}$.*

- *If there is a real value $\lambda < 0$ so that $A - \lambda I$ is not invertible, and also a real value $\lambda > 0$ so that $A - \lambda I$ is not invertible, then the origin is a saddle point of $\mathbf{y}' = A \cdot \mathbf{y}$.*

This looks a lot like our analysis of minima and maxima: when the λ's are all negative, one thing happens, when they're all positive, another thing happens, and when they're mixed, we get a saddle point.

Example 13.4. What kind of equilibrium is the origin in the differential equation $x' = -7x - y$, $y' = -6x - 8y$?

We can write out the corresponding matrix, $\begin{bmatrix} x' \\ y' \end{bmatrix} = \mathbf{y}' = \begin{bmatrix} -7 & -1 \\ -6 & -8 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$. We set $\det\left(\begin{bmatrix} -7 - \lambda & -1 \\ -6 & -8 - \lambda \end{bmatrix}\right) = 0$.

Expanding the polynomial, we have

$$(-7 - \lambda)(-8 - \lambda) - 6 = 0,$$

or

$$\lambda^2 + 15x + 50 = 0.$$

We can use the quadratic formula, but we might get lucky and notice that this factors nicely into $(\lambda + 5)(\lambda + 10) = 0$, so the values of λ which make $\begin{bmatrix} -7 & -1 \\ -6 & -8 \end{bmatrix} - \lambda I$ non-invertible are -5 and -10.

These are both negative, so the equilibrium is a sink.

Exercise 13.4. *What kind of equlibrium is the origin in the system of differential equations* $x' = x - 2y, y' = -x - y$?

We can extend this analysis to center points and spiral points, too. When A is a 2×2 matrix of real numbers, the equation $\det(A - \lambda I) = 0$ is a quadratic equation with real coefficients. We've already accounted for all the cases where the values of λ solving this are non-zero real numbers, but it's also possible to get complex solutions. Because the coefficients are real, we can only get complex solutions in pairs of the form $\lambda = a \pm bi$.

Theorem 13.5. *When A is a 2×2 matrix:*

- *if the two values of λ so that $A - \lambda I$ is not invertible are purely imaginary numbers (numbers of the form $b \cdot i$ for a real number b) then the origin is a circle point of* $\mathbf{y}' = A \cdot \mathbf{y}$,

- *if the two values of λ so that $A - \lambda I$ is not invertible are $a + bi$ and $a - bi$ with $a < 0$ then the origin is spiral sink of* $\mathbf{y}' = A \cdot \mathbf{y}$,

- *if the two values of λ so that $A - \lambda I$ is not invertible are $a + bi$ and $a - bi$ with $a > 0$ then the origin is spiral source of* $\mathbf{y}' = A \cdot \mathbf{y}$.

When we dealt with minima and maxima, we didn't have to worry about imaginary and complex numbers because the matrices we dealt with there were always symmetric—we were looking at the matrix of second derivatives, where $f_{xy} = f_{yx}$; it turns out that for symmetric matrices, the values of λ are always real. But linear systems of equations could have any matrix, and in general we can get complex values for λ.

13.3   Inhomogeneous Equations

What about inhomogeneous differential equations, where the equilibrium will be somewhere other than $\mathbf{0}$?

Example 13.6. Where does $x' = -7x - y - 6, y' = -6x - 8y + 2$ have an equilibrium, and what kind of equilibrium is it?

We have an equilibrium when $x' = y' = 0$, so when $-7x - y = 6$ and $-6x - 8y = -2$. This is a pair of linear equations in two variables, so we can solve them by a variety of ways to find the solution $x = -1$, $y = 1$. So the only equilibrium is at $(-1, 1)$.

Let us rewrite the differential equation in matrix form:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} -7 & -1 \\ -6 & -8 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} -6 \\ 2 \end{bmatrix}.$$

Now we'll "recenter" the equation on the equilibrium: instead of multiplying by $\begin{bmatrix} x \\ y \end{bmatrix}$, we'll consider the equation relative to the separation from the equilibrium $\begin{bmatrix} x+1 \\ y-1 \end{bmatrix}$:

$$\begin{bmatrix} -7 & -1 \\ -6 & -8 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} -6 \\ 2 \end{bmatrix} = \begin{bmatrix} -7 & -1 \\ -6 & -8 \end{bmatrix} \begin{bmatrix} x+1-1 \\ y-1+1 \end{bmatrix} + \begin{bmatrix} -6 \\ 2 \end{bmatrix}$$

$$= \begin{bmatrix} -7 & -1 \\ -6 & -8 \end{bmatrix} \begin{bmatrix} x+1 \\ y-1 \end{bmatrix} + \begin{bmatrix} -7 & -1 \\ -6 & -8 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} + \begin{bmatrix} -6 \\ 2 \end{bmatrix}$$

$$= \begin{bmatrix} -7 & -1 \\ -6 & -8 \end{bmatrix} \begin{bmatrix} x+1 \\ y-1 \end{bmatrix}$$

In other words, this will be just like the system of differential equations $\begin{bmatrix} -7 & -1 \\ -6 & -8 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$ except that we've shifted it left by 1 and up by 1. In particular, we already know, from Example 13.4, that the equilibrium is a sink.

Now that we've seen this done once, we can guess that this always works for inhomogeneous linear differential equations: we can rewrite them to look like homogeneous equations

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = A \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}$$

where (x_0, y_0) is the equilibrium, and then analyze the matrix A just like for a homogeneous equation.

Exercise 13.5. *Where is the equilibrium in the system of differential equations $x' = x - 2y, y' = -x - y + 3$? Rewrite this differential equation in the form $\begin{bmatrix} x' \\ y' \end{bmatrix} = A \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}$ and identify what kind of equuliribrium it is.*

We can summarize everything we've said in the following theorem:

Theorem 13.7. *When* $\mathbf{y}' = A \cdot (\mathbf{y} - \mathbf{y}_0)$ *is a system of two differential equations, the only equilibrium is at* $\mathbf{y}_0$. *If* λ_1 *and* λ_2 *are the two values such that* $A - \lambda I$ *is not invertible then:*

- *if both* λ_1 *and* λ_2 *are negative real numbers, this equilibrium is a sink,*

- *if both* λ_1 *and* λ_2 *are positive real numbers, this equilibrium is a source,*

- *if* λ_1 *and* λ_2 *are real numbers where one is positive and the other is negative, this equilibrium is a saddle point,*

- *if* λ_1 *and* λ_2 *are purely imaginary, this equilibrium is a center point,*

- *if* λ_1 *and* λ_2 *are complex numbers with negative real part, this equilibirum is a spiral sink, and*

- *if* λ_1 *and* λ_2 *are complex numbers with positive real part, this equilibirum is a spiral source.*

13.4   Approximating Non-Linear Differential Equations

Why do we care so much about linear homogeneous differential equations? It's not because the equations we most often want to deal with tend to have that form. Quite the opposite: the most important differential equations are often very complicated and non-linear.

The reason we care about linear equations is because of linear approximations: because all differentiable functions are approximated by linear ones, understanding the behavior of linear systems of differential equations lets us understand equilibria for almost any system of differential equations.

Example 13.8. Suppose we have a system of differential equations

$$x' = \cos(x) + \sin(y), y' = \cos(x) - \sin(y).$$

This system is autonomous (neither x' nor y' depends on t), but non-linear since we're applying trig functions to x and y. We can still look for equilibria solutions: if $x' = y' = 0$ then $\cos(x) + \sin(y) = \cos(x) - \sin(y) = 0$. Adding the equations, $2\cos(x) = 0$, and

subtracting the second from the first, $2\sin(y) = 0$. So the equilibria are where $\cos(x) = 0$ and $\sin(y) = 0$.

Take the equilibrium $x_0 = \pi/2$, $y_0 = 0$. Near this equilibrium, we can take a linear approximation using the increment theorem. Let us name the two functions, $f(x,y) = \cos(x) + \sin(y)$ and $g(x,y) = \cos(x) - \sin(y)$. The partial derivatives are $\frac{\partial f}{\partial x} = -\sin(x)$, $\frac{\partial f}{\partial y} = \cos(y)$, $\frac{\partial g}{\partial x} = -\sin(x)$, and $\frac{\partial g}{\partial y} = -\sin(y)$.

Then the increment theorem tells us that

$$x' = f(x,y) \approx f(\pi/2, 0) - (x - \pi/2)\sin(\pi/2) + y\cos(0) = -(x - \pi/2) + y$$

and

$$y' = g(x,y) \approx g(\pi/2, 0) - (x - \pi/2)\sin(\pi/2) - y\cos(0) = -(x - \pi/2) - y.$$

This suggests a new, linear system of differential equations

$$x' = -(x - \pi/2) + y, \; y' = -(x - \pi/2) - y.$$

This new system is is not the same as our original one, but as long as we stay near the equilibrium near $(\pi/2, 0)$, this new system is very similar to our original one. We can write it in matrix form,

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} x - \pi/2 \\ y \end{bmatrix}.$$

Suppose we pick a point near $(\pi/2, 0)$ and start solutions to both differential equations there. Since the two equations are close, the solutions should also stay close, at least for a little while. That's at least enough to tell us that we expect the equilibrium behavior to be the same.

We know how to find the behavior of the linear system of differential equations. The matrix is $\begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix}$. Then $\begin{bmatrix} -1 - \lambda & 1 \\ -1 & -1 - \lambda \end{bmatrix} = 0$ exactly when $(-1 - \lambda)^2 + 1 = 0$, so when $2 + 2\lambda + \lambda^2$. The roots are

$$\lambda = \frac{-2 \pm \sqrt{4 - 8}}{2} = -1 \pm \sqrt{-4}/2.$$

These are complex roots, so we get a spiral, and since the real part is negative, it's a spiral sink. That's exactly what we see when we draw a vector field:
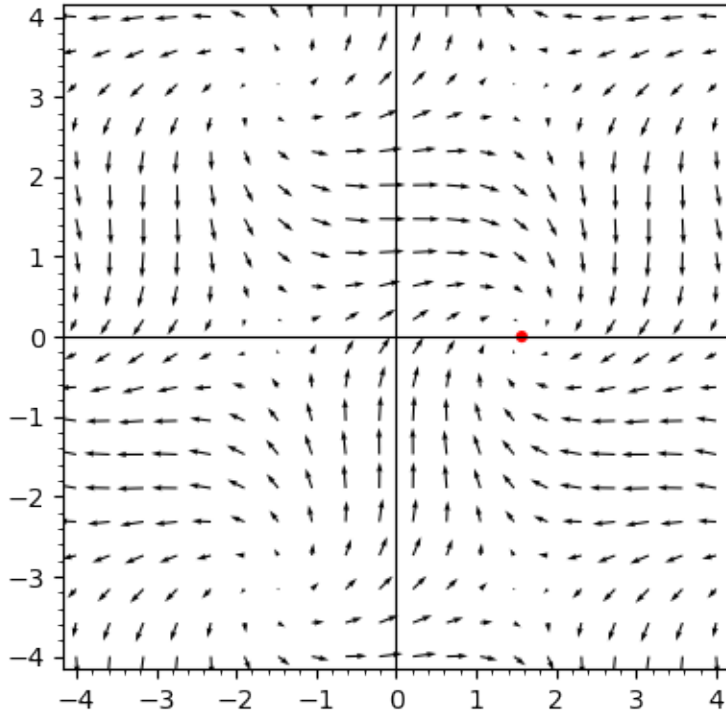
This idea works for any homogeneous system of differential equation. Suppose

$$x' = f(x, y), \ y' = g(x, y)$$

and that x_0, y_0 is an equilibrium point—that is, $f(x_0, y_0) = g(x_0, y_0) = 0$. Then we can rewrite these using the multivariable Taylor approximation: when x and y are near the equilibirum, we have

$$x' = f(x, y) \approx f(x_0, y_0) + \left.\frac{\partial f}{\partial x}\right|_{x_0, y_0} (x - x_0) + \left.\frac{\partial f}{\partial y}\right|_{x_0, y_0} (y - y_0)$$

and

$$y' = g(x, y) \approx g(x_0, y_0) + \left.\frac{\partial g}{\partial x}\right|_{x_0, y_0} (x - x_0) + \left.\frac{\partial g}{\partial y}\right|_{x_0, y_0} (y - y_0).$$

So

$$\begin{bmatrix} x' \\ y' \end{bmatrix} \approx \begin{bmatrix} \left.\frac{\partial f}{\partial x}\right|_{x_0, y_0} & \left.\frac{\partial f}{\partial y}\right|_{x_0, y_0} \\ \left.\frac{\partial g}{\partial x}\right|_{x_0, y_0} & \left.\frac{\partial g}{\partial y}\right|_{x_0, y_0} \end{bmatrix} \cdot \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}.$$

195

So, near an equilibrium, our differential equation has the form

$$\mathbf{y}' \approx A \cdot (\mathbf{y} - \mathbf{y}_0).$$

If this were an exact equality, Theorem 13.7 would tell us what happens. But the approximate equality is enough to tell us that, near the equilibrium, the behavior will be similar.

Theorem 13.9. *When $\mathbf{y}' = F(\mathbf{y})$ is a system of two differential equations with an equilibrium at $\mathbf{y}_0$, $A = (DF)(\mathbf{y}_0)$ is the matrix of derivatives evaluated at $\mathbf{y}_0$, and λ_1 and λ_2 are the two values such that $A - \lambda I$ is not invertible then:*

- *if both λ_1 and λ_2 are negative real numbers, this equilibrium is a sink,*

- *if both λ_1 and λ_2 are positive real numbers, this equilibrium is a source,*

- *if λ_1 and λ_2 are real numbers where one is positive and the other is negative, this equilibrium is a saddle point,*

- *if λ_1 and λ_2 are complex numbers with negative real part, this equilibirum is a spiral sink,*

- *if λ_1 and λ_2 are complex numbers with positive real part, this equilibrium is a spiral source.*

Exercise 13.6. *Consider the system of differential equations $x' = x^3 - y$, $y' = x - y^3$. Find an equilibrium other than the origin and identify what kind of equilibrium it is.*

You might notice that Theorem 13.9 is missing one of the cases from Theorem 13.7: the case where λ_1 and λ_2 are purely imaginary and the equilbrium is a center point. The reason is that, in that case, the equilibrium could actually be a spiral source or sink instead, and we don't have enough information to know whether that happens. After all, we only have the approximate equality $\mathbf{y}' \approx A \cdot (\mathbf{y} - \mathbf{y}_0)$; the small amount of error can't turn a sink into a source, but it can break the perfect circles we need to be a center point.

One way to see the difference is to notice that center points are the only case which was characterized by an equality: we need the real part of λ_1 and λ_2 to be *exactly equal* to 0. All the other cases are characterized by inequalities. A little error won't change an inequality:

if $\lambda_1 < 0$ then, when the error is small enough $\lambda_1 + \epsilon$ will still be negative. But if λ_1 is purely imaginary, even a very small ϵ could cause $\lambda_1 + \epsilon$ to have a positive real part.

We'll finish by noting the steps we used here.

- We analyzed a simplified case, the systems of linear differential equations, where we could use a combination of linear algebra and calculus to fully classify the qualitative behavior of our differential equations.

- We can't hope to solve general differential equations, but we can use Taylor's theorem to approximate them by linear differential equations.

- We then had to consider when the qualitative behavior is robust: in what situations is knowing what happens with the linear approximation enough to tell us about our original problem?

The real world rarely presents us problems with problems in the form of equations to be solved. Instead, steps like these are more typical of what it looks like to apply mathematical ideas: the search for mathematical models we can solve exactly, or exactly enough, but which still resemble the real situation we care about well enough to give us useful information.