Honors Math notes

mike miller eismeier

Contents

Do not try to remember theorems by the way I've numbered them here. I always try to remember a theorem by what it says, sometimes informally.

Chapter 1

Introduction to Honors Math

If I am being bold, the goal of the Honors Math sequence is to train you to *effectively learn and communicate mathematical understanding*. (We will do this by studying linear algebra and multivariable calculus, and hopefully get a good sense of those.) But what is *mathematical understanding*?

I can't pretend to give an answer that would satisfy every mathematician, but I can identify some recurring themes:

- Mathematicians like to find commonalities between similar examples and *abstract* them into a common statement.
- Mathematicians like to see *examples* which elucidate abstract statements.
- Mathematicians like to use different types of understanding to grasp at a particular idea or object (visual, computational, structural...)
- Mathematicians like to find patterns and *structure* to help them get a feel for how some ideas or objects 'work'.
- Mathematicians want to see *proofs* of claims and deductions.

These are all important, and we'll should see all of them over the next semester and year. The last, the idea of formal proof, is both math's bread and butter and its most subtle.

Question 1. What, exactly, is a *proof*?

Here are some guesses at a definition.

- A proof is a complete explanation for why something is true, using nothing that we don't already know.
- A proof is a sequence of logical deductions starting from a collection of axioms we take to be true.
- A proof is an argument that should be sufficiently detailed to convince any reader.

But it's hard for me to say that any one of these are the correct definition: the notion of 'complete explanation' depends on the reader; the second is transparently not how most math is communicated (we don't write math in pure symbols, and often steps are left implicit); the last is too broad (do we really mean any reader, even one who doesn't know the relevant terminology?)

I am not going to try to pin down an unambiguously correct definition of proof, which would be a major achievement in the philosophy of mathematics. While flawed, the above attempts at a definition give a good *sense* for what mathematicians mean when they talk about proofs, and a major part of this course is about learning how to write arguments that mathematicians will agree are 'complete proofs'.

Question 2. Why do we care about proving things?

Again, you are not going to get the same answer from every mathematician, but let me attempt to give a few answers.

- We are not satisfied by an argument that seems mildly convincing. We want to know that statements are **definitely, beyond true, without reasonable doubt**.
- A proof helps us communicate mathematical understanding to one another. Its formal nature allows one mathematician to write something which should be comprehensible to any other mathematician with sufficient background.
- When carefully writing down a proof, we find errors in our thinking. In correcting these, we enhance our thinking, and build better mental models of the objects we're working with.
- We want to know that a theorem is true, but we also want to know **why** it is true.

The last two bullet points hew closest to my own viewpoint.

If you're interested in seeing a professional mathematician's thoughts about the notions of "mathematics" and "proof", I strongly recommend Bill Thurston's On proof and progress in mathematics (link here). The first five sections are relatively non-technical; in section six he talks a bit about his early-career work (graduate and onwards) which can get somewhat technical.

Let me now briefly describe the structure of Honors A–B, and explain why it is structured that way.

- The first two weeks of the course will be a crash course in mathematical logic and proof-writing. I do not expect anyone to have mastered proof-writing by the end of these two weeks; the rest of the course (or perhaps the rest your mathematics education) is meant to help you get closer to that. But these will give you the tools you need to start your journey into writing formal arguments.
- In the third week, we will discuss the formal idea of sets and functions. On the one hand, we need these for everything that's to come. On the other hand, set theory is also a great place to get practice with your new formal argument skills.
- For the rest of Honors Math A, we'll learn about linear algebra. One can easily teach three or four very different linear algebra courses depending on your focus (say, abstract, geometric, or computational). We'll try to take a mix of all three of these, with the abstraction being important for getting comfortable with proofs. (It'll give us a playground to start off in.) This should cover roughly comparable material to UN2010 Linear Algebra.
- In the second semester, we learn multivariable calculus. To my mind, calculus is the study of linear approximation (though see Thurston's essay above for some other ideas of what differentiation is all about). There will be a small amount of ϵ-δ style analysis, but mostly the point of this term is to understand how the linear algebra we learned in the previous semester allows us to generalize the ideas and structure you're comfortable with from single-variable calculus. This contains the content of UN1205 Accelerated Multivariable Calculus or UN1201-1202 Calc III-IV, but remains proof-based.

Finally, let me make some remarks on collaboration. I think you should absolutely work together; in my own work I find my thinking better with collaborators, and I find that students who learn with other students do better with the material overall. I'll leave some time in class to help find groups of other students to work with. I'd advise between 2-4 people; too large a group and it's hard to communicate together much.

I often find that one learns things best when *explaining them to others*; it forces you to clarify and articulate your own understanding, and is a rather active learning process. If one of your classmates is explaining to you how something works, when they finish you should turn it on its head and *tell them your understanding*, from start to finish. This puts the shoe on the other foot — now you're forced to clarify and articulate your understanding!

Mathematical jargon

In the sections to come, you will see me use a number of words that might mean something different here than elsewhere in the English language. Let me try to pin down the relevant ones.

- An **axiom** is something we take for granted (assume to be true).
- A **definition** is a new name, together with a description of what properties must hold for an object to be called by that name. (A sentence like "We say the integer n is even if $n = m + m$ for some other integer m" is a definition; the clause after 'if' says what must be true for n to be called even.)
- A **proposition** is a claim which follows from a set of axioms, together with a proof that the proposition is true (meaning, that it follows from those axioms).
- A **lemma** is like a 'mini-proposition': it's a proposition you prove on the way to establishing some bigger, more important proposition.
- A **theorem** is like a 'mega-proposition': it's a proposition which is particularly important and worthy of note.
- A **proof** is (something like) an argument which derives the truth of a proposition from the assumed truth of some axioms, using only those axioms and other facts which we have established follow from those axioms.
- A **remark** is an off-hand comment which is not important for the main discussion.
- An **exercise** is something that I think it would be useful if you did on your own time before moving on to the next part of the notes.

If you feel I've missed some piece of math jargon which doesn't coincide with more standard English, let me know and I'll add it to this list.

Comments on TeX

I wrote these notes in LaTeX, which is the tool almost all professional mathematics is written in. LaTeX looks like a complicated programming language when you first see it, but it turns out to be relatively simple: "Write math expressions between dollar signs, like `$x^{2n} + 3$`, use the online tool 'DeTeXify' to find the codes for symbols I don't recognize, and google why things are going wrong when they are". I will introduce the TeX commands for new symbols as we go.

If you want to learn to write in TeX, I suggest starting by using a sample document on Overleaf, which is a great place to do TeX until you need a more serious TeX compiler. Play around a bit and see if you can get a knack for writing in this language. I can also give access to the TeX for my notes if you want to use it as a point of comparison.

I will give a small amount (up to 5%) of extra credit on homework assignments for writing them in TeX. This is just meant to incentivize you to try. It will not be a big difference grade-wise if you choose not to, or find TeX daunting.

Acknowledgements

Thanks to Shashank Choudhary, Peyton Chui, Lisa Faulkner, Jaylene Huang, Vishal Muthuvel, Aiden Sagerman, Benjamin Silverman, and Jazmyn Wang for finding errors in the course materials.

Chapter 2

Mathematical logic and formal proofs

In this chapter of the notes, we'll go over introductory mathematical logic and how to use it to manipulate mathematical statements (and write proofs). Some additional sources for relevant material (some of it disjoint from what we cover in these notes) are:

- There is the very, very talkative book "How to prove it" by Daniel Velleman. What we cover in the first chapter here he covers in about 80 pages. Still, if you find I'm too terse (or would like a more approachable reference), his book could be useful for you. I do not have a link, but ask me if you want to know how to find books online.
- Michael Hutchings' introduction to proofs notes (link here). About 27 pages.
- Michael Thaddeus' very brief cheat-sheet on mathematical logic (link here) and discussion of good proof-writing style (link here).

The assignments and the lectures assume that you have read my own notes, but you may find it useful to go over alternate sources (especially if you don't like my writing style!); some students find it useful to reference multiple sources for the same material during a single course. You may not even find my notes necessary, so long as you read something else that covers the right content.

I will attempt to give alternate sources for the material we cover at the beginning of each chapter.

2.1 Propositional logic

To start everything off, we need to understand the basic language of deductive logic (propositions and logical operations). This is not the most exciting thing we're going to do, but it's essential to have a firm grounding in mathematical logic to be able to think effectively about what a proof is.

2.1.1 Statements and logical operations

To discuss rules of logical inference, first we need to discuss what those rules are meant to refer to. The basic object of study is a *proposition* (or *logical statement*). I will be vague about what this means. As examples of things I would call propositions:

Example 1.

- P = "The number 9 is the square of an integer"
- Q = "Every injective map between sets is a bijection"
- A = "If a given day is a Monday or it's raining that day, then I'm going to be miserable that day"
- B = "There exists a smallest positive integer and it is 0"
- C = "October 3rd, 2022 is a Monday"

- D = "If today is a Monday or there is no lasagna, Garfield will be displeased"

$\Diamond$

To a proposition we can assign a truth value. In standard models of logic, a truth value is always either true or false. A proposition is one, and only one, of true or false. If one defines the terms above the way I'm used to, P, C, D are true, while Q, A, B are false (at least, if A is applied to me: I can enjoy a Monday just fine).

Remark 1. This is rather different from what I meant by 'proposition' when I discussed conventions on Page 7. There, I was saying that when you see a bolded, numbered proposition in these notes like "Proposition 14", it's a statement we've produced a proof of. This is an informal term, more or less what mathematicians call a theorem that isn't important enough to be named "theorem".

This is distinct from the mathematical logic statement of "proposition", which is just a statement with a truth-value (not necessarily true, and we have not necessarily written a proof). While this overuse of the word may be confusing, I am hoping it will be clear from context what is meant. $\Diamond$

Some of the statements above could be made out of smaller statements out of certain *logical operations*.

Definition 1. The following are the three basic logical operations.

a) If P is a statement, the statement "not P" (in symbols: $\neg P$; in TeX, `$\neg P$`) is a statement which is true if P is **false** and false if P is **true**. For instance, if P is the statement "October 3rd, 2022 is a Monday", then $\neg P$ is the statement "October 3rd, 2022 is not a Monday". Exactly one of these can be true. (In this case, P is true and $\neg P$ is false.)

b) If P and Q are statements, the statement "P or Q" (in symbols: $P \vee Q$; in TeX, `$P \vee Q$`) is true if at least one of P and Q is true. For instance, let P be the statement "The integer 207 is even" and Q be the statement "The integer 208 is even", while R is the statement "The integer 207 is odd". Then P is false, Q is true, and R is true. The statement $P \vee Q$ reads: "Either 207 is even or 208 is even" (true, since 208 is indeed even); the statement $P \vee R$ reads: "Either 207 is even or 207 is odd" (true, since 207 is odd). The statement $Q \vee R$ reads "Either 208 is even or 207 is odd" which is true **for two different reasons**: yes, 208 is even, and also 207 is odd. In mathematical writing, 'or' is the inclusive or, unless otherwise specified; if P and Q are both true, then $P \vee Q$ is still true, too.

c) If P and Q are statements, the statement "P and Q" (in symbols: $P \wedge Q$; in TeX, `$P \wedge Q$`) is true if both P is true and Q is true, and false if either P is false or Q is false. For instance, if P is the statement "Today is my birthday" and Q is the statement "I'm having a good day", then $P \wedge Q$ is the statement ("Today is my birthday and I'm having a good day"). It can only be true one day per year, and even then it's not always true — not only does it have to be my birthday, I have to actually be having a good one, too. $\Diamond$

You can combine these to create more complicated statements. For instance, "$Q \wedge \neg P$" can be read as "Q is true, and $\neg P$ is true", or it can be read more simply as "Q is true, and P is false". This statement holds whenever **both** Q is true and P is false, and this statement is false when Q is false or P is true (or both).

In some sense, these logical operations are enough to describe all possible logical operations you can think of.

Exercise. Try phrasing the relation of "Exclusive or" in terms of those above, where P xor Q is true if exactly one of P and Q is true. (I will do this in the next section, so you should try this exercise before moving on.)

2.1.2 Logical equivalence and truth tables

We say that two statements are *logically equivalent* if they share the same truth value. (This gets more interesting when you talk about infinite families of statements, which we will below.)

When we start with a handful of statements — say, just P and Q for now — and we form a new statement by applying some of our logical operations, the truth value of the resulting statement only depends on the truth values of the statements we started with, P and Q.

We can encode the dependence in the following diagram, a *truth table*. The left two columns indicate the possible truth values of P and Q (there are four possibilities in total), while the right columns indicate the truth value of some proposition made up out of P and Q. I include the three basic logical operations discussed above, as well as a couple of more complicated ones.

Example 2.

P	Q	$\neg P$	$P \vee Q$	$P \wedge Q$	P xor Q	$P \vee (Q \wedge \neg P)$	$(P \vee Q) \wedge \neg(P \wedge Q)$
F	F	T	F	F	F	F	F
T	F	F	T	F	T	T	T
F	T	T	T	F	T	T	T
T	T	F	T	T	F	T	F

To determine this table, I started with the truth-values on the left side and filled in what I know about how the statements are defined. For instance, to determine the truth-value of $P \vee (Q \wedge \neg P)$ when P is false and Q is true, I first notice that the statement "P or $(Q \wedge \neg P)$ is true" when at least one of its constitutient parts are true; P itself is false, so the only way this could be true is if $Q \wedge \neg P$ is true. For an 'and'-statement to be true, I need both of its parts to be true. Because we started by assuming Q is true and P is false (so that $\neg P$ is true), we see that $Q \wedge \neg P$ is indeed true. So the whole mess, $P \vee (Q \wedge \neg P)$, is true.

In symbols, this read $F \vee (T \wedge \neg F) = F \vee (T \wedge T) = F \vee T = T$. In the first part, I negated false to be true; in the second, I recognized that the statement "[true statement] and [true statement]" is true by definition of 'and', and in the last, I recognized that "[false statement] or [true statement]" is true by definition of 'or'. ◊

In the above example, notice that the truth values in the final column and in the third-to-final column (xor) are exactly the same. This proves:

Proposition 1. *The statements P xor Q and $(P \vee Q) \wedge \neg(P \wedge Q)$ are logically equivalent (no matter what statements P and Q are).*

In plain English, this states that the two claims "Exactly one of the statements P, Q are true" and "At least one of P or Q is true, but it is not the case thath both are true" are the same claims, which I hope is plausible even without the truth table.

2.1.3 Implications and equivalences

There are two more logical operations that are useful to us. We will see soon that they can be reframed in terms of the operations above, but they play a special role in the structure of mathematical logic itself.

Definition 2. Suppose P and Q are statements. The statements "$P \implies Q$" (in TeX, `$P \implies Q$`), which should be read as "If P is true, then Q is true" (or 'if P then Q'), is the statement defined by the following truth table.

P	Q	$P \implies Q$
F	F	T
T	F	F
F	T	T
T	T	T

That is, if P is false, then $P \implies Q$ is true; if P is true **and** Q is true, then $P \implies Q$ is true. The only way for $P \implies Q$ to be false is if P is true and Q is false. ◊

The truth table is kind of confusing, at least to me. Why should "$P \implies Q$" be true even when P is false? The setup in terms of truth tables kind of hides the way we usually think about implications.

This is partly because a statement like "$P \implies Q$" is usually proven *before one knows* the truth-value of P itself. It can be used as a tool to say something about Q when you later learn that P is true. For instance...

Example 3. Let P be the statement "It is raining today". This statement is really a bunch of statements, depending on what day it is (and where the proposition is being spoken). It isn't true at the time I'm writing it, but it may be true at the time you're reading it. I can't really ascertain the truth value.

Though you couldn't logically infer it from an earlier example, you might have guessed from above that I don't really like the rain. I get grumpy about having to hold an umbrella or walk home wet if I forgot one. Because one of those must happen if it's raining, the following statement is simply true:

If [it is raining today], then [I am going to be grumpy today].

This statement takes the form $P \implies Q$, where P is as above and Q is "I am going to be grumpy today." If I look out the window and it's raining today (so P is true), then this statement allows me to infer Q (that I am going to be grumpy about it).

If I look out the window and the sun is shining, then I can't infer anything at all (unfortunately, it's possible to be grumpy on sunny days, too). That doesn't mean that the statement "$P \implies Q$" is false, since it's a statement about what would happen if P were true! This statement just becomes boring if P is false, since then it has no predictive power. ◊

By comparing truth tables, we can see that the statement $P \implies Q$ can be describes in terms of our more simple logical operations. I'll leave verifying this to you as an exercise.

Proposition 2. *The statement $P \implies Q$ is logically equivalent to the statement $\neg P \vee Q$. That is, the statement "$P \implies Q$" is true exactly when the statement "P is false or Q is true" is true.*

Suppose you're trying to prove a statement of the form "$P \implies Q$". Because this is automatically true when P is false, this means you *get to assume* P and try to prove Q, using that piece of information.

Example 4. Let's actually prove a mathematical statement. If n is an integer (like $-1, 0, 1, \cdots$), we say n is *even* if $n = 2m$ for some other integer m. We say n is **odd** if $n = 2m + 1$ for some other integer m. [See footnote for a side comment which is not relevant to the rest of the discussion; I will occasionally put side comments in footnotes like this.][1]

Let's prove the statement "If n is even, then $n + 1$ is odd."

An implication is automatically true when the hypothesis is false, *so we can ignore that possibility.* We may as well assume n is even, meaning that $n = 2m$ for some integer m. Then $n + 1 = 2m + 1$, so that $n + 1$ is odd (by definition of odd!)

Now let's prove the statement "If n is odd, then $n + 1$ is even."

Again, when trying to prove an implication, we are allowed to **take for granted** the 'if' part of the statement. (That's why we like to prove conditionals! They give us something to work with.) So we may assume $n = 2m + 1$ for some integer m.

Then $n + 1 = (2m + 1) + 1 = 2m + 2 = 2(m + 1)$. Because m is an integer, $m + 1$ is an integer too. So by definition, $n + 1 = 2k$ for some integer k — the integer k being $m + 1$. ◊

There is one more logical operation that is particularly important for mathematicians. It turns out it's logically equivalent to something we can cook up out of things we already know, but it's so important it deserves its own name and symbol.

Definition 3. We say that P is true **if and only if** Q is true (written $P \iff Q$ or `$P \iff Q$`) for the statement with the following truth-table:

[1] At this point, it's not even clear that an integer has to be exactly one of even or odd. What if it is neither? What if it is both? You might be able to give me an argument that this is not true using, say, the division algorithm, but that's not something we've established in this example! If you defined the notion of 'even' and 'odd' the same way for rational numbers, every rational number would be both even and odd, since you could just take $m = n/2$ and $m = (n - 1)/2$, respectively.

P	Q	$P \iff Q$
F	F	T
T	F	F
F	T	F
T	T	T

That is, $P \iff Q$ is true if P is logically equivalent to Q. ◊

Exercise. Check that the statement $P \iff Q$ is logically equivalent to the statement $(P \implies Q) \wedge (Q \implies P)$, by comparing truth tables. In practice, you'll see that you usually prove $\iff$ statements by proving both statements $P \implies Q$ and $Q \implies P$, one at a time. In plain words, I would say that $P \iff Q$ means "P is true precisely when Q is true", while $(P \implies Q) \wedge (Q \implies P)$ means in plain words "If P is true, then Q is true, and if Q is true, then P is true." Informally, this seems like the same plain-language statement as the previous.

Remark 2. Why the phrase "if and only if"? Let's try to understand this in plain language. "If P, then Q" means that when P is true, Q has to be true as well. This could be rephrased as "P is true **only if** Q is true". That is, if P is true, then Q is forced to be as well. The "only if" suggests that we've ruled out a possibility: if Q were not true, then P couldn't be.

On the other hand, "P is true **if** Q is true" is the implication $Q \implies P$. (It's taken the sentence "If Q is true, then P is true", and moved that first clause to the end of the sentence.)

So we say "if and only if" to mean both of these implications hold: both $P \implies Q$ (P only if Q) and $Q \implies P$) (P if Q).

◊

Let's prove our first example of an 'if and only if' statement.

Proposition 3. *An integer n is even if and only if the integer $n+1$ is odd.*

Proof. We are trying to prove a statement of the form $P \iff Q$, which we do by proving that $P \implies Q$ and $Q \implies P$, one at a time.

We have already proven the implication $P \implies Q$, here

$$[\text{the integer } n \text{ is even}] \implies [\text{the integer } n+1 \text{ is odd}].$$

What we need to show is the 'reverse implication' (sometimes called the 'converse')

$$[\text{the integer } n+1 \text{ is odd}] \implies [\text{the integer } n \text{ is even}].$$

It's not so hard to do so, by the same technique — but it's still a new argument. If $n+1$ is odd, then $n+1 = 2m+1$ for some m, so

$$n = (n+1) - 1 = (2m+1) - 1 = 2m,$$

and thus n is even.[2] □

2.1.4 Standard equivalences

Suppose you wanted to show that the two statements $(P \vee Q) \wedge (P \vee (R \wedge S))$ and $P \vee (Q \wedge R \wedge S)$ are logically equivalent. In principle, you could write out the two truth tables and compare them, but these have *sixteen* rows — you're going to get bored pretty quick. In practice, there are a list of standard equivalences one can use to show two statements are equivalent (instead of writing out their truth tables).

[2]One reasonable objection might be: "This is the same as the previous argument. You're just reversing your steps." This is true! But you're using a non-trivial fact here, that the operations of addition and subtraction can be used to undo one another. Once you know that, you can turn a proof in one direction into a proof for the other direction. But the directions themselves remain logically distinct.

Proposition 4. *The following propositions are logically equivalent, as can be seen by comparing their truth tables:*

$$P \vee T \equiv T \tag{2.1}$$
$$P \vee F \equiv P \tag{2.2}$$
$$P \wedge T \equiv P \tag{2.3}$$
$$P \wedge F \equiv F \tag{2.4}$$

$$P \vee Q \equiv Q \vee P \tag{2.5}$$
$$P \wedge Q \equiv Q \wedge P \tag{2.6}$$

$$(P \vee Q) \vee R \equiv P \vee (Q \vee R) \tag{2.7}$$
$$(P \wedge Q) \wedge R \equiv P \wedge (Q \wedge R) \tag{2.8}$$

$$\neg\neg P \equiv P \tag{2.9}$$
$$\neg(P \vee R) \equiv \neg P \wedge \neg R \tag{2.10}$$
$$\neg(P \wedge R) \equiv \neg P \vee \neg R \tag{2.11}$$

$$P \vee (Q \wedge R) \equiv (P \vee Q) \wedge (P \vee R) \tag{2.12}$$
$$P \wedge (Q \vee R) \equiv (P \wedge Q) \vee (P \wedge R). \tag{2.13}$$

These break up into rather distinct-feeling groups.

- The first four describe the behavior of $\vee$ and $\wedge$: for the first, if one of the statements is true the 'or' statement is true, while if one of the statements is false the 'or' reduces to the other statement; similarly for 'and' statements.
- The next two assert that it doesn't matter what order we list statements when we say "or" or "and". For instance, "$P \vee Q$ being true means "at least one of P or Q is true", and that's the same logical idea as "at least one of Q or P is true"!
- The next two assert when we write strings of "or"s or strings of "and"s, we can do so without parentheses: for instance, the meaning of $P \wedge Q \wedge R \wedge S$ is unambiguous (this statement is true precisely when all of P, Q, R, S is true). Notice that here all the operations are the same: when I talk about the way the operations $\vee$ and $\wedge$ interact, we have to be careful about our bracketing.
- The next three talk about the behavior of the negation operation. The tenth and eleventh should be pretty plausible. For instance, for (6): what does it mean to say 'P or R' is false? For 'P or R' to be true, *at least one of* P, R should be true. If this is false, then **neither** can be true. So if $P \vee R$ is false, then both P and R are false — that is, $\neg(P \vee R)$ is true precisely when $(\neg P) \wedge (\neg R)$ is true.

  The ninth is rather special. It is sometimes called the *law of the excluded middle*. It asserts that a statement **has to be** either true or false, and it cannot be both of them. If P is not *not true*, that is, if P is not false, then P is true. Propositional logic is not fuzzy: a statement is true, or a statement is false, and there is no in-between. It is essentially equivalent to the statement that $P \wedge \neg P$ is false — you cannot have both P and $\neg P$ be true. On the other hand, $P \vee \neg P$ is simply true.

- The last couple are called *de Morgan's laws*, or sometimes just *distributivity*. You should think of them as somehow analogous to the distributivity laws for addition and multiplication:

$$p \cdot (q + r) = (p \cdot q) + (p \cdot r).$$

Exercise: Check de Morgan's laws by showing that both sides of the equivalence have the same truth tables. Then convince yourself, in plain language, that these 'should' be true.

Let's use these to show how you can show some statements are equivalent without just writing truth tables, using these common logical equivalences.

Example 5. Let's show

$$(P \vee Q) \wedge (P \vee (R \wedge S)) \equiv P \vee (Q \wedge R \wedge S).$$

My first thought is that this expression looks like $(P \vee Q) \wedge (P \vee R')$, where R' here is $R \wedge S$. That expression shows up in de Morgan's first law! We know that

$$(P \vee Q) \wedge (P \vee R') \equiv P \vee (Q \wedge R') = P \vee (Q \wedge (R \wedge S)).$$

Lastly, because $Q \wedge (R \wedge S) \equiv (Q \wedge R) \wedge S$, it doesn't matter how I group the 'and' terms: I can just write this as $P \vee (Q \wedge R \wedge S)$, as desired. $\Diamond$

Here's a much more intricate example. In this case, the fastest approach would probably be to compare truth tables — but I want to point out that we **could** still handle it using the standard equivalences above. I wanted to show how to use these equivalences in a couple of different ways.

Example 6. Let's prove that $P \iff Q$ is logically equivalent to the statement $(P \wedge Q) \vee (\neg P \wedge \neg Q)$. The final statement says: "This statement is true if P and Q are both true, or if P and Q are both false." That should track: to say that $P \iff Q$ holds means that P, Q have the same logical content (one is true when the other is, one is false when the other is).

Let's argue this formally. By definition,

$$P \iff Q = (P \implies Q) \wedge (Q \implies P) \equiv (\neg P \vee Q) \wedge (\neg Q \vee P).$$

Here all I did was write down that an $\iff$ means that both implications $P \implies Q$ and $Q \implies P$ hold, while the next step uses the equivalence from Proposition 2 between $P \implies Q$ and $\neg P \vee Q$.

Now we want to show that

$$(\neg P \vee Q) \wedge (\neg Q \vee P) \equiv (P \wedge Q) \vee (\neg P \wedge \neg Q).$$

The two sides look pretty different from one another. This is a good place to mention one of my favorite proof strategies: *mess around with the tools I have and see what happens.*

On the left side, I know from de Morgan's law how to "distribute" things across that wedge. Now, I don't know yet that this should be a good idea, but I'm going to apply it here and see what happens:

$$(\neg P \vee Q) \wedge (\neg Q \vee P) \equiv [(\neg P \vee Q) \wedge \neg Q] \vee [(\neg P \vee Q) \wedge P].$$

Here I thought of this statement as $R \wedge (S \vee T)$, where $R = (\neg P \vee Q)$ and $S = \neg Q$ and $P = T$; because

$$R \wedge (S \vee T) \equiv (R \wedge S) \vee (R \wedge T)$$

by de Morgan's law, that gives the expression above.

This last statement looks more complicated, but promising, because it looks like it can still be simplified further. Let's look at one of the pieces, $(\neg P \vee Q) \wedge \neg Q$. If I apply de Morgan's laws to **that**, it gives me

$$(\neg P \vee Q) \wedge \neg Q \equiv \neg Q \wedge (\neg P \vee Q) \equiv (\neg Q \wedge \neg P) \vee (\neg Q \wedge Q).$$

(In the first step I rearranged the terms around the 'and' so it looks more like the phrasing of de Morgan's law above.) Now this is very promising, because I see that $\neg Q \wedge Q$ term and I know: "FALSE!" This statement literally means "Q is false and Q is true", which is simply not possible. I can write $\neg Q \wedge Q \equiv F$. Then

$$(\neg Q \wedge \neg P) \vee (\neg Q \wedge Q) \equiv (\neg Q \wedge \neg P) \vee F \equiv \neg Q \wedge \neg P \equiv \neg P \wedge \neg Q.$$

The statement "$R \vee F$" is true precisely when at least one of R and F is true. But R is false. So it just simplifies to "R is true".

All in all, we've simplified

$$[(\neg P \vee Q) \wedge \neg Q] \vee [(\neg P \vee Q) \wedge P] \equiv (\neg P \wedge \neg Q) \vee [(\neg P \vee Q) \wedge P].$$

Now you could argue like I did above to see that

$$(\neg P \vee Q) \wedge P \equiv P \wedge (\neg P \vee Q) \equiv (P \wedge \neg P) \vee (P \wedge Q) \equiv F \vee (P \wedge Q) \equiv P \wedge Q,$$

so that our final statement is equivalent to

$$(\neg P \wedge \neg Q) \vee [(\neg P \vee Q) \wedge P] \equiv (\neg P \wedge \neg Q) \vee (P \wedge Q) \equiv (P \wedge Q) \vee (\neg P \wedge \neg Q),$$

which is what we wanted to show in the first place. ◊

Truth tables are certainly much more efficient in this argument — but less efficient in the first! *It's worth having both tools in your toolkit.*

2.2 Logical equivalences and proof strategies

In this section we're going to describe how logical equivalences give rise to different ways of proving things. (The idea: we take a statement, rewrite it in a logically equivalent but more-accessible form, and then prove the more accessible form.) In most of these cases the specific kinds of propositions we're trying to prove are *implications* $P \implies Q$.

2.2.1 Contrapositives

Take the statement "If it is raining, then I am sad". The only thing such an implication tells me is that **when** it is raining, I necessarily must be sad. We can turn this on its head: "If I am not sad, then it's not raining!" If I'm not sad, there's no possible way it could be raining. We know that if it's raining, I'm sad. But I'm not sad, so it must not be raining!

This particular logical equivalence is called the **contrapositive**. Let's record it and give a more formal proof.

Proposition 5 (An implication is the same as its contrapositive)**.** *If P and Q are logical statements, the local statements $P \implies Q$ and $\neg Q \implies \neg P$ are logically equivalent.*

Proof. You can write down the truth tables, if you want (that's basically what we did in the discussion above). Here's a proof in terms of the standard logical equivalences and Proposition 2. We have

$$[\neg Q \implies \neg P] \equiv [\neg(\neg Q) \vee \neg P] \equiv [Q \vee \neg P] \equiv [\neg P \vee Q] \equiv [P \implies Q].$$

In the first and last step we used Proposition 2, and in the middle we used that $\neg\neg Q \equiv Q$ and that the order of the two parts of an "or" statement doesn't matter — Proposition 4 (2.9) and (2.5), respectively. □

Example 7. Imagine n is an integer. Suppose you want to prove the statement: "If n^2 is even, then n is even." This is kind of hard. From what's given, I can say that $n^2 = 2m$ for some integer m, and I want to say $n = 2k$ for some integer k. But why should the square root of $2m$ be an integer?

This is a statement of the form $P \implies Q$. It seems like Q is the stronger of the two statements, and maybe easier to work with; I want to 'flip this around' and work with the contrapositive $\neg Q \implies \neg P$, which is logically equivalent to $P \implies Q$. That is, it suffices to prove: "If n is **not** even, then n^2 is **not** even."

This is suddenly a lot easier! Let's take for granted something you'll be able to prove after we cover induction: *every integer is either even or odd, but not both.* Now.

If n is not even, it must be odd. So $n = 2m+1$ for some integer m. Then $n^2 = (2m+1)^2 = 4m^2+4m+1 = 2(2m^2+2m)+1$. So n^2 is odd. Because an integer cannot be both even and odd, we see that n^2 is not even — which is what we wanted to show in our proof of the contrapositive $\neg Q \implies \neg P$.

Because the contrapositive $\neg Q \implies \neg P$ of the original statement is equivalent to the original statement, we're finished! ◊

Remark 3. The contrapositive $\neg Q \implies \neg P$ should not be confused with the **converse** $Q \implies P$. The contrapositive is equivalent to $P \implies Q$; the converse is a logically different statement.

As an example, take P to be the statement "x is an integer", and Q be the statement "x^2 is an integer". The table below shows the three different implications we wrote above.

$P \implies Q$	an implication	If x is an integer, then x^2 is an integer
$\neg Q \implies \neg P$	its contrapositive	If x^2 is **not** an integer, then x is **not** an integer.
$Q \implies P$	its converse	If x^2 is an integer, then x is also an integer.

The statement in the second row is equivalent to the one in the first row (and is also true). The statement in the last row is simply false; for instance, $x = \sqrt{2}$ is a counterexample.

When you try to prove a statement of the form $P \iff Q$, you have to prove both of the (independent) implications $P \implies Q$ and $Q \implies P$. If you want, using the contrapositive makes this equivalent to proving both $P \implies Q$ and $\neg P \implies \neg Q$. But you'd still have to write two proofs. ◊

2.2.2 Proving with 'or' and 'and'

There are four particular kinds of statements I'd like to look at in this section — two simpler than the others:

$$(P \wedge Q) \implies R \qquad P \implies (Q \wedge R) \qquad\qquad (P \vee Q) \implies R \qquad P \implies (Q \vee R).$$

Let's start with the somewhat simpler ones.

Example 8. Let's look at a statement of the form $P \implies (Q \wedge R)$ — for instance, "If n is the square of an integer, then $n \geqslant 0$ and $n \neq 2$."

Here, we're allowed to assume that $n = m^2$ is the square of an integer, and our goals are to prove **both** that $n \geqslant 0$ and $n \neq 2$. First, observe that either n is negative, zero, or positive; the product of two negative numbers is positive, as is the product of two positive numbers, and the product of zero with itself is zero. In any case, we've shown that the square of an integer has $n \geqslant 0$.

To see that $n \neq 2$, observe that the larger an integer is in absolute value the larger its square is, and $4 = (\pm 2)^2$ is larger than 2. The only integers which could square to 2 are $-1, 0, 1$, and those square to $0, 1$. So 2 is not the square of any integer, and in particular $n \neq 2$.

Punchline: The statement $P \implies (Q \wedge R)$ is logically equivalent to $(P \implies Q) \wedge (P \implies R)$. We used the hypothesis P to prove Q, and then separately we used the hypothesis P to prove R. That's how you'll always prove a statement of the form $P \implies (Q \wedge R)$: proving each of the statements $P \implies Q$ and $P \implies R$ separately.

The story is the same for statements of the form $(P \vee Q) \implies R$. To say "P or Q implies R" means you're allowed to assume *one of* P or Q is true, and from that conclude that R is true. But you don't know which one it is! If you're going to prove this in general (where maybe P is true but Q is false, or maybe P is false but Q is true, or maybe they're both true), you have to prove both that $P \implies R$ and that $Q \implies R$.

For instance, to prove "If n is odd **or** n is divisible by four, then n cannot be written as $4k + 2$ for any integer k," I'll just show the desired claim for both n odd and n divisible by four. If n is odd, then n is not $4k + 2 = 2(2k + 1)$ because this number is even, and a number can only be one of even or odd. (I promise you'll prove this soon!) On the other hand, if $n = 4m$ is divisible by four, then n cannot be $4k + 2$: this would give $4m = 4k + 2$ or $4(m - k) = 2$, but 2 is not divisible by four.

Punchline. The statement $(P \vee Q) \implies R$ is logically equivalent to $P \implies R$ and $Q \implies R$, and you'll prove the more complicated statement by proving the two simpler implications $P \implies R$ and $Q \implies R$ separately.

◊

Now let's look at the two slightly more intricate ways to include "and" and "or" statements in an implication.

Example 9. Let's look at a statement of the form "$(P \wedge Q) \implies R$" — for instance. In principle, this means that you may assume that P and Q are **both** true, and do your best to prove R from that information. For instance, if we wanted to prove "If n is an even integer **and** $n > 0$, then $n \geqslant 2$", we start with the knowledge that $n = 2m$ and (because we also know $n > 0$) that $m > 0$. Because m is an integer, this means $m \geqslant 1$, so that $n = 2m \geqslant 2$.

There are some statements where this direct approach is not sufficient, and the best approach is to use a sort of contrapositive. For instance, consider:

"If n is an even integer **and** n is not divisible by 4, then n is twice an odd integer."

In this statement, we can take for granted that $n = 2m$ for some integer m, and also that n is **not** $4k$ for any integer k. Because $n = 2m$, it seems like a reasonable goal to show that m is an odd number. I'm going to once again take for granted that an integer has to be exactly one of even or odd, so that this is the same as showing that m is not even.

Where we are is the statement: "Let $n = 2m$ for m an integer. If n is not divisible by 4, then m is not divisible by 2." With there being "nots" in front of both terms here, this seems like a good place to apply the contrapositive.

We still have that $n = 2m$ for m an integer. The contrapositive is the statement "Let $n = 2m$ for m an integer. If m is divisible by 2, then n is divisible by 4." Now we have way more leverage. If m is divisible by 2, then $m = 2k$ for some k, so $n = 2(2k) = 4k$, and we have shown that n is divisible by 4. We're finished!

Punchline 1. To prove a statement of the form $(P \wedge Q) \implies R$, you assume both P and Q are true, then do your best to use both of those pieces of information to prove R.

Punchline 2. A variant of the contrapositive holds: any statement $(P \wedge Q) \implies R$ is logically equivalent to $(P \wedge \neg R) \implies \neg Q$ or $(Q \wedge \neg R) \implies \neg P$. If you know — or take for granted — that P is true, then this literally is taking the contrapositive.

Here, we passed to the (equivalent) statement "If n is an even integer but n is not twice an odd number, then n is divisible by 4." Then $n = 2m$ and m is not odd, so it's even ($m = 2k$), and thus $n = 4k$.

$\Diamond$

The most interesting, I think, is the following; you really need to use logical equivalences to make any progress.

Example 10. How do we prove a statement of the form $P \implies (Q \vee R)$? For instance, let's try the statement "If x is a real number, then $x < 1$ or $x = y^2$ is the square of another real number." (Let me take for granted that every $x \geqslant 0$ has a square root; I just want to make a point about logic.)

I find this very difficult to do anything with, because **I don't know what I'm trying to prove**. I'm told to prove one thing, or possibly another thing, with no direction about which to choose or why. How do I prove something indefinite?

One way to think about this would be in terms of *case analysis*. If x is a real number which has $x < 1$, then we already know the conclusion holds: we just needed to show that $x < 1$ or that $x = y^2$, and we know that $x < 1$ is true by assumption.

In the other possible case $x \geqslant 1$, the first statement in the 'or' ($x < 1$) is false. If we want to show that "$x < 1$ or $x = y^2$" is true, we have to show that $x = y^2$ for some y. Because $x \geqslant 1 \geqslant 0$, we took this for granted already: we can take $y = \sqrt{x}$.

What we did amounted to the observation that our claim is equivalent to "If x is a real number **and** $x \geqslant 1$, then $x = y^2$ is the square of another real number". That is, $P \implies (Q \vee R) \equiv (P \wedge \neg Q) \implies R$. If we want to prove that some proposition P implies an "or" statement, we can assume that one of the "or" statements is **false** and try to prove the other one — because if the first "or" statement was true, we'd already be finished. $\Diamond$

Let's record the logical equivalences we used above as a proposition.

Proposition 6. *The following statements are logically equivalent.*

$$P \implies (Q \wedge R) \equiv (P \implies Q) \wedge (P \implies R). \tag{2.14}$$
$$(P \vee Q) \implies R \equiv (P \implies R) \wedge (Q \implies R). \tag{2.15}$$

$$P \implies (Q \vee R) \equiv (P \wedge \neg Q) \implies R \equiv (P \wedge \neg R) \implies Q. \tag{2.16}$$
$$(P \wedge Q) \implies R \equiv (P \wedge \neg R) \implies \neg Q. \tag{2.17}$$

Exercise: Verify these, either using truth tables or using standard logical equivalences.

2.2.3 Proof by contradiction

Suppose we're trying to prove the statement $P \implies Q$. First you try to write down a direct proof. Playing with the assumption P, you can't seem to really make progress; playing with $\neg Q$ you don't get anywhere. But P and $\neg Q$ seem somehow at odds with one another, and when *both of them* are true you can see things you couldn't from just one.

This situation is handled by the idea of proof by contradiction:

Proposition 7. *The statement $P \implies Q$ is logically equivalent to the statement $P \wedge \neg Q \implies F$. That is, if from P and $\neg Q$ you can derive something preposterous, the statement $P \implies Q$ is true.*

Let's see this done in a famous example.

Example 11. Let's prove the statement "If x is a rational number, then $x^2 \neq 2$." Neither the hypothesis nor the conclusion seems to get me far by itself: if $x^2 = 2$, great, but why does that mean x is not rational? And if x **is** a rational number p/q, how do I show $(p/q)^2$ is not 2?

If I assume **both** of them, I can make some progress. Suppose $x = p/q$ is a rational number written in lowest terms (that is, there is no $d > 1$ which divides both p and q); every rational number can be written that way. **Towards a contradiction**[3], we also assume that $x^2 = 2$.

Our goal is to show that this contradicts something from earlier, so that this new hypothesis must be false and in fact $x^2 \neq 2$.

Let's combine these. If $x = p/q$ (where p and q have no common divisor), then $x^2 = p^2/q^2$; so if $x^2 = 2$, after cross-multiplying we see that $p^2 = 2q^2$. Now, because p^2 is even, we must have p divisible by 2 (as otherwise p^2 would be odd). But then $p = 2m$ for some m, and $[2m]^2 = 2q^2$, so $4m^2 = 2q^2$, so $2m^2 = q^2$.

Now we see that q^2 is even, so that q must be divisible by 2! This **contradicts** the fact that (p, q) were chosen to have no common divisor $d > 1$. Thus our new assumption $x^2 = 2$ must be at fault.

We have given a proof (by contradiction) that if x is rational, we have $x^2 \neq 2$ (or, as people more often say, "$\sqrt{2}$ is irrational"). ◊

Remark 4. Usually, a majority of the proofs I see on an upper-level math assignment are proofs by contradiction, while a somewhat small minority of proofs in published mathematics are proofs by contradiction. What's the difference?

- I understand why students often prefer to write proofs by contradiction. They're attractive: they give you two pieces of information to play with (P and also $\neg Q$). For a student who's looking to make progress in any direction, this is great (and I encourage you to try it!) But oftentimes, the proof that ends up being produced is something along the lines of "Towards a contraduction, assume P is true and Q is false. Because [argument that only uses P being true], we can see that Q is true. This contradicts the assumption that Q is false, so Q must have been true," so the phrasing in terms of contradiction obfuscates the real argument you want to make. (Similarly common is really a proof by contrapositive

[3]This means that I think this assumption will be erroneous (impossible). I introduce it because I think that if I do, I will eventually be able to prove something false. Call the thing that I assume (but think is false) "P". If I prove $P \implies F$ (so assuming P is true leads me to a contradiction), then this gives a proof of $\neg P$. So the conclusion of this argument is going to be: "Contradiction! We must have $x^2 \neq 2$."

which is just phrased as a contradiction.) For instance, if you prove $P \implies R$ and $\neg Q \implies \neg R$ and say "Then assuming P true and Q false gives a contradiction, because R and $\neg R$ would both be true!" you've used a superfluous contradiction; because you proved $\neg Q \implies \neg R$, you also proved $R \implies Q$. Combining this with $P \implies R$ you see that $P \implies Q$.

- Mathematicians usually want to phrase their arguments so that each individual step is a Lemma they can now use elsewhere later. In a proof by contradiction, you start by assuming something erroneous — if you start by assuming $P \wedge \neg Q$, which turns out to be false. If on the way to deriving your contradiction, you prove something like R, **you can't use that outside the proof you're writing**. Maybe your proof of R used your erroneous hypotheses! On the other hand, if you were trying to show $P \implies Q$ directly by assuming P and attempting to prove Q, and on the way you establish R — then you've given a correct proof of $P \implies R$ which you can use later in other contexts.

  Mathematicians tend to prefer writing *direct* proofs, deriving contradictions as little as possible; it is often cleaner and more clear to simply say how to go from A to B, and you don't have to throw away your intermediate work when you're done. I like Joel David Hamkin's answer (link here) about this.

In practice, mathematicians may *find* a proof by starting a proof by contradiction, and then rewrite it into a more clear direct proof.

Interestingly, there are some theorems (for instance, the intermediate value theorem) which simply cannot be proved without applying proof by contradiction. Mathematicians have formalized this by developing a version of logic (called "constructive logic" among other names) for which $\neg\neg P$ is not equivalent to P; knowing that P is not false does not tell you that P is true, so a proof by contradiction (which establishes that P is not false) doesn't get you where you want to go. In constructive logic, the intermediate value theorem is false! $\Diamond$

2.3 Quantifiers and induction

In many cases above, the statements I talked about were not really a single statement, but rather a *family* of statements depending on some parameter. For instance, the statement "n is even" refers to an integer n; I might refer to the family of statements $P(n)$, where for each integer n we have the statement "n is even". So $P(0)$ is true (0 is even), but $P(1)$ is false (1 is not even), while $P(2)$ is true, and so on.

Given a family of statements, there are two important ways of constructing statements that refer to the *whole family*:

- If we have a family of statements $P(x)$, we can say "For all x, $P(x)$ is true." This asks that the statement $P(x)$ is **always true**.

- If we have a family of statements $P(x)$, we can also say "For at least one x, $P(x)$ is true." This asks that the claim is *sometimes true*, that there's at least one example of an isntance when the claim is true.'

These are so important they get their own notation.

Definition 4. Suppose we have a family of statements $P(x)$ which depend on some parameter x. The statement "$\forall x P(x)$ (in TeX, `$\forall x P(x)$`; in plain language, "for all x, $P(x)$") is the statement which is true precisely when **all of the statements** $P(x)$ **are true**.

The statement "$\exists x P(x)$ (in TeX, `$\exists x P(x)$`; in plain language, "There exists an x such that $P(x)$" is the statement which is true precisely when **at least one of the statements** $P(x)$ **are true**. $\Diamond$

I implicitly used these many times in the examples earlier in the notes. For instance, the statement $P(n)$ = "n is even" means "there exists some integer m so that $n = 2m$". If $Q(n, m)$ is the statement $n = 2m$ (where n and m are understood to be integers), then the statement $P(n)$ is

$$P(n) \equiv \exists_m Q(n, m).$$

(Make sure you understand this!)

The symbols $\forall$ and $\exists$ are called **quantifiers**. They quantify over some 'set' of propositions (more on sets next week). Sometimes when we want to specify more clearly exactly what exactly they quantify over, we write something like $\forall_{x\in\mathbb{Z}}$. The expression "$x \in \mathbb{Z}$" should be read as "x is in Z", or better "x is an integer". The symbol $\mathbb{Z}$ (TeX: `$\mathbb Z$`, plain language: "the integers", from German 'zahlen') refers to the set of integers, while $\in$ (TeX: `$\in$`, plain language: "is in") denotes membership — it says that x is a member of the integers (that is, it is an integer).

Above I defined a statement $Q(n, m)$ for integers n and m (the statement being $n = 2m$). Notice that this statement makes sense more generally — for instance, it makes sense for n and m in the real numbers (denoted $\mathbb{R}$, TeX `$\mathbb R$`) or the rationals (denoted $\mathbb{Q}$, TeX `$\mathbb Q$`). I will write $P_{\mathbb{R}}(x)$ for the statement "There exists a real number y so that $x = 2y$" — or in other words,

$$P_{\mathbb{R}}(x) = \exists_{y\in\mathbb{R}} Q(x, y).$$

Now this statement is **always true**: given any real number x, we can take $y = x/2$, which always gives another real number. (I wrote this in TeX as `$P_{\mathbb R}(x) = \exists_{y \in \mathbb R} Q(x,y)$\.`)

Thus $\forall_{x\in\mathbb{Z}} P_{\mathbb{Z}}(x) = \forall_{x\in\mathbb{Z}}\exists_{y\in\mathbb{Z}} Q(x, y)$ is false (for instance, $P_{\mathbb{Z}}(1)$ is false — 1 is not even) and

$$\forall_{x\in\mathbb{R}} P_{\mathbb{R}}(x) = \forall_{x\in\mathbb{R}}\exists_{y\in\mathbb{R}} Q(x, y)$$

is true. The domain we quantify over matters, and can change whether statements are true or not.

Remark 5. In practice, when proving a statement such as $\forall_x P(x)$, you do so by showing that the statement is true for some arbitrarily chosen x (which you use no special information about, so that your argument applies to any x whatosever). When the index set is the natural numbers ($\mathbb{N}$), we will discuss a particular proof strategy at the end of the day that helps you prove statements like $\forall_{n\in\mathbb{N}} P(n)$ indexed on the natural numbers $\mathbb{N} = \{0, 1, 2, \cdots\}$. ◊

2.3.1 Nested quantifiers

In the discussion above you saw your first example of *nested quantifiers*, statements which use one or more quantifiers of different types. These appear relatively often in math (the definition of continuity is a triply nested quantifier), so it's worth saying one or two things to be clear about them.

The statement $\forall_x\exists_y Q(x, y)$ means "For all x, you can find a y (which maybe depends on x!) so that $Q(x, y)$ is true." For instance, if $Q(x, y)$ is the statement $y - x = 1$, then $\forall_x\exists_y Q(x, y)$ is true: for each x, I can find a y so that $y - x = 1$ (take $y = x + 1$). Notice that the y I found depended on what x was, but that doesn't matter — the point is that for any fixed x, I can find some y for which the statement is true for that particular x.

If I reverse the order of these quantifiers, I get something differnt. The statement $\exists_y\forall_x Q(x, y)$ means "There exists some y so that, for this particular y and **all** x, the statement $Q(x, y)$ is true." This is **very, very different** from the previous statement. For example... If $Q(x, y)$ is the statement $y - x = 1$, then "$\exists_y\forall_x Q(x, y)$ reads: "There is some special number y so that for all numbers x, we have $y - x = 1$". But this is not true! It doesn't matter what y is, there's always some number x for which $y - x$ is not one. For instance, $Q(y, y)$ is always false (since $y - y = 0 \neq 1$), no matter what y is.

For an example of a true statement of the form $\exists_x\forall_y R(x, y)$, take $R(x, y)$ to be "$xy = 0$". Then $\exists_x\forall_y Q(x, y)$ means "There exists some number x so that $xy = 0$ is always true, no matter what number y is." And this is true: $x = 0$ is such a number (because $0y = 0$ is always true, no matter whath y is). Notice that in this case $\forall_y\exists_x R(x, y)$ is also true, and seems easier to prove. This says that for all numbers y, we can find some number x (which maybe depends on y) so that $xy = 0$. But we already know how to find one, and our choice doesn't depend on y: we can just take $x = 0$.

This is a general phenomenon.

Proposition 8 ($\exists\forall$ is stronger than $\forall\exists$)**.** *For any family of statements $Q(x,y)$, the statement*

$$\exists_x \forall_y Q(x,y) \implies \forall_y \exists_x Q(x,y)$$

is true.

Proof. We're trying to prove an implication. We're allowed to assume the hypothesis is true, and our goal is to prove the conclusion. If $\exists_x \forall_y Q(x,y)$ is true, this means that there is some special $x = x_0$ so that $Q(x_0, y)$ is always true, no matter what y is.

The statement $\forall_y \exists_x Q(x,y)$ means that for every y, there is some x which maybe depends on y (which we often denote by $x = x(y)$, to explain that it might depend on y) so that $Q(x,y)$ is true. But we know this, because $Q(x_0, y)$ is true for all y; we can take $x(y) = x_0$. □

Punchline: The order of these quantifiers matters!

Remark 6. We don't consider $\forall_x \forall_y P(x,y)$ or $\exists_x \exists_y P(x,y)$ to be nested quantifiers, because these can be rephrased as a single quantifier (over a different indexing set). For instance, "$\exists_{x\in\mathbb{Z}} \exists_{y\in\mathbb{Z}} \big[x^2 + xy + y^2 + x = 1\big]$ says "There exists an integer x, so that there exists a y, so that $x^2 + xy + y^2 + x = 1$ is true." But we could just rephrase this to: "There exists a pair of integers (x,y) so that $x^2 + xy + y^2 + x = 1$ is true." That is, this is the same as $\exists_{(x,y)\in\mathbb{Z}^2} \big[x^2 + xy + y^2 + x = 1\big]$, which is an existential quantifier indexed on pairs of integers. ♢

2.3.2 Quantifiers and logical operations

Before moving on, I want to discuss how quantifiers interact with the standard logical operations. The most important interaction is with negation.

Suppose you wanted to prove a claim like $\forall_x P(x)$ to be **false** (for instance, suppose the statement is "For all integers n, we have n^2 odd"). The only way this statement can be true is if $P(x)$ is true for every single x; so if we want to show that it's false, we just need to establish that $P(x)$ is false for *some particular* x. For instance, $\forall_n \big[n^2 \text{ is odd}\big]$ is false, because $2^2 = 4$ is not odd (so $P(2)$ fails).

What I'm asserting here is that there is a logical equivalence between $\neg\forall_x P(x)$ ("it is not true that $P(x)$ holds for all x") and $\exists_x \neg P(x)$ "there exists an x for which $P(x)$ is false"). This is mostly important when thinking about how to prove a "for-all" type statement false. You're going to **provide** a counterexample: an x for which $P(x)$ is false, together with a proof that $P(x)$ is false.

On the other hand, how would you prove a claim like $\exists_x P(x)$ to be false ("there does **not** exists an x for which $P(x)$' is true")? If there does not exist an x for which $P(x)$ is true, then $P(x)$ must be false for every single x! Counter to the above, we've now established that $\neg\exists_x P(x) \equiv \forall_x \neg P(x)$; to show that an "existence" statement is false, you must show that there is no such object.

For instance, take "There exists an integer n so that n is even and n is odd". To show that this is false amounts to showing "For any integer n, it is not the case that n is both even and odd". This (to me) sounds amenable to proof by contradiction: let's suppose $n = 2m$ and $n = 2k + 1$ and see if this results in nonsense. If this is the case, then $2m = n = 2k + 1$, so $2(m - k) = 1$. Because m and k are both integers, so is $m - k$, so this would imply that 1 is even; but that's not true. Contradiction!

We have established that for all natural numbers, n is not both even and odd. Equivalently, *there does not exist a natural number n which is both even and odd.*

To talk about the interactions with the and/or operations, I want to suggest two heuristics for understanding quantifiers.

Heuristic 1: "$\forall_x P(x)$" is a really big "and" operation, where the "and" runs over every single parameter x. (If there are four parameters $1, 2, 3, 4$, then it would say: "$P(1)$ and $P(2)$ and $P(3)$ and $P(4)$." With infinitely many parameters, it might be understood as "$P(1)$ and $P(2)$ and $P(3)$ and ..."

Heuristic 2: $\exists_x P(x)$ is a really big "or" operation, where the "or" runs over every single parameter x. If the parameter x is only $1, 2, 3$, I understand this as "$P(1)$ or $P(2)$ or $P(3)$."

These heuristics already explain the discussion above: $\neg\exists \equiv \forall\neg$ is related to the equivalence $\neg(P \vee Q) \equiv (\neg P) \wedge (\neg Q)$, while $\neg\forall \equiv \exists\neg$ corresponds to the equivalence $\neg(P \wedge Q) \equiv (\neg P) \vee (\neg Q)$.

Let me remind you from Proposition 4 that it doesn't matter what order you do a sequence of "ands" in, and it doesn't matter how you bracket them. From this perspective, if I have two families of statements $P(x)$ and $Q(x)$, and I want to prove $\forall_x[P(x) \wedge Q(x)]$, I recognize: "This means I'm trying to prove $P(x)$ and $Q(x)$ both, for every x." By moving the "ands" around, this statement is equivalent to showing $\forall_x P(x) \wedge \forall_x Q(x)$.

On the other hand, if I want to prove $\exists_x[P(x) \vee Q(x)]$, this says: "Find an example of an x for which $P(x)$ is true or $Q(x)$ is true." Any x is fine, and either statement is fine! Shuffling around the "or"s, this is equivalent to

$$\exists_x[P(x) \vee Q(x)] \equiv [\exists_x P(x)] \vee [\exists_x Q(x)].$$

Let me record the results of this discussion as a proposition below.

Proposition 9. *We have the logical equivalences*

$$\neg\exists_x P(x) \equiv \forall_x \neg P(x) \tag{2.18}$$
$$\neg\forall_x P(x) \equiv \exists_x \neg P(x) \tag{2.19}$$
$$\forall_x[P(x) \wedge Q(x)] \equiv [\forall_x P(x)] \wedge [\forall_x Q(x)] \tag{2.20}$$
$$\exists_x[P(x) \vee Q(x)] \equiv [\exists_x P(x)] \vee [\exists_x Q(x)]. \tag{2.21}$$

2.3.3 Mathematical induction

Let's finally talk about what we need to do to prove the statement "For all natural numbers n, either n is even or odd, but not both". More briefly, write $E(n)$ for the statement "n is even" and $O(n)$ for the statement "n is odd". We're asserting $\forall_n[E(n) \text{ xor } O(n)]$, where xor is the exclusive or. (And hidden in "$E(n)$" and "$O(n)$" are the existential quantifiers $\exists_m[n = 2m]$ and "$\exists_m[n = 2m + 1]$").

As I mentioned before, I don't know how to do this with anything we've mentioned so far: maybe there's some really large and complicated integer which is nowhere close to an even integer, much less one away from it. This seems wrong, because I know that when I add 1 to an even number it becomes odd and vice versa, but that doesn't get me to "for all n, an integer is exactly one of even or odd", it only tells me how to go from that statement for n to the corresponding statement for $n + 1$, and that's not quite a proof of the claim!

What we need to bridge the gap is precisely the principle of induction:

Theorem 10 (Principle of mathematical induction)**.** *If $P(n)$ is a family of statements indexed by the natural numbers $n = 0, 1, 2, \cdots$, and both*

(Base case) $P(0)$ is true,

(Inductive step) for all natural numbers n, the statement $[P(n) \implies P(n+1)]$ is true,

Then the statement $P(n)$ is true for all natural numbers n. That is, the following implication always holds:

$$\big[P(0) \wedge [\forall_n[P(n) \implies P(n+1)]]\big] \implies \forall_n P(n).$$

If I know both the statement $P(0)$ and "$\forall_n[P(n) \implies P(n+1)]$", here's how I'd try to prove $\forall_n P(n)$:

We know $P(0)$ is true, and we know $P(0) \implies P(1)$, so $P(1)$ is also true; we also know $P(1) \implies P(2)$ is $P(2)$ is also true; we also know that $P(2) \implies P(3)$ is $P(3)$ is also true...

As a person writing a proof, I can only do this finitely many times. (Up above, I explained why $P(0)$ through $P(3)$ are true). The principle of mathematical induction asserts that I can "just keep going" in this argument: that I can jump one step at a time from 0 up through any natural number n. While (to me) this is intuitive, it is not automatic: this is an **axiom** of the natural numbers. (If you're interested more in the axiomatization of the natural numbers, look into "Peano arithmetic"; the inductive axiom is the most complicated axiom needed.)

Let me try to show how this works by using it in practice.

Theorem 11 (Naturals are even or odd)**.** *If n is a natural number, then n is exactly one of even (a multiple of 2) or odd (one above a multiple of 2).*

Proof. As discussed above, let $P(n)$ be the estatement $E(n)$ xor $O(n)$, which is true when the natural number n is exactly one of even or odd. To

(Base case) We need to show that $P(0)$ is true: that is, that 0 is exactly one of even or odd. We know 0 is even, as $0 = 2(0)$. We also know that 0 is not odd, as $0 = 2m + 1$ would imply $-1 = 2m$; but -1 cannot be divisible by 2, as m would have to be between -1 and 0, and there is no integer between -1 and 0. Because 0 is even but not odd, $P(0)$ is true.

(Inductive step) We need to show that for any natural number n, the statement $P(n) \implies P(n+1)$ is true; that is, if we know that the integer n is exactly one of even or odd, then $P(n+1)$ is also exactly one of even or odd.

The trick here is that we proved the equivalence $E(n) \equiv O(n+1)$ in Proposition 3, and the argument that $O(n) \equiv E(n+1)$ is similar. Thus

$$P(n+1) = E(n+1) \text{ xor } O(n+1) \equiv O(n) \text{ xor } E(n) \equiv E(n) \text{ xor } O(n) = P(n).$$

Thus the statement $P(n)$ is equivalent to the statement $P(n+1)$, and in particular, $P(n) \implies P(n+1)$ for all n.

We know that $P(0)$ is true, and we know that $\forall_n[P(n) \implies P(n+1)]$ is true. By the principle of induction, this shows that $P(n)$ is true for all natural numbers n, which is what we wanted to show. □

Example 12. Consider the statement $P(n)$ which asserts the handy formula

$$\sum_{i=0}^{n} i = 0 + 1 + 2 + \cdots + n = \frac{n(n+1)}{2}$$

holds. You can check by hand that this is true for $n = 0, 1, 2, 3, \ldots$, at which point you might guess that it's true for all n. This is a perfectly reasonable place to try to prove $\forall_n P(n)$ by induction (that is, the formula above is always true, for any natural number n).

You've already verified this statement for n equal to the first few integers; in particular, you've already verified $P(0)$ — so the base casae is finished. What you need to show is that $P(n) \implies P(n+1)$.

Now (as with any time we prove an implication) we may assume $P(n)$ is true, so that $\sum_{i=0}^{n} i = n(n+1)/2$; your goal is to show that

$$\sum_{i=0}^{n+1} i = (n+1)(n+2)/2.$$

Now, using what we know, we have

$$\sum_{i=0}^{n+1} i = \left(\sum_{i=0}^{n} i\right) + (n+1) = \frac{n(n+1)}{2} + (n+1) = \frac{n(n+1) + 2(n+1)}{2} = \frac{(n+2)(n+1)}{2} = \frac{(n+1)(n+2)}{2},$$

so that $P(n+1)$ is true. Thus we've established $P(n) \implies P(n+1)$ for all natural numbers n, and the principle of induction tells us that $P(n)$ is simply true for all n.

This proves our desired formula $\forall_n P(n)$ by induction. But that doesn't mean it's the only way we could have proven that formula; we could also have just directly established $P(n)$ for each integer n without relying on knowledge about $P(n-1)$. (If you want to try, here's a hint: try adding terms in pairs, grouping 0 and n as well as 1 and $n-1$, and so on...) ♢

Remark 7. In the argument above, it would have made the calculations look a little bit nicer to prove $P(n-1) \implies P(n)$ for all natural numbers $n \geqslant 1$. This is equivalent to proving $P(n) \implies P(n+1)$ for all natural numbers $n \geqslant 0$, so it would have been fine to do so instead. ♢

Remark 8. When arguing by induction, always be sure you've proved the base case as well as the inductive step, otherwise you haven't completed a proof! If all you know is $\forall_n[P(n) \implies P(n+1)]$, then this tells you that if one $P(n)$ is true then all $P(m)$'s after it will also be true, but it doesn't tell you anything before that one, and it doesn't even tell you that any one of them are true (it could be that $P(n)$ is always false!)

As a concrete example, take $P(n)$ to be the statement "For every natural number n, the number $2n+1$ is even." (Preposterous: this is false for every integer n.)

The statement $\forall_n[P(n) \implies P(n+1)]$ is **true**, and you wouldn't see any issue with an inductive approach! If $2n+1$ really was even, so $2n+1 = 2m$, then $2(n+1)+1 = 2n+3 = (2n+1)+2 = 2m+2 = 2(m+1)$, and thus $2(n+1)+1$ would be even as well.

But $P(0)$ is false, as it asserts that 1 is even, and it certainly is not. In fact $P(n)$ is false for all n. $\Diamond$

Chapter 3

Sets and functions

In this part, we investigate the foundational notion in mathematics: sets and functions between them. There are two essentially different ways to talk about

- *Naive set theory* specifies a few reasonable axioms for the idea of 'set', quickly derives a list of basic facts about these axioms (and the notion of set), and uses these facts to formulate and do other mathematics. If the mathematics we want to do requires another axiom, we're fine with adding it without too much philosophical fuss, and we don't spend too much time thinking about whether our axioms are consistent with one another.
- *Axiomatic set theory* gives a complete list of axioms for set theory, and all reasoning is done with those axioms. The axioms are themselves objects of study, as are the resulting notions of set theory (which will depend on your precise axioms).

The latter is a major and active area of research in mathematical logic (the most common axiomatization is called ZFC, or "Zermelo-Fraenkel set theory with the axiom of choice"). However, most mathematicians think of sets along the lines of naive set theory: sets give a useful language to make mathematical statements precise, but we mostly don't worry about the technical details of the axiomatic framework. Most mathematicians, myself included, could not tell you the precise list of axioms in ZFC without looking them up.

We will go through a brief discussion of naive set theory, sufficient to do what the math we want to do for the rest of the course (and no more). If you're interested in axiomatic set theory, it's really neat stuff! But it's not in the purview of this course.

For the purposes of this course, it will suffice to know the material in Sections 1-9 of Halmos' book "Naive set theory". At some point in your studies, you should probably learn the rest of what's in that book (which primarily relates to some of the complications you can see with infinite sets). This is roughly what I will be covering in these notes, albeit in a more abridged fashion than Halmos.

Another good reference is the first chapter of Munkres' textbook "Topology" (Sections 1.1, 1.2, and 1.6 suffice for the purposes of this course, but it's a well-written chapter in general).

If you find the exposition here insufficient (or if you have the time to spare), I encourage you to look at one of those references as well. I'm going to be very brief!

3.1 Sets and operations

A 'set', informally, is simply a collection of things. We make essentially no constraint on what those "things" could be: maybe they're numbers, maybe functions, maybe sets themselves. To quote Halmos, "A pack of wolves, a bunch of grapes, or a flock of pigeons are all examples of sets of things." So as to be a foundation for many different areas of mathematics, it is useful to allow sets to be comprised of many different kinds of objects; we don't specify what, exactly, a set has to be comprised of.

What is important is that a set is comprised of things, its elements (a particular wolf, a particular grape, a particular pigeon), and that we say two sets are the same if they have the same elements. This is our first axiom.

Axiom 1 (Sets are specified by their elements)**.** *Sets are comprised of their members, or "elements". We write $x \in S$ to denote that x is an element of the set S (TeX:* `$x \in S$`*). Two sets are equal if they have precisely the same elements. Symbolically,*

$$S = T \iff \forall_x [x \in S \iff x \in T].$$

As an example, consider the sets

$$S = \{\text{a red fox}, 2, 7\}, \quad T = \{2, 7\}, \quad R = \{\text{a red fox}, 3\}.$$

The notation {list} is intended to denote "The set which consists of the elements enumerated in the braces". For instance, T is the set which consists of two elements, the numbers 2 and 7. (In TeX, `$T = \{2, 7\}$`.)

Here, the sets S, T, R are all different, because between each pair, one of them has an element the other does not. For instance, S is different from T because S contains a red fox as an element and T does not (so they do not contain the same elements). Curiously, $U = \big\{\{2, 7\}\big\}$ is **different** from T. While T is the set with two elements, the numbers 2 and 7, instead U is a set with **one element**: the set $\{2, 7\}$. Sets can themselves be elements in other sets, and this is an important feature (not a bug).

It should be mentioned that there is no such thing as some elements appearing more than once in a set. Something is either in the set or it is not. If someone told you that $S = \{0, 1, 1\}$, they are describing the set which consists of 0 and 1 (for some reason they mentioned 1 twice, but that's not relevant; they told you that 0 is an element and 1 is an element, and there aren't any other elements). We have $\{0, 1, 1\} = \{0, 1\}$, because both are ways of describing "the set which contains the numbers 0 and 1 and nothing else"; they have the same elements, and Axiom 1 tells us that this means they are the same set. (This confusion usually doesn't arise, because we rarely go out of our way to repeat elements when defining a set.)

In the previous section, we thought a lot about the natural numbers. If sets are going to be a foundation for mathematics, we should at least assume that there is a set which encodes the natural numbers. That will be our second axiom.

Axiom 2 (The natural numbers form a set)**.** *There is a set $\mathbb{N}$ of natural numbers, whose elements are $\{0, 1, 2, \cdots\}$.*

Very frequently we want to specify "subsets" (for instance, now that we have the natural numbers, how about the even natural numbers or the square natural numbers?). This is our third, very important, axiom.

Axiom 3 (We can define subsets with formulas)**.** *Suppose X is a set and that for all $x \in X$ we are given a proposition $P(x)$. Then*

$$S = \{x \in X \mid P(x) \text{ is true}\},$$

the set of x so that $x \in X$ and also $P(x)$ is true, is also a set.

The notation $S = \{x \in X \mid \text{property of } x\}$ is called 'set-builder notation'. The beginning of the string (before the bar, written `$\mid$` in TeX) gives a name to a generic element of your set, and states what larger set it comes from; the portion after the bar tells us what conditions the elements of X must satisfy to be in S.

Example 13. Here are some statements defined for natural numbers n:

$$E(n) = [\exists_{m \in \mathbb{N}} [n = 2m]], \quad S(n) = [\exists_{m \in \mathbb{N}} [n = m^2]], \quad P(n) = [n \text{ is prime}].$$

In plain language, $E(n)$ is true when n is even, and $S(n)$ is true when n is a perfect square. (I didn't want to formalize $P(n)$, but see 2(a) on Homework 1.)

Then Axiom 3 tells us that each of

$$E = \{n \in \mathbb{N} \mid E(n) \text{ is true}\} = \{n \in \mathbb{N} \mid n \text{ is even}\} = \{0, 2, 4, 6, \cdots\}$$

and

$$S = \{n \in \mathbb{N} \mid n \text{ is a perfect square}\} = \{1, 4, 9, \cdots\}, \quad P = \{n \in \mathbb{N} \mid n \text{ is prime}\} = \{2, 3, 5, 7, \cdots\}$$

are all sets. So is

$$\{n \in \mathbb{N} \mid E(n) \wedge P(n) \text{ is true}\} = \{2\},$$

the set of even primes, as well as

$$\{n \in \mathbb{N} \mid E(n) \vee P(n) \text{ is true}\} = \{0, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 22, \cdots\},$$

the set of natural numbers which are either even or prime ◊

Example 14. This axiom forces us to accept the existence of a set denoted $\varnothing$ which contains **no elements whatsoever**, called the 'empty set'. This is a subset of any set X, defined by the proposition $P(x) = \text{False}$; because $P(x)$ is never true, $\{x \in X \mid P(x) \text{ is true}\}$ is precisely this so-called empty set.

◊

This construction always produces sets which are in a sense 'smaller' (or at least 'no bigger than') the original set: all the elements in the new set were already in the old set. This is an important property which deserves its own name and notation.

Definition 5. If all elements of S are also elements of a larger set X, we say S is a subset of X and write $S \subset X$ (TeX: `$S \subset X$`). If we want to emphasize that it's possible that S could equal X, we write $S \subseteq X$ (TeX: `\subseteq`; this is synonymous with $\subset$. The only difference is emphasis). If X is strictly larger, meaning that there exists some $x \in X$ which is not in S ($x \notin S$), we write $S \subsetneq X$ (TeX: `$S \subsetneq X$`).

Symbolically,

$$S \subset X \text{ is the formal statement } \forall_s[s \in S \implies s \in X],$$

while

$$S \subsetneq X \text{ is the formal statement } \forall_s[s \in S \implies s \in X] \wedge [\exists_{x \in X} \neg[x \in S]].$$

◊

This notion is often useful. Notice that $S \subsetneq X$ can be written as "$[S \subset X] \wedge \neg[X \subset S]$" (the set S is contained in X, but X is not contained in S).

In fact, it's the primary logical basis we use for proving that two sets are equal.

Proposition 12 (To prove two sets are equal, you show each is contained in the other)**.** *Given two sets S and T, we have $S = T$ if and only if each of S and T is contained in the other (that is, $S \subset T$ and $T \subset S$). Symbolically,*

$$S = T \quad \iff \quad [S \subset T] \wedge [T \subset S].$$

Proof. Explicitly, remember that $S = T$ means "The elements of S are precisely the same as the elements of T"; that is,

$$\forall_x[x \in S \iff x \in T].$$

If we remember that $P \iff Q$ is the same statement as $[P \implies Q] \wedge [Q \implies P]$ (and that's how we prove iff statements, anyway — we prove each implication separately), we can rewrite $S = T$ as

$$\forall_x\big[[x \in S \implies x \in T] \wedge [x \in T \implies x \in S]\big] \equiv \big[\forall_x[x \in S \implies x \in T] \wedge \forall_x[x \in T \implies x \in S]\big] = [S \subset T] \wedge [T \subset S].$$

That is, to show "the elements of S are precisely the same as the elements of T" is the same as showing two statements: that every element of S is contained in T, and every element of T is also contained in S. □

We will come back to this proposition a lot. Showing that two sets are equal by showing that each is contained in the other is called a *double containment argument*. Virtually every proof you write that two sets are equal will be a double containment argument.

Remark 9. There is a useful guiding principle which appears in both the proposition above and elsewhere thusfar: when you're trying to show an equivalence (or an equality), it is often helpful to break that into two statements.

If I want to prove $P \iff Q$, I prove both $P \implies Q$ and $Q \implies P$, because each of those implications gives me somewhere to start (assume that P is true, or assume that Q is true); I have something to work with. Similarly, if I want to show that two sets S, T coincide, I show this in two steps: first I show that every $x \in S$ is also in T, and then I show that every $x \in T$ is also in S. Each of these hypotheses gives me something to work with (if $x \in S$, I can look at the definition of S and try to argue from that that x is also in T, and vice versa).

Oftentimes, the arguments you want to make to go from S to T will not be the same arguments you use to go from T to S! $\Diamond$

3.1.1 New sets from old

We will need a total of four constructions of new sets out of old ones: **products, power sets, unions, and intersections**.

Remark 10. There are more constructions one can define, and not all of them can be defined in terms of what I write below; the most crucial constructions not discussed below are **infinite** products, unions, and intersections. One further construction which we won't cover is the idea of 'quotient sets'. These will appear in your studies when they have to, and it's not worth learning them until you have a good reason to. $\Diamond$

The first kind of set we need to guarantee exists is the *Cartesian product*. You have already seen this used elsewhere, just implicitly.

Axiom 4 (Products exist)**.** *Given two sets X and Y, there is a set called the 'Cartesian product' $X \times Y$ (TeX:* `$X \times Y$`*), whose elements are ordered pairs $\{(x, y) \mid x \in X, y \in Y\}$.*

For instance, the set $\mathbb{Z} \times \mathbb{Z} = \mathbb{Z}^2$ is the set whose elements are pairs (n, m), where each n, m is an integer; the set $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ is the set of pairs (x, y) where each of x, y is a real number. You can iterate this construction: the set

$$\mathbb{R} \times \mathbb{Q} \times \mathbb{Z} = \{(x, y, z) \mid x \in \mathbb{R}, \quad y \in \mathbb{Q}, \quad z \in \mathbb{Z}\}$$

is the set of triples (x, y, z), where the first component is a real number, the second component is a rational number, and the final component is an integer. Linear algebra is, to some extent, about the study of vectors in $\mathbb{R}^n$ (whose elements are strings of n real numbers, or 'n-tuples' of real numbers: think triples, quadruples, etc).

You have certainly thought about **elements** of the next set before (we defined them in Definition 5!), but you have probably not thought about it as a set in its own right.

Axiom 5 (There is a set of all subsets of X)**.** *Given any set X, there is a set called $\mathcal{P}(X)$ (the **power set** of X, in TeX* `$\mathcal P(X)$`*) whose elements are the subsets $S \subset X$.*

Example 15. If $X = \{0, 1\}$ is a 2-element set, then

$$\mathcal{P}(X) = \big\{\varnothing,\ \{0\},\ \{1\},\ \{0, 1\}\big\}$$

is a four-element set, consisting of: the empty set; the two one-element subsets of X; the single two-element subset of X. If $X = \{0, 1, 2\}$, then $\mathcal{P}(X)$ has eight elements

$$\mathcal{P}(X) = \big\{\varnothing,\ \{0\},\ \{1\},\ \{2\},\ \{0, 1\},\ \{0, 2\},\ \{1, 2\},\ \{1, 2, 3\}\big\}.$$

There are too many elements of $\mathcal{P}(\mathbb{N})$ to list them out (this is a theorem, proved using what is called Cantor's diagonalization argument). Some famous ones you know: the set of even naturals is an element of $\mathcal{P}(\mathbb{N})$; the set of odd naturals is an element of $\mathcal{P}(\mathbb{N})$; the set of primes is an element of $\mathcal{P}(\mathbb{N})$.

For a more arcane example, the set S of naturals which appear somewhere in the decimal expansion of π — which includes, for instance, all of $1, 3, 4, 5, 314, 159, 926$ — is an element of $\mathcal{P}(\mathbb{N})$ (though it would be a

shocking advance if you could describe this set explicitly; it is probably all of $\mathbb{N}$, but I do not expect someone to prove this in my lifetime).

In general, the set $\mathcal{P}(X)$ is always much larger than the set X, in a sense we will not make precise. $\Diamond$

The power set is a very important construction, because in many cases we will understand mathematical objects in terms of certain kinds of simpler subsets (elements of the power set). It is used in virtually every construction of important sets, and you will see it used in the second homework to formally define a set of functions using these axioms.

Definition 6. Let $U, V \subset X$ be a pair of subsets of a larger set X. We define three more subsets of X as follows:

- The **complement** U^c (sometimes written $X \backslash U$) is the set of points $x \in X$ **which are not in** U. That is,
$$x \in U^c \iff x \in X \text{ and } x \notin U.$$
- The **union** $U \cup V$ is the set of points x which are in either U or V. That is,
$$x \in U \cup V \iff [x \in U] \vee [x \in V].$$
- The **intersection** $U \cap V$ is the set of points x which are in **both** U and V. That is,
$$x \in U \cap V \iff [x \in U] \wedge [x \in V]. \quad \Diamond$$

I'm going to list out a handful of facts about these operations. I'm only going to prove **one of them**, so that you can see how one would write a proof about these things. But here is a list of facts which can generate any other fact about these operations.

Proposition 13. *Let X be a set. For any subsets U, V, W of X, the following sets are equal:*

$$U \cup X = X \tag{3.1}$$
$$U \cup \varnothing = U \tag{3.2}$$
$$U \cap X = U \tag{3.3}$$
$$U \cap \varnothing = \varnothing \tag{3.4}$$

$$U \cup V = V \cup U \tag{3.5}$$
$$U \cap V = V \cap U \tag{3.6}$$
$$(U \cup V) \cup W = U \cup (V \cup W) \tag{3.7}$$
$$(U \cap V) \cap W = U \cap (V \cap W) \tag{3.8}$$

$$(U^c)^c = U \tag{3.9}$$
$$(U \cup V)^c = U^c \cap V^c \tag{3.10}$$
$$(U \cap V)^c = U^c \cup V^c \tag{3.11}$$

$$U \cup (V \cap W) = (U \cup V) \cap (U \cup W) \tag{3.12}$$
$$U \cap (V \cup W) = (U \cap V) \cup (U \cap W). \tag{3.13}$$

Does this resemble anything you've seen before? Can you give a reasonable guess as to why?

Proof of (3.10). Recall from Axiom 1 that to tell whether two sets are equal, we need to check that they have the same elements. That means that I need to show **two things**. First, I need to show that every

element of $(U \cup V)^c$ is contained in $U^c \cap V^c$; secondly, I need to show that every element of $U^c \cap V^c$ is contained in $(U \cup V)^c$. This will establish a 'double containment'

$$(U \cup V)^c \subset U^c \cap V^c \subset (U \cup V)^c,$$

which shows that these two sets are equal: any element in one is in the other, so they have precisely the same elements. **This is, more or less, always how you will reason to show that two sets are equal.**

So first, what does it mean to talk about "an element $x \in (U \cup V)^c$"? Because this lies in a complement, this means $x \notin (U \cup V)$. Then we remember what it means to be in a union. If $y \in U \cup V$, we have[1] either $y \in U$ or $y \in V$. If x is **not** in $U \cup V$, then this is false — so $x \notin U$ **and** $x \notin V$. This means precisely that $x \in U^c$ and $x \in V^c$; because x lies in both of these, we have $x \in U^c \cap V^c$. This shows the first containment.

For the other containment, we start with $x \in U^c \cap V^c$. This means $x \in U^c$ and $x \in V^c$, or in other words, $x \notin U$ and $x \notin V$. Because x is not in U **and** it is not in V, it cannot sit in the union $U \cup V$ (which consists of points lying in at least one of U, V). Therefore $x \notin U \cup V$, so that $x \in (U \cup V)^c$. This shows the second containment, and we are finished. □

Remark 11. It's certainly possible to write the proof above in a more concise manner, but doing so wouldn't serve my purpose: explaining how to write a double containment argument and going through all the details carefully and explicitly. You will be writing **a lot** of double containment arguments. I want you to get some practice. ◊

3.1.2 The connection to logic

Here is a brief observation, which I'm not going to refer back to later.

You may have noticed that all of the relations satisfed by the logical operations $\neg, \wedge, \vee$ are also satisfied by the operations of complement, intersection, and union; for instance,

$$(S \cup T)^c = S^c \cap T^c \quad \text{compared to} \quad \neg(P \vee Q) \equiv (\neg P) \wedge (\neg Q)$$

or

$$S \cap (T \cup R) = (S \cap T) \cup (S \cap R) \quad \text{compared to} \quad P \wedge (Q \vee R) \equiv (P \wedge Q) \vee (P \wedge R).$$

Here is a way to relate the two stories. Suppose you have some big set X of all possible parameters to your logical statements ($X = \mathbb{N}$ was common in previous examples, especially when discussing induction). A family of propositions $P(x)$ indexed by $x \in X$ gives rise to a subset $S(P) \subset X$, the set of $x \in X$ so that $P(x)$ is true: in symbols,

$$x \in S(P) \iff P(x) \text{ is true.}$$

This subset satisfies $S(\neg P) = S(P)^c$:

$$x \in S(\neg P) \iff \neg P(x) \text{ is true} \iff P(x) \text{ is false} \iff x \notin S(P) \iff x \in S(P)^c.$$

Similarly, you can check that $S(P \wedge Q) = S(P) \cap S(Q)$, while $S(P \vee Q) = S(P) \cup S(Q)$. The various relations between the logical operations (like de Morgan's laws above) then translate immediately to the corresponding set-theoretic statements about $S(P)$.

3.2 Functions and cardinality

Probably the single most important foundational idea in mathematics is the idea of a **function**. Most people have seen functions presented to them as pictures on a page (the 'graph' of the function; more on this in the homework) or as given by explicit formulas, but that's not what a function is! Let me give a definition and then a long slew of examples and non-examples.

[1] I refer to y here instead of x because I want to make a general observation about elements which lie in $U \cup V$; the particular element x we were discussing does not.

Definition 7. Let X and Y be two given sets. A **function** f from X to Y (written $f : X \to Y$; in TeX `$f: X \to Y$`) is a rule which, for any element of x, produces a unique (unambiguous) output $f(x) \in Y$.

◊

Remark 12. A function is **three** pieces of data: a domain X, a 'codomain' Y, and a machine which takes inputs from X and produces unambiguous outputs in Y.

◊

Example 16. You already know a truly preposterous number of functions $f : \mathbb{R} \to \mathbb{R}$. For instance, you can take the functions given by $f(x) = \cos x$ or $f(x) = x^2$ or $f(x) = e^x$ or $f(x) = x + 17.3$, or compositions of these, or products of these.

One can construct more functions by changing the domain or codomain. For instance, the formula $f(x) = e^x$ can also be used to define a function $f : (0, \infty) \to (1, \infty)$, or $f(x) = x + 17.3$ can be used to define a function $(0, \infty) \to \mathbb{R}$. The "codomains" do not need to include every output value (this would massively limit our ability to talk about functions effectively); all that matters is that for every x in the domain X, every output value $f(x)$ actually lies in the codomain Y.

Similarly, the input domains do not have to be "as large as possible" (whatever that means). **To define a function, both domain and codomain must be specified.** ◊

Example 17. Suppose I wanted to discuss "the function $f(x) = 1/x$" in this context. There are two issues. The first is that one needs to specify a domain and codomain to define a function: what should those be here? You might reasonably guess I want my domain and codomain to be $\mathbb{R}$, but $f(0)$ is **not defined**.

One reasonable way to replace this would be to define this function with domain

$$X = \mathbb{R}\backslash\{0\} = (-\infty, 0) \cup (0, \infty)$$

and the same codomain $Y = \mathbb{R}\backslash\{0\}$. This does define a function. For each nonzero $x \in \mathbb{R}$, the number $1/x$ makes sense, is unambiguous, and is again a nonzero real number (so an element of Y). ◊

Remark 13. Here is an important notational point. When I tell you "let f be the function given by $f(x) = x^2$, I would not say that "x^2" is the name of the function. The name of the function is f. However, to specify this function, I need to tell you what it does to an arbitrary input real number x. What I have given above is a formula which tells you how to compute f when given some arbitrary input; when I write $f(x)$, it represents the output of the function on the input x.

Here, x is the input (an element of $X = \mathbb{R}$), while f is the name of the function, and its output when applied to x is called $f(x) \in \mathbb{R} = Y$. ◊

Example 18. There is a function $f : \mathbb{N} \to \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ which sends an integer n to the n'th decimal digit of π. For instance,

$$f(0) = 3, \quad f(1) = 1, \quad f(2) = 4, \quad f(3) = 1, \quad f(4) = 5, \quad f(5) = 9, \quad \cdots$$

This is well-defined (the number π has an unambiguous decimal expansion), but good luck computing it, much less graphing it! For instance, what is $f(10^{32})$? Computing this would be very time-consuming and I doubt anyone here has the computation power available to do so. The point here is that **just because you can tell me how a function is defined doesn't mean that should tell you a way to compute it.**

A similar example is the following. Let $f : \mathbb{R} \to \mathbb{R}$ be the function defined as follows.

$$f(x) = \begin{cases} 0 & \text{if the Collatz conjecture is true} \\ 1 & \text{if the Collatz conjecture is false} \end{cases}.$$

One of those is true, so this is indeed a function: it gives an unambiguous output for any given x. But you will never be able to tell me what it is until someone proves or disproves the Collatz conjecture, so I can't actually tell you what the value $f(3)$ is... ◊

Example 19. It is a theorem that there are too many functions $f : \mathbb{N} \to \mathbb{N}$ to count (basically also 'Cantor's diagonalization argument'). Every single object which is describable by human language can be listed off (counted), so this means there must be some function which cannot possibly be described in its entirety by any human-communicable sentence. In fact, most are indescribable and uncomputable in this way. This is

in stark contrast to the fact that every function you've ever worked with in your mathematical life so far has most likely been given to you by an explicit formula. The horrifying truth is that you absolutely cannot give an explicit formula for just about any function that actually exists, and we cannot analyze arbitrary functions by thinking in terms of formulas.

◊

Example 20. Consider the following attempted definition. "Let $f : [0, \infty) \to \mathbb{R}$ be the function which sends x to a number $f(x)$ with $f(x)^2 = x$." This is not a definition of a function. As long as $x > 0$, there are **two** numbers which square to x: one negative, and one positive. The definition above is ambiguous about which I prefer, so it does not define a function.

However, "Let $f : [0, \infty) \to \mathbb{R}$ be the function which sends x to a **non-negative** number $f(x)$ with $f(x)^2 = x$" is indeed a function. Such a number $f(x)$ always exists, and there is always exactly one such number. So this does produce an unambiguous real number, given any real number $x \geqslant 0$. ◊

3.2.1 Images and preimages

I mentioned above that the codomain of a function $f : X \to Y$ does not need to include all 'output values'. It would still be useful to be able to discuss what those output values *are*. We record two definitions. The first records *the output values of f on a subset S of the domain*; it shows what values f can take on over the subset S. The second records *the input values of f which land in a certain subset T of the codomain*; it shows where the inputs of f have to lie if we want them to get sent into a particular set of points.

Definition 8. If $f : X \to Y$ is a function, we can use this to produce new subsets of X and Y:

a) If $S \subset X$ is a subset, we say the **image** (or **forward image**) of S under the function f is

$$f(S) = \{y \in Y \mid y = f(s) \text{ for some } s \in S\}.$$

That is,

$$y \in f(S) \iff \exists_{s \in S}[y = f(s)];$$

the image is the set of points in Y which are mapped to by some point(s) in S.

b) If $T \subset Y$ is a subset, we say the **preimage** of T under the function f is

$$f^{-1}(T) = \{x \in X \mid f(x) \in T\}.$$

That is, the preimage is the set of points in X which map to some point in T.

The forward image of the domain, $f(X)$, is sometimes called 'the image of f', and is sometimes written $\mathrm{Im}(f)$ or $\mathrm{image}(f)$.
The inverse image of a singleton set $\{y\}$ is usually denoted $f^{-1}(y)$. ◊

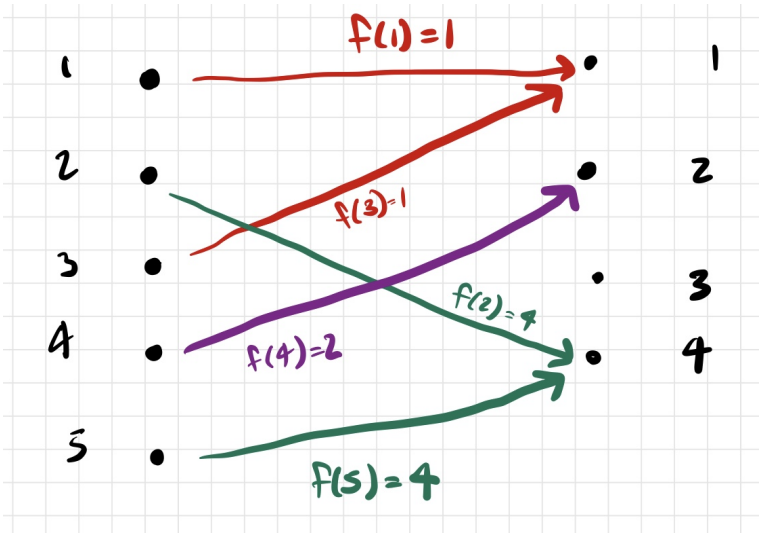
Remark 14. **By far**, the most important of these for us are going to be the image $f(X)$ of the whole domain (which give the set of all possible output values of f) and the preimage $f^{-1}(y)$ of a singleton, which gives the set $\{x \in X \mid f(x) = y\}$ of points which map to y. However, I find it useful to talk about the general notion to get practice with double-containment arguments. ◊

Exercise. Show that $f^{-1}(Y) = X$ no matter what X, f, and Y are. This implies that there is no useful notion of "the inverse image of f", unlike the idea of $\mathrm{image}(f) = f(X)$.

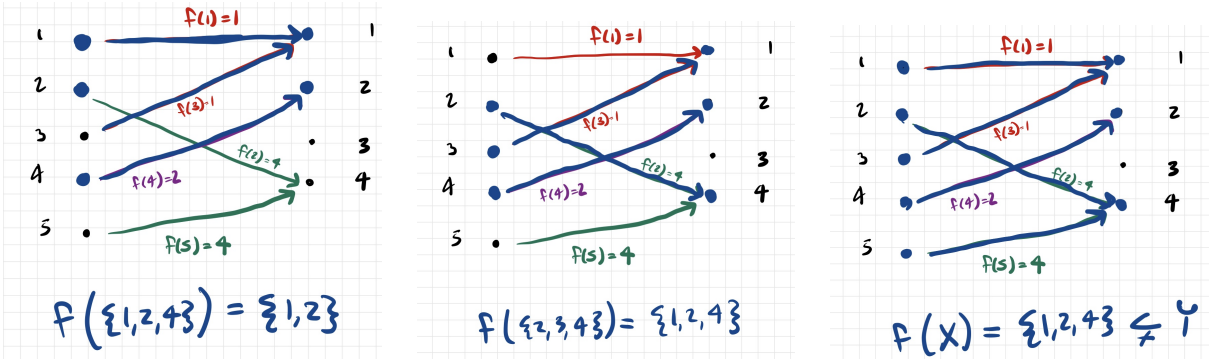
Below, I drew a function $f : \{1, 2, 3, 4, 5\} \to \{1, 2, 3, 4\}$, defined by

$$f(1) = 1, \quad f(2) = 4, \quad f(3) = 1, \quad f(4) = 2, \quad f(5) = 4.$$

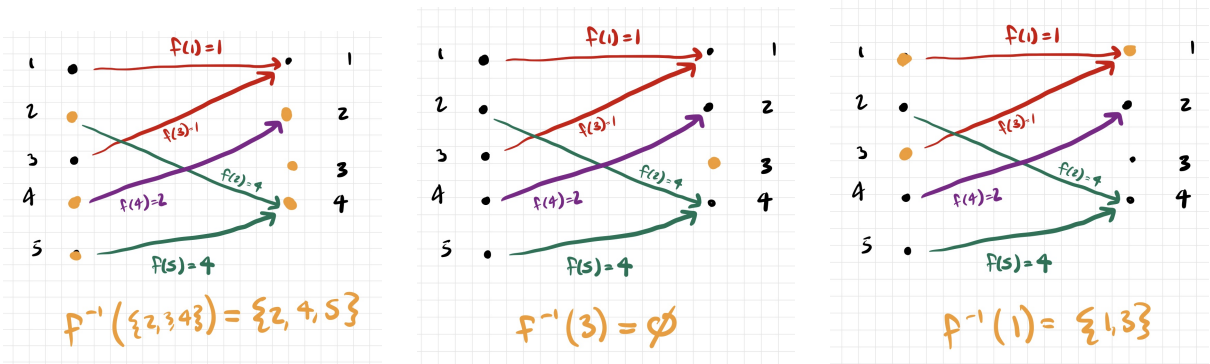
To depict this, I drew all the elements of the domain in a column on the left, and all the elements of the codomain in a column on the right. If $y = f(x)$, I drew an arrow from x to y.



Next, I drew the forward image of three different subsets of the domain X. I drew the elements of the given subset of X as blue dots on the left; the forward image is every point in the right column with an arrow connecting it to a blue dot (I colored these blue as well).



Last, I drew the inverse image of three different subsets of the codomain Y. I drew the elements of the given subset of Y as orange dots on the right; theh inverse image is every point in the left column with an arrow connecting it to an orange dot (which I colored orange as well).



Example 21. Let's use an example you're comfortable with to see how these work in some specific cases: suppose $f : \mathbb{R} \to \mathbb{R}$ is the function defined by $f(x) = x^2$. I encourage you to work through this example carefully. If my arguments don't make sense, try to write your own.

First, let's calculate a small handful of forward images.

- We have $f(\mathbb{R}) = [0, \infty)$. To see this, we need to establish a double containment. The statement $f(\mathbb{R}) \subset [0, \infty)$ means: "If $x \in \mathbb{R}$, then $x^2 \geqslant 0$." You have been using this for a very long time; you can prove it by splitting into cases $x > 0$, $x = 0$, $x < 0$. The other containment $[0, \infty) \subset f(\mathbb{R})$ says: "For every real $x \geqslant 0$, there exists some $y \in \mathbb{R}$ so that $y^2 = x$." In fact, you can take y to be the unique non-negative number with this property, which we call $\sqrt{x}$. (A careful proof that this number exists requires some analysis of the real numbers; you need to know the 'least upper bound property'. I am fine taking the statement that $\sqrt{x}$ exists for granted.)

- We have $f\big([-1, 1]\big) = [0, 1]$. To prove this, we have to show two containments. First, if $-1 \leqslant x \leqslant 1$, then $0 \leqslant x^2 \leqslant 1$, so $f\big([-1, 1]\big) \subset [0, 1]$. For the other direction, notice that each $x \in [0, 1]$ can be written as the square of $\sqrt{x} \in [0, 1]$. Because $\sqrt{x} \in [-1, 1]$ and $x = f(\sqrt{x})$, this shows that $[0, 1] \subset f\big([-1, 1]\big)$. This proves the other containment, so these two sets are equal.

- In fact, this is true even for a smaller input set: $f\big([-1, -1/2) \cup [0, 1/2]\big) = [0, 1]$. The first containment follows as above: if $x \in [-1, -1/2) \cup [0, 1/2]$ then in particular $-1 \leqslant x \leqslant 1$, so $0 \leqslant x^2 \leqslant 1$, which establishes the containment $f(S) \subset [0, 1]$. As for the second containment, we have to be a little more careful. For $x \in [0, 1]$, it is no longer necesarily true that $\sqrt{x} \in S$ (for instance, $x = 4/9$ has $\sqrt{x} = 2/3$ which is not in S). However, either $\sqrt{x}$ is in S or $-\sqrt{x}$ is. (Proof: if $\sqrt{x} \leqslant 1/2$ it lies in the $[0, 1/2]$ term; if $1 \geqslant \sqrt{x} > 1/2$ then $-1 \leqslant -\sqrt{x} < -1/2$, which lies in the $[-1, -1/2)$ term.) This shows that $[0, 1] \subset f(S)$. We have established two containments, so these sets are equal. ◊

Next let's look at some inverse images. I'm going to focus particular attention on the inverse images of points, since these are the most useful for us in this course.

Example 22. Again, let's consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$, and compute the inverse image of different points.

- Let's look at $f^{-1}(0)$. By definition, this is the set of points
$$\big\{x \in \mathbb{R} \mid f(x) \in \{0\}\big\} = \{x \in \mathbb{R} \mid x^2 = 0\}.$$
There is only one real number which squares to zero: $x = 0$. Thus
$$f^{-1}(0) = \{x \in \mathbb{R} \mid x^2 = 0\} = \{0\}.$$

- Next, let's look at $f^{-1}(1)$. This is the set of points
$$\big\{x \in \mathbb{R} \mid f(x) \in \{1\}\big\} = \{x \in \mathbb{R} \mid x^2 = 1\} = \{-1, 1\}.$$
There are two numbers which map to 1: we have $1^2 = (-1)^2 = 1$.

- On the other hand,
$$f^{-1}(-1) = \{x \in \mathbb{R} \mid x^2 = -1\} = \varnothing.$$
No real numbers map to -1.

In general, if $x > 0$ is positive, $f^{-1}(x) = \{\sqrt{x}, -\sqrt{x}\}$ is a two-element set; if $x < 0$ is negative, $f^{-1}(x) = \varnothing$ is empty; if $x = 0$ then $f^{-1}(0) = \{0\}$ is a one-element set. ◊

I mentione above that only some special cases of these operations will be really useful to us, but they do give us a good place to practice set containment / double containment arguments. Let me prove three (kind of random) facts about the way these operations behave in terms of other operations we already know; you'll prove some more like this on your homework.

Proposition 14. *We have the following set inclusions/equalities. (There are many more such relations; these are only a few.)*

$$S \subset f^{-1}(f(S)) \tag{3.14}$$
$$f(f^{-1}(T)) \subset T \tag{3.15}$$
$$f(A \cap B) \subset f(A) \cap f(B). \tag{3.16}$$
$$f(A \cup B) = f(A) \cup f(B) \tag{3.17}$$
$$\tag{3.18}$$

Exercise. Find examples of functions f and sets S, T, A, B so that the first, second, and third inclusions are **not** equalities (that is, the sets on the right-hand-side are strictly larger: there is some element in the set on the right-hand-side which is not in the set on the left-hand-side).

Proof. Let's go down the list.

- If I want to show that $S \subset f^{-1}(f(S))$, this means I want to show that if $x \in S$, we also have $x \in f^{-1}(f(S))$. What does this second statement mean? **By definition**,

  $$f^{-1}(f(S)) = \{x \in X \mid f(x) \in f(S)\}.$$

  So if we want to show $[x \in S] \implies [x \in f^{-1}(f(S))]$, this amounts to showing: "if $x \in S$, then $f(x) \in f(S)$." OK, how do I show this? I'd better write out the definition of $f(S)$! An element $y \in Y$ lies in $f(S)$ if and only if there is some $s \in S$ so that $y = f(s)$. So if $x \in S$, then $y = f(x)$ is certainly in $f(S)$! Here, our 's' is the element $x \in S$.

- This will be similar to the previous argument, but I am going to be more brief this time. If $y \in f(f^{-1}(T))$, then $y = f(x)$ for some $x \in f^{-1}(T)$ by definition of forward image. Next, if $x \in f^{-1}(T)$, this means by definition of inverse image that $f(x) \in T$. Thus $y = f(x) \in T$. This proves that $f(f^{-1}(T)) \subset T$.

- If $y \in f(A \cap B)$, this means there is some $x \in A \cap B$ with $f(x) = y$. Because $x \in A \cap B$, we have both $x \in A$ and $x \in B$. By definition of forward image, this means $y = f(x) \in f(A)$ and also $y = f(x) \in f(B)$. It follows from the definition of intersections that $y \in f(A) \cap f(B)$.

- Here we finally have an equality, so it's time to write out a double inclusion argument! First, let's show $f(A \cup B) \subset f(A) \cup f(B)$. This is very similar to the previous part. If $x \in A \cup B$, our goal is to show that $f(x) \in f(A) \cup f(B)$. The assumption means that either $x \in A$ or $x \in B$. If we're in the first case $x \in A$, we know by definition that $f(x) \in f(A)$, so in particular $f(x) \in f(A) \cup f(B)$. Alternatively, if in fact $x \in B$, then $f(x) \in f(B)$, so in particular $f(x) \in f(A) \cup f(B)$. Since this exhausts the two possibilities, we have shown that $A \cup B \subset f(A) \cup f(B)$.

  For the reverse inclusion $f(A) \cup f(B) \subset f(A \cup B)$, suppose $y \in f(A) \cup f(B)$. Thus either $y = f(a)$ for some $a \in A$ or $y = f(b)$ for some $b \in B$. Because $A \subset A \cup B$ and $B \subset A \cup B$, we have both $a \in A \cup B$ and $b \in A \cup B$; in either case we see that $y = f(x)$ for some $x \in A \cup B$ (our element x will either be $a \in A$ or $b \in B$, depending on which of the two cases we're in).

  This completes the double containment, and establishes $f(A) \cup f(B) = f(A \cup B)$, as desired. □

Remark 15. The statements above are probably never going to be used in this course. However, the **arguments** above are crucial. These are called 'definition-pushing' arguments. How do I show $S \subset f^{-1}(f(S))$? First I write out the definition of subset (each $x \in S$ is in $f^{-1}(f(S))$). How do I show that? To show it, I have to understand what it means, so I write out a definitition of $f^{-1}(f(S))$. Then to show membership in this set, I have to spell out what $f(S)$ means. Once I have spelled everything out completely, it all falls like a line of dominoes.

In some sense, this sounds like it should easy (algorithmic, even: just keep unwrapping definitions until they are as clear as possible!), but these kinds of arguments are consistently difficult for early mathematics students, and are worth practicing — especially because a great many arguments in early undergraduate mathematics are precisely such 'definition-pushing' arguments.

The best advice I have for them is: **don't stop working**. If you're writing an argument but you don't see what to do next, look for anything you can work with — any definition you haven't expanded out, any relevant formula/theorem you haven't used, and see if you can put it to work. Time when you're spent waiting for the idea to come to you is usually very unproductive! ◊

3.2.2 Injectivity, surjectivity, and bijectivity

There are three properties of functions which we will find extremely important when discussing the 'size' of various sets (in particular, when discussing stuff about the idea of dimension in linear algebra). They deserve to be named explicitly.

Definition 9. Suppose $f : X \to Y$ is a function.

a) If $f(X) = Y$ (that is, the image of f is all of Y), we say that f is **surjective** (or we say f is a surjection).

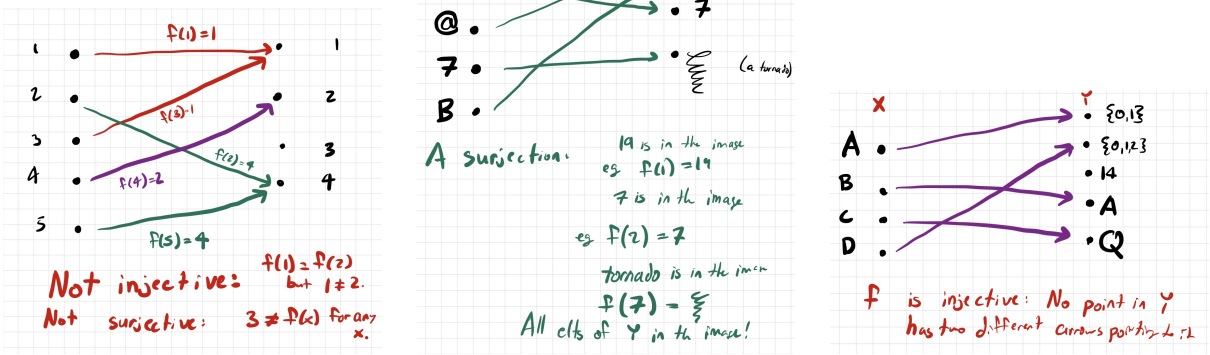
b) Suppose we know that for any two elements $x_1, x_2 \in X$, the implication $f(x_1) = f(x_2) \implies x_1 = x_2$ holds. (That is, whenever two elements x_1, x_2 have the same output value, they were actually the same element to begin with!) Then we say f is **injective** (or we say f is an injection).

c) If f is both injective and surjective, we say it is **bijective** (or we say f is a bijection). ◊

Remark 16. Sometimes, in introductory math textbooks, you will see the term "one-to-one" used synonymously with injective, and the term "onto" synonymously with surjective. Sometimes they will refer to a bijection as a function which is "one-to-one and onto" (injective and surjective).

Mathematicians do not regularly use these terms (despite perhaps seeming more accessible), and I will not either. ◊

Here are three pictures of functions like the examples above. The example I gave earlier of a function $f : \{1, 2, 3, 4, 5\} \to \{1, 2, 3, 4\}$ was neither injective nor surjective; I also give a picture of an injective function and a picture of a surjective function.



In a bit, we'll try to rephrase these conditions to more clearly see what they mean.

Example 23. Consider the function $f : \mathbb{Z} \to \mathbb{N}$ given by

$$f(n) = \begin{cases} 2n & n \geqslant 0 \\ -1 - 2n & n < 0 \end{cases}.$$

I claim that f is a bijection. To prove this, we need to show that f is both an injection and a surjection.

- To see that f is injective, start with two (a priori distinct) integers $m, n \in \mathbb{Z}$ so that $f(m) = f(n)$. Our goal is to show that $m = n$. Notice that when $m \geqslant 0$, we have $f(m)$ even, while if $m < 0$ we have $f(m)$ odd. Because odd numbers cannot be even and vice versa, we see that if $f(m) = f(n)$, we must either have both $m, n \geqslant 0$ or we must have both $m, n < 0$. In the first case, $f(m) = f(n)$ implies $2m = 2n$, so that $m = n$ as desired. In the second case, $f(m) = f(n)$ implies $-1 - 2m = -1 - 2n$, so adding 1's and cancelling the factor of -2 we see that $m = n$, as desired. Thus in either case, we see that $f(m) = f(n)$ implies $m = n$. Thus f is injective.

- To see that f is surjective, we nee to show that every natural number $m \in \mathbb{N}$ is in the image of f. Let us show this in cases, first proving this when m is even, and then proving this when m is odd. Remember that every natural number is non-negative. If $m = 2k$ is even, then $k \geqslant 0$ so $f(k) = 2k = m$ by definition. If m is odd, then $m = 2k - 1$ for some $k > 0$. Because $-k < 0$, we have $f(-k) = -1 - 2(-k) = 2k - 1 = m$. Thus every m — whether even or odd — appears in the image of f. This means that f is surjective, as desired.

This completes the proof that f is a bijection.

◊

The next proposition is how I actually *think* about injectivity and surjectivity.

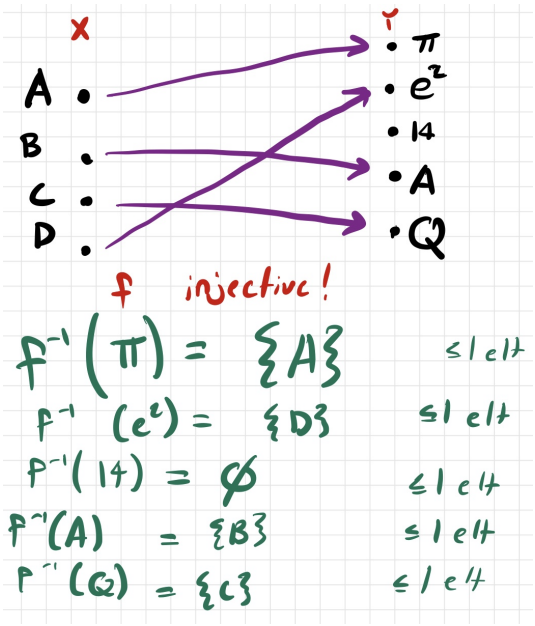
Proposition 15. *Let $f : X \to Y$ be a function.*

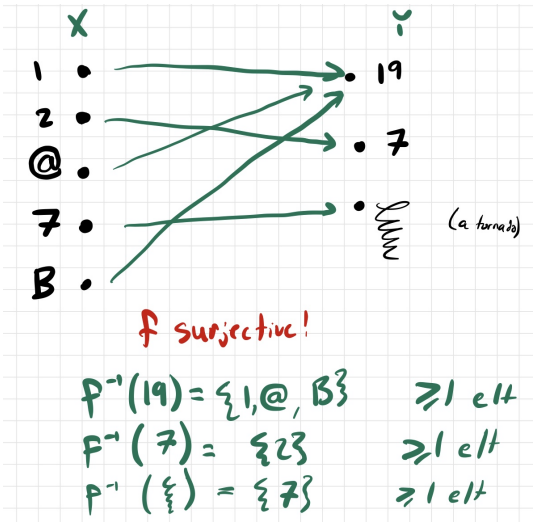
*a) f is an injection if and only if, for all $y \in Y$, the set $f^{-1}(y)$ contains **at most** one element.*

*b) f is a surjection if and only if, for all $y \in Y$, the set $f^{-1}(y)$ contains **at least** one element.*

*c) f is a bijection if and only if, for all $y \in Y$, the set $f^{-1}(y)$ contains **exactly** one element.*

Before giving the proof, here are some pictures of the idea this proposition is trying to encode.





Proof of Proposition 15. Let's show these one by one. There are a lot of steps here: each "if and only if" is secretly two implications!

a) Suppose f is an injection. Let's try to show that each $f^{-1}(y)$ contains **at most** one element. Suppose I have two (possibly identical!) elements $x_1, x_2 \in f^{-1}(y)$. This means that $f(x_1) = f(x_2)$. By the definition of injectivity, this tells us that $x_1 = x_2$. Therefore, any two elements in $f^{-1}(y)$ are actually equal! There cannot be more than one distinct element in $f^{-1}(y)$. (It is possible that there are zero elements in this set! All we can confidently rule out is that there are not 2 or more.)

Now suppose that for all $y \in Y$, the set $f^{-1}(y)$ contains at most one element; let's show that f is an injection. If $f(x_1) = f(x_2)$, our goal is to show $x_1 = x_2$. If $y = f(x_1)$, then by definition we have $x_1, x_2 \in f^{-1}(y)$. Because $f^{-1}(y)$ contains at most one element, we see that in fact $x_1 = x_2$ — they must have been the same element of X all along.

This completes the proof of the $\iff$ statement.

b) Now I am going to speed up my exposition a bit. Suppose f is a surjection. This means that $f(X) = Y$; in particular, $Y \subset f(X)$. Thus for each $y \in Y$ we have $y \in f(X)$. Recalling the definition of $f(X)$, this means there exists some $x \in X$ so that $f(x) = y$. But then $x \in f^{-1}(y)$. Thus, for all $y \in Y$, the set $f^{-1}(y)$ contains at least one element (it is not empty).

Conversely, suppose that for all $y \in Y$, the set $f^{-1}(y)$ is nonempty. This means there is some $x \in f^{-1}(y)$, which means there is some x with $f(x) = y$. Thus, for all $y \in Y$, we have $y \in f(X)$. Thus $Y \subset f(X)$. Because f is a function from X to Y, we have $f(X) \subset Y$, so that $f(X) = Y$ as desired.

c) If f is a bijection, it is both an injection and a surjection, so that each set $f^{-1}(y)$ has $\leqslant 1$ elements (by injectivity) and $\geqslant 1$ elements (by surjectivity). Hence each $f^{-1}(y)$ contains exactly one element. Conversely, suppose each set $f^{-1}(y)$ has exactly one element. In particular, each $f^{-1}(y)$ contains $\geqslant 1$ elements (so f is surjective) and each $f^{-1}(y)$ contains $\leqslant 1$ elements (so f is injective). Thus f is bijective. We have proved both implications; this establishes the $\iff$ claim. □

We can now use this to prove our first major theorem.

Theorem 16. *Let $X = \{1, \cdots, n\}$ denote the set of integers between 1 and n (so X has exactly n elements). Let $Y = \{1, \cdots, m\}$ denote the set of integers between 1 and m (so Y has exactly m elements).*

a) If $f : X \to Y$ is an injection, then $n \leqslant m$.

b) If $f : X \to Y$ is a surjection, then $n \geqslant m$.

c) If $f : X \to Y$ is a bijection, then $n = m$.

This suggests a good intuition. If there exists an injection $f : X \to Y$, you can think of this as exhibiting the claim that Y has 'at least as many elements as X'. If there exists a surjection $f : X \to Y$, you can think of this as exhibiting the claim that Y has 'at most as many elements as X'. And a bijection can be understood as exhibiting the claim 'X and Y have exactly the same number of elements.'

Correspondingly, we often write $|X|$ for the number of elements of X. (This gets subtle when X is infinite, but you can make sense of it.) The statement above says that if $f : X \to Y$ is injective, surjective, or bijective, then we have respectively $|X| \leqslant |Y|$ or $|X| \geqslant |Y|$ or $|X| = |Y|$.

Proof. Notice that every single element of X lies in some $f^{-1}(j)$ for some $1 \leqslant j \leqslant m$; in fact, it lies in *exactly one* such $f^{-1}(j)$ — we have $x \in f^{-1}(j) \iff f(x) = j$. This 'partitions' the set X into a bunch of disjoint subsets $f^{-1}(j)$, as j runs between 1 and m. Thus we have

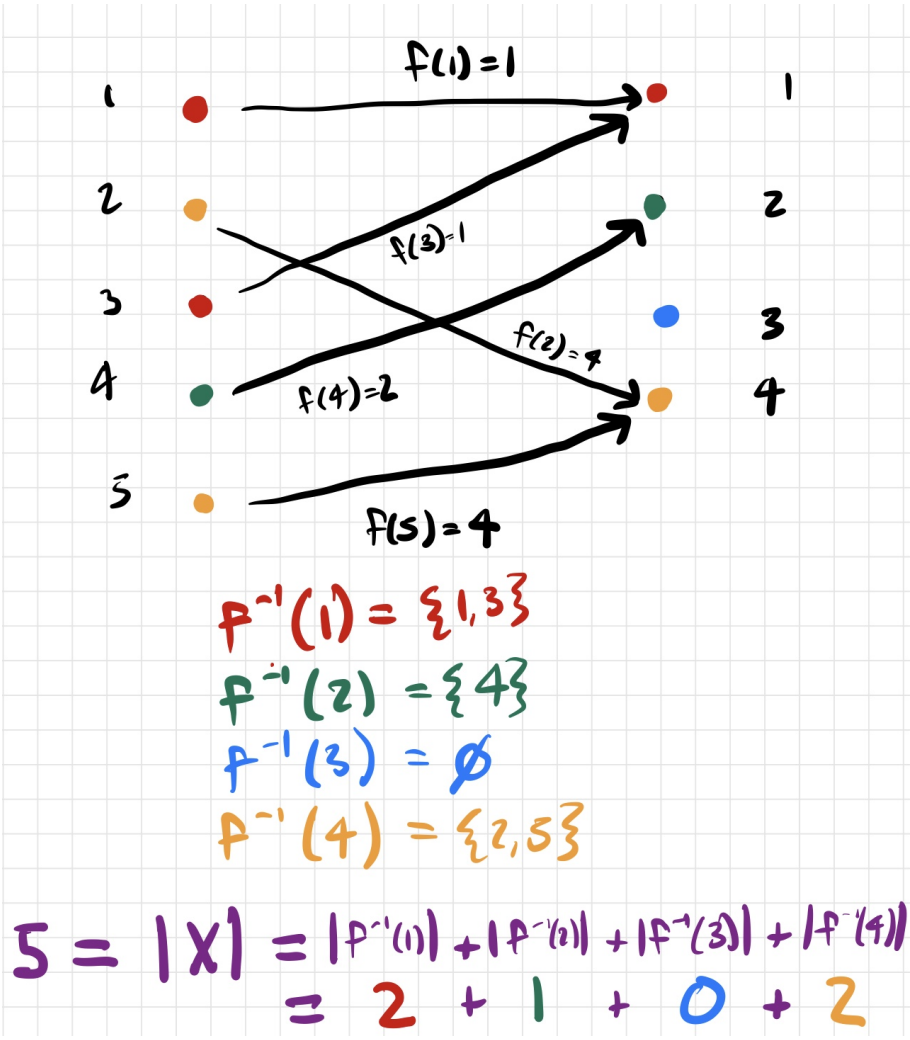
$$n = |X| = \sum_{i=1}^{m} |f^{-1}(j)|.$$

On the page is a picture which should help you intuitively justify this claim.

If f is injective, each term in the sum above is $\leqslant 1$, which implies $n \leqslant m$:

$$n = |X| = \sum_{i=1}^{m} |f^{-1}(j)| \leqslant \sum_{i=1}^{m} 1 = m.$$

On the other hand, if f is surjective, each term in the sum above is $\geqslant 1$, which implies $n \geqslant m$. Finally, if f is bijective, then each term in the sum above is $= 1$, so $n = m$. □


f(1)=1
1
2
3
4
5
1
2
3
4
f(3)=1
f(2)=4
f(4)=2
f(5)=4
$f^{-1}(1) = \{1,3\}$
$f^{-1}(2) = \{4\}$
$f^{-1}(3) = \emptyset$
$f^{-1}(4) = \{2,5\}$
$5 = |X| = |f^{-1}(1)| + |f^{-1}(2)| + |f^{-1}(3)| + |f^{-1}(4)|$
$= 2 + 1 + 0 + 2$

Chapter 4

Foundations of linear algebra

Alternate references

Linear algebra takes many forms and can be taught in as many different ways as there are linear algebra professors. (Maybe this is an exaggeration, but I can think of at least four very different ways to teach a linear algebra course).

We will be looking at linear algebra from a number of different perspectives, and no standard textbook fits all of our course goals. Still, as we cover each 'unit' in linear algebra, I will recommend alternate sources **which match the perspective of that unit.**

The next two or three weeks of this course match closely with the contents of the first chapter of Axler's textbook on linear algebra ('Linear Algebra Done Right'); we will begin to veer away from his perspective as we move forward towards the study of linear maps. If you're feeling bold, everything we'll do in these first few weeks is covered in the first 18-20 pages of Halmos's textbook 'Finite-dimensional vector spaces'. His exercises are also exceptional, and no doubt I will be using some of them on occasion. Formulated appropriately, almost all of the results of this chapter apply to infinite-dimensional vector spaces as well; precisely one towards the end is special to finite-dimensional vector spaces. These generalizations are discussed in a curio.

4.1 Introduction to linear algebra

4.1.1 What is linear algebra?

This is a difficult question to answer. Once again, I will give multiple answers.

- **Geometry of flat objects in Euclidean space.** Linear algebra is the study of objects which include lines and planes in 3-dimensional space $\mathbb{R}^3$, as well as the study of areas and volumes of flat objects (like rectangles or parallelgrams sitting in 3D space). It also includes generalizations that we cannot visualize so easily to flat objects in high-dimensional Euclidean spaces $\mathbb{R}^n$. This is the perspective that motivated the early history of Grassmann's[1] study of linear algebra, nowadays called "exterior algebra" or "geometric algebra." This is about half of the perspective taught in calculus classes, and it is a perspective which is important for physicists.

  When you study linear algebra geometrically, you will study **linear transformations of space**, such as rotations, shearing maps, reflections and scaling maps. These 'linear transformations' carry out visual operations that we see and do in our own real lives; for instance, rotation is what happens when we turn our head. This visual understanding of linear algebra is invaluable in computer graphics; a friend of mine in this field once asked me to "if nothing else, make sure your students know linear transformations are geometric things that **do stuff you can see**."

[1] Hermann Grassmann was a German mathematician in the 1800's. His work was not celebrated by his colleagues at the time, though history has redeemed him. He later moved on to being an incredibly influential linguist, most known for "Grassmann's law."

- **The abstract study of vector spaces and linear transformations.** After extensive experience with linear algebra in $\mathbb{R}^n$, mathematicians begin to notice that the use of the real numbers specifically is not crucial (for instance, one may carry out almost all basic theorems in linear algebra over either the rationals or complex numbers, and the latter fact is useful even if you're mainly interested in the real numbers). Similarly, the 'linear transformations' described above can be defined in terms of two simple properties.

  The definition of 'abstract vector space' and 'linear transformation' was written down in the early 20th century, though at the time was considered rather pointlessly abstract. Since then, mathematicians have found a great many places where this abstraction is useful (and I think the next decade or so of mathematical research convinced them, but my history on this topic is weak): any time you can fit some example into the paradigm of 'vector spaces and linear maps', you immediately get all the results we'll prove over the course of the term.

- **A computational tool.** This perspective you've probably seen a lot in high school (possibly disguised): "We like to solve systems of linear equations, and we'll show you how to do Gaussian elimination to solve them!" It's not a very honest picture of linear algebra: Gaussian elimination is a **basic tool**, not the fundamental point of linear algebra. The majority of applications of linear algebra do not come from solving some system of linear equations, but rather from the study of linear maps (or, relatedly, matrices). This shows up a lot when studying the evolution of systems that evolve by some 'transition matrix' over time (see: predator-prey models or the old widely publicized version of the PageRank algorithm), and understanding the behavior of these transition matrices helps us understand how our system changes over time, and what the "limiting behavior" is as time goes to infinity.

- **The numerical study of the behavior of finite-dimensional arrays (matrices)**. This is the perspective you would see in a 'computational linear algebra' class: the goal is to find well-behaved ('numerically stable') algorithms to quickly compute objects of interest in linear algebra: bases for kernels and images; diagonalizations; eigenvalues; inverses; ...

  This is a mind-bogglingly important part of linear algebra. On the other hand, we're still not going to talk about it in any detail! If you want to learn that, you'll have to take a class on numerical linear algebra (sometimes: "linear algebra for computation"). We don't have the time!

All of these perspectives are useful. My own perception is: one should understand linear algebra first by seeing some basic geometric examples. This should motivate one into studying the idea of vector spaces and linear transformations. When you study linear transformations, you should see the abstract definition, but also many concrete examples, and you should do concrete computations to get a sense of how it all works in practice.

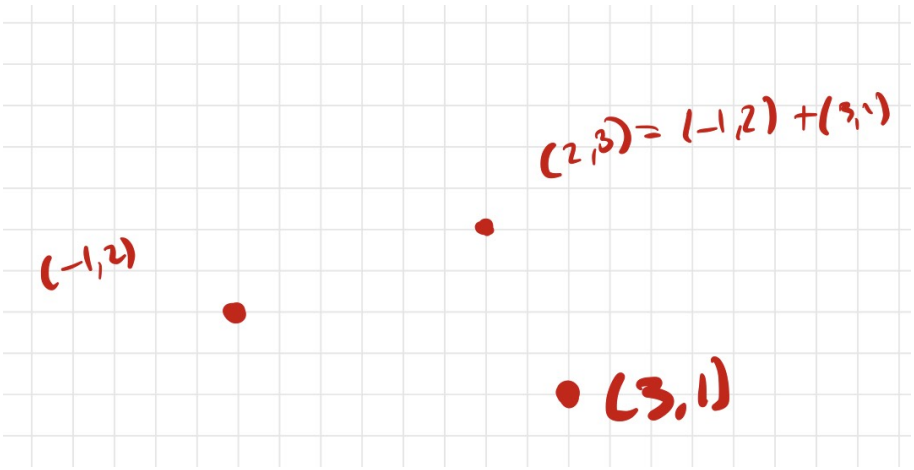
Today, we'll start off by getting that "geometric picture".

4.1.2 Coordinatewise addition

Here is an **unhelpful** first sentence for a class in linear algebra. *"We are already comfortable with addition and multiplication of numbers. In linear algebra, we move on to addition and multiplication of pairs or triples or sequences of numbers. For instance, we add in each coordinate separately: we would define* $(3,1)+(-1,2)=(2,3)$*."*

I think this is unhelpful for two reasons. First, it's not clear to me why I care about adding pairs of numbers. Secondly, and perhaps most importantly, I cannot make heads or tails of what this addition operation is supposed to "represent". What am I trying to model by defining addition this way?

Here is a picture of the three points $(3,1)$, $(-1,2)$, $(2,3)$ (for the sake of discussion, I'm not displaying the origin):

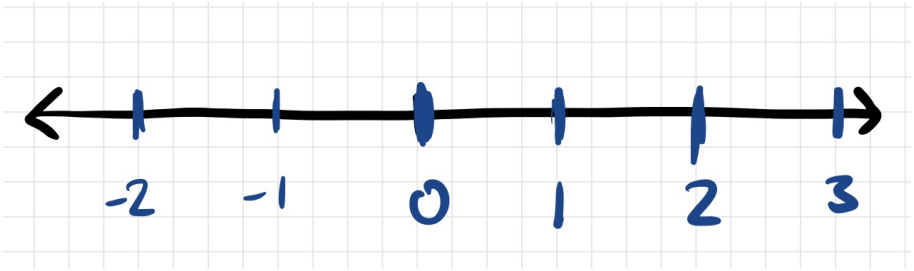


In what sense does the first and second point add to the third? These appear to me to be completely and totally unrelated points. Yes, it's true, I can define addition this way. But why would I do that? What is it supposed to mean? And how would I ever use it?

This makes me confused about the very beginning of the discussion above. **What does it even mean to add numbers?**

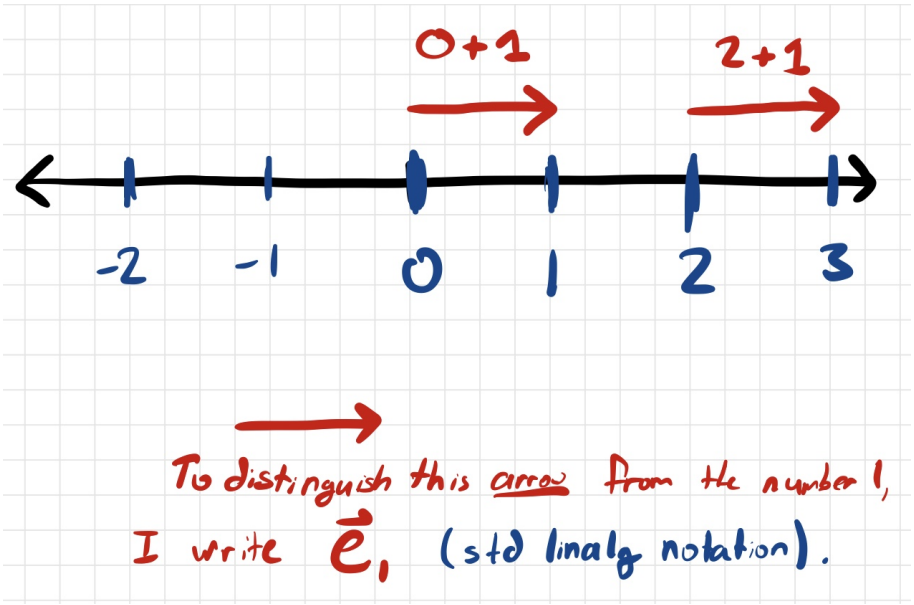
4.1.3 Vectors in $\mathbb{R}$

Let's go back to second grade. You just learned what a 'number line' is: it's a bunch of numbers on a line, each spaced out the same distance from one another.

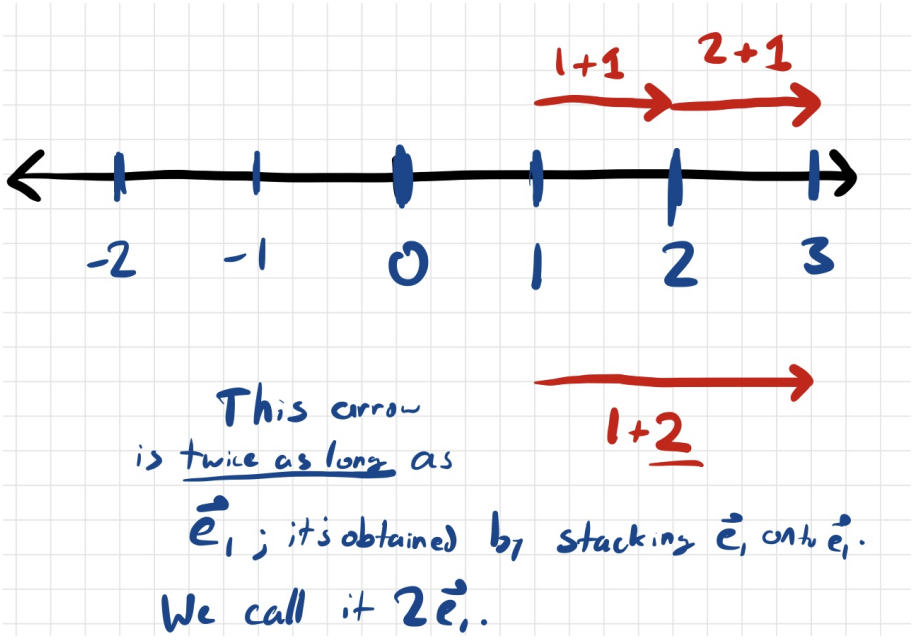


But even in terms of this picture, it is not clear to me what addition is supposed to mean. Why is $2 + 1 = 3$? What does that mean? When a grade school teacher discusses this, they will often draw arrows for the operation "add 1". To add 1 to an integer means **to move to the right**.

Write $\vec{e}_1$ for the little arrow I drew which starts at 0 and ends up at 1 on the number line. When I write that $2 + 1 = 3$, what I mean is: "If I slide that arrow so that its foot/base sits at 2, then its head sits at 3." Notice that we're freely sliding this arrow around the number line, but we always keep its length and direction (right or left) fixed. This is our first example of a 'vector'.



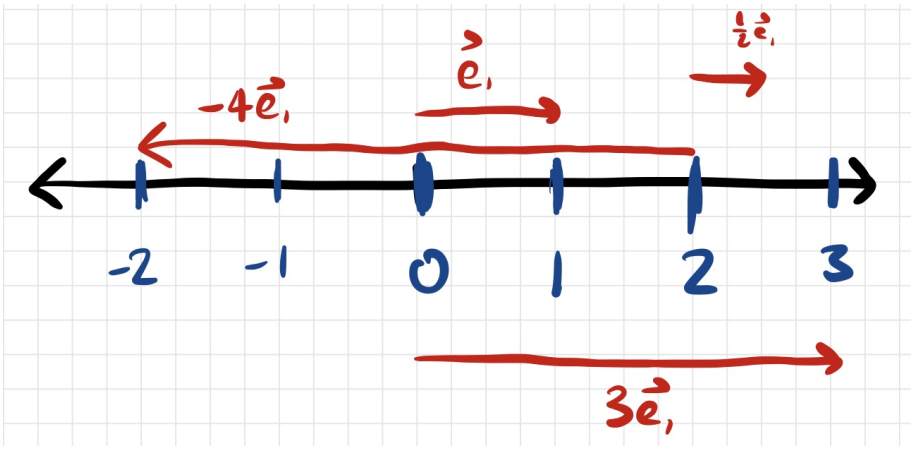
This gives us a geometric explanation for what "adding 1" should mean. But what is adding 2? What does it mean that $1 + 2 = 3$? We know that $2 = 1 + 1$. When we add 2, we should be able to do this in two steps, as $(1 + 1) + 1$. (I just used that addition should be **associative.**) So this should be the same as starting at 1 and moving to the right twice, moving one unit along $\vec{e}_1$ each time.



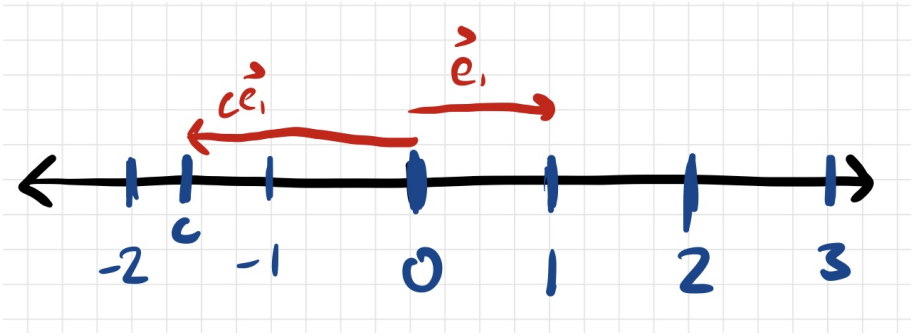
This leads us to two ideas. The first is *adding numbers/vectors*. To add $\vec{e}_1$ to itself, I took two copies of this arrow and placed them head-to-tail, and looked at the arrow I got by concatenating them. We wrote $2\vec{e}_1 = \vec{e}_1 + \vec{e}_1$ to mean the vector I got as the result. It still points right, but it moves twice the distance in total.

This leads us to the idea of *multiplying numbers*, or *scaling numbers*. This new vector $2\vec{e}_1 = \vec{e}_1 + \vec{e}_1$ moved twice as far as the original $\vec{e}_1$. We can carry out the same idea for other numbers on the number line: $3\vec{e}_1$ moves three units to the right, while $\frac{1}{2}\vec{e}_1$ moves half a unit right, and $-4\vec{e}_1$ moves four units **left** (the minus indicating that we do the **opposite** of what we wanted to do at the beginning; instead of moving 4

units right, we undo that in moving 4 units left).



If c is a number, then $c\vec{e}_1$ will be represented by the arrow which starts at 0 and ends at c.



Thus addition is given by stacking arrows together, and scaling is given by making vectors longer or shorter (or possibly moving in the opposite direction). These operations satisfy some simple relations; they're associative, distributive, and commutative (doesn't matter whether I move 3 right first and then 1 right or if I move 1 right then 3 right; I still end up moving 4 units right in total).

This in hand, I can try to give a sense of what vectors in $\mathbb{R}^2$ are meant to be.

4.1.4 Vectors in $\mathbb{R}^2$

What is a vector?

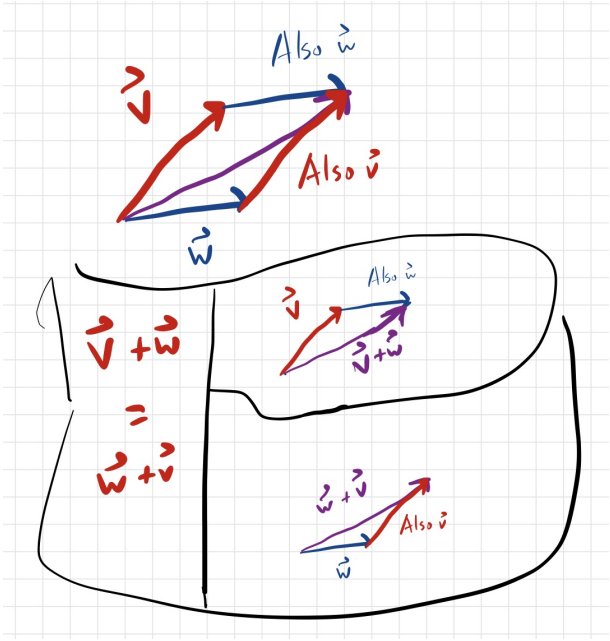
First, let me try to formalize the previous discussion.

Definition 10. A vector $\vec{v} \in \mathbb{R}^2$ is an arrow between two points in the plane, where we consider two arrows $\vec{v}$ and $\vec{w}$ to be **equal** if we can translate (slide without changing angle or length) $\vec{v}$ to obtain $\vec{w}$. ◊

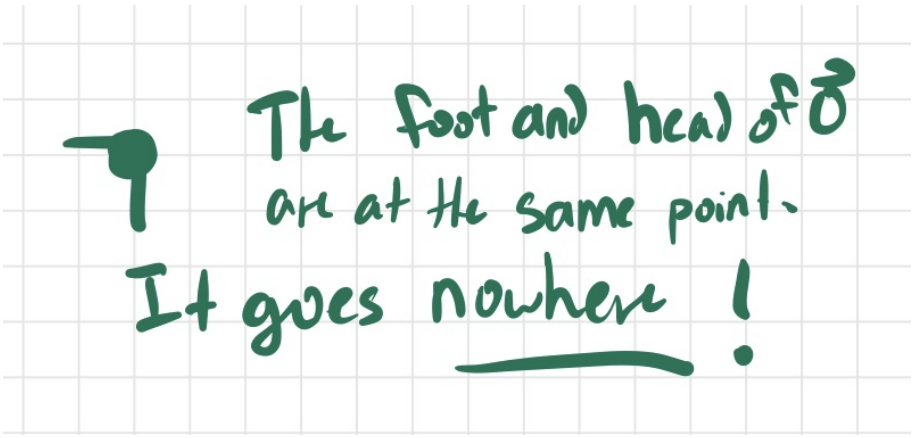
Notice that there are no numbers involved whatsoever. All of this is completely geometric. Vectors are just arrows! And the same is true for the main operations on vectors.

What are vector operations?

Definition 11. Suppose we have two vectors $\vec{v}, \vec{w} \in \mathbb{R}^2$. Their **sum** $\vec{v} + \vec{w}$ is the vector obtained as follows: translate $\vec{w}$ so that its foot is placed at the head of $\vec{v}$. Then $\vec{v} + \vec{w}$ is the vector whose foot lies at $\vec{v}$'s foot, and whose head lies at $\vec{w}$'s head. ◊



This operation is commutative: $\vec{v} + \vec{w} = \vec{w} + \vec{v}$. Both of these represent the diagonal of a parallelogram with sides $\vec{v}$ and $\vec{w}$:



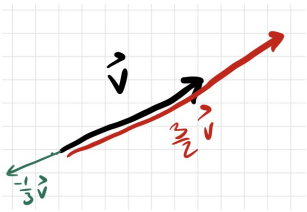
Similarly (but more simply), this operation is also associative.

Remark 17. It is sometimes helpful to think of the vector $\vec{v}$ as an instruction. If I sit at a point P, then I can cook up a new point $P + \vec{v}$ by placing the vector so that its foot is at P, and defining $P + \vec{v}$ to be the location of the head of $\vec{v}$. In this way, a vector tells us an instruction: it tells us how to move from one point to another.

In this perspective, the vector sum $\vec{v} + \vec{w}$ is the simple instruction "First follow the instruction $\vec{v}$ gives you, then the instruction $\vec{w}$ gives you."

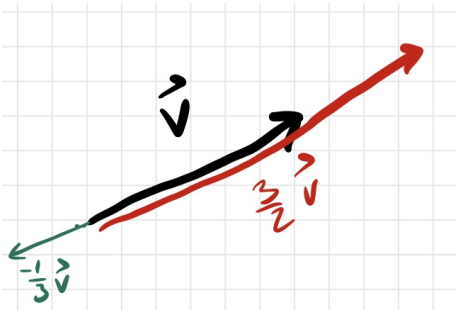
If this remark doesn't make sense, don't worry about it. ◊

Definition 12. The **zero vector** $\vec{0}$ is the vector which starts at some point P and ends at that same point P; it does not move anywhere at all. ◊



This vector has the property that $\vec{0} + \vec{v} = \vec{v} + \vec{0} = \vec{v}$. Adding the zero vector to another vector doesn't change anything at all. (From the perspective of 'instructions', the zero vector is the instruction: "Stay put!")

Definition 13. If $\vec{v} \in \mathbb{R}^2$ is a vector and $c \in \mathbb{R}$ is a number, we say the **scaled vector** $c\vec{v} \in \mathbb{R}^2$ is the vector parallel to $\vec{v}$, which is $|c|$ times as long, and points in the same direction as $\vec{v}$ if $c > 0$ but the opposite direction if $c < 0$. ◊



Scaling is a simple operation (it is associative in the sense that $(cd)\vec{v} = c(d\vec{v})$, it distributes over addition, and scaling by 1 changes nothing whatsoever).

One crucial fact is true precisely because we can divide by real numbers:

If $\vec{v}$ and $\vec{w}$ are parallel, in the sense that $c\vec{v} = d\vec{w}$ for some $c, d \neq 0$, then $\vec{w}$ can be written as a multiple of $\vec{v}$.

(Proof: We have $\vec{w} = \frac{c}{d}\vec{v}$ and similarly $\vec{v} = \frac{d}{c}\vec{w}$.) This may sound trivial, but it ends up being crucial to a lot of the theory later!

What are coordinates?

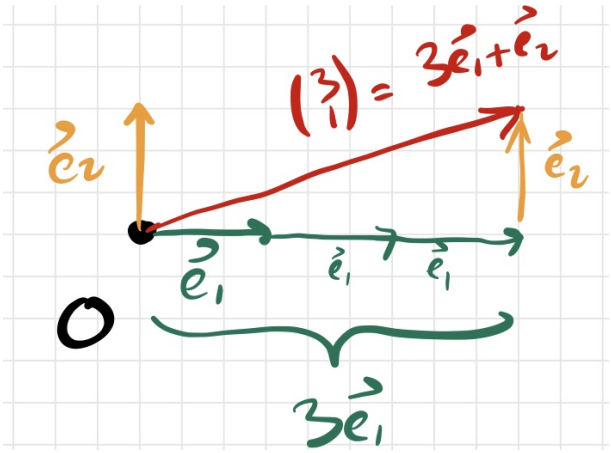
This has all been pictorial so far. Let me conclude by going back to the original question: **what does it mean to say that** $(-1, 2) + (3, 1) = (2, 3)$**?**

To make sense of this, we need to make some choices. Somewhere on the plane:

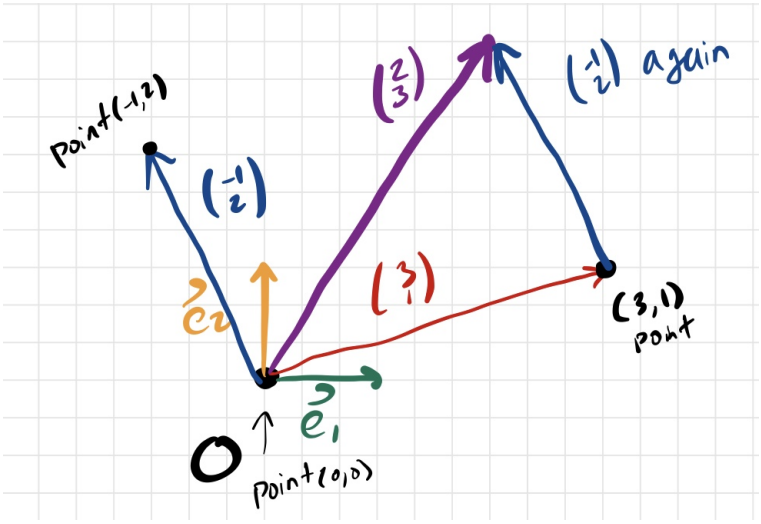
- Pick a point O to represent the origin.
- Pick a vector $\vec{e}_1$ starting from the origin to represent "moving one unit right".
- Pick a vector $\vec{e}_2$ (we usually choose the one perpendicular and counterclockwise to this, of the same length) to represent "moving one unit up".

Then for any vector $\vec{v}$, we can represent this vector as $x\vec{e}_1 + y\vec{e}_2$, or: "move x units right, and y units up." We write this vector as $\vec{v} = \begin{pmatrix} x \\ y \end{pmatrix}$.

Here is a picture of $\begin{pmatrix} 3 \\ 1 \end{pmatrix}$:



Now let's draw the vectors $\begin{pmatrix} 3 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} -1 \\ 2 \end{pmatrix}$ and $\begin{pmatrix} 2 \\ 3 \end{pmatrix}$:



Finally, our original picture has some meaning! The operation we defined completely geometrically gives us precisely the notion of coordinate-wise addition:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = (x_1\vec{e}_1 + x_2\vec{e}_2) + (y_1\vec{e}_1 + y_2\vec{e}_2) = (x_1 + y_1)\vec{e}_1 + (x_2 + y_2)\vec{e}_2 = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \end{pmatrix}.$$

Similarly, the notion of coordinate-wise scaling is given by

$$c\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = c(x_1\vec{e}_1 + x_2\vec{e}_2) = (cx_1)\vec{e}_1 + (cx_2)\vec{e}_2 = \begin{pmatrix} cx_1 \\ cx_2 \end{pmatrix}.$$

I hope I've convinced you that some standard operations on numbers or pairs of numbers — addition and scaling — can be represented entirely geometrically. Whenever a geometric intuition is useful (for instance, in any physics problem), this notion of vector becomes useful. On the other hand, we will see that the entire apparatus of linear algebra is useful far outside the context of physics (or geometry).

Using the discussion today as motivation for what 'linear algebra' ought to be, over the next few weeks we will set up the axioms and some of the basic theory.

4.2 Fields: where scalars live

Last time, we looked at an important motivating example for the study of linear algebra: vectors in $\mathbb{R}^2$ and the algebraic structure on them. The most important things were:

- Given two vectors in $\mathbb{R}^2$, we can *add* them to obtain another vector;
- Given a vector in $\mathbb{R}^2$ and a number ('scalar') in $\mathbb{R}$, we can scale the vector by the given quantity to obtain a new vector.

What I want to do in this section is go from this to an axiomatization of the relevant object of study: a *vector space*. This is precisely what is outlined in the bullet points above: something where we know how to add and we know how to scale, where the two notions of 'addition' and 'scaling' satisfy some reasonable axioms which are satisfied for $\mathbb{R}^2$.

The first thing to axiomatize is *what we scale by*. The obvious answer is: "We scale a vector by a number!" But what kind of number, and what axioms do these have to satisfy? It turns out we can do linear algebra with scalars in any of $\mathbb{Q}, \mathbb{R}$, or $\mathbb{C}$, as well as more exotic examples (which do appear in applications). The reason we come up with the axiom scheme below is that it allows us to see *exactly* the generality we can use our work — no need to repeat the arguments again the moment I study linear algebra with a new kind of scalar.

Mathematicians call a set of scalars with the required properties a "field". (The name is an accident of history; it doesn't carry much meaning.) I'm going to write the axioms below, and then we'll start to reason about it.

To make it easy to refer back to, I'm going to put the definition of a field on its own page. Notice that I divide the definition into two parts. This is common in abstract algebra; you will have a structure which is defined in terms of a collection of operations, and a list of axioms those operations must satisfy. Here we have two: addition and multiplication.

Definition 14. A **field** is three pieces of data $(\mathbb{F}, +, \cdot)$
The data are:

(D1) A set **F** ('the scalars', or 'the elements of the field'),

(D2) An addition map $+ : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$, meaning for every pair of scalars $a, b \in \mathbb{F}$, we define a sum $a + b \in \mathbb{F}$;

(D3) A product map $\cdot : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$, meaning for every pair of scalars $a, b \in \mathbb{F}$, we define a product $a \cdot b \in \mathbb{F}$.

These are required to satisfy the following axioms.

(F1) **Associativity.** Addition and multiplication are 'associative', meaning that for all $a, b, c \in \mathbb{F}$, we have

$$(a + b) + c = a + (b + c) \quad \text{and} \quad (a \cdot b) \cdot c = a \cdot (b \cdot c).$$

(F2) **Commutativity.** Addition and multiplication are 'commutative', meaning that for all $a, b \in \mathbb{F}$, we have

$$a + b = b + a \qquad \text{and} \qquad a \cdot b = b \cdot a.$$

(F3) **Distributivity.** Multiplication distributes over addition, meaning that for all $a, b, c \in \mathbb{F}$, we have

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

(F4) **Identities.** There exist elements $0, 1 \in \mathbb{F}$ which serve as 'additive and multiplicative identities', meaning that for all $a \in \mathbb{F}$, we have

$$0 + a = a \qquad \text{and} \qquad 1 \cdot a = a.$$

(F5) **Additive inverses.** For all $a \in \mathbb{F}$, there exists an element $b \in \mathbb{F}$ with $a + b = 0$. We call b the 'additive inverse' of a, and denote it as $-a$.

(F6) **Multiplicative inverses.** For all **nonzero** scalars $a \in \mathbb{F}$, there exists an element $b \in \mathbb{F}$ so that $ab = 1$. We call b the 'multiplicative inverse' of a, and denote it $1/a$ or a^{-1}.

(F7) **Nontriviality.** There are at least two elements of $\mathbb{F}$.

$\Diamond$

In symbols, the axioms are

$$\begin{aligned}
&\forall_{a,b,c\in\mathbb{F}}\big[\big[(a+b)+c = a+(b+c)\big] \wedge \big[(a\cdot b)\cdot c = a\cdot(b\cdot c)\big]\big].\\
&\quad \forall_{a,b\in\mathbb{F}}\big[\big[a+b=b+a\big] \wedge \big[a\cdot b = b\cdot a\big]\big].\\
&\forall_{a,b,c\in\mathbb{F}}\big[a\cdot(b+c) = a\cdot b + a\cdot c\big].\\
&\quad \exists_{0,1\in\mathbb{F}}\forall_{a\in\mathbb{F}}\big[\big[0+a=a\big] \wedge \big[1\cdot a = a\big]\big].\\
&\qquad \forall_{x\in\mathbb{F}}\exists_{y\in\mathbb{F}}\big[x+y=0\big].\\
&\qquad \forall_{a\in\mathbb{F}}\big[\neg\big[a=0\big] \implies \big[\exists_{b\in\mathbb{F}} ab = 1\big]\big].\\
&\quad \exists_{a,b\in\mathbb{F}}\big[a\neq b\big].
\end{aligned}$$

I want to give a handful of examples and non-examples, but first, let me make some remarks on the axioms. First, the 'data' just says "we know how to multiply and we know how to add;" nothing more. The first three axioms about these operations are things you use constantly and implicictly when doing arithmetic without even thinking about it; for instance,

$$(22 + z)(3x + 4y) = 66x + 88y + 3xz + 4yz.$$

uses all three of them. We'd like to be able to continue doing that.

The next three axioms are more interesting. they assert the *existence* of certain elements which behave well with respect to the algebraic operations; every one of these assumptions is crucial one way or another when studying vector spaces. First I'd like to list a few examples (some you know, some you may not) of fields. Afterwards, I'll talk about some algebraic structures which satisfy some (but not all) of the axioms above, and what we lose in doing so.

4.2.1 Some things which are fields

There are three fields which we will use most often throughout the course. (Frankly, most of our geometric thinking happens over $\mathbb{R}$, but the others can be useful on occasion — there are some things which $\mathbb{R}$ just can't do.)

Example 24. The rational numbers $\mathbb{Q}$ of fractions p/q (where p, q are integers, $q \neq 0$) form a field, with the usual addition

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}$$

and multiplication

$$\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}.$$

If p/q is a rational number, its additive inverse is $-p/q$; if p/q is a nonzero rational number (so $p \neq 0$) its multiplicative inverse is q/p.

I won't verify the axioms, but they're not too hard to check by hand if you believe that (F1)-(F3) are true for the integers. ♢

Example 25. The real numbers $\mathbb{R}$ form a field, too, with the usual addition and multiplication. ♢

Example 26. One example which might be more (or less) familiar to you is the complex numbers $\mathbb{C}$. We think of elements of $\mathbb{C}$ as being $x + iy$, where x, y are real numbers and i is some new number with $i^2 = -1$. Formally, as a set, we have

$$\mathbb{C} = \{x + iy \mid (x, y) \in \mathbb{R}^2\}.$$

We define the addition operation as

$$(a + ib) + (c + id) = (a + c) + i(b + d),$$

while the multiplication operation is more intricate:

$$(a + ib)(c + id) = (ac - bd) + i(ad + bc).$$

To understand where this comes from, just expand out the product as

$$ac + iac + ibc + i^2bd,$$

and remember that we want to have $i^2 = -1$. (This is how I compute the product anyway; I don't memorize the formula.)

It is not obvious that $\mathbb{C}$ is a field, but it is, and you'll prove it as a homework exercise. ♢

The three fields above are by far the most useful to us. However, let me point out that there are more examples than just these. Here is one particularly strange one.

Example 27. The **field with two elements** $\mathbb{F}_2 = \{0, 1\}$ has addition and multiplication defined as follows:

$$0 + 1 = 1 + 0 = 1, \qquad 0 + 0 = 0 = 1 + 1, \qquad 0 \cdot 0 = 0 \cdot 1 = 1 \cdot 0 = 0, \qquad 1 \cdot 1 = 1.$$

The only operation here which should be surprising is $1 + 1 = 0$. In this field, we have $2 = 0$! This field appears in two contexts. First, this is related to understanding "modular arithmetic" over the integers $\mathbb{Z}$ with respect to the modulus 2. This is a fancy way of saying that the number 0 represents "even", while 1 represents "odd". The equation $1 + 1 = 0$ corresponds to the fact that the sum of two odd integers is even. One can generalize this to define fields $\mathbb{F}_p$ with p elements $\{0, 1, \cdots, p-1\}$ for any **prime number** p, where the addition and multiplication are given by taking the usual sum or product, then subtracting off multiples of p until the result is between 0 and $p - 1$. Once again, it is not obvious that this is a field; proving it requires some background knowledge about number theory you may not have.

Another place the field $\mathbb{F}_2$ occurs is in mathematical logic itself. We can think of the elements of $\mathbb{F}_2$ as being $\{T, F\}$, with 1 representing 'true' and 0 representing 'false'. We take xor as our addition operation and $\wedge$ as our multiplication operation. ♢

This field may come back once or twice, but mostly to point out assumptions we make that don't apply in full generality.

4.2.2 Things which aren't fields

Now that we've seen some examples, let's see the ways in which a number system can fail to be a field, and get a sense for why these issues are undesirable.

Non-example 1. The set of even integers $2\mathbb{Z}$ carries an addition and multiplication operation; we set $2k+2m = 2(k+m)$ and $2k \cdot 2m = 2(2km)$. These operations are associative, commutative, and multiplication distributes over addition. There is also an additive identity $2(0) = 0 \in 2\mathbb{Z}$ and additive inverses exist, as $2k + 2(-k) = 0$ for all $k \in \mathbb{Z}$.

However, $2\mathbb{Z}$ does not include a multiplicative identity, because the integer 1 is odd! So $2\mathbb{Z}$ is not a field. ♢

We need both additive and multiplicative identities when we talk about scaling, because we'll find need for both operations "don't scale at all" and "scale everything down to zero". (They're also crucial in stating (F5) and (F6), which are even more important.)

Non-example 2. The set of non-negative real numbers $[0, \infty)$ carries addition and multiplication operations which satisfy (F1)-(F3). The real number 0 serves as an additive identity, while the real number 1 serves as a multiplicative identity. There are even multiplicative inverses: if $t \neq 0$ is a real number, there is a real number $1/t$ so that $t \cdot 1/t = 1$.

However, $[0, \infty)$ is not a field. It fails axiom (F5), the existence of additive inverses: there's no non-negative real number I can add to 1 so that $1 + x = 0$. ♢

We need additive inverses so we can talk about *subtraction*. If I asked you what the solutions to $x + 2y = 2x + y$ are, you might say "If you subtract x and y from both sides, the resulting equation reads $y = x$". When you did so, you implicitly used all of axioms (F1), (F2), (F4), and (F5): first, in subtracting x, you assert that there is some number $-x$ with $x + (-x) = 0$. Then the left side simplifies to

$$(x + 2y) + (-x) = (2y + x) + (-x) = 2y + (x + (-x)) = 2y + 0 = 0 + 2y = 2y;$$

spot where I use each of (F1), (F2), and (F4) here. Simplifying the right-hand side even uses (F3) and (F4) to rewrite $2x + (-1)x = (2 - 1)x = 1x = x$!

The last axiom is the most subtle.

Non-example 3. The integers $\mathbb{Z}$ with the usual addition and multiplication satisfy all axioms (F1)-(F5). However, they fail (F6): multiplicative inverses usually do not exist. There is no integer x so that $2x = 1$; that variable x wants to be $1/2$, which is not an integer. ♢

Without multiplicative inverses, it becomes much harder to describe solutions to simple equations like $2x = 3y$. Over a field,[2] I can tell you right away "The solutions are pairs $(\frac{3}{2}y, y)$, where $x \in \mathbb{F}$": the expression "3/2" makes sense, because it's $3 \cdot (1/2)$, where $1/2$ is the multiplicative inverse of $2 \neq 0$. This is because I can scale both sides by $1/2$, and

$$1/2 \cdot (2x) = (1/2 \cdot 2) \cdot x = 1 \cdot x = x, \quad \text{while} \quad 1/2 \cdot (3y) = (1/2 \cdot 3)y = (3/2) \cdot y.$$

Over the integers, however, the answer is that (x, y) is a solution if it takes the form $(3n, 2n)$ for some integer $n \in \mathbb{Z}$. This feels to me like a clunkier answer, and one that's harder to find. It's possible to make more complicated examples, too.

Non-example 4. There is only one $\mathbb{F}$ which satisfies axioms (F1)-(F6) but fails axiom (F7): $\mathbb{F} = \{0\}$, with $0 + 0 = 0$ and $0 \cdot 0 = 0$. In this not-quite-a-field, 0 is the multiplicative identity; in some sense this means $0 = 1$ in this not-quite-a-field.

There is nothing interesting to say about this $\mathbb{F}$, and using it makes a bunch of different actually interesting theorems false. So we add (F7) to exclude it from the set of fields. ◊

4.2.3 Our first facts about fields

Now that I've tried to justify why these axioms are good ideas (and seen a few examples), let me go through some standard consequences of the field axioms: things which serve as a sanity check that our field axioms make sense and nothing too weird happens.

Lemma 17. *The additive identity in a field is unique.*

To be precise, suppose 0 *and* $0'$ *both satisfy the following property: for all* $a \in \mathbb{F}$*, we have*

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

Then $0 = 0'$.

Proof. Apply $0 + a = a$ to $a = 0'$ to see that $0 + 0' = 0'$. On the other hand, apply $0' + a = a$ to $a = 0$ to see that $0' + 0 = 0$. Combining these (and axiom A2, that addition is commutative), we see that

$$0 = 0' + 0 = 0 + 0' = 0',$$

so that $0 = 0'$. □

Thus we are justified in saying **the** additive identity 0.

Next, remember that for every $a \in \mathbb{F}$, we know that there *exists* an additive inverse b with $a + b = 0$. However, this doesn't really justify the notation "$-a$", which suggests we've pinned down *a specific* additive inverse. Our next goal is to show that these are unique, and then to give a formula for the additive inverse.

Lemma 18. *Additive inverses in a field are unique.*

To be precise, fix an element $a \in \mathbb{F}$*. If* b, b' *are two elements so that* $a + b = 0 = a + b'$*, then* $b = b'$.

Proof. I want to cancel out a from both sides of the equation $a + b = a + b'$. To do that, I'll use one of these additive inverses. Add b to both sides and simplify; we see

$$b = 0 + b = (a + b') + b = a + (b' + b) = a + (b + b') = (a + b) + b' = 0 + b' = b',$$

so that $b = b'$; that proves what we wanted to show.

The middle equality is where we "add b to both sides"; really, I pass from $(a + b) + b$ to $(a + b') + b$ because I happen to know that $a + b = a + b'$. Everything else is either (F1), (F2), (F4), or (F5).

As you've seen in practice by now, it doesn't really matter how I arrange the brackets when I'm adding different elements of $\mathbb{F}$, nor does it matter what order we write them in. If we internalize this, we could write the previous string of equalities more simply as

$$b = 0 + b = (a + b) + b = (a + b') + b = (a + b) + b' = 0 + b' = b',$$

where we skipped all the juggling needed to go from $(a + b') + b$ to $(a + b) + b'$. □

[2] Technically, I'm assuming here that $1 + 1 \neq 0$, which is not true in every field! It's true in every case we'll cover in this class, though.

From now on, as discussed above, I will be more brief when passing between two expressions like $(a+b')+b$ and $(a+b)+b'$.

In the next theorem, we show that the additive identity behaves like you expect in terms of multiplication:

Lemma 19. *Let $\mathbb{F}$ be a field, and write $0 \in \mathbb{F}$ for the additive identity. Then $0 \cdot x = 0$ for all $x \in \mathbb{F}$.*

Proof. This uses axioms (F1), (F3), (F4), and (F5): we actually need subtraction to show this!

The trick is to use the fact that $0 + 0 = 0$, and multiply by x; this will give us "two copies" of $0 \cdot x$ on one side, and one copy on the other; cancelling one copy out shows that $0 \cdot x = 0$.

We have

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

Adding $-(0 \cdot x)$ to both sides, we see that

$$(0 \cdot x + 0 \cdot x) + (-(0 \cdot x)) = 0 \cdot x + (-(0 \cdot x));$$

the left-hand side simplifies to $0 \cdot x$, while the right-hand side simplifies to 0; thus we see that

$$0 \cdot x = 0.$$

□

We needed the previous lemma to show

Lemma 20. *If $a \in \mathbb{F}$, then the additive inverse is given by $-a = (-1) \cdot a$.*

Proof. Here we'll use (F3), plus the stuff that came before. Because additive inverses in a field are unique (Lemma 18), it suffices to show that $(-1) \cdot a$ *is* an additive inverse of a. For this, we use distributivity:

$$a + (-1) \cdot a = 1 \cdot a + (-1) \cdot a = (1 + (-1)) \cdot a = 0 \cdot a = 0.$$

In the last step, we used Lemma 19, that the product of any scalar with 0 is again 0. □

That about covers everything there is to say about addition and additive inverses. I want to move on to an important property of multiplication.

Proposition 21. *If $a, b \in \mathbb{F}$, then $a \cdot b = 0$ if and only if $a = 0$ or $b = 0$.*

Proof. Here is where multiplicative inverses make their first (and crucial!) appearance.

First, let's show

$$[a = 0] \text{ or } [b = 0] \implies [a \cdot b = 0].$$

If $a = 0$, then we can use Lemma 19 to see that $0 \cdot b = 0$, as desired. On the other hand, if $b = 0$, then notice that $a \cdot 0 = 0 \cdot a = 0$, first using commutativity and then applying the same lemma.

The other direction is more interesting. We're trying to show

$$a \cdot b = 0 \implies [a = 0] \vee [b = 0].$$

Because this is logically equivalent to

$$[a \cdot b = 0] \wedge \neg[a = 0] \implies b = 0,$$

we can rephrase this as follows: suppose $a \cdot b = 0$ and $a \neq 0$. Prove that $b = 0$.

Why is this good? Because $a \neq 0$ promises us the existence of a multiplicative inverse $c \in \mathbb{F}$ with $c \cdot a = 1$. Now, because $a \cdot b = 0$, we see that

$$c \cdot (a \cdot b) = c \cdot 0;$$

the left-hand side simplifies to

$$c \cdot (a \cdot b) = (c \cdot a) \cdot b = 1 \cdot b = b,$$

while the right-hand side simplifies to

$$c \cdot 0 = 0 \cdot c = 0$$

by Lemma 19. This proves that $b = 0$. Thus $ab = 0$ and $a \neq 0$ implies $b = 0$, which is what we wanted to show. □

The property above can hold even when you don't have multiplicative inverses (for instance, it's true in $\mathbb{Z}$), but it's usually harder to establish.

4.3 Vector spaces: where vectors live

Last time, we talked about the kind of object the "set of scalars" should be. Today we'll get into the meat of the subject: what is a vector space, and what properties should it satisfy?

When studying vector spaces, we'll always be studying them with respect to a **fixed** field of scalars. There's no good way to discuss the relationship between vector spaces where your scalars are in $\mathbb{Q}$ and vector spaces where your scalars are in $\mathbb{F}_5$. This isn't really an issue, and there is very rarely a good reason to **want** to think about multiple kinds of scalars at once.

Definition 15. A vector space $(V, +, \cdot)$ (over the field $\mathbb{F}$) consists of the following data, satisfying the following axioms.

(D1) A set V of vectors.

(D2) An addition map $+ : V \times V \to V$; that is, for each $v, w \in V$, we define a sum $v + w \in V$;

(D3) A scalar multiplication map $\cdot : \mathbb{F} \times V \to V$; that is, for each scalar $a \in \mathbb{F}$ and each vector $v \in V$, we define a scalar multiplication $a \cdot v \in V$. We sometimes abbreviate this to av, without the $\cdot$.

These are required to satisfy the following axioms.

(V1) **Associativity.** Addition and scalar multiplication in V are associative: for all $v, w, u \in V$, we have

$$(v + w) + u = v + (w + u),$$

and for all $a, b \in \mathbb{F}$ we have

$$(ab) \cdot v = a \cdot (bv).$$

(V2) **Commutativity.** Addition is commutative: for all $v, w \in V$, we have

$$v + w = w + v.$$

(V3) **Distributivity.** Scalar multiplication distributes over addition: for all $a \in \mathbb{F}$ and all $v, w \in V$, we have

$$a(v + w) = av + aw.$$

Similarly, for all $a, b \in \mathbb{F}$ and all $v \in V$, addition distributes over scalar multiplication: we have

$$(a + b)v = av + bv.$$

(V4) **Additive identity.** There exists an additive identity, the "zero vector" $\vec{0} \in V$, so that for all $v \in V$ we have

$$\vec{0} + v = v.$$

(V5) **Additive inverse.** For any $v \in V$, there exists a $w \in V$ so that $v + w = \vec{0}$. We usually denote w by the symbol $-v$.

(V6) **Multiplicative identity.** If $1 \in \mathbb{F}$ is the multiplicative identity, then for any $v \in V$, we have $1 \cdot v = v$.

$\Diamond$

Remark 18. Notice that **it does not make sense** to assert that scalar multiplication is commutative. The inputs of scalar multiplication are a scalar $a \in \mathbb{F}$ and a vector $v \in V$, and the output is the vector av. There is no way to "swap the two inputs"; va is not something we've defined, nor would it be terribly interesting to talk about it.

For similar reasons, there is no notion of a "multiplicative inverse" for scalar multiplication. The only multiplicative identity in sight is $1 \in \mathbb{F}$, which is a scalar; on the other hand, scalar multiplication produces a vector. It does not make sense to write $av = 1$. ◊

Remark 19. This time, I separated out the additive identity and multiplicative identity axioms into two separate axioms. That's because here they're actually of a different character: here (V4) asserts the **existence** of a certain vector, while (V6) prescribes the way **the scalar** $1 \in \mathbb{F}$ behaves; this scalar is already known to exist from the field axioms. ◊

Proposition 22. *Let V be a vector space over a field $\mathbb{F}$. Then the following are true.*

(P1) Additive identities in V are unique: if $\vec{0}$ and $\vec{0}'$ are both additive identities, then $\vec{0} = \vec{0}'$.

(P2) Additive inverses in V are unique: if $v \in V$ is a vector and $w, w' \in V$ have $v + w = \vec{0} = v + w'$, then $w = w'$.

(P3) If $a \in \mathbb{F}$ and $v \in V$, then $a \cdot v = \vec{0}$ if and only if either $a = 0$ or $v = \vec{0}$.

(P4) The additive inverse to $v \in V$ is given by $(-1) \cdot v$, the scalar multiplication of v by $-1 \in \mathbb{F}$.

The proofs of these propositions mirror those in the previous section, and I will not repeat them (though I will ask you to prove at least one of these on the homework).

4.3.1 Examples

Last week we discussed the example of $\mathbb{R}^2$ (and more generally $\mathbb{R}^n$). This same example makes sense — though becomes much less visual — over any field.

Example 28. Fix a field $\mathbb{F}$. The vector space $\mathbb{F}^n$ has as elements n-tuples of elements in $\mathbb{F}$. For instance, $(3, \pi, 2)$ is an element of $\mathbb{R}^3$, while $(1, 1/2, 1/3, 1/4, 1/5)$ is an element of $\mathbb{Q}^5$.

For good reasons I will justify later, we usually write elements of $\mathbb{F}^n$ as *vertical lists*

$$\vec{v} = \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} \in \mathbb{F}^n \iff \text{For all } 1 \leqslant i \leqslant n \text{ we have } a_i \in \mathbb{F}.$$

For instance, I will write $\begin{pmatrix} 1 \\ 2 \end{pmatrix} \in \mathbb{R}^2$ to refer to the vector which points one unit right and two units up. The different terms $a_1, \cdots, a_n$ are usually called the *components* or *coordinates* of $\vec{v}$.

To describe a vector space V over a field $\mathbb{F}$, I need to give you three things: a set V (which I have given you), as well as an addition map and a scalar multiplication map, which I still have to describe. The addition map on $\mathbb{F}^n$ is defined coordinatewise:

$$\begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} + \begin{pmatrix} b_1 \\ \cdots \\ b_n \end{pmatrix} = \begin{pmatrix} a_1 + b_1 \\ \cdots \\ a_n + b_n \end{pmatrix}$$

(notice that this makes sense because we already have a notion of addition in $\mathbb{F}$, which is what we're using in each coordinate!)

The scalar multiplication on $\mathbb{F}^n$ is also defined coordinatewise. Given $c \in \mathbb{F}$ and $\vec{v} \in \mathbb{F}^n$, we define the scalar multiplication by

$$c \cdot \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} = \begin{pmatrix} ca_1 \\ \cdots \\ ca_n \end{pmatrix}.$$

It is straightforward to verify that the vector space axioms follow from the field axioms. The additive identity is the zero vector

$$\vec{0} = \begin{pmatrix} 0 \\ \cdots \\ 0 \end{pmatrix};$$

the additive inverse is

$$-\begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} = \begin{pmatrix} -a_1 \\ \cdots \\ -a_n \end{pmatrix},$$

and for instance, we have

$$\begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} + \begin{pmatrix} b_1 \\ \cdots \\ b_n \end{pmatrix} = \begin{pmatrix} a_1 + b_1 \\ \cdots \\ a_n + b_n \end{pmatrix} = \begin{pmatrix} b_1 + a_1 \\ \cdots \\ b_n + a_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \cdots \\ b_n \end{pmatrix} + \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix},$$

which establishes the vector space axiom (V2). Here I used the field axiom (F2) to assert that $a_i + b_i = b_i + a_i$ for each i.

The justifications for (V1), (V3), and (V5)-(V6) are similar: they reduce to the corresponding claims for fields. ◊

The preceding example is the foundational example — in some sense, every example can be artificially made to look like the example above (more on this later).

This example naturally generalizes to an 'infinite-dimensional version' (not that we know what 'dimension' means!)

Example 29. Fix a field $\mathbb{F}$, and let X be any set whatsoever. We write $\mathrm{Map}(X, \mathbb{F})$ for the set of functions $f : X \to \mathbb{F}$. That is, an element of $\mathrm{Map}(X, \mathbb{F})$ is a function $f : X \to \mathbb{F}$.

The addition is termwise. We define a function $f + g$ by the formula

$$(f + g)(x) = f(x) + g(x),$$

which makes sense because $f(x)$ and $g(x)$ are unambiguously defined elements of $\mathbb{F}$, and we know how to add two elements of $\mathbb{F}$.

The scalar multiplication is termwise, too. If $c \in \mathbb{F}$ and $f \in \mathrm{Map}(X, \mathbb{F})$, we define a new function by

$$(cf)(x) = cf(x).$$

Once again, this makes sense because $f(x) \in \mathbb{F}$, and I know how to take the product of two elements of $\mathbb{F}$.

The "zero vector" is the function $f_0(x) = 0$ which takes every input x and spits out the number $0 \in \mathbb{F}$. If $f : X \to \mathbb{F}$ is a function, its additive inverse is $(-f)(x) = -f(x)$. The vector space axioms here follow quickly once more from the field axioms. ◊

Remark 20. Example 28 is more or less a special case of Example 29: take $X = \{1, \cdots, n\}$. A function $f : \{1, \cdots, n\} \to \mathbb{F}$ is the same as the data of n elements $f(1), \cdots, f(n) \in \mathbb{F}$, or in other words is the same data as an n-tuple $(f(1), \cdots, f(n)) \in \mathbb{F}^n$. The novelty of Example 29 is it allows for possibly infinite sets X, which give rise to "infinite-dimensional vector spaces".

Formally, there is a bijection between the set $\mathrm{Map}(\{1, \cdots, n\}, \mathbb{F})$ and the set $\mathbb{F}^n$, and this bijection "preserves" the addition and scaling. This is called an isomorphism, and we will be spending a lot of time thinking about these soon enough. ◊

Example 30. There are a family of examples from analysis, written $C^k(\mathbb{R})$ for $0 \leqslant k \leqslant \infty$.

The first of these, $C^0(\mathbb{R})$, is the set of **continuous functions** $f : \mathbb{R} \to \mathbb{R}$. I would like to say that this has a sum operation defined as follows: if f, g are continuous functions $\mathbb{R} \to \mathbb{R}$, then we define $f + g$ to be the function $(f + g)(x) = f(x) + g(x)$. To say that $f + g \in C^0(\mathbb{R})$ means that this function is again continuous. This is something which is usually asserted in a standard Calculus class, and is not hard to prove **as soon as you have a rigorous definition of continuity**. I will take it for granted.

Similarly, if $c \in \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$ is continuous, the function cf defined by $(cf)(x) = cf(x)$ is also continuous, so this defines a scalar multiplication map on $C^0(\mathbb{R})$.

The next set, $C^1(\mathbb{R})$, consists of those functions which are continuous, and for which $f : \mathbb{R} \to \mathbb{R}$ has a well-defined derivative f', for which f' is also a continuous function. Again, this has a sum and scalar multiplication defined by the same formula, and the key observation is that if f, g are differentiable so is $f+g$ with $(f+g)' = f'+g'$, and therefore if f', g' are continuous so is $(f+g)'$; similarly with scalar multiplication.

In general, $C^k(\mathbb{R})$ is the set of functions for which the first k derivatives are defined everywhere, and so that the first k derivatives are all continuous functions. The set $C^\infty(\mathbb{R})$ is the set of functions for which all derivatives are defined and continuous. ◊

Let me conclude with a sort of silly example.

Example 31. In Example 28, take $n = 0$. The vector space $\mathbb{F}^0$ is supposed to be the set of 0-tuples of elements of $\mathbb{F}$. It's not clear what this should mean (I take... no elements of $\mathbb{F}$?) By convention, we usually set $\mathbb{F}^0 = \{\vec{0}\}$ to consist of a single element, named 0, for which $c \cdot \vec{0} = \vec{0}$ and $\vec{0} + \vec{0} = \vec{0}$. This rather trivially satisfies all of the vector space axioms, and this is usually called the **trivial vector space**. You can visualize it as a dot, a single point. It is more important than it seems. ◊

4.3.2 Non-examples

Just like with fields, the first three axioms of vector spaces — (V1)–(V3) — are purely arithmetical. They usually hold automatically in most examples of interest, and are necessary to do anything even remotely interesting. What is more interesting is things that fail because of axioms (V4)–(V6) (or places where you fail to provide the necessary data in (D1)-(D3).)

Non-example 5. The empty set $\varnothing$ is not a vector space over any field. The issue is Axiom (V4), which asserts that **there exists** a vector $\vec{0} \in V$ — yes, with a particular property, but in particular there is at least one vector! ◊

Non-example 6. Here is a bizarre example where additive identities exist but additive inverses do not. If $\mathbb{F}$ is a field, write $\mathbb{F}_o = \mathbb{F} \cup \{o\}$ for the set $\mathbb{F}$ together with a new element named o. We define a sum operation on $\mathbb{F}_o$ by setting, for $a, b \in \mathbb{F}_o$,

$$a + b = \begin{cases} a + b \in \mathbb{F} & a, b \in \mathbb{F} \\ b & a = o \\ a & b = o \\ o & a = b = o \end{cases}.$$

We also define scalar multiplication as usual on $\mathbb{F}$, but set $a \cdot o = o$ for all $a \in \mathbb{F}$. It is irritating but not difficult to establish (V1)-(V3) for this set. Further, $o \in \mathbb{F}_o$ is an additive identity, as $a + o = a$ for all $a \in \mathbb{F}_o$. It's a bit like we added a new copy of 0 to our set, except that $0 + o = o + 0 = o \neq 0$, so 0 is no longer the additive identity.

As a result, there are no additive inverses! The only two elements that add to o are $o + o = o$. Every other sum spits out an element in $\mathbb{F}$. So no element except for o itself has an additive inverse. (What goes wrong in the proof that $-1 \cdot v$ should be an additive inverse to v?) ◊

Non-example 7. The following extremely artificial example satisfies every axiom but (V6), but is also totally worthless. Fix your favorite field $\mathbb{F}$. Let $V = \mathbb{F}$ with the usual addition, but with the new multiplication

$$\text{For all } c \in \mathbb{F} \text{ and } v \in V, \text{ set } cv = \vec{0}.$$

Then $1 \cdot v = \vec{0} \neq v$ for most v, so (V6) fails. In this vector space, there is no way to "unscale", because scaling is a silly operation that crushes everything to zero. ◊

4.4 Subspaces of vector spaces

In studying the geometry of Euclidean space, one often wants to consider lines and planes in that Euclidean space, such as

$$V = \left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3 \mid x + y + z = 0 \right\}.$$

Exercise. Verify that the usual coordinatewise addition and scalar multiplication make V into a vector space. (Among other things, this includes verifying that if $\vec{v}, \vec{w} \in V$, then $\vec{v} + \vec{w} \in V$ as well, and similarly for scalar multiplication.)

This is a common phenomenon: we start with a vector space, and then want to look at another vector space contained in it. It's sufficiently common that we encode it as a definition.

Definition 16. Suppose V is a vector space over the field $\mathbb{F}$. A subset $W \subset V$ is said to be a **linear subspace** (or more briefly 'a subspace') of V if the following three conditions hold:

(S1) For all $w_1, w_2 \in W$, we have $w_1 + w_2 \in W$. We say 'W is closed under addition.'

(S2) For all $c \in \mathbb{F}$ and all $w \in W$, we have $cw \in W$. We say 'W is closed under scalar multiplication.'

(S3) We have $\vec{0} \in W$.

$\Diamond$

This includes the example I just mentioned:

Example 32. Let's verify that the set $V = \{(x, y, z) \in \mathbb{R}^3 \mid x + y + z = 0\}$ is indeed a subspace of $\mathbb{R}^3$. First, we should verify that if $\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ and $\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$ are both in V, then their sum $\begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ x_3 + y_3 \end{pmatrix}$ is as well. To show this element is in V means precisely showing that

$$(x_1 + y_1) + (x_2 + y_2) + (x_3 + y_3) = 0.$$

But this follows from the assumption that our first two vectors were in V, as

$$(x_1 + y_1) + (x_2 + y_2) + (x_3 + y_3) = (x_1 + x_2 + x_3) + (y_1 + y_2 + y_3) = 0 + 0 = 0;$$

first I rearranged, and in the second step I used that we knew $x_1 + x_2 + x_3 = 0$ and $y_1 + y_2 + y_3 = 0$ by hypothesis.

Secondly, we should show the same claim for scalar multiplication: if $c \in \mathbb{R}$ and $\vec{v} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \in V$, then we have $c\vec{v} \in V$ as well. Because $c\vec{v} = \begin{pmatrix} cx_1 \\ cx_2 \\ cx_3 \end{pmatrix}$, this amounts to the fact that

$$cx_1 + cx_2 + cx_3 = c(x_1 + x_2 + x_3) = c(0) = 0.$$

Finally, we should verify that $\vec{0} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \in V$. This amounts to the claim $0 + 0 + 0 = 0$, which is certainly true. $\Diamond$

In this example, the subspace $V \subset \mathbb{R}^3$ was a plane sitting inside 3D space. This is a good visual intuition for what a subspace should look like in general. You will show in your homework that the subspaces of $\mathbb{F}^2$ are $\{\vec{0}\}, \mathbb{F}^2$ itself, and the lines through the origin. A similar statement is true in $\mathbb{F}^3$: the subspaces are all $\{\vec{0}\}, \mathbb{F}^3$, lines through the origin, or planes through the origin.

In fact, the 1-dimensional case of this is interesting, and the first place that the existence of multiplicative inverses is really used.

Proposition 23. *Consider $\mathbb{F}$ as a vector space over itself. If $V \subset \mathbb{F}$ is a linear subspace, then $V = \{\vec{0}\}$ or $V = \mathbb{F}$.*

Proof. We are trying to prove a statement of the form $P \implies [Q \vee R]$. P is "If V is a linear subspace", while Q is "$V = \{\vec{0}\}$" and R is "$V = \mathbb{F}$". Whaht I am actually going to prove is the equivalent statement:

"If $V \subset \mathbb{F}$ is a linear subspace and $V \neq \{\vec{0}\}$, then $V = \mathbb{F}$." So we are going to start with a subspace $V \neq \{\vec{0}\}$ and show that $V = \mathbb{F}$ by a double containment argument. The containment $V \subset \mathbb{F}$ is automatic from the definition of subspace, so our goal is to show $\mathbb{F} \subset V$.

Notice that $\vec{0} \in V$ by (S3) in the definition of linear subspace, so $\{\vec{0}\} \subset V$. The claim $V \neq \{\vec{0}\}$ therefore means that there exists some **nonzero** $c \in V$. Because $c \in V \subset \mathbb{F}$ and $c \neq 0$, the multiplicative inverse axiom (F6) for fields guarantees that there exists some $d \in \mathbb{F}$ so that $dc = 1$.

Now suppose $a \in \mathbb{F}$ is arbitrary. By (S2) in the definition of linear subspace, we have $(ad)c \in V$, because $c \in V$ and V is closed under scalar multiplication. But $(ad)c = a(dc) = a(1) = a$ by distributivity and the multiplicative identity axiom.

Therefore for any $a \in \mathbb{F}$, we have $a \in V$. This proves the reverse inequality, and thus $V = \mathbb{F}$, as claimed. □

Remark 21. The subspace $\{\vec{0}\} \subset V$ is always a subspace of any vector space whatsoever. It is not very interesting, and is often called the "trivial subspace". By (S3), the trivial subspace is contained in every other subspace. ◊

The examples discussed above are very visual (lines and planes in Euclidean spaces!), but some of our more abstract examples come with natural subspaces, too:

Example 33. Each $C^k(\mathbb{R})$ is a subspace of $C^0(\mathbb{R})$. This amounts to the following claims:

(S1) If f, g are continuous and their first k derivatives exist and are continuous, the same is true of $f + g$.

(S2) If f is continuous and its first k derivatives exist and are continuous, the same is true of cf, for any $c \in \mathbb{R}$.

(S3) The function $f_0(x) = 0$ is continuous, and its first k derivatives exist and are continuous. (True: all derivatives of f_0 are still f_0, and f_0 is continuous.)

(S2) and (S3) reduce to the facts that $(f + g)' = f' + g'$ and $(cf)' = cf'$. ◊

As is the case with all of these examples, in general, a subspace is a vector space in its own right:

Proposition 24. *If $W \subset V$ is a linear subspace of the vector space V, then W is in a natural way once again a vector space.*

Proof. If $w_1, w_2 \in W$, we define their sum to be $w_1 + w_2$ using the addition operation from V; by (S1) this is indeed an element of W. Similarly if $c \in \mathbb{F}$ and $w \in W$, we define the scalar multiplication to be cw, which is again an element of W by (S2).

Axioms (V1)-(V3) and (V6) follow from the corresponding axiom for V. For instance, (V2) for V asserts that for all $v, w \in V$, we have $v + w = w + v$, while (V2) for W asserts this only for $v, w \in W$, which is a smaller set of vectors!

As for (V4), notice that (S3) says that $\vec{0} \in W$, so we at least have a candidate for the additive identity. But because the addition in W is th same as that of V, we have $w + \vec{0} = \vec{0} + w = w$ (because $\vec{0}$ is an additive identity for V). Thus (V4) is true for W.

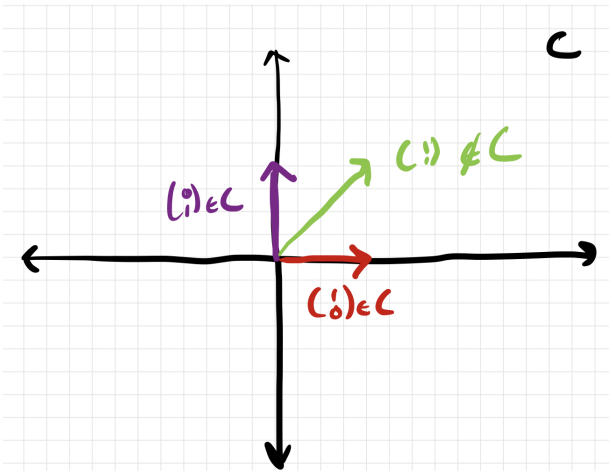
The existence of additive inverses follows from the fact that additive inverses in V are given by $-1 \cdot v$ and the fact that W is closed under scalar multiplication. □

Let us conclude with some examples of subsets which are **not** subspaces.

Example 34. Consider the set

$$C = \{(x, y) \in \mathbb{R}^2 \mid xy = 0\} = \{(x, y) \in \mathbb{R}^2 \mid x = 0 \text{ or } y = 0\}.$$

This is the union of the *coordinate axes*, drawn as follows:



C satisfies (S2) (proof: if $(x, y) \in C$, then $c(x, y) = (cx, cy)$ has

$$(cx)(cy) = c^2(xy) = c^2(0) = 0,$$

so $c(x, y) \in C$ as well) and (S3) (proof: $(0, 0) \in C$ because $0 \cdot 0 = 0$). However, it fails (S1): there exist vectors $v, w \in C$ so that $v + w \notin C$. For instance,

$$(1, 0) \in C \text{ and } (0, 1) \in C \text{ but } (1, 0) + (0, 1) = (1, 1) \notin C.$$

◊

Example 35. Consider the set

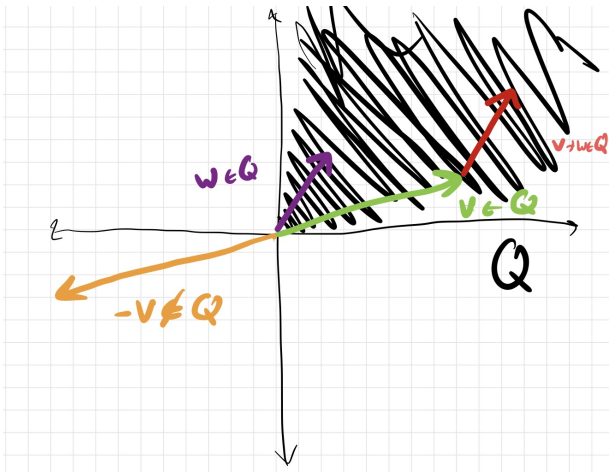
$$Q = \{(x, y) \in \mathbb{R}^2 \mid x \geqslant 0 \text{ and } y \geqslant 0\}.$$

Q stands for "quadrant", as this is the first quadrant of the plane. Then Q satisfies (S1) — if $(x_1, x_2) \in Q$ and $(y_1, y_2) \in Q$, so that $x_i, y_i \geqslant 0$ for $i = 1, 2$, then

$$(x_1, x_2) + (y_1, y_2) = (x_1 + y_1, x_2 + y_2) \in Q$$

as well, because $x_1 + y_1 \geqslant 0$ (being the sum of two non-negative reals), and similarly for $x_2 + y_2 \geqslant 0$.

Because $(0, 0) \in Q$, the set Q also satisfies (S3). But it fails (S2). To see this, I need to give you a single example of a scalar c and a vector $v \in Q$ so that $cv \notin Q$. We can take $c = -1$ and $v = (1, 1)$: we have $cv = (-1, -1) \notin Q$. (In fact, any $c < 0$ and any $v \neq (0, 0)$ will be a counterexample to (S2).)



◊

4.5 Spans

Last Tuesday, I ended the lecture by taking about the idea of coordinates in $\mathbb{R}^2$. The essential point is that I had two vectors, $\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\vec{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ which I can use to describe any vector in $\mathbb{R}^2$, in a unique way. For instance, the vector which moves three units right and five units up can be written $3\vec{e}_1 + 5\vec{e}_2$, and it cannot be written as $a\vec{e}_1 + b\vec{e}_2$ for any other a, b.

We should encode the idea of "making a vector out of other vectors" into a definition.

Definition 17. Suppose V is a vector space. If $a_1, \cdots, a_n \in \mathbb{F}$ are elements of the underlying field, and $v_1, \cdots, v_n \in V$ are elements of V, then

$$a_1 v_1 + \cdots + a_n v_n$$

is also an element of V, called a **linear combination** of the vectors $v_1, \cdots, v_n$. ◊

Hidden in this definition is a mathematical statement (a linear combination of elements of V is a meaningful description of another element of V), and this deserves a proof, our first proof by induction.

Proof that a linear combination of elements of V defines another element of V. The idea is that each new step in forming a linear combination involves only adding and scaling, operations we know how to do. The reason induction comes into play is that we are carrying out these operations n times (where n can be any integer), and induction is precisely the tool that lets us argue claims about all integers.

Let $P(n)$ be the claim: "For all $a_1, \cdots, a_n \in \mathbb{F}$, and all $v_1, \cdots, v_n \in V$, we can cogently define an element $a_1 v_1 + \cdots + a_n v_n \in V$." Let's prove this claim by induction.

- **Base case** $P(1)$. Our goal is to show thaht if $a_1 \in \mathbb{F}$ and $v_1 \in V$, then $a_1 v_1 \in V$. This is part of the definition of a vector space: if we have an element of the field and an element of our vector space, we have a well-defined scalar multiplication $a_1 v_1 \in V$.
- **Inductive step** $P(n) \implies P(n+1)$**.** Suppose we know that, for all $a_1, \cdots, a_n \in \mathbb{F}$ and for all $v_1, \cdots, v_n \in V$, we know how to define a sum $a_1 v_1 + \cdots + a_n v_n \in V$. I want to know that, for all $a_1, \cdots, a_n, a_{n+1} \in \mathbb{F}$ and for all $v_1, \cdots, v_n, v_{n+1} \in V$, we know how to define a sum $a_1 v_1 + \cdots + a_n v_n + a_{n+1} v_{n+1} \in V$.

  We already know what $v = a_1 v_1 + \cdots + a_n v_n$ is, by the inductive hypothesis that we know how to take a linear combination of n vectors. We're trying to define $v + a_{n+1} v_{n+1}$. But we already know how to define $a_{n+1} v_{n+1}$ (this is part of the data in a vector space: we know how to scale any given element by any given scalar), and we already know how to add two vectors (again, this is part of the data of a vector space), so we know how to define this element $v + a_{n+1} v_{n+1}$, as claimed.

□

The content of the above is: "Just scale each new vector and add it, one at a time." The phrasing in terms of induction is just to justify that "one at a time" makes sense.

When we discussed $\mathbb{R}^2$ above, what we were saying is the statement: "Every vector $\vec{v} \in \mathbb{R}^2$ can be written in a unique way as a linear combination $\vec{v} = a_1\vec{e}_1 + a_2\vec{e}_2$." Whenever I have an idea like the above of the form "there exists a unique" (there exists a unique way to write $\vec{v}$ as a linear combination of $\vec{e}_1$ and $\vec{e}_2$) I like to break this up into two parts. In this section, we'll focus on existence.

Definition 18. Suppose V is a vector space over the field $\mathbb{F}$, and suppose $S \subset V$ is a subset (with no other conditions). The **span** of S is the set

$$\text{span}(S) = \{a_1 v_1 + \cdots + a_n v_n \mid a_1, \cdots, a_n \in \mathbb{F} \text{ and } v_1, \cdots, v_n \in S\} \subset V.$$

That is, the span is the set of all possible linear combinations we can make out of elements of S. ◊

Remark 22. The definition above does not suppose that S is a finite set. If you look at, for instance, Axler's textbook, he only refers to spans of finite *lists* of vectors, and many other authors only look at spans of finite sets. In this case, if $S = \{v_1, \cdots, v_n\}$, then the span is precisely

$$\text{span}(v_1, \cdots, v_n) = \{a_1 v_1 + \cdots + a_n v_n \mid a_1, \cdots, a_n \in \mathbb{F}\} \subset V.$$

That is, the span of $\{v_1, \cdots, v_n\}$ is the set of all possible linear combinations of these vectors.

When we work with possibly infinite sets S, it does not make sense to take infinite sums, so the definition must be modified as above: it's the set of all possible vectors I can make by picking a finite collection of vectors from S, a finite collection of scalars, and then scaling and adding everything up. ◊

Remark 23. This remark can safely be skipped on a first reading unless you feel rather comfortable with what we've done so far.

The definition above does not appear to work correctly for $S = \varnothing$; what is the span of no vectors? What is a linear combination of no vectors?

We will see in a moment that $\text{span}(S)$ is always a subspace, at least for S nonempty. So that this holds true when S is the empty set, we usually set $\text{span}(\varnothing) = \{0\} \subset V$ by assumption. But there is actually a good reason this edge case should work like this.

Suppose I have two finite, non-empty, non-intersecting sets S and T of vectors, so I have a vector $v_i \in V$ for every $i \in S$. I can define $\sum_{i\in S} v_i$ to be the sum of all of these finitely mant vectors. For instance, $\sum_{i\in\{1,2\}} v_i = v_1 + v_2$, while $\sum_{i\in\{1,4,5,18\}} v_i = v_1 + v_4 + v_5 + v_{18}$.

When $S \cap T = \varnothing$, we have the relation $\sum_{i\in S} v_i + \sum_{i\in T} v_i = \sum_{i\in S\cup T} v_i$. For instance,

$$\sum_{\{1,2\}} v_i + \sum_{\{3,4,17\}} v_i = v_1 + v_2 + v_3 + v_4 + v_{17} = \sum_{\{1,2,3,4,17\}} v_i.$$

If I want this to hold for all sets S and T, including the empty set, I am **forced** to say that the empty sum is zero: the above relation should say

$$\sum_{\varnothing} v_i + \sum_{i\in S} v_i = \sum_{i\in S} v_i,$$

because $\varnothing \cup S = S$. That is, if $v_\varnothing = \sum_\varnothing v_i$ and $v_S = \sum_{i\in S} v_i$, then $v_\varnothing + v_S = v_S$. Subtracting v_S from both sides, this says the only reasonable definition of $v_\varnothing$ is $v_\varnothing = 0$!

This tells me that *it is reasonable to define the sum of an empty set of vectors to be zero.* Therefore, I would say that $\text{span}(\varnothing)$ is not the empty set. It includes only one element: the sum over the empty set of vectors, and this element is zero. The discussion above, I hope, justifies the convention that $\text{span}(\varnothing) = \{0\}$. ◊

Examples

Let's look at a small handful of examples of spans.

Example 36. Our first example was discussed above. Write $e_i \in \mathbb{F}^n$ for the vector

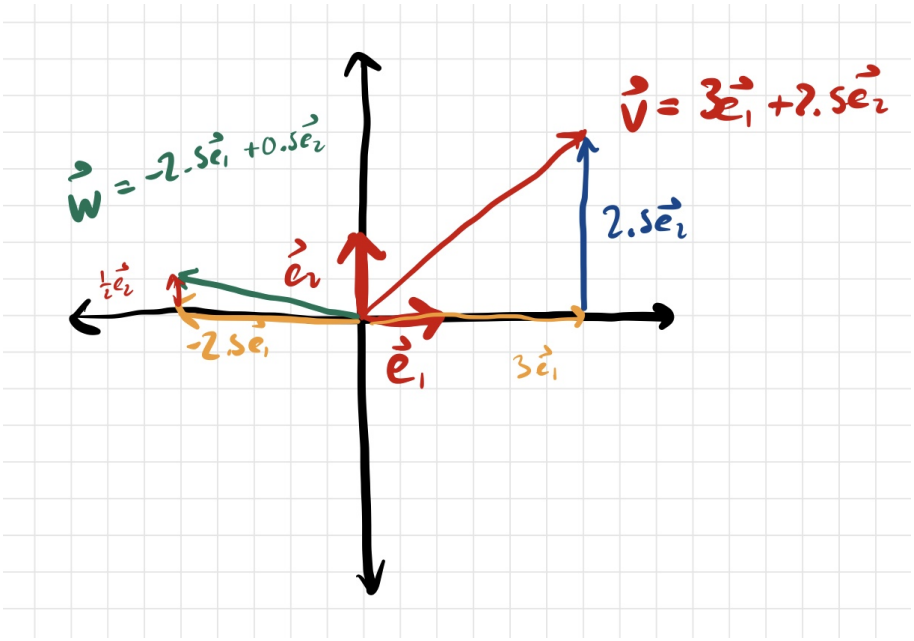
$$e_i = \begin{pmatrix} 0 \\ \cdots \\ 1 \\ \cdots \\ 0 \end{pmatrix}, \quad \text{where the only nonzero entry is in the } i\text{'th coordinate.}$$

For instance, in $\mathbb{F}^3$, we have $e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$.

Then the set $\{e_1, \cdots, e_n\}$ spans the whole of $\mathbb{F}^n$. This is because I can write a vector in $\mathbb{F}^n$ as

$$\begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1 \\ 0 \\ \cdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ x_2 \\ \cdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \cdots \\ x_n \end{pmatrix} = x_1e_1 + x_2e_2 + \cdots + x_ne_n.$$

We saw this for $\mathbb{R}^2$ in picture the other day:



◊

Example 37. Another good example of a vector space is

$$V = \left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{F}^3 \;\middle|\; x + y + z = 0 \right\}.$$

I claim that this is the span of the set $\{v_1, v_2\}$, where v_1 and v_2 are the vectors

$$v_1 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \qquad v_2 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}.$$

To carry this out, let's start with an arbitrary vector $v \in V$, so $v = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$ where $x + y + z = 0$. My idea is to use x of the first vector (as this is the only vector with a nonzero x component) and y of the second vector (as this is the only vector with a nonzero y component). Notice that

$$xv_1 + yv_2 = \begin{pmatrix} x \\ 0 \\ -x \end{pmatrix} + \begin{pmatrix} 0 \\ y \\ -y \end{pmatrix} = \begin{pmatrix} x \\ y \\ -x - y \end{pmatrix}.$$

But because $x + y + z = 0$, we see that $z = -x - y$, so in fact the last component of $xv_1 + yv_2$ agrees with the last component of v. This proves that our arbitrary vector v can be written as $xv_1 + yv_2$ for some $x, y \in \mathbb{F}$, so that

$$V \subset \operatorname{span}(v_1, v_2).$$

On the other hand, because $v_1, v_2 \in V$ are elements of this subspace, and subspaces are closed under addition and scalar multiplication, we also have $xv_1 + yv_2 \in V$. This proves the reverse containment $\operatorname{span}(v_1, v_2) \subset V$.

In lecture I will present a picture of this space for $\mathbb{F} = \mathbb{R}$, as well as a visual description of what it means to say these vectors span V. (Unfortunately, a still image is not really sufficient to "see" this space.)

◊

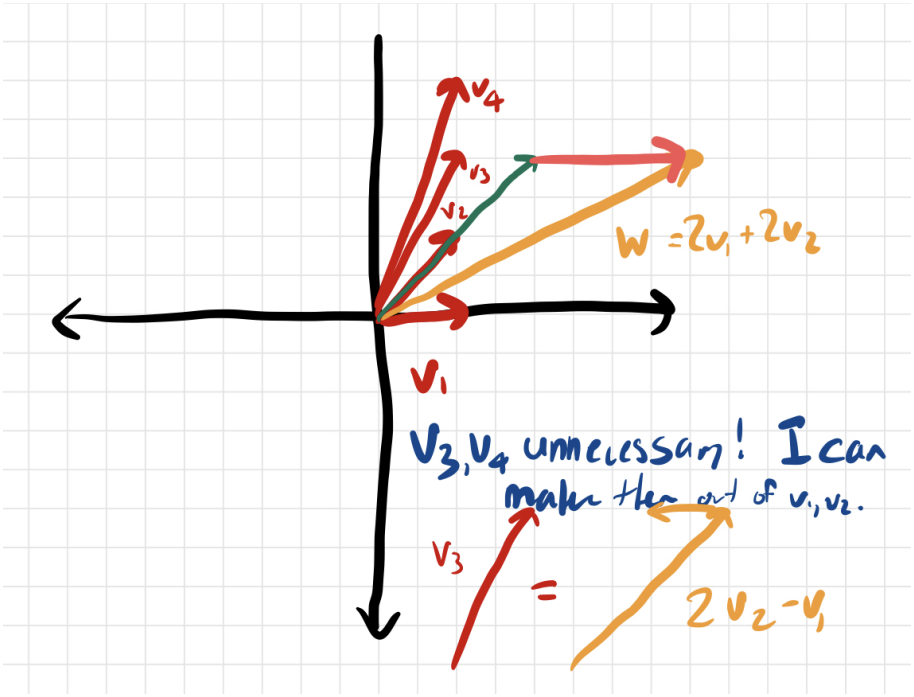
Example 38. There can be huge amounts of redundancy in spanning sets. For instance, the span of the set of vectors

$$\left\{\begin{pmatrix}1\\0\end{pmatrix}, \quad \begin{pmatrix}1\\1\end{pmatrix}, \quad \begin{pmatrix}1\\2\end{pmatrix}, \quad \begin{pmatrix}1\\3\end{pmatrix}\right\}$$

is all of $\mathbb{R}^2$: for an arbitrary $\begin{pmatrix}x\\y\end{pmatrix} \in \mathbb{R}^2$, we have for instance

$$\begin{pmatrix}x\\y\end{pmatrix} = (x-y)\begin{pmatrix}1\\0\end{pmatrix} + y\begin{pmatrix}1\\1\end{pmatrix}.$$

In this description of the vectors in $\mathbb{R}^2$, we didn't use either of the last two vectors!



Another example with redundancy is given by, say,

$$\left\{\begin{pmatrix}1\\0\\0\end{pmatrix}, \quad \begin{pmatrix}1\\1\\0\end{pmatrix}, \quad \begin{pmatrix}3\\2\\0\end{pmatrix}, \quad \begin{pmatrix}0\\1\\1\end{pmatrix}\right\}.$$

The span of this set is all of $\mathbb{R}^3$, as before:

$$\begin{pmatrix}x\\y\\z\end{pmatrix} = (x-y+z)\begin{pmatrix}1\\0\\0\end{pmatrix} + (y-z)\begin{pmatrix}1\\1\\0\end{pmatrix} + z\begin{pmatrix}0\\1\\1\end{pmatrix}.$$

In this case, the third vector on the list was totally unnecessary, but this time it wasn't at the end of the list. In fact, there is nothing we can "reach" with the third vector that we can't reach by using the first and second vectors. ◊

Example 39. Let's try an infinite example. Consider the set $\mathrm{Map}(\mathbb{N}, \mathbb{F})$, whose elements are functions $f : \mathbb{N} \to \mathbb{F}$ (that is, an element of this vector space is a choice, for each natural number n, of an element $f(n) \in \mathbb{F}$.) For instance, $(1, 3, 5, 7, 9, 11, 13, \cdots)$ is an element of $\mathrm{Map}(\mathbb{N}, \mathbb{Q})$; writing this as a function, this corresponds to $f : \mathbb{N} \to \mathbb{Q}$ defined by $f(n) = 2n + 1$.

Let's write $f_i : \mathbb{N} \to \mathbb{F}$ for the function

$$f_i(n) = \begin{cases} 1 & i = n \\ 0 & i \neq n \end{cases}.$$

In terms of sequences, f_0 corresponds to the sequence $(1, 0, 0, \cdots)$ while f_1 corresponds to $(0, 1, 0, \cdots)$ and so on: each f_i corresponds to the sequence which is nonzero in exactly the i'th term. Thus I have a set $S = \{f_0, f_1, f_2, \cdots\} \subset \mathrm{Map}(\mathbb{N}, \mathbb{F})$. What is the span of this set?

I claim that the span of this set is the subspace $\mathrm{Map}_{\mathrm{fin}} \subset \mathrm{Map}(\mathbb{N}, \mathbb{F})$, where $f \in M_{\mathrm{fin}}$ if and only if f is eventually zero:

$$f \in M_{\mathrm{fin}} \iff \exists_{N \in \mathbb{N}} \text{ such that } \forall_{m>N} f(m) = 0.$$

Explicitly, the sequence $(1, 3, 1, 4, 1, 5, 0, 0, 0, 0, \cdots)$ is an element of M_{fin}, as is the sequence $(0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, \cdots)$ which starts as zero, becomes nonzero for a bit, but eventually goes back to always being zero. On the other hand, $(1, 3, 5, 7, \cdots)$ is not in M_{fin}, as this sequence is never zero![3]

In the next Proposition, I will prove that this is indeed the span of S. ◊

Proposition 25. *In Example 39, we have* $\mathrm{span}(S) = \mathrm{Map}_{\mathrm{fin}}$.

Proof. We have two containments to prove. First, observe that each $f_i \in S$ is an element of $\mathrm{Map}_{\mathrm{fin}}$: it is only nonzero at one point, so is certainly eventually zero. Next, notice that $\mathrm{Map}_{\mathrm{fin}}$ is a linear subspace: it is closed under scalar multiplication and addition (if f has $f(n) = 0$ for all $n > M$, the same is true for cf; if f has $f(n) = 0$ for all $n > N$, and g has $g(n) = 0$ for all $n > M$, then so long as $n > \max(M, N)$, we have

$$(f + g)(n) = f(n) + g(n) = 0 + 0 = 0.$$

Thus sums of elements of $\mathrm{Map}_{\mathrm{fin}}$ are also in this set, so this is a subspace of $\mathrm{Map}(\mathbb{N}, \mathbb{F})$.)

It follows that any linear combination of the f_i are also in $\mathrm{Map}_{\mathrm{fin}}$, because subspaces are closed under linear combinations. This proves the containment $\mathrm{span}(S) \subset \mathrm{Map}_{\mathrm{fin}}$.

For the other containment $\mathrm{Map}_{\mathrm{fin}} \subset \mathrm{span}(S)$, pick an arbitrary function $g : \mathbb{N} \to \mathbb{F}$ which is eventually zero, so $f(n) = 0$ for all $n > N$. I claim

$$g = \sum_{i=0}^{N} g(i) f_i.$$

The expression on the right is given by

$$\left(\sum_{i=0}^{N} g(i) f_i \right)(n) = \sum_{i=0}^{N} g(i) f_i(n) = \begin{cases} g(n) & n \leqslant N \\ 0 & n > N \end{cases}.$$

But this is precisely the same as g, beause $g(n) = 0$ for all $n > N$. Therefore we've shown that every $g \in \mathrm{Map}_{\mathrm{fin}}$ can be represented as a linear combination of elements in $\{f_0, f_1, \cdots\} = S$, showing the reverse containment. This completes the proof. □

A crucial property

We implicitly saw the following fact in the previous argument.

Proposition 26. *Let V be a vector space, and let $S \subset V$ be an arbitrary subset.*

The set $\mathrm{span}(S)$ *satisfies the following three properties:*

(a) The set $\mathrm{span}(S) \subset V$ *is a linear subspace.*

(b) We have $S \subset \mathrm{span}(S)$.

(c) If $W \subset V$ is any linear subspace which contains S (so $S \subset W \subset V$), then we have $\mathrm{span}(S) \subset W$.

[3] I chose the name $\mathrm{Map}_{\mathrm{fin}}$ to mean "maps with finite support", as in, they are only nonzero in finitely many terms.

Remark 24. We usually interpret this as meaning that span(S) is the 'smallest linear subspace containing S'. The first two parts assert that span(S) is indeed a linear subspace containing S, while the last part asserts that any other set with the same property contains span(S), vis a vis, the span is the smallest such set. In fact, we can precisely state

$$\text{span}(S) = \bigcap_{\substack{S \subset W \subset V \\ W \text{ a linear subspace}}} W.$$

This latter formulation is not very useful; the formulation of Proposition 26 says the same thing, but more explicitly. ◊

Proof of Proposition 26. I am going to present a proof under the additional assumption that $S = \{v_1, \cdots, v_n\}$ is finite, because the notation is mildly irritating for infinite S (and we are by far more interested in the finite case anyway). In a remark after the proof, I will explain how to modify the same proof to work when S is infinite.

a) To prove that span(S) is a linear subspace, we need to argue three things. For (S1), we need to show it is closed under addition. Suppose $v, w \in \text{span}(S)$. This means

$$\begin{aligned} v &= a_1 v_1 + \cdots + a_n v_n \quad \text{for some } a_1, \cdots, a_n \in \mathbb{F} \\ w &= b_1 v_1 + \cdots + b_n v_n \quad \text{for some } b_1, \cdots, b_n \in \mathbb{F}. \end{aligned}$$

Their sum is

$$(a_1 v_1 + \cdots + a_n v_n) + (b_1 v_1 + \cdots + b_n v_n) = (a_1 + b_1) v_1 + \cdots + (a_n + b_n) v_n,$$

which is by definition again an element of span(S). Similarly for (S2), if $v = a_1 v_1 + \cdots + a_n v_n \in \text{span}(S)$, then we have

$$cv = c(a_1 v_1 + \cdots + a_n v_n) = (ca_1) v_1 + \cdots + (ca_n) v_n \in \text{span}(S).$$

Finally, for (S3), we should argue that the zero vector is in span(S). When S is nonempty (so $v_1 \in S$, say), this follows because $0 v_1 = \vec{0} \in V$ is a linear combination of elements of S. When S is empty, one must either say that this is true by convention or argue that $\vec{0}$ should be called the empty linear combination, as in Remark 23.

b) For any $v_i \in S$, we have $v_i = 0v_1 + 0v_2 + \cdots + 1v_i + \cdots + 0v_n \in \text{span}(S)$ — that is, take the linear combination with exactly one coefficient equal to 1, and all others equal to zero. Therefore $S \subset \text{span}(S)$.

c) Suppose $v \in \text{span}(S)$; our goal is to show $v \in W$. By definition of the span, we have

$$v = a_1 v_1 + \cdots + a_n v_n \text{ for some } a_1, \cdots, a_n \in \mathbb{F}.$$

Because we assumed $S \subset W$, we know that $v_1, \cdots, v_n \in W$. Because W is a subspace, it is closed under scalar multiplication, so $a_i v_i \in W$ for all $1 \leqslant i \leqslant n$. Finally, because W is closed under addition, we can argue inductively that $a_1 v_1 + \cdots + a_n v_n \in W$ as well. Thus $v = a_1 v_1 + \cdots + a_n v_n \in W$. □

Exercise: Write out the inductive proof alluded to above.

Remark 25. When S is infinite, the proof above is made more irritating by the fact that when I define a linear combination of elements of S, I have to choose the finitely many terms I'm adding (whereas when S is finite, I can just say "add all the elements, with some weights"). This is resolved if I instead define a linear combination in S to be

$$\sum_{i \in S} a_i v_i \quad \text{where all but finitely many of the } a_i \text{ are zero.}$$

Thus this infinite-looking sum is secretly a finite sum. When this is the case, the above argument goes through without change:

$$\left(\sum_{i\in S} a_i v_i\right) + \left(\sum_{i\in S} b_i v_i\right) = \sum_{i\in S}(a_i + b_i)v_i,$$

and because all but finitely many a_i are zero and all but finitely many b_i are zero, this implies for all but finitely many i we have **both** a_i and b_i zero. Thus all but finitely many $a_i + b_i$ are zero, so the sum described above is again a finite sum, hence an element of $\text{span}(S)$.

The rest of the proof proceeds with minimal change. ◊

Earlier I said a lot of times "The span of this set is the whole space." This is a common enough notion that it is worth naming.

Definition 19. If $S \subset V$ is a subset for which $\text{span}(S) = V$, we say that 'S spans V'. If there is a **finite** set S which spans V, we say that V is **finite-dimensional**. ◊

This course is the study of finite-dimensional vector spaces, and how they relate to each other.

Let me conclude this section by spelling out the definition above. By definition, we have $\text{span}(S) \subset V$ for any set $S \subset V$ whatsoever, so the content of 'S spans V' must be the other containment $V \subset \text{span}(S)$. This containment asserts that

> For all $v \in V$, there exist some natural number n and scalars $a_1, \cdots, a_n \in \mathbb{F}$ and vectors $v_1, \cdots, v_n \in S$ so that $v = a_1 v_1 + \cdots a_n v_n$.

4.6 Linear independence

Before getting into definitions, let me return to an example of spans. The following example about redundancy in spans is similar to Example 38.

Example 40. Recall the notation $e_1 = \begin{pmatrix}1\\0\end{pmatrix}$ and $e_2 = \begin{pmatrix}0\\1\end{pmatrix}$ for the two standard vectors in $\mathbb{F}^2$ (which we will soon call 'the standard basis vectors'). We saw that $\text{span}(e_1, e_2) = \mathbb{F}^2$, and in fact every vector $v = \begin{pmatrix}x\\y\end{pmatrix} \in \mathbb{F}^2$ can be written in a unique way as a linear combination of these two vectors: $v = xe_1 + ye_2$.

We also have $\text{span}(e_1, e_2, e_1 + e_2) = \mathbb{F}^2$, but that last vector was unnecessary. I already know that this space is the span of the first two vectors; this third vector is a linear combination of the first two. Now there are many ways to write a given vector as a linear combination of the three of these; for instance,

$$v = \begin{pmatrix}3\\2\end{pmatrix} = (3-x)e_1 + (2-x)e_2 + x(e_1 + e_2)$$

for any $x \in \mathbb{F}$. I might say that there are many *representations* of v as a linear combination of the vectors in the set $\{e_1, e_2, e_1 + e_2\}$.

Here is a useful trick. Take two such representations; for instance, $v = 3e_1 + 2e_2 + 0(e_1 + e_2)$ and $v = 2e_1 + 1e_2 + 1(e_1 + e_2)$. "Subtracting the second equation from the first", we find that

$$\vec{0} = v - v = (3e_1 + 2e_2 + 0(e_1 + e_2)) - (2e_1 + 1e_2 + 1(e_1 + e_2)) = e_1 + e_2 - 1(e_1 + e_2).$$

That is, there is a non-trivial linear combination of the vectors $\{e_1, e_2, e_1 + e_2\}$ which gives us **the zero vector**. This is not true for the pair of vectors $\{e_1, e_2\}$, as if

$$\begin{pmatrix}x\\y\end{pmatrix} xe_1 + ye_2 = \vec{0} = \begin{pmatrix}0\\0\end{pmatrix},$$

then $x = 0$ and $y = 0$. ◊

There were three issues enumerated above with the spanning set $\{e_1, e_2, e_1 + e_2\}$ of $\mathbb{F}^2$: there was an unnecessary vector in the list; there's more than one way to write a vector as a linear combination of these; you can make the zero vector out of these in a non-trivial way. These all turn out to be equivalent, and the last of these is easiest to check, and we encode it as a definition.

Definition 20. Let V be a vector space over a field $\mathbb{F}$. Given a set $S = \{v_1, \cdots, v_n\} \subset V$ of vectors, a **linear relation** between them is a sequence $a_1, \cdots, a_n \in \mathbb{F}$ so that

$$a_1 v_1 + \cdots + a_n v_n = \vec{0}.$$

No matter what $v_1, \cdots, v_n$ is, $(0, \cdots, 0)$ always defines a linear relation

$$0v_1 + \cdots + 0v_n = \vec{0}$$

between them, and is called the **trivial linear relation**. Any other linear relation, a linear relation $(a_1, \cdots, a_n)$ for which **there exists an** i so that $a_i \neq 0$, is called a **nontrivial linear relation**. ◊

Remark 26. The set of linear relations

$$\mathrm{Rel}(S) = \{(a_1, \cdots, a_n) \in \mathbb{F}^n \mid a_1 v_1 + \cdots + a_n v_n = \vec{0}\}$$

is a linear subspace of $\mathbb{F}^n$: it contains the zero vector $(0, \cdots, 0)$ (the trivial relation!); if $(a_1, \cdots, a_n)$ and $(b_1, \cdots, b_n)$ are linear relations, then $(a_1 + b_1, \cdots, a_n + b_n)$ is too:

$$(a_1 + b_1)v_1 + \cdots + (a_n + b_n)v_n = (a_1 v_1 + \cdots + a_n v_n) + (b_1 v_1 + \cdots + b_n v_n) = \vec{0} + \vec{0} = \vec{0}.$$

The set of linear relations is closed under scalar multiplication by a similar argument.

The subspace $\mathrm{Rel}(S)$ is related to the subspace $\mathrm{span}(S)$ in an interesting way that we will explore when we learn about linear transformations (the space $\mathrm{Rel}(S)$ can be understood as the 'kernel' of a certain linear map, and $\mathrm{span}(S)$ the image of that same linear map). ◊

Remark 27. You can make sense of this definition when S is infinite: a linear relation is a choice of vectors $v_1, \cdots, v_n \subset S$ and $a_1, \cdots, a_n \in \mathbb{F}$ so that $a_1 v_1 + \cdots + a_n v_n = \vec{0}$. ◊

The first example of the section suggests that the property I want is that there are *no non-trivial linear relations* between the vectors in S. I'll write this in a logically equivalent form.

Definition 21. We say that a set $S = \{v_1, \cdots, v_n\} \subset V$ of vectors is **linearly independent** if the only linear relation between them is the trivial linear relation. That is, $\{v_1, \cdots, v_n\}$ is linearly independent when

$$\forall_{a_1, \cdots, a_n \in \mathbb{F}} a_1 v_1 + \cdots + a_n v_n = \vec{0} \implies a_1 = \cdots = a_n = 0.$$

On the other hand, if there exists a non-trivial linear relation between the vectors in S, we say S is **linearly dependent**. ◊

Remark 28. "S is linearly indepedent" as saying that the subspace $\mathrm{Rel}(S) = \{\vec{0}\}$ is the trivial subspace of $\mathbb{F}^n$. ◊

Example 41. The set $\{e_1, \cdots, e_n\} \subset \mathbb{F}^n$ is linearly independent:

$$a_1 e_1 + \cdots + a_n e_n = \vec{0} \implies \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} = \begin{pmatrix} 0 \\ \cdots \\ 0 \end{pmatrix};$$

to say these vectors are equal in $\mathbb{F}^n$ precisely means all their coordinates are equal, so $a_i = 0$ for all i.

This is hardly the only linearly independent set in $\mathbb{F}^n$. It is easy to come up with more. For instance, for any choice of $x \in \mathbb{F}$, the pair $\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} x \\ 1 \end{pmatrix} \right\}$ is a linearly independent set in $\mathbb{F}^2$: if

$$\begin{pmatrix} a_1 + a_2 x \\ a_2 \end{pmatrix} = a_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + a_2 \begin{pmatrix} x \\ 1 \end{pmatrix} = \vec{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

then by comparing coefficients we see that $a_1 + a_2 x = 0$ and $a_2 = 0$. Applying the latter to the first, we see that $a_1 = 0$ as well. Thus any linear relation (a_1, a_2) between these vectors must have $(a_1, a_2) = (0, 0)$, and the vectors are linearly independent. ◊

Example 42. We saw earlier that the set $\{e_1, e_2, e_1 + e_2\}$ is linearly dependent, because $(1, 1, -1)$ gives a non-trivial linear relation between them:

$$1e_1 + 1e_2 + (-1)(e_1 + e_2) = \vec{0}.$$

◊

Example 43. Any set $\{v_1, \cdots, v_n\}$ which contains the zero vector (so $v_i = \vec{0}$ for some i) is linearly dependent, because

$$0v_1 + 0v_2 + \cdots + 1v_i + \cdots + 0v_n = v_i = \vec{0};$$

here the only non-zero coefficient is the coefficient of $v_i = \vec{0}$. ◊

4.6.1 Redundancy

I mentioned at the beginning that another way the list of vectors $(e_1, e_2, e_1 + e_2)$ was inefficient (for the sake of taking spans) is that I could write the last term as a linear combination of the previous terms. Notice that this requires talking about *ordered lists* of vectors (to refer to 'the previous terms') as opposed to sets, whose elements are not ordered (the set $\{1, 2\}$ is the same as the set $\{2, 1\}$, as they have the same elements). Let's record this in the following definition.

Definition 22. Let V be a vector space. Given an ordered list of vectors $(v_1, \cdots, v_n)$, with $v_i \in V$ for all $1 \leqslant i \leqslant n$, we say that a vector v_i on this list is **redundant** if

$$v_i \in \text{span}(v_1, \cdots, v_{i-1}).$$

That is, v_i is redundant if and only if there are scalars $a_1, \cdots, a_{i-1} \in \mathbb{F}$ so that

$$a_1v_1 + \cdots + a_{i-1}v_{i-1} = v_i.$$

In the case that $i = 1$, we say v_1 is redundant if and only if $v_1 = \vec{0}$. ◊

Remark 29. If you accept that 'the empty sum' is $\vec{0}$ or that $\text{span}(\varnothing) = \{\vec{0}\}$, the case $i = 1$ is subsumed into the case above. ◊

The notion of redundancy depends on the order I list the vectors. For instance, in $(e_1, e_2, e_1 + e_2)$, the first two vectors are non-redundant (e_1 is nonzero and e_2 is not a multiple of e_1), but the third vector on the list is redundant, as $e_1 + e_2 = 1e_1 + 1e_2$.

On the other hand, in the list $(e_1, e_1 + e_2, e_2)$, the first two vectors are still non-redundant — e_1 is nonzero and $e_1 + e_2$ is not a multiple of e_1 — but e_2 is now the redundant vector, as

$$e_2 = (-1)e_1 + 1(e_1 + e_2).$$

Lemma 27. *Suppose V is a vector space, and $(v_1, \cdots, v_n)$ a list of vectors in V. If v_i is redundant (for some $1 \leqslant i \leqslant n$), then $\text{span}(v_1, \cdots, v_n) = \text{span}(v_1, \cdots, v_{i-1}, v_{i+1}, \cdots, v_n)$; that is, the span does not change by excluding v_i.*

Proof. We suppose v_i is redundant and try to show these two spans are equal. The containment

$$\text{span}(v_1, \cdots, v_{i-1}, v_{i+1}, \cdots, v_n) \subset \text{span}(v_1, \cdots, v_n)$$

is automatic (and has nothing to do with the assumption that v_i is redundant): all the vectors in the first list are contained in the second list, so all linear combinations of vectors in the first list are among the linear combinations of vectors in the second list.

The other containment

$$\text{span}(v_1, \cdots, v_n) \subset \text{span}(v_1, \cdots, v_{i-1}, v_{i+1}, \cdots, v_n)$$

has more content. First, let me spell out what it means to say that v_i is redundant: this means there exist $b_1, \cdots, b_{i-1} \in \mathbb{F}$ so that $v_i = b_1v_1 + \cdots + b_{i-1}v_{i-1}$. Now given a vector in the first span, so some vector of the form

$$v = a_1v_1 + \cdots + a_nv_n,$$

notice that we can use the expression above to rewrite this without v_i whatsoever; it's

$$\begin{aligned} v &= (a_1v_1 + \cdots + a_{i-1}v_{i-1}) + a_iv_i + (a_{i+1}v_{i+1} + \cdots + a_nv_n) \\ &= (a_1v_1 + \cdots + a_{i-1}v_{i-1}) + a_i(b_1v_1 + \cdots + b_{i-1}v_{i-1}) + (a_{i+1}v_{i+1} + \cdots + a_nv_n) \\ &= (a_1 + a_ib_1)v_1 + \cdots + (a_{i-1} + a_ib_{i-1})v_{i-1} + a_{i+1}v_{i+1} + \cdots + a_nv_n \in \operatorname{span}(v_1, \cdots, v_{i-1}, v_{i+1}, \cdots, v_n), \end{aligned}$$

as desired. (If $i = 1$ so that v_1 redundant means $v_1 = \vec{0}$, we just ignore the $a_1\vec{0}$ term, as this contributes nothing to the sum anyway.) □

Corollary 28. *Let V be a vector space. If $(v_1, \cdots, v_n)$ is a list of vectors in V, there is a smaller list $(v_{i_1}, \cdots, v_{i_k})$ of vectors, obtained by removing the redundant vectors of $(v_1, \cdots, v_n)$, so that*

$$\operatorname{span}(v_1, \cdots, v_n) = \operatorname{span}(v_{i_1}, \cdots, v_{i_k})$$

but $(v_{i_1}, \cdots, v_{i_k})$ contains no redundant vectors whatsoever.

To make the statement clear, suppose $(v_1, v_2, v_3, v_4, v_5)$ is a list of vectors where v_1, v_3, v_4 are all redundant. Then $\operatorname{span}(v_1, v_2, v_3, v_4, v_5) = \operatorname{span}(v_2, v_5)$, but the list (v_2, v_5) has no redundant vectors whatsoever.

Proof. This is a proof by induction, where we induct on the number of vectors in the list which are redundant. Precisely, let $P(m)$ be the statement: "If $(v_1, \cdots, v_n)$ has exactly m redundant vectors, then there exists a smaller list $(v_{i_1}, \cdots, v_{i_{n-m}})$ of vectors, obtained by removing the m redundant vectors, so that this smaller list has the same span as the original list."

The base case, $P(0)$, states: if this list has no redundant vectors, then removing no vectors gives us a list with the same span and with no redundant vectors. This is a tautology.

Now suppose the statement $P(m)$ is true (the inductive hypothesis). Let's prove that $P(m+1)$ is true. So suppose we have a list $(v_1, \cdots, v_n)$ that has exactly $m+1$ redundant vectors. If the first of these is v_i, then the list $(v_1, \cdots, v_{i-1}, v_{i+1}, \cdots, v_n)$ is a list of $n-1$ vectors of which exactly m are redundant: if v_j is redundant in the original list for $j > i$, it can be written as a linear combination of the previous vectors; by replacing v_i with a linear combination of terms before it, this gives an expression for v_j as a linear combination of vectors before it excluding v_i, so it is redundant in the new list as well. By Lemma 27, this new list has the same span as the previous, because v_i is redundant.

By the inductive hypothesis, there is a sublist $(v_{i_1}, \cdots, v_{i_{n-m-1}})$ of this list with the same span for which none of these vectors are redundant. Because $\operatorname{span}(v_1, \cdots, v_{i-1}, v_{i+1}, \cdots, v_n) = \operatorname{span}(v_1, \cdots, v_n)$, this sublist has the same span as our original list, and is obtained by removing all the redundant vectors (v_i, and all the redundant vectors after it), and has no more redundat vectors. This completes the proof of $P(m+1)$.

By the principle of induction, $P(m)$ is true for all m; that is, the claim is true no matter how many vectors there are. □

To conclude this discussion, let me observe that the notion of linear dependence is the same as the notion of having a redundant vector in the list.

Proposition 29. *Let V be a vector space. Given a finite list of vectors $(v_1, \cdots, v_n)$ in V, these vectors are linearly dependent if and only if there is some $1 \leqslant i \leqslant n$ so that v_i is redundant.*

Proof. Suppose these vectors are linearly dependent. This means there exist $a_1, \cdots, a_n \in \mathbb{F}$ so that

$$a_1v_1 + \cdots + a_nv_n = \vec{0}$$

so that at least one a_j is nonzero. Let $1 \leqslant i \leqslant n$ be the **largest** number so that a_i is nonzero, so that this relation reads

$$a_1v_1 + \cdots + a_iv_i + 0v_{i+1} + \cdots + 0v_n = \vec{0}, \quad \text{or} \quad a_1v_1 + \cdots + a_iv_i = \vec{0}.$$

Subtracting a_iv_i from both sides, we have

$$a_1v_1 + \cdots + a_{i-1}v_{i-1} = -a_iv_i.$$

By the assumption that $a_i \neq 0$ we also have $-a_i \neq 0$ (as $-0 = 0$), so there exists a multiplicative inverse $-1/a_i \in \mathbb{F}$. Scaling both sides by this quantity, we see that

$$\frac{-a_1}{a_i}v_1 + \cdots + \frac{-a_{i-1}}{a_i}v_{i-1} = v_i,$$

so that v_i is redundant.

Conversely, if v_i is redundant — so that $a_1v_1 + \cdots + a_{i-1}v_{i-1} = v_i$ for some $a_1, \cdots, a_{i-1} \in \mathbb{F}$ — then subtracting v_i from both sides gives a non-trivial linear relation between our list of vectors, given by

$$a_1v_1 + \cdots + a_{i-1}v_{i-1} + (-1)v_i + 0v_{i+1} + \cdots + 0v_n = \vec{0};$$

this relation is non-trivial because $-1 \neq 0$ (as $1 \neq 0$). □

It's worth pointing out that linear dependence is a notion which does not depending on the ordering of the vectors in the list. If we reorder a list of vectors, this will change which vectors are considered redundant. On the other hand, it will not change whether or not there *is* a redundant vector: there will be a redundant vector in the list if and only if the list is linearly dependent.

Corollary 30. *Given any set $S = \{v_1, \cdots, v_n\}$ of vectors in a vector space V, there is a subset $S' = \{v_{i_1}, \cdots, v_{i_k}\}$ with $S' \subset S$ of these vectors so that* $\mathrm{span}(S') = \mathrm{span}(S)$ *and S' is linearly independent.*

Proof. Order these vectors into a list $(v_1, \cdots, v_n)$ and apply Corollary 28 to produce a smaller list of vectors with the same span and no redundancy, then apply the contrapositive of Proposition 29 to see that this list is linearly independent. □

4.6.2 Linearly independent sets and spans

We can use what we've proved so far to establish a really powerful result, which will be the foundation of an important idea in the next section. It's important enough to call it a theorem. The argument largely follows the presentation in Axler's book.

Theorem 31 (Size of linearly independent set $\leqslant$ size of spanning set)**.** *Let V be a* ***finite-dimensional vector space****, so that there exists a finite spanning set for V. Given any set $\{v_1, \cdots, v_n\}$ of linearly independent vectors in V, and any set $\{w_1, \cdots, w_m\}$ of vectors which spans V, we have $n \leqslant m$.*

Remark 30. This result holds for infinite-dimensional vector spaces as well, but the proof here does not directly apply, as our proof is by a sort of induction which doesn't make sense in infinite-dimensional vector spaces. ◊

We will prove this by first proving a lemma, which will amount to the inductive step in a proof.

Lemma 32. *Let V be a vector space. If $(v_1, \cdots, v_n)$ is a linearly independent list of vectors in V, and $(v_1, \cdots, v_k, w_1, \cdots, w_{m-k})$ is a spanning list of vectors for V (where $k < n$ and $k \leqslant m$), there is also a spanning list $(v_1, \cdots, v_{k+1}, w'_1, \cdots, w'_{m-k-1})$ with the same number of vectors total but one more vector from $(v_1, \cdots, v_n)$. In particular, we have $k + 1 \leqslant m$ as well.*

Proof. Consider the list $(v_1, \cdots, v_k, v_{k+1}, w_1, \cdots, w_{m-k})$ with $m + 1$ vectors and one more vector from the list of v_i. This list is linearly dependent: because $v_{k+1} \in \mathrm{span}(v_1, \cdots, v_k, w_1, \cdots, w_{m-k})$ there is a linear relation of the form

$$a_1v_1 + \cdots + a_kv_k + (-1)v_{k+1} + b_1w_1 + \cdots + b_{m-k}w_{m-k} = \vec{0}.$$

Because this list is linearly dependent, there must be a redundant vector on this list. But it cannot be among the first $(k+1)$, as $\{v_1, \cdots, v_n\}$ is linearly independent. Thus there must be a later vector on this list (so $m - k > 0$, or $k + 1 \leqslant m$) which is redundant, and in particular, w_i is redundant for some $1 \leqslant i \leqslant w_{m-k}$. Removing it, we obtain a list

$$(v_1, \cdots, v_{k+1}, w_1, \cdots, w_{i-1}, w_{i+1}, \cdots, w_{m-k})$$

of precisely m vectors which has the same span as the previous list, as desired. □

Iterating this process (a proof by induction!) we eventually obtain a new spanning list of the same size which starts with the list of linearly independent vectors.

Corollary 33. *If $(v_1, \cdots, v_n)$ is a linearly independent list of vectors in V, and $(w_1, \cdots, w_m)$ is a spanning list for V, then there exists a spanning list $(v_1, \cdots, v_n, w_1', \cdots, w_{m-n}')$ which starts with the linearly independent list but has exactly the same number of vectors.*

Proof of Theorem 31. By Corollary 33, there is a spanning set of size m which contains the linearly independent set $\{v_1, \cdots, v_n\}$. Thus the linearly independent set has no more vectors than the spanning set: $n \leqslant m$. □

Notice that $\mathbb{F}^n$ has a set $\{e_1, \cdots, e_n\}$ of vectors which is linearly independent and also spans $\mathbb{F}^n$. It follows that every set of linearly independent vectors in $\mathbb{F}^n$ has at most n vectors — for instance, no list of four vectors in $\mathbb{F}^3$ can be linearly independent — and every spanning set for $\mathbb{F}^n$ has at least n vectors — so no list of two vectors in $\mathbb{F}^3$ can span it.

Corollary 34. *A subspace W of a finite-dimensional vector space V is finite-dimensional.*

Proof. Passing to the contrapositive, the claim is: if W is infinite-dimensional, then V is infinite-dimensional. You will prove on your homework that if W is infinite-dimensional, then there exists an infinite set $(w_1, w_2, \cdots)$ of linearly independent vectors in W, which are also linearly independent when considered as vectors in V. Thus, there are linearly independent sets $\{w_1, \cdots, w_n\}$ in V of size any natural number n. Applying Theorem 31, there is no finite set of vectors which span V, because for any natural number m there is a linearly independent set of size larger than m.

Thus V is infinite-diimensional. □

4.7 Bases and dimension

We are now going to combine the ideas from the last two sections into one extremely powerful idea.

Definition 23. Let V be a vector space. A **basis** for V is a linearly independent set $S \subset V$ for which $\text{span}(S) = V$. ◊

Example 44. The standard basis for $\mathbb{F}^n$ is the set $\{e_1, \cdots, e_n\}$, where

$$e_i = \begin{pmatrix} 0 \\ \cdots \\ 1 \\ \cdots \\ 0 \end{pmatrix}, \quad \text{where the only nonzero coordinate is the } i'\text{th coordinate.}$$

We have already seen both that these span $\mathbb{F}^n$ and that these vectors are linearly independent. ◊

Example 45. While $\mathbb{F}^n$ comes equipped with a canonical basis you can see and start computing with right away, most vector spaces don't have a canonical such basis. Moreover, it is hardly the only basis for $\mathbb{F}^n$; in fact, there are a humongous number of bases for any vector space (and this is a good thing: we will later want to work with a basis suited for the problem at hand). Because most vector spaces you can think of have infinitely many elements, it is difficult to give a good sense of how large this is, so let me refer briefly to the finite fields $\mathbb{F}_q$ with q elements (where $q = p^m$ is some prime power).

Then the set

$$\{(v_1, \cdots, v_n) \in (\mathbb{F}_q^n)^n \mid (v_1, \cdots, v_n) \text{ is a basis for } \mathbb{F}_q^n\}$$

has exactly

$$(q^n - 1)(q^n - q) \cdots (q^n - q^{n-1}) \approx q^{n^2}$$

elements. (Idea: You have exactly $q^n - 1$ choices for the first term, because you just need to make sure it's not zero; you have $q^n - q$ choices for the second term, since you need to make sure it's not in the span of the first; and so on, until you have a list of n, and we will see by the end of the section that a list of n linearly independent vectors in $\mathbb{F}^n$ is necessarily a basis.)

That is, there are approximately $2^{50} \approx 10^{15}$ basis for the vector space $\mathbb{F}_2^{10}$; the vector space itself, by contrast, has $2^{10} = 1024 \approx 10^3$ elements. The number of bases for a vector space grows at an insane speed as you increase the 'dimension' of the vector space.

(By the end of this section, you will have enough information to prove that this calculation is correct, but I will not ask you to do so on homework.) ◊

As before, I am going to focus on finite-dimensional vector spaces, even though this discussion extends with minimal change to the infinite-dimensional setting. A great many of the results below apply in the infinite-dimensional setting (and I will say when they don't), but the proofs use the axiom of choice in an essential way; this is something you can pursue in the curios, if you want.

The following lemma gives a useful interpretation of bases, and shows how we'll use them in the future (in particular, when we discuss matrices). Still, when actually verifying whether some set is a basis, you should return to the definition.

Lemma 35. *Let V be a vector space. A set $S = \{v_1, \cdots, v_n\}$ is a basis for V* ***if and only if*** *for all $v \in V$, there exists a* ***unique*** *way to express v as a linear combination*

$$a_1 v_1 + \cdots a_n v_n$$

of the elements of S. That is,

$$\{v_1, \cdots, v_n\} \text{ a basis for } V \iff \forall_{v \in V} \exists!_{a_1, \cdots, a_n \in \mathbb{F}} v = a_1 v_1 + \cdots + a_n v_n.$$

Proof. Let's start with the backward direction, which is easier. Given any vector $v \in V$, the assumption says there exists a way to represent v as a linear combination $v = a_1 v_1 + \cdots + a_n v_n$ of the vectors in S, so S spans V. Further, because each vector v can be represented in an *unique* way as a linear combination of vectors in S, applying this to $v = \vec{0}$ asserts that there exists a unique linear relation

$$a_1 v_1 + \cdots + a_n v_n = \vec{0}.$$

Because $(a_1, \cdots, a_n) = (0, \cdots, 0)$ is such a relation, and there is a unique such relation, this implies every relation between these vectors is trivial — so S is a linearly independent set. Because S is a linearly independent spanning set, it is by definition a basis.

Conversely, suppose S is a basis for V. For every vector $v \in V$, there *exists* a representation $v = a_1 v_1 + \cdots + a_n v_n$ as a linear combination of the elements of S, and we need to argue that it's unique. That is, if

$$v = a_1 v_1 + \cdots + a_n v_n \quad \text{and also} \quad v = b_1 v_1 + \cdots + b_n v_n \quad \text{then } a_i = b_i \text{ for all } i.$$

The information we have is that the only linear relation is the trivial relation, so we want to use that fact somehow. Subtracting the two equations from one another, we see that

$$\vec{0} = v - v = (a_1 - b_1) v_1 + \cdots + (a_n - b_n) v_n.$$

Thus $(a_1 - b_1, \cdots, a_n - b_n)$ are the coefficients in a linear relation between the vectors $\{v_1, \cdots, v_n\}$. Because the only such relation is the trivial relation, we have $a_i - b_i = 0$ for all i, or equivalently, $a_i = b_i$ for all i. This is what we wanted to prove. □

Next, I want to prove a handful of technical facts about bases which are useful in constructing them in practice. I will use these so often I'm going to name them.

Lemma 36 (Basis reduction lemma)**.** *If $S = \{v_1, \cdots, v_n\}$ is any spanning set for V, there is a subset $S' \subset S$ which is a basis for V.*

Proof. This is the content of Corollary 30: given any finite set of vectors, you can order them into a list and remove all the redundant vectors to obtain a linearly independent set with the same span. In this case, a linearly independent set with the same span has $\text{span}(S') = \text{span}(S) = V$, so is a linearly independent spanning set for V — that is, it's a basis. □

Proposition 37. *Every finite-dimensional vector space V has a finite basis.*

Proof. Choose a finite spanning set $\{v_1, \cdots, v_n\}$ for V and apply the basis reduction lemma to find a basis contained in this set. □

We can slightly extend this result.

Lemma 38 (Basis extension lemma)**.** *Suppose V is a finite-dimensional vector space. If $S = \{v_1, \cdots, v_n\}$ is any linearly independent set for V, there is a larger set $S' = \{v_1, \cdots, v_n, v_1', \cdots, v_k'\}$ which is a basis for V (where $k = 0$ if $\text{span}(S) = V$).*

Proof. Let $\{w_1', \cdots, w_m'\}$ be any spanning set for V, and consider the list $(v_1, \cdots, v_n, w_1', \cdots, w_m')$. This is a spanning list for V whose first n terms are linearly independent. It follows that all redundant vectors are among the vectors w_i'. Discarding these, we obtain a set $\{v_1, \cdots, v_n, w_1, \cdots, w_k\}$ of linearly independent vectors which span V. If $(v_1, \cdots, v_n)$ already span V, then all the added vectors are redundant and thrown out. □

Proposition 39. *Suppose $W \subset V$ is a subspace of a finite-dimensional vector space. Then there exists a basis $(w_1, \cdots, w_n)$ for W and a basis $(w_1, \cdots, w_n, v_1, \cdots, v_k)$ for V which begins with the given basis for W (where $k = 0$ if $W = V$).*

Proof. First, choose a basis $(w_1, \cdots, w_n)$ for W: one exists by Proposition 37 and the fact that W is finite-dimensional (because subspaces of finite-dimensional spaces are finite-dimensional, Corollary 34). In particular, this set is linearly independent. Next, apply the basis extension lemma to extend this linearly independent set to a basis for V. □

So we finally know that every vector space has a basis, and we can study vector spaces by studying their bases.

4.7.1 Dimensions

Let's use the results from the previous section to show that you can in fact extract information from a basis for a vector space.

Corollary 40 (Dimension is well-defined)**.** *Any two bases for a finite-dimensional vector space V have the same number of elements. If $\{v_1, \cdots, v_n\}$ and $\{w_1, \cdots, w_m\}$ are bases for V, then $n = m$.*

Proof. If $\{v_1, \cdots, v_n\}$ and $\{w_1, \cdots, w_m\}$ are both bases for V, notice that in particular the first set is linearly independent while the second spans V; applying Theorem 31 we see that $n \leqslant m$. On the other hand, the second set is also linearly independent and the first spans V, so that $m \leqslant n$ as well. Because $n \leqslant m \leqslant n$, we must in fact have $m = n$. □

This quantity — the number of elements in a basis — is crucial. It tells us *exactly how much information we need to describe an arbitrary element of V.* If $(v_1, \cdots, v_n)$ is a basis, then there is a bijection $A : \mathbb{F}^n \to V$ given by sending

$$A \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} = a_1 v_1 + \cdots + a_n v_n.$$

(This is our first example of a 'linear transformation.) Surjectivity is the fact that $(v_1, \cdots, v_n)$ spans V, while injectivity follows from the fact that $(v_1, \cdots, v_n)$ is linearly independent.

I need n coordinates to describe an arbitrary element of V. If $V = \text{span}(v)$ (where $v \neq \vec{0}$) is a line, this means I need exactly one coordinate to describe a point on a line; if $V = \text{span}(v, w)$ (where $v \neq \vec{0}$ and $w \notin \text{span}(v)$) is a plane, this means I need exactly two coordinates to describe a point on a plane. To my reckoning, a line is 1-dimensional, and a plane is 2-dimensional, so this inspires the following definition.

Definition 24. Let V be a finite-dimensional vector space. We define $\dim V \in \mathbb{N}$ to be the number of elements in a basis for V. ◊

The fact that this actually gives an unambiguous number is precisely because of the previous corollary: any two bases have the exact same number of elements.

Example 46. Because $\{e_1, \cdots, e_n\}$ is a basis for $\mathbb{F}^n$, we have $\dim(\mathbb{F}^n) = n$. ◊

Remark 31. As a special case of the previous example, we have $\dim\{\vec{0}\} = 0$, because $\varnothing$ is a basis for $\{\vec{0}\}$. ◊

Example 47. Let's find a basis for the space

$$V = \left\{ \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} \in \mathbb{F}^4 \;\middle|\; a + 3c + 4d = 0, \;\; a = c \right\},$$

where $\mathbb{F}$ is a field in which $4 \neq 0$ (for instance, $\mathbb{Q}$; a field where $2 = 0$ is called 'a field of characteristic two', and $4 = 0$ in a field in fact implies $2 = 0$).

My first thought is to find a spanning set (maybe with some redundant vectors) and remove any redundant vectors as necessary. I notice that $v_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ -1 \end{pmatrix}$ is in this set, because $1 + 3(1) + 4(-1) = 0$ and $1 = 1$; I also notice that $v_2 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$ is in this set. Now, if $\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}$ is an arbitrary vector in V, notice that because $a = c$ we have $4a + 4d = 0$, so that (because $4 \neq 0$ so that we can divide by 4) we have $d = -a$. Thus thus vector takes the form

$$\begin{pmatrix} a \\ b \\ a \\ -a \end{pmatrix} = a \begin{pmatrix} 1 \\ 1 \\ 1 \\ -1 \end{pmatrix} + (b - a) \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = av_1 + (b - a)v_2.$$

Thus these two vectors span V. On the other hand, neither is redundant, as $v_1 \neq \vec{0}$ and $v_2 \notin \operatorname{span}(v_1)$, so they are linearly independent, and thus $\{v_1, v_2\}$ forms a basis for V.

Thus $\dim V = 2$. (If we work in a field where $2 = 0$, interestingly enough $\dim V = 3$ instead: the two defining equations simplify to just the one equation $a = c$.) ◊

Example 48. If I want, I can extend the previous basis to a basis for the whole of $\mathbb{F}^4$. One way to do this is to stick on a set I already know spans $\mathbb{F}^4$ and remove redundant vectors: $(v_1, v_2, e_1, e_2, e_3, e_4)$ is a spanning set for $\mathbb{F}^4$. The first two vectors were shown to be non-redundant above. The vector e_1 is not redundant because it does not lie in $\operatorname{span}(v_1, v_2) = V$ (eg, we do not have $a = c$ for the vector e_1). On the other hand, the vector e_2 already appeared earlier in this list (in fact, $v_2 = e_2$), so it is redundant.

Next we should determine if e_3 is redundant. The general form of a vector in $\operatorname{span}(v_1, v_2, e_1)$ is

$$a_1v_1 + a_2v_2 + a_3e_1 = \begin{pmatrix} a_1 \\ a_1 \\ a_1 \\ -a_1 \end{pmatrix} + \begin{pmatrix} 0 \\ a_2 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} a_3 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} a_1 + a_3 \\ a_1 + a_2 \\ a_1 \\ -a_1 \end{pmatrix}.$$

In particular, the third coordinate is the negative of the fourth coordinate. The vector e_3 does not take this form, as its third component (1) is not the negative of its fourth component (0).

Lastly, e_4 is indeed redundant, as

$$e_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \\ -1 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} = -1v_1 + 1v_2 + 1e_1 + 1e_3.$$

Thus (v_1, v_2, e_1, e_3) is a basis for $\mathbb{F}^4$ whose first two terms give a basis for V.

◊

The second part of the next theorem is the first result that legitimately uses finite dimensionality. (The point is that for finite sets $S \subset S'$, if S is a proper subset of S', then $|S| < |S'|$. This is no longer true for infinite sets, where you can have proper subsets of nonetheless the same cardinality, such as the naturals $\mathbb{N}$ and the even naturals $2\mathbb{N}$).

Theorem 41. *If $W \subset V$ is a subspace of a finite-dimensional vector space, we have $\dim W \leqslant \dim V$. If $\dim W = \dim V$, then in fact $W = V$.*

Note the useful contrapositive statement: if $W \subset V$ is a proper subspace of a finite-dimensional vector space ($W \neq V$), then $\dim W < \dim V$.

Proof. Use Proposition 39 (that there exists a basis for V which starts with a basis for W) to find a basis $(w_1, \cdots, w_n, v_1, \cdots, v_k)$ for V so that $(w_1, \cdots, w_n)$ is a basis for W. This gives $\dim V = n+k$ and $\dim W = n$; as $k \geqslant 0$, we have $\dim W \leqslant \dim V$.

If $\dim W = \dim V$, then we must have $k = 0$, so in fact $(w_1, \cdots, w_n)$ is a basis for V. In particular, as it spans both W and V, we have $W = V$. □

Remark 32. There is a sort of arithmetic for infinite cardinalities, and in it this argument fails, because for 'infinite numbers' you can have $n + k = n$ even though $k > 0$. For instance, $|\mathbb{N}| + |\mathbb{N}| = |\mathbb{N}|$: two copies of the naturals are in bijection with the naturals themselves by sending one copy to the even integers and one copy to the odd integers.) You still have $n + k \geqslant n$, so the first part is fine. ◊

This is our first result which truly requires the use of finite-dimensional vector spaces, and it will be the foundation for many future results that truly require finite-dimensionality. The next — which uses it — often appears in computational linear algebra courses (in a different guise, which we will see explicitly later).

Corollary 42. *If $\{v_1, \cdots, v_n\}$ is a set of n linearly independent vectors in an n-dimensional vector space V, then in fact $\{v_1, \cdots, v_n\}$ spans V.*

So if you can find n linearly independent vectors and you happen to know that's enough (say, in $\mathbb{F}^n$), you now know that's a basis without doing any span computations.

Proof. Suppose $\{v_1, \cdots, v_n\}$ is linearly independent. Consider $W = \operatorname{span}(v_1, \cdots, v_n) \subset V$. Because $\{v_1, \cdots, v_n\}$ is a linearly independent spanning set for W, we have $\dim(W) = n = \dim V$. By Theorem 41, this implies $W = V$, so $\{v_1, \cdots, v_n\}$ spans V. □

Chapter 5

Foundations of linear maps

Alternate references

We now move on to the study of linear transformations (sometimes called linear maps or linear functions). Chapter 3 of Axler's book "Linear Algebra Done Right" remains somewhat adjacent to our approach, but starts to diverge; in particular, he has a strong desire to mention matrices as little as possible, whereas I want to try to make the matrix picture comprehensible to you (as if you had also taken a computational linear algebra class).

Chapter 1.3-1.6, Chapter 2 of Sergei Treil's book "Linear Algebra Done Wrong" may also serve as a nice supplemnt (you may find Chapter 1.8 interesting, though it won't help you understand the material). He is less interested in removing matrices (and later, determinants) from the exposition, which more closely matches my own view.

5.1 Linear maps

Finite-dimensional vector spaces are *a useful language*, but as it turns out there is not much to study about them. In a very important sense, every finite-dimensional vector space is "equivalent to" $\mathbb{F}^n$ for some n (technically, we will say every vector space is isomorphic to $\mathbb{F}^n$), and even if we want to talk about subspaces, we can prove that every pair (W, V) where $W \subset V$ is a subspace of the vector space V is "equivalent to" some pair $(\mathbb{F}^k, \mathbb{F}^n)$, where $\mathbb{F}^k \subset \mathbb{F}^n$ is the subspace

$$\mathbb{F}^k = \left\{ \begin{pmatrix} a_1 \\ \cdots \\ a_k \\ 0 \\ \cdots \\ 0 \end{pmatrix} \in \mathbb{F}^n \quad \middle| \quad a_1, \cdots, a_k \in \mathbb{F}; \quad a_{k+1} = \cdots = a_n = 0 \right\} \subset \mathbb{F}^n.$$

Thus as objects of study, vector spaces and their subspaces by themselves are not very interesting: we will later say that they are determined up to isomorphism by their dimensions.

In reality, vector spaces themselves are used as homes for more interesting things of actual importance. The actual fundamental object of study is a *linear map between vector spaces*: a function which takes us from one vector space to another, in a way compatible with the operations of linear algebra (addition and scaling).

Definition 25. Let V and W be vector spaces over the field $\mathbb{F}$. A **linear map** from V to W (sometimes "linear transformation" or "linear operator") is a function $A : V \to W$ which satisfies

(L1) For all $v_1, v_2 \in V$, we have $A(v_1 + v_2) = A(v_1) + A(v_2)$ ("A respects addition"),

(L2) For all $c \in \mathbb{F}$ and $v \in V$, we have $A(cv) = cA(v)$ ("A respects scaling").

◊

Example 49. The function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by

$$f\begin{pmatrix} x \\ y \\ z \end{pmatrix} = 3x + 7y - z$$

is a linear function: it satisfies (L1) because

$$\begin{aligned} f\left(\begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} + \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix}\right) &= f\begin{pmatrix} x_1 + x_2 \\ y_1 + y_2 \\ z_1 + z_2 \end{pmatrix} \\ &= 3(x_1 + x_2) + 7(y_1 + y_2) - (z_1 + z_2) = (3x_1 + 7y_1 - z_1) + (3x_2 + 7y_2 - z_2) \\ &= f\begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} + f\begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix}, \end{aligned}$$

and (L2) because

$$f\begin{pmatrix} cx \\ cy \\ cz \end{pmatrix} = 3(cx) + 7(cy) - (cz) = c(3x + 7y - z).$$

In fact, much more generally, pick $a_1, \cdots, a_n \in \mathbb{F}$. The function $f : \mathbb{F}^n \to \mathbb{F}$ given by

$$f\begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = a_1x_1 + \cdots + a_nx_n$$

is linear, by essentially the same argument: explicitly compute both sides of the equalities in (L1) and (L2) and see that they are equal. In fact, every linear function is of this form. ◊

Lemma 43. *Every linear function $f : \mathbb{F}^n \to \mathbb{F}$ takes the form $f\begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = a_1x_1 + \cdots + a_nx_n$ for some $a_1, \cdots, a_n \in \mathbb{F}$.*

Proof. Notice that $\begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = x_1e_1 + \cdots + x_ne_n$. If $f : \mathbb{F}^n \to \mathbb{F}$ is a linear map, we have

$$f(x_1e_1 + \cdots + x_ne_n) = f(x_1e_1) + \cdots + f(x_ne_n) = x_1f(e_1) + \cdots + x_nf(e_n).$$

Set $a_i = f(e_i)$; because multiplication in a field is commutative, we have just shown that

$$f\begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = x_1f(e_1) + \cdots + x_nf(e_n) = a_1x_1 + \cdots + a_nx_n,$$

as claimed. □

Exercise. Before moving on, show that for any vector space V over $\mathbb{F}$, a linear map $T : \mathbb{F} \to V$ takes the form $T(c) = cv$ for some vector $v \in V$. I think of the map T as tracing out a line in V (if $v \neq \vec{0}$). Then show that T is injective if and only if $v \neq \vec{0}$.

Non-example 8. The function $f : \mathbb{F}^2 \to \mathbb{F}$ defined by $f(x, y) = x + y + 1$ is not a linear function, because $f(1, 0) = 2$ and $f(0, 1) = 2$, while

$$f(1, 1) = 3 \neq 4 = f(1, 0) + f(0, 1).$$

You may be used to functions $f(x) = ax + b$ being called linear; they are not, according to the definition above of linear map. In linear algebra I would call $f(x) = ax + b$ for $b \neq 0$ an **affine** function. ◊

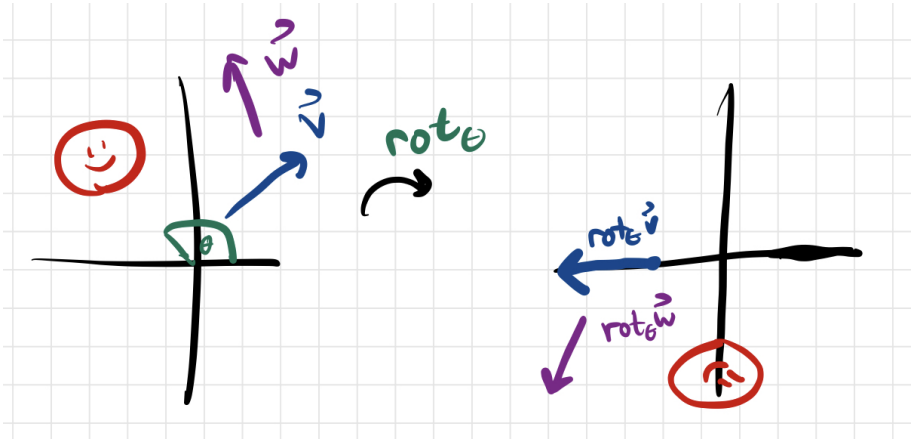
The above point is important: linear maps do not have 'constant terms'! You could axiomatize that, but the axiom system is much more complicated and unpleasant (for instance, the axiom corresponding to (L1) would be $A(v_1 + v_2 - v_3) = A(v_1) + A(v_2) - A(v_3)$). The claim that linear maps do not have 'constant terms' is formally the following statement:

Lemma 44. *If $A : V \to W$ is a linear map, we have $A(\vec{0}) = \vec{0}$.*

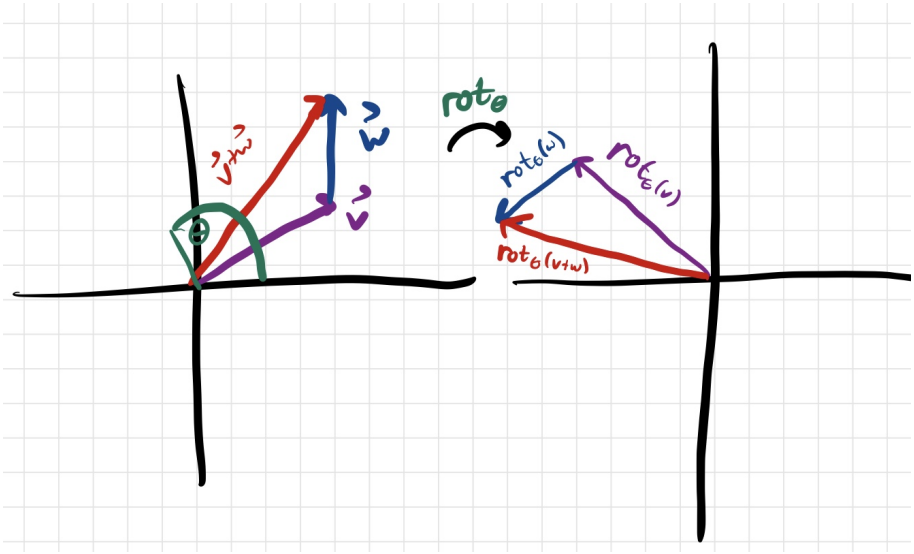
Proof. We have $A(\vec{0}) = A(0 \cdot \vec{0}) = 0 \cdot A(\vec{0}) = \vec{0}$. □

So far, the linear maps we have seen have not been very interesting (they take the form: "scale the coefficients and add them up"). It turns out that already in two dimensions linear maps can get very intricate.

Example 50. We are going to work with $\mathbb{R}^2$ in a way which is, for once, actually special to the real numbers. Pick a point in the plane and an angle $\theta \in \mathbb{R}$ (though two angles which differ by an integer multiple of 2π will lead to the same result). Define a map $\mathrm{rot}_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ by saying that $\mathrm{rot}_\theta(v)$ is the result of rotating the vector v by θ radians counter-clockwise around the circle.

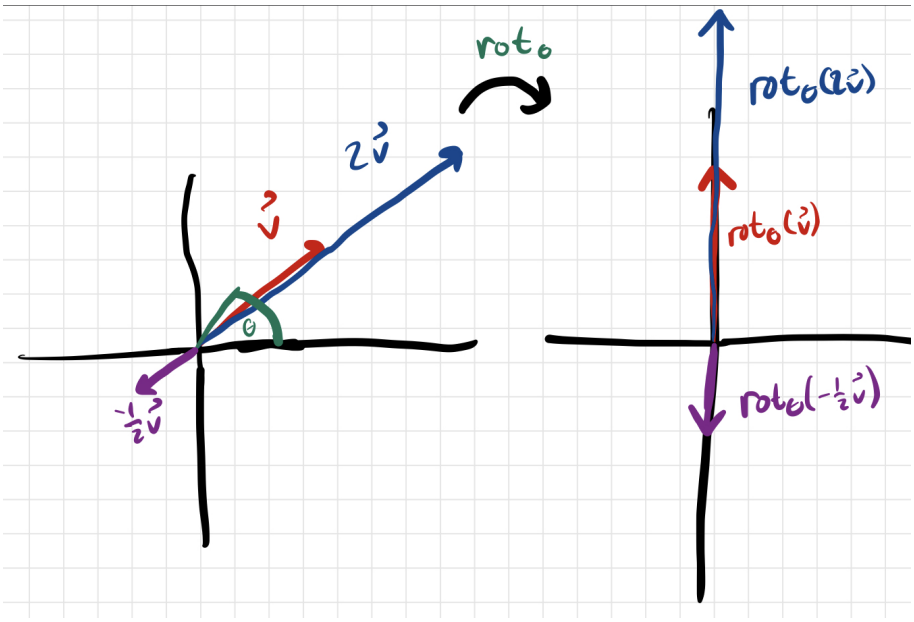


Even without giving a formula for this function, we can prove that it's linear. If v and w are vectors in $\mathbb{R}^2$, then $v + w$ is the third vector in a triangle whose sides are v and w (where the foot of w is at the head of v). If we rotate this triangle by angle θ, the result is again a triangle, now with sides $r_\theta(v), r_\theta(w)$, and $r_\theta(v + w)$.



But recall that $r_\theta(v) + r_\theta(w)$ is by definition the third side in a triangle whose first two sides are $r_\theta(v)$ and $r_\theta(w)$ (with the latter's foot at the head of the former).

The argument that $r_\theta(cv) = cr_\theta(v)$ is similar: cv is the vector parallel to v, which is $|c|$ times as long, and points in the same direction if $c > 0$ and in the opposite direction if $c < 0$. Because rotation takes parallel vectors to parallel vectors, preserves length, and preserves whether or not two vectors on the same line point in the same direction or not, $\text{rot}_\theta(cv)$ is also parallel to $\text{rot}_\theta(v)$, $|c|$ times as long, and points in the same direction (or opposite) as is appropriate.



On your homework, you will find a formula for this linear transformation using little more than basic linear algebra facts and the definition that $\cos(\theta), \sin(\theta)$ are the (x, y)-coordinates of a point θ radians counterclockwise from $(1, 0)$ on the unit circle. ◊

One example of a linear map between more complicated vector spaces (which you have seen before!) is the derivative.

Example 51. Recall that I write $C^0(\mathbb{R})$ for the set of continuous functions $f : \mathbb{R} \to \mathbb{R}$, and $C^1(\mathbb{R})$ for the set of differentiable functions $f : \mathbb{R} \to \mathbb{R}$ whose derivative is continuous. These are both vector spaces with addition defined by taking the sum of two functions $f + g$ to be the function $(f + g)(x) = f(x) + g(x)$, and the scalar multiple cf of a function f to be $(cf)(x) = c \cdot f(x)$. The vector space $C^1(\mathbb{R})$ is a subspace of $C^0(\mathbb{R})$ (a differentiable function is continuous).

Let $\frac{d}{dt} : C^1(\mathbb{R}) \to C^0(\mathbb{R})$ be the operation sending a function $f : \mathbb{R} \to \mathbb{R}$ to its derivative $\frac{d}{dt}f : \mathbb{R} \to \mathbb{R}$, sometimes written f'. By definition, a function $f \in C^1(\mathbb{R})$ is differentiable (so f' exists) and its derivative is continuous (so $f' \in C^0(\mathbb{R})$), so this does indeed define a function $C^1(\mathbb{R}) \to C^0(\mathbb{R})$.

That $\frac{d}{dt}$ is a linear map amounts to the claims (L1)

$$\frac{d}{dt}(f + g) = \frac{d}{dt}f + \frac{d}{dt}g$$

(the "addition rule") and (L2)

$$\frac{d}{dt}(cf) = c\frac{d}{dt}f$$

(the "product rule" where one of the functions is a constant function c, for which $c' = 0$).

Note that this does *not* mean that I'm claiming the derivative of a linear map is a linear map (whatever that means). For instance, the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = 2x$ is linear, but its derivative $f'(x) = 2$ is not. What I'm asserting is that *the differentiation operator* is a linear map, where "linear" just means:

you can add and scale either before or after applying the operation; whether you do so before or afterwards won't change the result. ◊

Here are some rather tautological examples.

Example 52. Let V be any vector space. The identity map $1_V : V \to V$ is the map which sends $1_V(v) = v$ for all $v \in V$; that is, it sends any v to itself. It changes nothing. This is a linear map because $1_V(v + w) = v + w = 1_V(v) + 1_V(w)$ and $1_V(cv) = cv = c1_V(v)$.

On the other hand, let V and W be any vector spaces. Then the map $0_{V,W} : V \to W$ which has $0_{V,W}(v) = \vec{0}$ for all $v \in V$, is also linear, because $0_{V,W}(v+w) = \vec{0} = \vec{0}+\vec{0} = 0_{V,W}(v)+0_{V,W}(w)$, and similarly with scaling. ◊

The next example will later let us analyze the relations of spans and linear independence in the language of linear transformations.

Example 53. Let V be a vector space, and $(v_1, \cdots, v_n)$ a list of vectors in V. Then there is a linear map $A : \mathbb{F}^n \to V$ defined by

$$A \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} = a_1 v_1 + \cdots + a_n v_n.$$

This is linear because

$$\begin{aligned} A \begin{pmatrix} a_1 + b_1 \\ \cdots \\ a_n + b_n \end{pmatrix} &= (a_1 + b_1)v_1 + \cdots + (a_n + b_n)v_n \\ &= (a_1 v_1 + \cdots + a_n v_n) + (b_1 v_1 + \cdots + b_n v_n) \\ &= A \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} + A \begin{pmatrix} b_1 \\ \cdots \\ b_n \end{pmatrix}, \end{aligned}$$

while

$$A \begin{pmatrix} ca_1 \\ \cdots \\ ca_n \end{pmatrix} = (ca_1)v_1 + \cdots + (ca_n)v_n = c(a_1 v_1 + \cdots + a_n v_n).$$

Notice, in particular, that $Ae_i = v_i$. ◊

I think the main thing that is interesting in this example is that I can tell you the entire map A just by telling you what happens to the vectors $e_1, \cdots, e_n$ (that is, I specified the outputs $Ae_i = v_i$, and this automatically gave me the linear map $A \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} = a_1 v_1 + \cdots + a_n v_n$.

The theorem below says that this is true for general linear transformations: *know what they do on a basis, and you know what they do everywhere.*

Theorem 45 (Linear maps are determined by their values on a basis)**.** *Suppose V and W are vector spaces, and that $\{v_1, \cdots, v_n\}$ is a basis for V. For any list of n vectors $(w_1, \cdots, w_n)$ in W, there exists a unique linear map $A : V \to W$ for which $Av_i = w_i$ for all $1 \leqslant i \leqslant n$.*

Proof. Let's start with uniqueness. Suppose $A : V \to W$ and $B : V \to W$ are linear maps with $Av_i = w_i = Bv_i$ for all $1 \leqslant i \leqslant n$. Let's prove that $Av = Bv$ for all $v \in V$, so that $A = B$. To see this, recall that $\{v_1, \cdots, v_n\}$ spans V (by definition of a basis); for any $v \in V$, there exist $a_1, \cdots, a_n \in \mathbb{F}$ so that $v = a_1 v_1 + \cdots + a_n v_n$. Then

$$\begin{aligned} A(v) &= A(a_1 v_1 + \cdots + a_n v_n) = a_1 A(v_1) + \cdots + a_n A(v_n) \\ &= a_1 B(v_1) + \cdots + a_n B(v_n) = B(a_1 v_1 + \cdots + a_n v_n) = B(v). \end{aligned}$$

Thus for any $v \in V$ we have $A(v) = B(v)$, so these functions are equal.

As for existence, recall that in fact for all $v \in V$ there is a **unique** way to write $v = a_1v_1 + \cdots + a_nv_n$, by Lemma 35. Define the map $A : V \to W$ by

$$A(a_1v_1 + \cdots + a_nv_n) = a_1w_1 + \cdots + a_nw_n.$$

Because every vector in V can be written uniquely as a linear combination of $v_1, \cdots, v_n$, this gives an unambiguous definition of A on every element of V. Further, it's linear (by the same argument as in Example 53 above). Because $Av_i = w_i$ by definition, this proves the existence of such a map. □

Remark 33. One small upshot of this is that if you're trying to show $A, B : V \to W$ are the same map, it suffices to check they have the same values on some basis of V; you don't need to check them on every vector. ◊

Example 54. One example that I think is somewhat nice to think of in terms of the way it acts on a basis — though admittedly infinite-dimensional — is the derivative operator on polynomials. Let $\mathbb{F} = \mathbb{R}$, and let $V = \mathbb{R}[t]$, the set of polynomials with real-valued coefficients. This vector space has the infinite basis $\{1, t, t^2, t^3, \cdots\}$, in the sense that every polynomial may be written as $a_0 + \cdots + a_nt^n$ for a unique sequence of coefficients $a_0, \cdots, a_n$.

The derivative map $\frac{d}{dt} : \mathbb{R}[t] \to \mathbb{R}[t]$ sends polynomials to polynomials, and on this basis we have $\frac{d}{dt}(t^n) = nt^{n-1}$ for all $n \geqslant 0$ (where in particular $\frac{d}{dt}t = 1$ and $\frac{d}{dt}1 = 0$), by the power rule. Knowing this computation is enough to tell us the value of the derivative on every polynomial by the scaling and addition rules:

$$\frac{d}{dt}(a_0 + \cdots + a_nt^n) = a_0\frac{d}{dt}(1) + \cdots + a_n\frac{d}{dt}(t^n) = a_1 + \cdots + na_{n-1}t^{n-1}.$$

For instance,

$$\frac{d}{dt}(2 + 3t + 4t^2 + 7t^3) = 3 + 8t + 21t^2;$$

you implicitly use that the derivative is a linear operator when expanding this, and you just needed to know what it did to each t^n to determine the whole result. ◊

5.2 Subspaces from linear maps: image and kernel

We've spent a long time discussing vector spaces and their subspaces, and I want to use these to help us get access to linear maps.

Definition 26. Let $A : V \to W$ be a linear map. Its **kernel** is the subspace $\ker(A) \subset V$ defined by

$$\ker(A) = \{v \in V \mid Av = \vec{0}\}.$$

Its **image** is the subspace $\mathrm{im}(A) \subset W$ defined by

$$\mathrm{im}(A) = \{Av \mid v \in V\} = \{w \in W \mid \exists_{v \in V} Av = w\}.$$

◊

That is, the kernel[1] is the set of all vectors A crushes to nothing (the set of all vectors A sends to zero). If I think a vector $v \in V$ carries some sort of "information", then the elements of the kernel of A are those vectors whose information is lost as we pass to W (they are sent to the zero vector, which surely contains no information. I cannot recover anything about vectors in $\ker(A)$ from the behavior of Av, because they are all sent to zero. In terms of inverse images, $\ker(A) = A^{-1}(\vec{0})$; it is the set of all vectors which A sends to the zero vector.

The image is precisely the same as the set-theoretic notion of image from Chapter 2: it's the set of all vectors a linear transformation spits out.

Before moving on, I should verify my claim that these are linear subspaces.

[1]Kernel is perhaps an unexpected word here, familiar to most as kernels of corn. The word 'kernel' refers to a seed of certain plants (think the center of a chestnut) or more broadly the core or central idea of something. It is difficult to find a good motivated history of the use of the term in linear algebra, but it's not hard to make up something that sounds plausible: we will see later that the kernel governs the set of solutions to equations like $Ax = b$, so the behavior of the kernel is the *core part* of understanding how to solve general systems of linear equations. Whether or not this is what was meant when the word was invented is a different question, harder to answer.

Lemma 46. *If $A : V \to W$ is linear, then* $\ker(A)$ *and* $im(A)$ *are indeed both subspaces (of V and W, respectively).*

Proof. (S1) If $v_1, v_2 \in \ker(A)$, then by (L1) we have $A(v_1 + v_2) = Av_1 + Av_2 = \vec{0} + \vec{0} = \vec{0}$, so that $v_1 + v_2 \in \ker(A)$ as well. If $w_1, w_2 \in \mathrm{im}(A)$, then by definition of image there exist $v_1, v_2 \in V$ so that $Av_i = w_i$. Then by (L1) $A(v_1 + v_2) = Av_1 + Av_2 = w_1 + w_2$, so $w_1 + w_2 \in \mathrm{im}(A)$ as well.

(S2) Fix $c \in \mathbb{F}$. If $v \in \ker(A)$ by (L2) we have $A(cv) = c(Av) = c \cdot \vec{0} = \vec{0}$, so $cv \in \ker(A)$ as well. If $w \in \mathrm{im}(A)$, then $w = Av$ for some v; by (L2) we have $A(cv) = c(Av) = cw$, so $cw \in \mathrm{im}(A)$.

(S3) By Lemma 44 we have $A(\vec{0}) = \vec{0}$, which proves both $\vec{0} \in \ker(A)$ and $\vec{0} \in \mathrm{im}(A)$.

□

To get a sense for these subspaces, let's see them in the examples we went through before.

Example 55. For a linear function $f : \mathbb{F}^n \to \mathbb{F}$, given by $f\begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = a_1x_1 + \cdots + a_nx_n$, by definition

$$\ker A = \left\{ \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} \;\middle|\; a_1x_1 + \cdots + a_nx_n = 0 \right\}.$$

This represents a line in $\mathbb{F}^2$, or a plane in $\mathbb{F}^3$, and the proof above very concisely explains why this is a subspace. More complicated subspaces of a similar form, such as

$$V = \left\{ \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \;\middle|\; 4x_1 - 3x_2 + x_4 = 0, \;\; x_1 + 3x_3 + 4x_4 = 0 \right\}$$

are also automatically seen to be subspaces, because V is the kernel of the linear transformation $A : \mathbb{F}^4 \to \mathbb{F}^2$ defined by

$$A\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 4x_1 - 3x_2 + x_4 \\ x_1 + 3x_3 + 4x_4 \end{pmatrix}.$$

The image of f in the example above is very simple. If $a_1 = \cdots = a_n = 0$, so that $f(x) = 0$ for all x, then $\ker(f) = \mathbb{F}^n$ is everything and $\mathrm{im}(f) = \{0\}$ consists only of zero. On the other hand, if a_i is nonzero for some i, then $\ker(f)$ is a proper subspace of $\mathbb{F}^n$ and $\mathrm{im}(f) = \mathbb{F}$ is everything. To see that $\mathbb{F} \subset \mathrm{im}(f)$, observe that

$$f\left(\frac{c}{a_i}e_i\right) = \frac{c}{a_i} \cdot a_i = c$$

for any $c \in \mathbb{F}$ (where here I used that $a_i \neq 0$ to divide by it). ◊

Example 56. Consider the rotation operator $\mathrm{rot}_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ by angle θ. We have $\ker(\mathrm{rot}_\theta) = \{\vec{0}\}$ because rotation preserves the length of vectors; if $\mathrm{rot}_\theta(v) = \vec{0}$, then v must have had length zero, so v must have been the zero vector. On the other hand, $\mathrm{im}(\mathrm{rot}_\theta v) = \mathbb{F}^2$: every vector $w \in \mathbb{F}^2$ can be written as $\mathrm{rot}_\theta v$ for some $v \in \mathbb{F}^2$ (meaning that w is the vector v rotated θ degrees around the origin). To see this, observe that I can take v to be $\mathrm{rot}_{-\theta} w$: if I rotate w clockwise θ degrees, then rotating that counterclockwise by θ degrees gets me back to w. ◊

Example 57. The derivative map $\frac{d}{dt} : C^1(\mathbb{R}) \to C^0(\mathbb{R})$ is an interesting example. **Exercise:** Compute $\ker \frac{d}{dt}$ and $\mathrm{im}\ \frac{d}{dt}$.

Did you complete the exercise? I'm going to answer it below, but it's useful as practice in understanding the definitions of im , ker, and $\frac{d}{dt}$.

I claim that $\ker \frac{d}{dt} = \mathbb{R} \subset C^1(\mathbb{R})$ consists of the constant functions. This is because $\ker \frac{d}{dt}$ is the set of functions $f : \mathbb{R} \to \mathbb{R}$ whose derivative is zero. Suppose $f(x)$ is such a function. By the fundamental theorem of calculus, we have $f(x) - f(0) = \int_0^x f'(t)dt$, so because $f'(t) = 0$ for all t, the right-hand side is zero for all x; thus $f(0) = f(x)$ for all $x \in \mathbb{R}$, and f is a constant function.

Next, I claim that $\text{im } \frac{d}{dt} = C^0(\mathbb{R})$ is all of $C^0(\mathbb{R})$. Once again, this is the fundamental theorem of calculus. If $f \in C^0(\mathbb{R})$ is a continuous function, set $F(t) = \int_0^t f(x)dx$ to be its definite integral (with $F(0) = 0$). Then the fundamental theorem of calculus asserts $f(t) = F'(t) = \frac{d}{dt}\int_0^t f(x)dx$.

Exercise. Show that the map $\int : C^0(\mathbb{R}) \to C^1(\mathbb{R})$ defined by $(\int f)(t) = \int_0^t f(x)dt$ is a linear map, has $\ker \int = \{0\}$ and

$$\text{im } \int = \{F \in C^1(\mathbb{R}) \mid F(0) = 0\}.$$

♢

Example 58. If $1_V : V \to V$ is the identity map, then $\ker 1_V = \{\vec{0}\}$ and $\text{im } 1_V = V$. On the other hand, if $0_{V,W} : V \to W$ is the zero map, we have $\ker 0_{V,W} = V$ and $\text{im } 0_{V,W} = \{\vec{0}\}$. ♢

Example 59. Suppose V is a vector space and $(v_1, \cdots, v_n)$ a list of n vectors in W. We defined in Example 53 a linear map $A : \mathbb{F}^n \to V$ defined by

$$A \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} = a_1 v_1 + \cdots + a_n v_n.$$

Then $\text{im } A = \text{span}(v_1, \cdots, v_n)$, because

$$\{Ax \mid x \in \mathbb{F}^n\} = \left\{ A \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} \;\middle|\; a_1, \cdots, a_n \in \mathbb{F} \right\} = \{a_1 v_1 + \cdots + a_n v_n \mid a_1, \cdots, a_n \in \mathbb{F}\} = \text{span}(v_1, \cdots, v_n).$$

On the other hand, $\ker A = \text{Rel}(v_1, \cdots, v_n)$ is the set of linear relations between this list of vectors. Precisely, we have

$$\ker A = \{x \in \mathbb{F}^n \mid Ax = \vec{0}\} = \left\{ \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} \in \mathbb{F}^n \;\middle|\; A \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} = \vec{0} \right\} = \left\{ \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} \in \mathbb{F}^n \;\middle|\; a_1 v_1 + \cdots + a_n v_n = \vec{0} \right\} = \text{Rel}(v_1, \cdots, v_n).$$

The two fundamental notions associated to a list of vectors $(v_1, \cdots, v_n)$ can be phrased in the language of linear maps and their associated subspaces. ♢

If $A : V \to W$ is a linear map, then A is surjective if and only if $\text{im } A = W$ (this is the definition of surjectivity). An analogous statement is true for $\ker A$ (and we saw an avatar of this when proving that there is a unique way to write a given vector as a linear combination of basis vectors).

Lemma 47. *If $A : V \to W$ is a linear map, then A is injective if and only if $\ker A = \{\vec{0}\}$.*

Proof. Recall that $A(\vec{0}) = \vec{0}$ by Lemma 44. If A is injective (so $Ax = Ay$ implies $x = y$) then if $Av = \vec{0} = A(\vec{0})$, this implies $v = \vec{0}$. Thus $\ker A = \{v \in V \mid Av = \vec{0}\} = \{\vec{0}\}$.

Conversely, suppose $\ker A = \{\vec{0}\}$; let's show A is injective. If $v_1, v_2 \in V$ have $Av_1 = Av_2$, then we also have $Av_1 - Av_2 = \vec{0}$. By linearity (and the fact that $-Av = (-1) \cdot Av$) we have $Av_1 - Av_2 = A(v_1 - v_2)$, so that $A(v_1 - v_2) = \vec{0}$. Because $\ker A = \{\vec{0}\}$, this implies $v_1 - v_2 = \vec{0}$, so that $v_1 = v_2$ as claimed. □

Thus the kernel measures precisely why and how a linear map A fails to be injective. In fact, if $A : V \to W$ is any linear map and $Av_1 = Av_2$, then there exists a $w \in \ker A$ so that $v_2 = v_1 + w$; any two elements which map to the same thing differ by an element of $\ker A$. (The proof is the same; here $w = v_2 - v_1$.)

Maybe it deserves convincing you that "a linear map is injective" is something worth knowing in terms of the objects we've studied so far.

Lemma 48. *If $A : V \to W$ is injective, and $\{v_1, \cdots, v_n\}$ is linearly independent in V, then $\{Av_1, \cdots, Av_n\}$ is linearly independent in W.*

Proof. Suppose $a_1Av_1 + \cdots + a_nAv_n = 0$. Then by linearity we have

$$A(a_1v_1 + \cdots + a_nv_n) = \vec{0}.$$

Because A is injective, it in particular has $\ker A = \{\vec{0}\}$, so $a_1v_1 + \cdots + a_nv_n = \vec{0}$. Now we can apply linear independence of $\{v_1, \cdots, v_n\}$ to see that $a_1 = \cdots = a_n = 0$. □

We conclude by extracting numberical information out of the image and kernel.

Definition 27. If $A : V \to W$ is a linear transformation, its **rank** is $\text{rank}(A) = \dim \text{im}(A)$ the dimension of its image, while its **nullity** is $\text{null}(A) = \dim \ker(A)$, the dimension of its kernel. ◊

Example 60. The map $A : \mathbb{F}^3 \to \mathbb{F}$ given by $A \begin{pmatrix} x \\ y \\ z \end{pmatrix} = x + y + z$ has $\ker A$ of dimension 2 (find a basis), so $\text{null}(A) = 2$, whereas $\text{im } A = \mathbb{F}$ is 1-dimensional, so $\text{rank}(A) = 1$.

The zero map $A : \mathbb{F}^3 \to \mathbb{F}$, given by $Av = \vec{0}$, has $\ker A = \mathbb{F}^3$ and $\text{im } A = \{\vec{0}\}$, so that $\text{null}(A) = 3$ and $\text{rank}(A) = 0$. ◊

These quantities figure into the most powerful theorem in finite-dimensional linear algebra. (Stated appropriately, the theorem and proof go through just as well for infinite-dimensional vector spaces, but it is much more powerful in finite dimensions.)

Theorem 49 (Rank-nullity theorem)**.** *If $A : V \to W$ is a linear map between finite-dimensional vector spaces, we have rank(A) + null(A) = $\dim V$.*

The idea here is relatively simple. When I follow a linear map A from V to W, I kill off a chunk of vectors in V (the vectors in $\ker(A)$), and what survives lives on in W as $\text{im}(A)$. The statement above says, in some sense, that V itself can be thought of as put together out of the vectors that die and vectors that survive. This strikes me as reasonable, and the abstract proof below tries to make it precise.

I strongly encourage you to try to not just understand each step below, but how I came up with it. This style of argument is quite useful, and at the end of the term, you may well look back on this proof and think "This is the only reasonable way one could have written that argument."

Abstract proof. Choose a basis $\{w_1, \cdots, w_m\}$ for $\text{im}(A)$, so that $\text{rank}(A) = m$. Because these vectors are all in the image of A, there exist vectors $v_1, \cdots, v_m \in V$ so that $Av_i = w_i$ for all $1 \leqslant i \leqslant m$.

Next, choose a basis $\{u_1, \cdots, u_k\}$ for $\ker(A)$, so that $\text{null}(A) = k$. I claim that $\{v_1, \cdots, v_m, u_1, \cdots, u_k\}$ is a basis for V, so that $\dim V = m + k = \text{rank}(A) + \text{null}(A)$; this will complete the proof. (The idea is that $\{u_1, \cdots, u_k\}$ spans "the part of V which is sent to zero", whereas $\{v_1, \cdots, v_m\}$ spans "a part of V which maps identically onto $\text{im}(A)$".

First let's show that this forms a linearly independent set. Suppose $a_1v_1 + \cdots + a_mv_m + b_1u_1 + \cdots + b_ku_k = \vec{0}$. Our goal is to show $a_i = b_j = 0$ for all i, j. Then

$$\vec{0} = A(\vec{0}) = A(a_1v_1 + \cdots + b_ku_k) = a_1A(v_1) + \cdots + a_mA(v_m) + b_1A(u_1) + \cdots + b_kA(u_k).$$

By definition of the v_i, we have $Av_i = w_i$ for all $1 \leqslant i \leqslant m$. Further, because $u_i \in \ker(A)$ by definition, we have $A(u_i) = \vec{0}$ for all $1 \leqslant i \leqslant k$. Therefore the above expression simplifies to

$$\vec{0} = a_1w_1 + \cdots + a_mw_m.$$

Now $\{w_1, \cdots, w_m\}$ was a basis for $\mathrm{im}(A)$, so in particular linearly independent; this implies $a_1 = \cdots = a_m = 0$.

Thus our original expression reduces to $b_1 u_1 + \cdots + b_k u_k = \vec{0}$. Because $\{u_1, \cdots, u_k\}$ formed a basis for $\ker(A)$, they were in particular linearly independent, so that $b_1 = \cdots = b_k = 0$. We have now shown that $\{v_1, \cdots, v_m, u_1, \cdots, u_k\}$ is a linearly independent set.

What remains is to verify that these vectors span V, which we will also do in two steps. Pick an arbitrary $v \in V$. First, observe that $Av \in \mathrm{im}(A) = \mathrm{span}(w_1, \cdots, w_m)$, so that $Av = a_1 w_1 + \cdots + a_m w_m$. Write $v' = v - a_1 v_1 - \cdots - a_m v_m$. By linearity, we have

$$Av' = Av - a_1 w_1 - \cdots - a_m w_m = \vec{0}.$$

Therefore, $v' \in \ker(A)$. Now because $\{u_1, \cdots, u_k\}$ is a basis for $\ker A$, we have $v' = b_1 u_1 + \cdots + b_k u_k$ for some $b_1, \cdots, b_k$. Altogether, this establishes that

$$v = v' + a_1 v_1 + \cdots + a_m v_m = a_1 v_1 + \cdots + a_m v_m + b_1 u_1 + \cdots + b_k u_k,$$

so that v is in the span of our list of vectors. As $v \in V$ was an arbitrary element, this shows that $\{v_1, \cdots, u_k\}$ spans V, and completes the proof that it is a basis for V. □

On the other hand, most linear algebra texts give a more concrete and computational argument. The details are more irritating, I feel, but it also has some insight to offer, so I will include a summary of the idea.

Sketch of the concrete proof. Give V the basis $\{v_1, \cdots, v_n\}$. The proof you will usually see in most linear algebra texts gives an algorithm to determine the redundant vectors among the list $(Av_1, \cdots, Av_n)$ (ultimately reducing to "Gauss–Jordan elimination", which we will talk about next week). They prove (essentially using the HW4 bonus, usually spelled out explicitly) that the redundant vectors in this list give rise to an explicit basis for $\ker(A)$. This explicit argument is what takes the most time, and uses the Gauss–Jordan elimination algorithm in an essential way.

Removing redundant vectors gives a basis for $\mathrm{span}(Av_1, \cdots, Av_n) = \mathrm{im}(A)$. Thus $\mathrm{rank}(A)$ is the number of non-redundant vectors in this list, while $\mathrm{null}(A)$ is the number of redundant vectors in this list. It follows that $\dim V = n = \mathrm{rank}(A) + \mathrm{null}(A)$. □

Remark 34. In fact, the conclusion in the concrete argument follows from the abstract argument, too. If $(v_1, \cdots, v_n)$ is a list of vectors in V, we can define the map $A : \mathbb{F}^n \to V$ by sending $Ae_i = v_i$ and extending linearly to all of $\mathbb{F}^n$. For this linear map, $\mathrm{im}(A) = \mathrm{span}(v_1, \cdots, v_n)$, and the dimension of this set is precisely $n - \#\text{redundant vectors}$. On the other hand, $\ker(A) = \mathrm{Rel}(v_1, \cdots, v_n)$ is the space of all linear relations among $v_1, \cdots, v_n$. Then the rank-nullity theorem asserts that $\mathrm{Rel}(v_1, \cdots, v_n)$ has dimension equal to the number of redundant vectors among $v_1, \cdots, v_n$; in fact, you can use those redundant vectors to form a basis for this space of relations, using each "redundancy relation"

$$a_1 v_1 + \cdots + a_{i-1} v_{i-1} + (-1) v_i + 0 v_{i+1} + \cdots + 0 v_n = \vec{0}$$

as one of the basis vectors. ◊

Corollary 50. *If $A : V \to W$ is a linear map, then*

- *If A is injective, we have $\dim V \leqslant \dim W$.*
- *If A is surjective, we have $\dim V \geqslant \dim W$.*
- *If A is bijective, we have $\dim V = \dim W$.*

Proof. These all follow from the rank-nullity theorem. If A is injective, then $\ker A = \{\vec{0}\}$, so $\mathrm{null}(A) = 0$ and by the rank-nullity theorem $\dim V = \mathrm{rank}(A) = \dim \mathrm{im}(A)$. Because $\mathrm{im}(A) \subset W$ is a subspace, Theorem 41 implies $\dim V = \mathrm{rank}(A) \leqslant \dim W$.

If $A : V \to W$ is surjective, now we have $\mathrm{rank}(A) = \dim \mathrm{im}(A) = \dim W$. By the rank-nullity theorem, this gives $\mathrm{null}(A) + \dim W = \dim V$. Because $\mathrm{null}(A) \geqslant 0$, this implies $\dim W \leqslant \dim V$.

The final claim follows by combining the previous two. □

5.3 Compositions and invertibility

We can string together linear maps, doing one and then the next. That is, we can take *composites* of linear maps (where the codomain of the first is the domain of the second), and the result is again a linear map.

Lemma 51. *If $A : V \to W$ and $B : W \to U$ are linear maps, then the composite $(BA) : V \to U$, defined by $(BA)(v) = B(Av)$, is also linear.*

Proof. For all $v \in V$, we have

$$(BA)(v + w) = B(A(v + w)) = B(Av + Aw) = B(Av) + B(Aw) = (BA)v + (BA)w.$$

In the first and last step we used the definition of the map BA; in the second step we used that A is linear, and in the third step that B is linear. Similarly, for all $c \in \mathbb{F}$ and all $v \in V$, we have

$$(BA)(cv) = B(A(cv)) = B(c(Av)) = c(B(Av)) = c(BA)v,$$

by the exact same steps. □

Composition of linear maps has some reasonable arithmetic properties.

Lemma 52. *Composition of linear maps has the following properties.*

(i) If $A : V \to W, B : W \to U$, and $C : U \to X$ are linear maps between vector spaces, then $C(BA) = (CB)A$.

(ii) If $A : V \to W$ is a linear map, then $A1_V = A = 1_W A$.

(iii) If $A : V \to W$ is a linear map and U is any other vector space, then $0_{W,U}A = 0_{V,U}$ and $A0_{U,V} = 0_{U,W}$.

Proof. For (i), spell out the definitions:

$$C(BA)v = C((BA)v) = C(B(A(v)) = (CB)(A(v)) = ((CB)A)v.$$

For (ii), we have

$$(A1_V)v = A(1_V v) = Av = 1_W(Av) = (1_W A)v.$$

For (iii), the only subtlety is making sure the domains and codomains match up. $A : V \to W$ and $0_{W,U} : W \to U$ compose to a map $0_{W,U}A : V \to U$, for which

$$(0_{W,U}A)v = 0_{W,U}(Av) = \vec{0}$$

for all $v \in V$; it is thus the zero map from V to U, denoted $0_{V,U}$. The same argument applies for $A0_{U,V} = 0_{U,W}$. □

Remark 35. When it is clear what the vector spaces V and W are from context, I will later start dropping the subscripts from $0_{V,W}$, and call it the "zero map". ◊

Example 61. If $\mathrm{rot}_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ and $\mathrm{rot}_\psi : \mathbb{R}^2 \to \mathbb{R}^2$ are rotations of the plane counterclockwise about the origin by angles θ and ψ, respectively, then the composite $\mathrm{rot}_\theta \mathrm{rot}_\psi$ is the map which first rotates by ψ, and then rotates by θ, for a total rotation of $\theta + \psi$ radians counterclockwise. That is, we have

$$\mathrm{rot}_\theta \mathrm{rot}_\psi = \mathrm{rot}_{\theta+\psi}.$$

◊

Example 62. The map $\left(\frac{d}{dt}\right)^2 : \mathbb{F}[t] \to \mathbb{F}[t]$ which takes the second derivative of a polynomial is linear, because it is the composite of the linear map $\frac{d}{dt}$ with itself. (Since this map has the same domain and codomain, the composite is defined.)

On basis vectors, we have

$$\left(\frac{d}{dt}\right)^2 t^n = \left(\frac{d}{dt}\right)(nt^{n-1} = n\left(\frac{d}{dt}\right)t^{n-1} = n(n-1)t^{n-2},$$

where this should be interpreted as zero for $n = 0, 1$. That this map is linear tells us that we can write down the value of $\left(\frac{d}{dt}\right)^2$ on any polynomial:

$$\left(\frac{d}{dt}\right)^2 (a_0 + \cdots + a_n t^n) = 2a_2 + 6a_3 t + \cdots + n(n-1)t^{n-2}.$$

$\Diamond$

Example 63. If $A : V \to W$ and $B : W \to U$ are linear maps, the only composite that makes sense in general is BA; to define the map AB I would need to demand that $U = V$ (so B goes where A starts from).

Even when this is the case, it is usually not true that $AB = BA$. Composition of linear maps is not remotely commutative, even though it was in the exampe of rot_θ above. If I hand you two generic linear maps $V \to V$, it is unlikely that they commute. For instance, take $A, B : \mathbb{F}^2 \to \mathbb{F}^2$ defined by

$$A\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x+y \\ y \end{pmatrix}$$

and

$$B\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -y \\ x \end{pmatrix}.$$

Then

$$AB\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x-y \\ x \end{pmatrix}, \qquad BA\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -y \\ x+y \end{pmatrix}.$$

These are simply not the same map. For instance, $(AB)e_2 = -e_1$ whereas $(BA)e_2 = -e_1 + e_2$, and these are different outputs. $\Diamond$

Suppose you have a linear map $A : V \to W$ and you want to 'undo it'. To do so, you would need to find a linear map $B : W \to V$ which 'reverses the original process', but there are a few things you could mean.

Definition 28. If $A : V \to W$ is a linear map, we say that

- A is **right invertible** if there exists a linear map $B : W \to V$ so that $AB = 1_W$.
- A is **left invertible** if there exists a linear map $B : W \to V$ so that $BA = 1_V$.
- A is **invertible** if there exists a linear map $B : W \to V$ so that $AB = 1_W$ and $BA = 1_V$. In this case, we write $B = A^{-1}$.

$\Diamond$

Remark 36. If A is right invertible and also left invertible, possibly with different inverses $AB = 1_W$ and $CA = 1_V$, then in fact $B = C$ and A is invertible. To see this, write $C = C1_W = C(AB) = (CA)B = 1_V B = B$.

Further, if A is invertible, it has only one inverse by the same argument. $\Diamond$

Let's go back to two earlier examples to get a sense for these conditions.

Example 64. The rotation maps $\text{rot}_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ are invertible, with inverse $\text{rot}_{-\theta}$, as

$$\text{rot}_\theta \text{rot}_{-\theta} = \text{rot}_{\theta-\theta} = \text{rot}_0 = 1_{\mathbb{R}^2} = \text{rot}_0 = \text{rot}_{-\theta+\theta} = \text{rot}_{-\theta}\text{rot}_\theta.$$

Less symbolically, if I rotate counterclockwise by θ radians and then clockwise by θ radians, ultimately I've moved zero radians, so I've done nothing; similarly if I first rotate clockwise by θ radians and afterwards counterclockwise by θ radians. $\Diamond$

Example 65. Consider the differentiation map $\frac{d}{dt} : C^1(\mathbb{R}) \to C^0(\mathbb{R})$. I defined another linear map $\int : C^0(\mathbb{R}) \to C^1(\mathbb{R})$ going the other way. By the fundamental theorem of calculus, we have

$$\left(\frac{d}{dt}\int f\right)(x) = \frac{d}{dt}\int_0^x f(t)dt = f(x),$$

so that $\frac{d}{dt}\int f = f$. That is, $\int$ is a right inverse to $\frac{d}{dt}$ (and $\frac{d}{dt}$ is a left inverse to $\int$). If I integrate and then take the derivative, I end up back where we started!

However, integration is not actually an inverse. We have

$$(\int \frac{d}{dt} f)(x) = \int_0^x f'(t)dt = f(x) - f(0),$$

so $\int \frac{d}{dt}$ sends f to the same function minus $f(0)$. That is, if I take the derivative and then the integral, it doesn't quite get me back where I started (the new function always has $(\int g)(0) = 0$, whereas $f(0)$ could have been anything). ♢

Let me remind you of some set-theoretic consequences of the existence of the various types of inverses.

Proposition 53. *Suppose $A : V \to W$ is a linear map.*

- *If A is right invertible, then A is surjective.*
- *If A is left invertible, then A is injective.*
- *If A is invertible, then A is bijective.*

Proof. If $AB = 1_W$ then for all $w \in W$ we have $w = A(Bw)$, so A is surjective. If $BA = 1_V$ then if $Av_1 = Av_2$, we have

$$v_1 = 1_V v_1 = (BA)v_1 = B(Av_1) = B(Av_2) = (BA)v_2 = 1_V v_2 = v_2,$$

so A is injective. If A is invertible then it is both left and right invertible, so A is both injective and surjective, hence bijective. □

Remark 37. Correspondingly, differentiation is surjective (every continuous function is the derivative of its integral) and integration is injective (if a function has integral zero everywhere, it must be zero). On the other hand, differentiation is not injective (constant functions get sent to zero) and integration is not surjective (since $(\int g)(0) = \int_0^0 g(x)dx = 0$ by definition, $\int g$ can never be a nonzero constant function, or any function with $g(0) \neq 0$). ♢

What is more remarkable is that these are biconditionals. That a bijective linear map is invertible is a definition push:

Proposition 54. *If $A : V \to W$ is a linear map which is also a bijection, then A is invertible.*

Proof. Define $B : W \to V$ by

$$Bw = \text{the unique } v \in V \text{ such that } Av = w.$$

More briefly, $Bw \in V$ is the unique vector with the property that $A(Bw) = w$.

This is defined (and unambiguous) for all $w \in W$ by the assumption that A is bijective: there always exists a unique such v. By definition, we have $A(Bw) = w$, as $Bw \in V$ is the unique vecctor with $A(Bw) = w$. On the other hand, we have $B(Av) = v$, because $B(Av)$ is the unique vector u so that $Au = Av$. Because there is a unique such vector, and $u = v$ is such a vector, we may conclude $B(Av) = v$.

Thus B is indeed an inverse to A. (This is the usual definition of an inverse function to a bijection; this discussion has nothing to do with linearity. It is the definition we use to define $\sqrt{x}$ or $\ln x$, for instance: the latter is the unique real number for which $e^{\ln x} = x$.)

The only novelty is proving that B is linear. To see this, observe that $B(v + w)$ is characterized by the property that $A(B(v + w)) = v + w$. Now

$$A(Bv + Bw) = A(Bv) + A(Bw) = (AB)v + (AB)w = v + w$$

by linearity of A, so that $B(v+w) = Bv+Bw$. Similarly, $B(cv)$ is the unique vector for which $A(B(cv)) = cv$. But $A(cBv) = cA(Bv) = cv$ by linearity of A, so that cBv is such a vector, and hence $B(cv) = cB(v)$. □

Whereas the other two require we do a little work. The first is easier, as it amounts to understanding that we can define linear maps in terms of a basis.

Proposition 55. *If $A : V \to W$ is a surjective linear map, then A is right invertible.*

Proof. Choose a basis $\{w_1, \cdots, w_m\}$ for W. By definition of surjectivity, there exist vectors $v_i \in V$ so that $Av_i = w_i$. (There may be many such vectors; pick one for each w_i.) By Theorem 45, I can define a linear map $B : W \to V$ by saying what it does to each basis vector in a given basis. So let's define B to be the unique linear map for which $Bw_i = v_i$. Then

$$(AB)w_i = A(Bw_i) = Av_i = w_i$$

for all basis vectors w_i. By Remark 33, this implies $AB = 1_W$ is the identity map. □

The second is the first place you really see the power of our recent tools, in particular the basis extension lemma.

Proposition 56. *If $A : V \to W$ is an injective linear map, then A is left invertible.*

Proof. The idea is: split W up into $\mathrm{im}(A)$ and another subspace which complements it. Define the map backwards to undo A on $\mathrm{im}(A)$, and "crush the rest to zero"; it doesn't really matter what happens away from $\mathrm{im}(A)$.

Let $\{v_1, \cdots, v_n\}$ be a basis for V; it is in particular a linearly independent set. Because A is injective, Lemma 48 guarantees that $\{Av_1, \cdots, Av_n\}$ is a linearly independent set in W. By Lemma 38 (the basis extension lemma) we may extend this to a basis $\{Av_1, \cdots, Av_n, w_1, \cdots, w_m\}$ of W. By Theorem 45, I can define a linear map $B : W \to V$ by saying what it does to each basis vector in a given basis. So let's define B to be the unique linear map for which

$$B(Av_i) = v_i \text{ for all } 1 \leqslant i \leqslant n, \text{ while } Bw_i = \vec{0}.$$

Because $(BA)v_i = v_i$ for all basis vectors v_i, we have $BA = 1_V$ by Remark 33: since every vector v is a linear combination of the v_i, that $(BA)v = v$ for all v follows from linearity of (BA). □

These are perhaps somewhat convenient (I mainly list them to exhibit the idea of using basis extension), but nowhere near as powerful as the following much-vaunted and mysticized fact which represents the conclusion of the first part of an elementary course in linear algebra, the following characterization of linear maps between vector spaces of the same dimension (in a traditional linear algebra course, this would be phrased as a characterization of invertible matrices). This is a finite-dimensional phenomenon which is completely and totally false in infinite dimensions.

Corollary 57 (Invertible matrix theorem)**.** *Let $A : V \to W$ be a linear map between two finite-dimensional vector spaces* ***of the same dimension****. Then all of the following are equivalent:*

(i) A is invertible.

(ii) A is injective, aka $\ker A = \{\vec{0}\}$.

(iii) A is surjective.

Proof. If A is invertible, it is bijective, so both injective and surjective. We will show that if A is injective it is also surjective (whence bijective, whence invertible), and that if A is surjective it is also injective (whence bijective, whence invertible).

Suppose A is injective. Then $\mathrm{null}(A) = \dim \ker(A) = \dim\{\vec{0}\} = 0$. By the rank-nullity theorem, we have $\dim V = \mathrm{rank}(A) = \dim \mathrm{im}(A)$. Thus $\mathrm{im}(A) \subset W$ is a subspace of dimension $\dim \mathrm{im}(A) = \dim V = \dim W$. By Theorem 41 (a theorem which only applied to finite-dimensional vector spaces!), any subspace of W of the same dimension is all of W, so this implies $\mathrm{im}(A) = W$, so that A is in fact surjective.

Suppose now that A is surjective. Then $\mathrm{rank}(A) = \dim \mathrm{im}(A) = \dim W$. The rank-nullity theorem implies that $\mathrm{null}(A) + \mathrm{rank}(A) = \dim V$, so that $\mathrm{null}(A) = \dim V - \dim W = 0$. Thus $\ker A$ is a 0-dimensional vector space. The only 0-dimensional vector space is $\{\vec{0}\}$, so that $\ker A = \{\vec{0}\}$ and A is injective. □

5.4 Matrices

So far, we've studied the abstract theory of linear maps. We had a few examples (some from calculus, some from geometry, some from algebra) and we found some useful general principles:

- A linear map $A : V \to W$ is injective if and only if $\ker(A) = \{\vec{0}\}$, and when this is true $\dim V \leqslant \dim W$;
- A linear map $A : V \to W$ is surjective if and only if $\mathrm{im}(A) = W$, and when this is true $\dim V \geqslant \dim W$;
- We have $\mathrm{rank}(A) + \mathrm{null}(A) = \dim\ \mathrm{im}(A) + \dim \ker(A) = \dim V$;
- Knowing what A does to a particular basis for V uniquely determines A; we can define a linear map by delclaring where it sends each element of this basis, and we can check two linear maps are equal by checking they agree on all the basis elements.
- The most powerful result of all is that if $\dim V = \dim W$, then $A : V \to W$ is invertible if and only if $\ker(A) = \{\vec{0}\}$ if and only if $\mathrm{im}(A) = W$. (The version I use most often is the first biconditional, since in many cases one can compute whether or not the kernel is zero by hand without too much trouble.)

Which is all well and good, except that

- I only have a small handful of examples and no way to quickly produce examples to play around with.
- Because of this lack of examples, we don't have a good visual sense for what a linear map 'is'. We might have an algebraic sense, because we've been working with the algebra of linear maps — but with very few pictures.
- It seems very difficult in practice to take a list of many vectors (say, ten vectors in $\mathbb{F}^{12}$) and determine whether one of the later entries in that list is redundant, so it seems hard to compute dimensions of the relevant subspaces. Similarly, while in many cases $\ker(A)$ is not too hard to compute, it would be nice to have an *algorithmic way* to do so.

In the next section, we will use the language of *matrices* to see a huge number of examples, and use this to help us vizualize linear maps better than we have been able to thusfar. Next week, we will cover the *Gauss–Jordan algorithm* which allows one to efficiently compute kernels and images.

Before then, I ought to tell you what a matrix is.

Definition 29. We say that an $m \times n$ **matrix over the field** $\mathbb{F}$ is an array M of elements of $\mathbb{F}$ with m rows and n columns, whose entries are elements of $\mathbb{F}$. We depict this as

$$M = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix},$$

where $a_{ij} \in \mathbb{F}$ is the entry in row i, column j of the matrix. If we want to write M in a compact way while emphasizing the names of its entries, we will write $M = (a_{ij})_{\substack{1 \leqslant i \leqslant m \\ 1 \leqslant j \leqslant n}}$, or when m and n are also clear from context, merely $M = (a_{ij})$. Occasionally we will simply write M_{ij} for the entry in row i and column j. If $m = n$ we say that M is a **square matrix**. We say two matrices M, M' are equal if they have the same shape/size ($m \times n$) and all of the entries are equal $M_{ij} = M'_{ij}$ for all i, j. ◊

Example 66.

$$M = \begin{pmatrix} 1 & 2 & 4 & 11 \\ 3 & 2 & -1 & \pi \end{pmatrix}$$

is a 2×4 matrix over $\mathbb{R}$, with $a_{14} = 11$ and $a_{23} = -1$.

$$M = \begin{pmatrix} 1 & 3 & 4 \\ 1/3 & -4/5 & 3/11 \\ 2 & 2 & 2 \\ 1 & 0 & -1/13 \end{pmatrix}$$

is a 4×3 matrix over $\mathbb{Q}$, where $a_{21} = 1/3$ and $a_{42} = 0$. ◊

Remark 38. To remember which is m and which is n, always think: "Rows before columns." Similarly, when remembering how to write the subscripts on a_{ij}, the first term is the row a_{ij} is in, and the second term is the column a_{ij} is in. This mnemonic will come up multiple times. This is one of the few things I think one should memorize. ◊

Definition 30. Suppose

$$M = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

is an $m \times n$ matrix over $\mathbb{F}$. There is an associated linear map $A_M : \mathbb{F}^n \to \mathbb{F}^m$ defined by

$$A_M \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \cdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{pmatrix}.$$

◊

Example 67.

$$M = \begin{pmatrix} 1 & 2 & 4 & 11 \\ 3 & 2 & -1 & \pi \end{pmatrix}$$

corresponds to the linear map $A_M : \mathbb{R}^4 \to \mathbb{R}^2$ given by

$$A_M \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} x + 2y + 4z + 11w \\ 3x + 2y - z + \pi w \end{pmatrix}.$$

$$M = \begin{pmatrix} 1 & 3 & 4 \\ 1/3 & -4/5 & 3/11 \\ 2 & 2 & 2 \\ 1 & 0 & -1/13 \end{pmatrix}$$

corresponds to the linear map $A_M : \mathbb{Q}^3 \to \mathbb{Q}^4$ defined by

$$A_M \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x + 3y + 4z \\ x/3 - 4y/5 + 3z/11 \\ 2x + 2y + 2z \\ x - z/13 \end{pmatrix}. \quad ◊$$

It is an elementary computation that A_M is a linear map. I choose to denote M and A_M by different names to distinguish between two concepts which are distinct (but ultimately turn out to be equivalent): one is a box with numbers in it; the other is a function which takes elements of $\mathbb{F}^n$ to elements of $\mathbb{F}^m$. These are simply not the same thing, *even if I can use one to obtain the other*. This is comparable to the fact that a function is not the same concept as its graph, even though I can recover the function from the graph and vice versa.

Remark 39. Notice that an $m \times n$ matrix corresponds to a linear map $\mathbb{F}^n \to \mathbb{F}^m$ (m and n seem to "flip"). Watch out for this! It is a persistent source of confusion. As we will see, the columns correspond to certain vectors in the codomain, so if the columns have length m (that is, if there are m rows) then the codomain ought to be $\mathbb{F}^m$. Similarly, the rows correspond to components of the output, so if there are m rows the output lands in $\mathbb{F}^m$. ◊

Let me try to give a sense of what these linear maps A_M actually *mean*. There are three useful perspectives (the third being a computation of the previous two).

5.4.1 The column perspective

This perspective is most like what we have discussed so far. We know from Theorem 45 that a linear map is uniquely determined by knowing what it does to a basis. Because $\mathbb{F}^n$ comes with a canonical / standard basis, this suggests to me that to get a sense for what A_M does, we should compute $A_M e_j$ for the different basis vectors e_j.

Lemma 58. *Let M be the $m \times n$ matrix $M = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$. Then $A_M e_j$ is the j'th column of the matrix M. That is,*

$$A_M e_j = \begin{pmatrix} a_{1j} \\ \cdots \\ a_{mj} \end{pmatrix}.$$

Proof. This is a direct computation using the definition of A_M. If I write a vector x as $x = \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix}$, then e_j is the vector for which $x_j = 1$ and $x_i = 0$ for $i \neq j$. Because

$$A_M \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \cdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{pmatrix},$$

substituting $x_j = 1$ and $x_i = 0$ otherwise, this expression simplifies to

$$A_M e_j = \begin{pmatrix} a_{1j} \\ \cdots \\ a_{mj} \end{pmatrix},$$

as claimed: this is the j'th column of M. □

So for the matrix $M = \begin{pmatrix} 1 & 3 & 4 \\ 1/3 & -4/5 & 3/11 \\ 2 & 2 & 2 \\ 1 & 0 & -1/13 \end{pmatrix}$, we have

$$Me_1 = \begin{pmatrix} 1 \\ 1/3 \\ 2 \\ 1 \end{pmatrix}, \quad Me_2 = \begin{pmatrix} 3 \\ -4/5 \\ 2 \\ 0 \end{pmatrix}, \quad Me_3 = \begin{pmatrix} 4 \\ 3/11 \\ 2 \\ -1/13 \end{pmatrix}.$$

Here's how this suggests I think about the map $A_M : \mathbb{F}^n \to \mathbb{F}^m$. Let us write $v_j = A_M e_i$. Some authors will draw the picture

$$M = \begin{pmatrix} | & \cdots & | \\ v_1 & \cdots & v_n \\ | & \cdots & | \end{pmatrix},$$

where the vertical bars $|$ are meant to indicate that the term v_1 (and so on) constitutes the whole column. If $x = \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix}$ is a general vector, then

$$A_M x = A_M(x_1 e_1 + \cdots + x_n e_n) = x_1 A_M e_1 + \cdots + x_n A_M e_n = x_1 v_1 + \cdots + x_n v_n.$$

That is, $A_M x$ is a linear combination of the columns of A, where the 'weights' in the linear combination are provided by the entries of the vector x. (Thus A_M is an instance of Example 53, applied to $V = \mathbb{F}^m$; the columns of M are the vectors used in that construction.) This gives us a few upshots; the first is immediate from the discussion of Example 59.

Corollary 59. *If M is an $m \times n$ matrix with columns $v_1, \cdots, v_n$, the associated linear map $A_M : \mathbb{F}^n \to \mathbb{F}^m$ has $\mathrm{im}(A_M) = \mathrm{span}(v_1, \cdots, v_n)$ equal to the span of the columns and $\ker(A_M) = \mathrm{Rel}(v_1, \cdots, v_n)$ is the set of linear relations between the columns of M.*

As for the second, it will be a quick application of Theorem 45.

Corollary 60. *If $A : \mathbb{F}^n \to \mathbb{F}^m$ is any linear map, then there exists a unique $m \times n$ matrix M so that $A = A_M$.*

Proof. Uniqueness is easy. If $A_M = A = A_N$, then in particular $A_M e_j = A_N e_j$ for all $1 \leqslant j \leqslant n$. It follows that the columns of M and the columns of N are precisely the same, and hence that M and N are equal (they have the same shape and the same entries); so if $A = A_M$ for some matrix M, then there is a unique such matrix M.

If I claim there exists such a matrix M, first I have to give you a matrix, then I have to prove $A = A_M$. **Given A, define the matrix M to be the matrix whose j'th column is Ae_j.** That is, let

$$M = \begin{pmatrix} | & \cdots & | \\ Ae_1 & \cdots & Ae_n \\ | & \cdots & | \end{pmatrix}.$$

Because $A_M e_j$ is the j'th column of M, and the j'th column of M is Ae_j by definition, we have $A_M e_j = Ae_j$ for all $1 \leqslant j \leqslant m$. Because $\{e_1, \cdots, e_n\}$ is a basis for $\mathbb{F}^n$, by Theorem 45 we have $Av = A_M v$ for all $v \in \mathbb{F}^n$; that is, $A = A_M$. □

Example 68. The identity map $1_{\mathbb{F}^n} : \mathbb{F}^n \to \mathbb{F}^n$ corresponds to the $n \times n$ matrix

$$I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

with 1's down the diagonal and 0's elsewhere. This is because $1_{\mathbb{F}^n} e_i = e_i$ for all i, so the corresponding matrix has i'th column equal to e_i. This means in the i'th column, the only nonzero entry is in the i'th row, so $a_{ii} = 1$ for all i and $a_{ij} = 0$ for $i \neq j$. This matrix is called the "identity matrix of size n", or when n is clear from context, "the identity matrix". You can check by direct computation using the definition of A_M that $A_{I_n} v = v$ is the identity for all $v \in V$, if you want. ♢

Example 69. The zero map $0_{\mathbb{F}^n \to \mathbb{F}^m} : \mathbb{F}^n \to \mathbb{F}^m$ has $0_{\mathbb{F}^n \to \mathbb{F}^m} v = \vec{0}$ for all $v \in \mathbb{F}^n$. In particular, $0_{\mathbb{F}^n \to \mathbb{F}^m} e_j = \vec{0}$ for all $1 \leqslant j \leqslant m$. It follows that the corresponding matrix is

$$0_{m \times n} = \begin{pmatrix} 0 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & 0 \end{pmatrix}$$

the all-zeroes matrix, because each column should be the zero vector. ♢

Exercise. If you have finished HW5 #3, write down the 2×2 matrix corresponding to the linear map $\mathrm{rot}_\theta : \mathbb{R}^2 \to \mathbb{R}^2$.

5.4.2 The row perspective

Let's focus first on the case of $1 \times n$ matrices, sometimes called *row vectors*. If M is a $1 \times n$ matrix,

$$M = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix},$$

then the corresponding linear map $A_M : \mathbb{F}^n \to \mathbb{F}$ is defined by

$$A_M \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = a_1 x_1 + \cdots + a_n x_n,$$

and we often abbreviate this to

$$\begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = a_1x_1 + \cdots + a_nx_n.$$

This is pure notation — I have an array M, I have a corresponding linear map A_M, and if I literally write the matrix M in place of A_M I see what I've depicted above: a row vector, placed next to a column vector of the same length, and wrote that the "result" was "equal to" $a_1x_1 + \cdots + a_nx_n$.

However, inspired by this notation, I'll define an actual "product-like" operation between rows and columns: a way to 'pair' a row vector and a column vector to get a number, by pairing up terms, multiplying them together, and adding them all up.

Definition 31. Suppose $\begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix}$ is a row vector corresponding to the linear function $f : \mathbb{F}^n \to \mathbb{F}$, and $\begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix}$ is a column vector corresponding to the vector $v \in \mathbb{F}^n$. We say that the output $f(v)$, explicitly given by

$$\begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = a_1x_1 + \cdots + a_nx_n,$$

is the **pairing** of the row vector and the column vector. ◊

Remark 40. People often visualize this operation by thinking of rotating the vector $\begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix}$ to be vertical, multiplying the matching entries a_i and x_i, and adding the result up. Try visualizing this by standing up the row vector in the following formula, and then doing the multiplications and adding them up:

$$\begin{pmatrix} 1 & 3 & -2 \end{pmatrix} \begin{pmatrix} 3 \\ 5 \\ 33 \end{pmatrix} = 1 \cdot 3 + 3 \cdot 5 + (-2) \cdot 33 = 3 + 15 - 66 = -48.$$

◊

Philosophy: A column vector is just a vector in $\mathbb{F}^n$ in the usual sense. A row vector, however, is a linear function $f : \mathbb{F}^n \to \mathbb{F}$. Its entries are the outputs $f(e_j)$ for each j. This is a fundamentally different creature than a column vector. Because functions eat vectors and spit out numbers, you can "multiply" a row vector by a column vector and get a number. **ROW BEFORE COLUMN:** just as $f(v)$ makes sense but $v(f)$ does not, when we talk abou this pairing operation, *the first vector is a row vector, the second is a column vector.* The other order will mean something completely different.

We can move on from this to an understanding of matrices in general. Before that, let me point out that just as there are special vectors $e_1, \cdots, e_n \in \mathbb{F}^n$, there are also special linear functions $p_1, \cdots, p_n : \mathbb{F}^n \to \mathbb{F}$. Precisely, define

$$p_i \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = x_i.$$

The map p_i is called *projection to the i'th coordinate*; it forgets about all information except for the information contained in the i'th coordinate itself. If I write it as a row vector, we have

$$p_i = \begin{pmatrix} 0 & \cdots & 1 & \cdots & 0 \end{pmatrix}, \qquad \text{where } a_i = 1 \text{ but } a_j = 0 \text{ for } j \neq i.$$

As a matrix, p_i is like the basis vector e_i but 'laid flat on the ground'. But it should not be confused for e_i. The column vector $e_i \in \mathbb{F}^n$ is honestly a vector. The column vector p_i does not represent a vector. It represents a function $p_i : \mathbb{F}^n \to \mathbb{F}$, the function which gives the i'th entry of a vector.

Lemma 61. *Suppose $A_M : \mathbb{F}^n \to \mathbb{F}^m$ is the linear map associated to an $m \times n$ matrix M. Then for $1 \leqslant i \leqslant m$, the i'th row of M is the row vector corresponding to the linear map $p_iA_M : \mathbb{F}^n \to \mathbb{F}$.*

Proof. Explicitly, we have

$$p_i A_M \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = p_i \begin{pmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \cdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{pmatrix} = a_{i1}x_1 + \cdots + a_{in}x_n.$$

But this is precisely the linear function $\mathbb{F}^n \to \mathbb{F}$ with row vector

$$\begin{pmatrix} a_{i1} & \cdots & a_{in} \end{pmatrix},$$

the i'th row of M. □

If the rows of M represent linear functions $f_1, \cdots, f_m : \mathbb{F}^n \to \mathbb{F}$, it will occasionally be useful to write

$$M = \begin{pmatrix} — & f_1 & — \\ \cdots & \cdots & \cdots \\ — & f_m & — \end{pmatrix}.$$

When this is the case, the i'th component of $A_M v$ is precisely $f_i(v)$; that is to say, $f_i(v) = p_i A_M v$.

5.4.3 The entry perspective

We have a perspective on matrices in terms of their columns, and a perspective on matrices in terms of their rows. Combining these, we can explain the meaning of the particular entries in a matrix.

Lemma 62. *If $M = (a_{ij})$ is an $m \times n$ matrix and $A_M : \mathbb{F}^n \to \mathbb{F}^m$ is the associated linear map, the entries a_{ij} are equal to $p_i A_M e_j \in \mathbb{F}$; that is, the entry a_{ij} is the i'th component of $A_M e_j$.*

Proof. Because $A_M e_j$ is the j'th column of M, and $p_i v$ is the i'th component of the vector v, we see that $p_i(A_M e_j)$ is the i'th component of $A_M e_j$. □

Example 70. As an example, take $M = \begin{pmatrix} 1 & 2 & 4 & 11 \\ 3 & 2 & -1 & \pi \end{pmatrix}$ with corresponding linear map $A_M : \mathbb{R}^4 \to \mathbb{R}^2$ given by

$$A_M \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} x + 2y + 4z + 11w \\ 3x + 2y - z + \pi w \end{pmatrix}.$$

Then $a_{23} = -1$ and indeed $A_M e_3 = \begin{pmatrix} 4 \\ -1 \end{pmatrix}$, whose second component is -1, so $p_2 A_M e_3 = -1 = a_{23}$. ◊

This perspective is, in my opinion, usually not very helpful (it's hard to see any useful information from such little data as a particular entry of a particular vector); most of the time I think about matrices, I'm thinking in either the row or column perspective (whichever is more useful at the moment). However, I have one specific application in mind.

5.4.4 Composition and matrix multiplication

Suppose I have an $(m \times n)$ matrix N and an $(\ell \times m)$ matrix M. These correspond to linear maps $A_N : \mathbb{F}^n \to \mathbb{F}^m$ and $A_M : \mathbb{F}^m \to \mathbb{F}^\ell$. We know the composite of linear maps is also a linear map, so their composite defines a linear map

$$A_M A_N : \mathbb{F}^n \to \mathbb{F}^\ell.$$

We know every linear map $\mathbb{F}^n \to \mathbb{F}^\ell$ is associated to some $\ell \times n$ matrix — what matrix is associated to the composition $A_M A_N$? Can I understand it entirely in terms of the matrices themselves?

Yes.

Definition 32. Suppose M and N are $\ell \times m$ and $m \times n$ matrices

$$M = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \cdots & \cdots & \cdots \\ a_{\ell 1} & \cdots & a_{\ell m} \end{pmatrix}, \qquad N = \begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \cdots & \cdots & \cdots \\ b_{m1} & \cdots & b_{mn} \end{pmatrix}.$$

Their **matrix product** $M * N$ is the $\ell \times n$ matrix whose entries $(M * N)_{ij}$ are obtained by pairing the i'th row of M to the j'th column of N. That is,

$$(M * N)_{ij} = \begin{pmatrix} a_{i1} & \cdots & a_{im} \end{pmatrix} \begin{pmatrix} b_{1j} \\ \cdots \\ b_{mj} \end{pmatrix} = a_{i1}b_{1j} + \cdots + a_{im}b_{mj}.$$

◊

Example 71. Let $M = \begin{pmatrix} 1 & 3 & 1 \\ 2 & 1 & 0 \end{pmatrix}$ and let $N = \begin{pmatrix} 1 & 5 \\ 2 & 4 \\ 3 & 7 \end{pmatrix}$. Their matrix product is

$$M * N = \begin{pmatrix} \begin{pmatrix} 1 & 3 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} & \begin{pmatrix} 1 & 3 & 1 \end{pmatrix} \begin{pmatrix} 5 \\ 4 \\ 7 \end{pmatrix} \\ \begin{pmatrix} 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} & \begin{pmatrix} 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 5 \\ 4 \\ 7 \end{pmatrix} \end{pmatrix}$$
$$= \begin{pmatrix} 1 \cdot 1 + 3 \cdot 2 + 1 \cdot 3 & 1 \cdot 5 + 3 \cdot 4 + 1 \cdot 7 \\ 2 \cdot 1 + 1 \cdot 2 + 0 \cdot 3 & 2 \cdot 5 + 1 \cdot 4 + 0 \cdot 7 \end{pmatrix} = \begin{pmatrix} 10 & 24 \\ 4 & 14 \end{pmatrix}.$$

For instance, the entry in row 2, column 1 of $M * N$ is equal to the result of pairing row 2 of M with column 1 of N.

Notice that while $M * N$ is a 2×2 matrix, on the other hand $N * M$ is a 3×3 matrix! Matrix multiplication is not commutative. In fact, even if $M * N$ makes sense, there is no reason for $N * M$ to: if M is 2×4 and N is 4×3, then $M * N$ is a 2×3 matrix, but $N * M$ is not defined. ◊

Remark 41. If

$$M = \begin{pmatrix} — & f_1 & — \\ \cdots & \cdots & \cdots \\ — & f_\ell & — \end{pmatrix}, \quad \text{and} \quad N = \begin{pmatrix} | & \cdots & | \\ v_1 & \cdots & v_n \\ | & \cdots & | \end{pmatrix},$$

then

$$M * N = \begin{pmatrix} f_1 v_1 & \cdots & f_1 v_n \\ \cdots & \cdots & \cdots \\ f_\ell v_1 & \cdots & f_\ell v_n \end{pmatrix}.$$

If $f_i = \begin{pmatrix} a_{i1} & \cdots & a_{im} \end{pmatrix}$ and $v_j = \begin{pmatrix} b_{1j} \\ \cdots \\ b_{mj} \end{pmatrix}$, then this is just restating the definition

$$f_i v_j = \begin{pmatrix} a_{i1} & \cdots & a_{im} \end{pmatrix} \begin{pmatrix} b_{1j} \\ \cdots \\ b_{mj} \end{pmatrix} = a_{i1}b_{1j} + \cdots + a_{im}b_{mj} = (M * N)_{ij}.$$

◊

This is relevant because, as might be clear from context, the matrix product is the matrix representing the composite of two linear maps.

Theorem 63 (Matrix multiplication is composition)**.** *If M is an $\ell \times m$ matrix and N is an $m \times n$ matrix, then we have $A_{M*N} = A_M A_N$. That is, the matrix associated to the composite of A_M and A_N is precisely the matrix product $M * N$.*

Exercise. Verify this theorem for the matrices given in the preceding example. That is, compute the composite of $A_M A_N$, and check that it is indeed A_{M*N} for the matrix $M * N$ we computed in the example.

Proof of Theorem 63. We have an excuse to think about the entry perspective. We want to understand the entries in the matrix associated to $A_M A_N$. By Lemma 62, the entry in row i and column j of the associated matrix is $p_i A_M A_N e_j \in \mathbb{F}$. Now $p_i A_M : \mathbb{F}^m \to \mathbb{F}$ is the function represented by row i of M, so $p_i A_M$ has associated matrix $\begin{pmatrix} a_{i1} & \cdots & a_{im} \end{pmatrix}$. On the other hand, $A_N e_j$ is the vector given by the j'th column of N, explicitly

$$A_N e_j = \begin{pmatrix} b_{1j} \\ \cdots \\ b_{mj} \end{pmatrix}.$$

The quantity $(p_i A_M)(A_N e_j)$ is the result of plugging in the vector $A_N e_j$ into the function $p_i A_M$, but this gives precisely

$$(p_i A_M)(A_N e_j) = \begin{pmatrix} a_{i1} & \cdots & a_{im} \end{pmatrix} \begin{pmatrix} b_{1j} \\ \cdots \\ b_{mj} \end{pmatrix} = a_{i1} b_{1j} + \cdots + a_{im} b_{mj} = (M * N)_{ij}.$$

□

It follows that if you're interested in studying linear maps and their compositions, you can largely get away with studying matrices and their matrix products instead. So far we have discussed how to associate matrices to linear maps $\mathbb{F}^n \to \mathbb{F}^m$. Next week we will explain how linear maps $V \to W$ *together with a choice of basis for each of* V *and* W give rise to matrices, too; this will allow us to go back and forth between the language of linear maps and bases, and the language of matrices.

All this said and done, in the future if I have an $m \times n$ matrix $M = (a_{ij})$ and a vector $v \in \mathbb{F}^n$, I will often write $A_M v$ as

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11} x_1 + \cdots + a_{1n} x_N \\ \cdots \\ a_{m1} x_1 + \cdots + a_{mn} x_n \end{pmatrix},$$

writing everything in matrix notation; notice that this is precisely the matrix product of the $m \times n$ matrix M and the $n \times 1$ matrix v. If matrix notation is convenient, I will work entirely in matrix notation. If it is important to emphasize the distinction between M and the associated linear map (and this will be relevant when we move on to linear maps between vector spaces *other than* $\mathbb{F}^n, \mathbb{F}^m$), I will do so, referring to the matrix as M and the associated linear map as A_M.

5.5 Examples of linear maps and their matrix representatives

In this section I'm going to focus on linear maps $A : \mathbb{R}^2 \to \mathbb{R}^2$. We can identify such maps with 2×2 matrices with real entries $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, and I claim that these are very visualizable. Before getting into the visualizations, let me point out that you already determined in your homework when these matrices are invertible:

Lemma 64. *If* $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, *then* A_M *is invertible if and only if* $ad - bc \neq 0$.

Proof. You checked in your homework that $\left\{ \begin{pmatrix} a \\ c \end{pmatrix}, \begin{pmatrix} b \\ d \end{pmatrix} \right\}$ is a basis for $\mathbb{R}^2$ if and only if $ad - cb \neq 0$. Thus the columns of M are a basis for $\mathbb{R}^2$ if and only if $ad - bc \neq 0$.

Now $\ker(A_M)$ can be identified with the set of linear relations between the columns of M, so A_M is injective if and only if the columns of M are linearly independent. Similarly, $\mathrm{im}(A_M)$ can be identified with the span of the columns of M, so A_M is surjective if and only if the columns of M span $\mathbb{R}^2$. It follows that the columns of M form a basis for $\mathbb{R}^2$ if and only if A_M is bijective, hence invertible. □

In the next few sections, I'm going to give some examples of 2×2 matrices, discuss how to visualize them, and discuss how this generalizes to higher dimensions. Here I will not make a big effort to distinguish between matrices and the corresponding linear maps; the conceptual difference will only become very important next week.

5.5.1 Scaling maps: diagonal matrices

The simplest linear map $\mathbb{R}^2 \to \mathbb{R}^2$ is the identity map $1_{\mathbb{R}^2}(v) = v$ for all $v \in \mathbb{R}^2$. The matrix representing this linear map is $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

The next simplest notion is that of an *diagonal matrix*. Write

$$D_{\lambda_1,\lambda_2} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}.$$

The associated linear map is

$$D_{\lambda_1,\lambda_2} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \lambda_1 x \\ \lambda_2 y \end{pmatrix}.$$

This map stretches the two coordinate axes; we have

$$D_{\lambda_1,\lambda_2} e_1 = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \lambda_1 \\ 0 \end{pmatrix} = \lambda_1 e_1,$$

$$D_{\lambda_1,\lambda_2} e_2 = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ \lambda_2 \end{pmatrix} = \lambda_2 e_2.$$

Visualization exercise. Click here. This applet displays a before-and-after picture of applying the linear transformation D_{l_1,l_2}; it shows a "before" grid in grey, as well as an "after" grid in green. It also depicts Ae_1, Ae_2 as blue and red line segments.

Play around with it.

- Do you see that l_1, l_2 correspond to how much the axes are stretched (or, if you like, how much a square is stretched in each direction)?
- What happens as you change l_1 and l_2? What happens to the image when you set one (or both) of these quantities equal to zero? What happens when l_1 changes from positive to negative?
- What is this map doing to the x-axis when l_1 is negative?

This notion makes sense in any dimension (and over any field). If $\lambda_1, \cdots, \lambda_n \in \mathbb{F}$, the **diagonal matrix** with entries $\lambda_1, \cdots, \lambda_n$ is the $n \times n$ matrix

$$D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix},$$

which is only nonzero along the diagonal, where its entries are $\lambda_1, \cdots, \lambda_n$. This matrix is determined by the property that $De_i = \lambda_i e_i$ for all $1 \leqslant i \leqslant n$.

Note that $\mathrm{null}(D)$ is the number of λ's which are equal to zero, as $\ker(D)$ is spanned by those e_i for which $\lambda_i = 0$. Correspondingly, $\mathrm{rank}(D)$ is the number of nonzero λ's. The map $D : \mathbb{F}^n \to \mathbb{F}^n$ is invertible if and only if **all the diagonal entries are nonzero,** in which case its inverse is

$$D^{-1} = \begin{pmatrix} 1/\lambda_1 & 0 & \cdots & 0 \\ 0 & 1/\lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1/\lambda_n \end{pmatrix}.$$

Just as I vizualize a 2×2 diagonal matrix as stretching the two axes in $\mathbb{R}^2$, sending the unit square to an appropriate rectangle, I can visualize a 3×3 diagonal matrix as stretching the three different axes in $\mathbb{R}^3$ by different amounts, sending a unit cube to some box with side-lengths $\lambda_1, \lambda_2, \lambda_3$.

5.5.2 Shearing maps: strictly upper-triangular matrices

The next special class of transformations I would like to introduce are called *shearing maps*. In 2D, these are especially simple (and a shearing transformation in n dimensions can, in a sense, be built out of 2D shearing maps).

Before going into a visual discussion of the special case, let me do a bit of analysis in general.

Definition 33. A **upper-triangular** $n \times n$ matrix is a matrix of the form

$$T = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix},$$

which has zeroes below the main diagonal.

We say that T is **strictly upper-triangular** if $a_{11} = a_{22} = \cdots = a_{nn} = 0$, whereas we say that M is **upper unitriangular** if $a_{11} = \cdots = a_{nn} = 1$. ♢

Remark 42. At the level of linear maps, an upper-triangular matrix is one so that for all i, the associated linear map has $A_M e_i \in \text{span}(e_1, \cdots, e_i)$. ♢

You will prove the following on your homework.

Lemma 65. *An upper-triangular matrix is invertible if and only if the diagonal entries $a_{11}, \cdots, a_{nn}$ are all nonzero.*

In general, the formula for the inverse is not very clean, but it is very easy to compute algorithmically. I will ask you to do this in some simple cases.

We will focus on the unitriangular matrices, because a general invertible upper-triangular matrix can be obtained from these by scaling:

Lemma 66. *Let T be an invertible upper-triangular matrix. Then there exists a unique diagonal matrix D and upper unitriangular matrix U so that $T = DU$.*

Proof. If $U = \begin{pmatrix} 1 & b_{12} & \cdots & b_{1n} \\ 0 & 1 & \cdots & b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$ and $D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$, then their matrix product is

$$DU = \begin{pmatrix} \lambda_1 & \lambda_1 b_{12} & \cdots & \lambda_1 b_{1n} \\ 0 & \lambda_2 & \cdots & \lambda_2 b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}.$$

If $T = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix}$ is an upper-triangular matrix which can be written as $T = DU$ for some diagonal matrix D and some upper unitriangular matrix U, it follows by comparing the definition of T to the formula above the entries of D are $\lambda_i = a_{ii}$. By assumption that T is invertible and Lemma 65, we see that λ_i are all nonzero, hence invertible in $\mathbb{F}$. Because the entries of T are related to the entries of U by $\lambda_i b_{ij} = a_{ij}$, we must have $b_{ij} = a_{ij}/\lambda_i$.

Thus T determines D and U. Conversely, for the matrices D and U determined above, we have $T = DU$. This proves existence and uniqueness of the claimed decomposition:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} 1 & a_{12}/\lambda_1 & \cdots & a_{1n}/\lambda_1 \\ 0 & 1 & \cdots & a_{2n}/\lambda_2 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

□

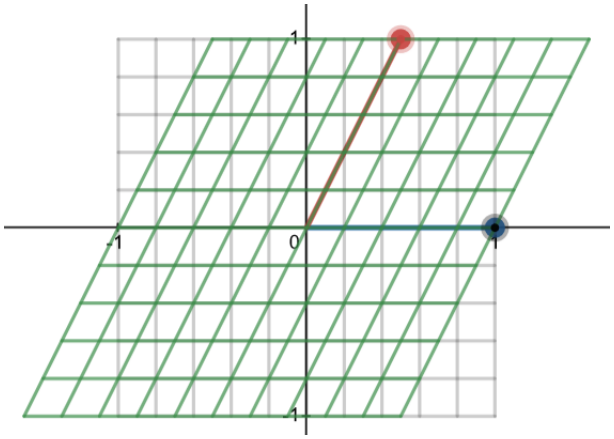
Now let's try to understand what these transformations *do*. In the simplest non-trivial case of 2×2 matrices, an upper unitriangular matrix takes the form $S_t = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$ for some $t \in \mathbb{F}$. The associated linear map has

$$S_t e_1 = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = e_1,$$

while

$$S_t e_2 = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} t \\ 1 \end{pmatrix} = te_1 + e_2.$$

If I draw a picture of this for $\mathbb{F} = \mathbb{R}^2$, I see that when applying S_t, the x-axis is left unchanged but the y-axis 'tilts', with the slope of the tilted line being $1/t$ (the line $x = 0$ is sent to the line $x = ty$).



Here is a before-and-after picture for the transformation $S_{1/2}$. The grey lines are the 'before' lines. Notice that the horizontal lines are still sent to horizontal lines (though they're shifted right or left, depending on the height of the horizontal line). However, the vertical lines are all tilted down. This is called an *shearing transformation*, for the following reason. Imagine that the plane consists of flexible material, like fabric. Put one of your hands above the x-axis, and one below the x-axis. Move your upper hand to the right, and your bottom hand to the left. Then the top part of the plane will get pushed right, and the bottom left, like in the picture above.

In engineering, one way to bend objects is to exert force on them in two different directions; this process is called shearing, whence the name of the linear transformation above.

Visualization exercise. Click here. This applet displays a before-and-after picture of applying the linear transformation S_t (though here t is denoted as a); it shows a "before" grid in grey, as well as an "after" grid in green. It also depicts Ae_1, Ae_2 as blue and red line segments.

Play around with it. Notice, in particular, that the map is always invertible.

If I want to understand shearing in higher dimensions, I have to work harder. I like to think of these transformations as happening step-by-step. For instance,

$$\begin{pmatrix} 1 & a_{12} & a_{13} \\ 0 & 1 & a_{23} \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & a_{13} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & a_{23} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & a_{12} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Notice that the three matrices on the right behave in a simpler way. For instance, for the last of these matrices (the first to get applied to a vector), we have

$$\begin{pmatrix} 1 & a_{12} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x + a_{12}y \\ y \\ z \end{pmatrix}.$$

This means that the z coordinate is totally untouched! This transformation only affects the x and y coordinates. For the other two transformations, one affects the x and z coordinates, and the last affects the y and z coordinates.

Visualization exercise. Click here. This applet shows a before-and-after picture of applying a **3D matrix**. The image of e_1 is depicted as a red vector; the image of e_2 is depicted as a green vector; the image of e_3 is depicted as a blue vector. You can move the camera around 3D space by clicking and dragging the screen.

Adjust the entries a_{12}, a_{13}, a_{23}, one at a time and a small amount at a time (I suggest small numbers like $a_{12} = 0.3$ and seeing what changes as you change these by small amounts).

- Do you see what happens as you make just a slight change to one of these three entries?
- What happens, in particular, as you change a_{12}? How about a_{13} and a_{23}?

5.5.3 Rotations and reflections

The final piece of today's puzzle is different than the previous parts.

- This discussion is special to $\mathbb{R}$. One can make sense of some of the relevant algebraic properties more generally, but they are most useful for $\mathbb{R}$, and most geometrically meaningful there.
- While you can make sense of the discussion in higher dimensions — and I hope to towards the end of the term — it is necessarily much more intricate. In particular, it is much harder to write down an explicit description of the general case of a 'rotation' or a 'reflection' than it is to write down the general notion of a shearing map (a unitriangular matrix).

In your homework, you will find the formula $\mathrm{rot}_\theta \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x\cos(\theta) - y\sin(\theta) \\ x\sin(\theta) + y\cos(\theta) \end{pmatrix}$. This means that rot_θ corresponds to the matrix $R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$.

Visualization exercise. Click here. This applet displays a before-and-after picture of applying the linear transformation R_θ (though here θ is denoted as b); it shows a "before" grid in grey, as well as an "after" grid in green. It also depicts Ae_1, Ae_2 as blue and red line segments.

Play around with θ; watch what happens to the picture as you increase or decrease it, and what happens to the picture as you increase θ from 0 to 2π.

Another interesting family of matrices is instead the *reflection matrices*. Given a line $L \subset \mathbb{R}^2$ passing through the origin, write $0 \leqslant \psi < \pi$ for the angle it makes with the positive x-axis. Reflection across this line is a linear transformation, and one can compute that the corresponding matrix is

$$\mathrm{ref}_\psi = \begin{pmatrix} \cos(2\psi) & \sin(2\psi) \\ \sin(2\psi) & -\cos(2\psi) \end{pmatrix}.$$

Visualization exercise. Click here. This applet displays a before-and-after picture of applying the linear transformation ref_ψ (though here ψ is denoted as b); it shows a "before" grid in grey, as well as an "after" grid in green. It also depicts Ae_1, Ae_2 as blue and red line segments.

- What happens as you increase ψ?
- The maps rot_θ and $\mathrm{ref}_{\theta/2}$ are not the same, but the output of the 'grids' looks the same. To tell the difference, pull both apps up and compare them (with $b = 1$ in the first app and $b = 1/2$ in the second). You should notice that Ae_1 is the same in both cases, but Ae_2 is its opposite.
- Observe that the red axis Ae_2 is always counter-clockwise from Ae_1 in the first picture, but always clockwise from Ae_2 in the second picture. Later this will be relevant to our notion of *orientation* and our analysis of the idea of determinants.

5.5.4 Invertible 2×2 matrices

We have now defined three families of 2×2 matrices.

- There are the diagonal matrices $D_{\lambda_1,\lambda_2} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$, which correspond to the linear map $\mathbb{R}^2 \to \mathbb{R}^2$ which scales the first coordinate by λ_1 and the second coordinate by λ_2. These are invertible if and only if λ_1, λ_2 are both nonzero.
- There are the shearing matrices $S_t = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$, which shear the plane by a factor of t (so the vertical line $x = 0$ is sent to the line $x = ty$). These are invertible regardless of what t is.
- There are the rotation matrices
$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$
corresponding to the linear map which rotates the plane θ radians counter-clockwise around the origin.

It turns out that every linear map $\mathbb{R}^2 \to \mathbb{R}^2$ can be made out of these, and (so long as you state this carefully) in an essentially unique way. I will prove this for the invertible matrices; you will handle rank-1 matrices on your homework.

Proposition 67. *Suppose $A : \mathbb{R}^2 \to \mathbb{R}^2$ is an invertible linear map. There is a unique choice of*
$$\theta \in [0, 2\pi), \quad \lambda_1 \in (0, \infty) \subset \mathbb{R}, \quad \lambda_2 \in \mathbb{R}\backslash\{0\}, \quad t \in \mathbb{R}$$
so that $A = R_\theta D_{\lambda_1,\lambda_2} S_t$.

The expression $\lambda_2 \in \mathbb{R}\backslash\{0\}$ denotes that λ_2 is an element of the real numbers other than zero; so $\lambda_2 \in \mathbb{R}$ and $\lambda_2 \neq 0$.

Exercise. If ref_ψ is the map which reflects along the line of angle ψ, for $0 \leqslant \psi < \pi/2$, determine how to decompose ref_ψ as in the statement of the Proposition.

Proof of Proposition 67. Here is the essential idea. We're going to determine what some of these values $\theta, \lambda_1, \lambda_2, t$ would *have to be* for $A = R_\theta D_{\lambda_1,\lambda_2} S_t$ to be true. Then we'll multiply by an appropriate inverse to 'cancel them out' and move on to the remaining values, until we've determined what all of the values have to be. Because A determines what the values have to be, this proves uniqueness. Then we'll verify that A actually is what we say it is. Our first step is to figure out what θ and λ_1 must be.

Let's see what the transformation $R_\theta D_{\lambda_1,\lambda_2} S_t$ does to the basis vector e_1, to start. (This seems easier than e_2, because $S_t e_1$ is simpler than $S_t e_2$.) We'll compute it by applying each transformation one-by-one. We have
$$\begin{aligned} R_\theta D_{\lambda_1,\lambda_2} S_t e_1 &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \lambda_1 \\ 0 \end{pmatrix} = \begin{pmatrix} \lambda_1 \cos\theta \\ \lambda_1 \sin\theta \end{pmatrix}. \end{aligned}$$

If $R_\theta D_{\lambda_1,\lambda_2} S_t = A$, then our calculation implies
$$Ae_1 = \begin{pmatrix} \lambda_1 \cos\theta \\ \lambda_1 \sin\theta \end{pmatrix}.$$

Because we assumed A is invertible, in particular $\ker(A) = \{\vec{0}\}$, so $Ae_1 \neq \vec{0}$. I claim that if $v \in \mathbb{R}^2$ is a nonzero vector, there exists a unique $\lambda_1 > 0$ and $\theta \in [0, 2\pi)$ so that $v = \begin{pmatrix} \lambda_1 \cos\theta \\ \lambda_1 \sin\theta \end{pmatrix}$.

This follows from two facts:

- For a point $\begin{pmatrix} x \\ y \end{pmatrix}$ on the unit circle — so $x^2 + y^2 = 1$ — there is a unique $\theta \in [0, 2\pi)$ so that $\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$. This is the most common *definition* of these trigonometric functions: they are the x and y coordinates of the point on the unit circle which lies θ radians around the unit circle, starting from e_1.
- For a nonzero vector $v = \begin{pmatrix} a \\ b \end{pmatrix}$, we have $a^2 + b^2 > 0$, and

$$v' = \begin{pmatrix} a/\sqrt{a^2+b^2} \\ b/\sqrt{a^2+b^2} \end{pmatrix}$$

  has unit length.

So if $Ae_1 = \begin{pmatrix} a \\ b \end{pmatrix}$, then $\lambda_1 = \sqrt{a^2+b^2}$ and θ is the angle Ae_1 lies around the unit circle (how many radians Ae_1 is counter-clockwise from the positive x-axis).

It seems hard to apply a similar analysis to Ae_2, because in the very first step I shear Ae_2, and then it goes through more complicated transformations like scaling and rotation. Instead of studying A, I will study a new linear transformation in which we 'undo the last steps'.

Consider a new transformation, $B = D_{\lambda_1,1}^{-1} R_\theta^{-1} A$. Because $Ae_1 = R_\theta D_{\lambda_1,1} e_1$, we have

$$Be_1 = D_{\lambda_1,1}^{-1} R_\theta^{-1} R_\theta D_{\lambda_1,1} e_1 = D_{\lambda_1,1}^{-1} D_{\lambda_1,1} e_1 = e_1.$$

Therefore the matrix form of B is $B = \begin{pmatrix} 1 & c' \\ 0 & d' \end{pmatrix}$ for some scalars $c', d' \in \mathbb{R}$. If we take $\lambda_2 = d'$ and $t = c'$ then we can explicitly see

$$D_{1,2} S_t = \begin{pmatrix} 1 & 0 \\ 0 & d' \end{pmatrix} \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & c' \\ 0 & d' \end{pmatrix} = B.$$

Because A is invertible, so is B, as B is a composition of invertible matrices. Therefore $\lambda_2 = d' \neq 0$ — otherwise the image of this matrix would be one-dimensional (it would be the x-axis).

Let me summarize what we have done so far. We started with an arbitrary linear transformation, A. We showed that if we take λ_1 to be the length of Ae_1 and θ to be the number of radians it lies counterclockwise around the positive x-axis, then the linear transformation $B = D_{\lambda_1,1}^{-1} R_\theta^{-1} A$ satisfies $Be_1 = 1$. Further, we have $B = D_{1,\lambda_2} S_t$, where t is the first component of Be_2 and λ_2 is its second component. Therefore

$$A = R_\theta D_{\lambda_1,1} B = R_\theta D_{\lambda_1,1} D_{1,\lambda_2} S_t = R_\theta D_{\lambda_1,\lambda_2} S_t,$$

proving existence. Because we could recover $\theta, \lambda_1, \lambda_2, t$ starting from information about A, this proves uniqueness. □

While this proof was long and complicated, I think the idea can be made completely visual.

Visualization exercise. Click here. This is a link to a visualization tool which demonstrates the output of the linear map $A = R_b D_{l_1,l_2} S_a$; that is, l_1, l_2 denote the stretch factors we called λ_1, λ_2 above; a denotes the shearing factor we called t above; b denotes the rotation angle we called θ above. While the applet allows you to make l_1 zero or negative, for the sake of this discussion, keep it positive.

If you choose l_1, l_2, a, b, the applet shows the output of A applied to a green "grid" of vectors and in particular to the vectors e_1 and e_2 (their outputs Ae_1 and Ae_2 are indicated by a blue line segment ending at a blue dot and a red line segment ending at a red dot, respectively).

By default, the matrix this app displays is the identity matrix. What happens as you move the different sliders? Play around with the sliders in this applet. Try to observe the following.

- No matter what l_1, l_2, a are, the angle b is always how far around the unit circle the blue dot Ae_1 is. This was used in the argument above (it is how we recovered θ). In particular, if you keep the last slider the same but you change the first three sliders (with $l_1 > 0$!), the blue dot doesn't change its angle around the circle.
- No matter what l_2, a, b are, the number l_1 is always how far the blue dot is from the circle; if you change the last three sliders, this length doesn't change, even if the blue dot moves. This was used in the argument above to recover λ_1 as the length of the vector Ae_1.
- When you change l_1 or l_2 from positive to negative or vice versa the linear transformation becomes non-invertible on the way (you "shrink everything down to zero" before reversing and expanding things in the negative direction).
- If $l_1 = 1$ and $b = 0$ — this is the situation we were in when we passed to B — the terms l_2 and a just tell you the coordinates of the red dot.
- In general, however, changing any of the four coordinates will change the position of the red dot (not just changes to l_2 or a). Contrast this with the first two bullet points above, where changing l_2 or a didn't change the position of the blue dot whatsoever.
- Try to get a visual sense of what each slider governs. If you start at some complicated choice of (l_1, l_2, a, b), what happens to the picture as you change each of those? Can you get a sense for what each of these factors "means"?

There is, in fact, a unique way to decompose any invertible map $A : \mathbb{R}^n \to \mathbb{R}^n$ into a product RDS of three matrices, where R is an $n \times n$ rotation matrix, D is a diagonal matrix with all entries positive except possibly one negative entry, and S is a strictly upper-triangular matrix, so every invertible transformation of $\mathbb{R}^n$ can be thought of as shearing, scaling, and rotation.

We cannot prove such a claim right now, because I cannot tell you rigorously what a "rotation matrix" is in dimensions larger than 2. I hope to return to this topic towards the end of the term.

5.6 Algorithmically solving systems of equations

Now, half-way through the term, I'm going to talk about the topic usually mentioned at the *very beginning* of a course on linear algebra.

Thusfar, we have frequently had occasion to wonder two questions:

- Suppose I have $v_1, \cdots, v_n \in V$, and another vector $v \in V$. Is it the case that $v \in \text{span}(v_1, \cdots, v_n)$, or not? (We think about this question when determining if vectors are redundant, and — applied to $v = e_1, \cdots, e_n$ — this can be used to determine if a set spans V.)
- Suppose I have a linear map $A : V \to W$. Is A injective? Alternatively, are there non-zero vectors $v \in \ker(A)$? (We think about this when determining if a set of vectors is linearly independent or not.)

These questions are usually written at the start of most linear algebra books, in the following form. Suppose

$$M = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

is an $m \times n$ matrix, and consider the associated linear map $A_M : \mathbb{F}^n \to \mathbb{F}^m$, given by

$$A_M \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \cdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{pmatrix}.$$

Question. For a fixed vector $b \in \mathbb{F}^m$, find the set of vectors $x \in \mathbb{F}^n$ for which $A_M x = b$. That is, determine the set $A_M^{-1}(\{b\})$. This includes the questions of whether or not this equation has *any* solutions, and also

whether or not this equation has *unique* solutions, or if there is more than one solution.

If I spell out what this question is asking, writing $b = \begin{pmatrix} b_1 \\ \cdots \\ b_m \end{pmatrix}$, we are looking to find the set of solutions $(x_1, \cdots, x_n)$ to the system of equations

$$\begin{aligned} a_{11}x_1 + \cdots a_{1n}x_n &= b_1 \\ &\cdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m. \end{aligned}$$

This is what you would call in algebra a *system of m equations in n unknowns*. We're going to develop an algorithm (a process that a computer can run, which always follows the same steps, with no external input) for solving such systems of equations.

Example 72. The system of equations

$$\begin{aligned} 2x_2 + x_3 - 5x_5 &= 3 \\ x_1 - 4x_4 + x_5 &= 1 \\ x_1 - 2x_2 + 3x_3 - x_5 &= 3 \end{aligned}$$

corresponds to the equation $A_M x = b$, where $M = \begin{pmatrix} 0 & 2 & 1 & 0 & -5 \\ 1 & 0 & 0 & -4 & 1 \\ 1 & -2 & 3 & 0 & -5 \end{pmatrix}$ and $b = \begin{pmatrix} 3 & 1 & 3 \end{pmatrix}$. One can rewrite the equations above as

$$\begin{pmatrix} 0 & 2 & 1 & 0 & -5 \\ 1 & 0 & 0 & -4 & 1 \\ 1 & -2 & 3 & 0 & -5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ 3 \end{pmatrix}.$$

(Do the matrix multiplication to see that the product on the left-hand side gives precisely the left-hand-side of the equations above.) ◊

5.6.1 Echelon forms

Before developing the algorithm, let me present two kinds of matrices for which it is easy to solve this equation (in two different senses).

Definition 34. An $m \times n$ matrix M is said to be in **reduced row echelon form** if the following conditions hold:

- The first nonzero entry in each row of M (if there is one) is equal to 1. This is usually called a 'leading 1' or 'pivot entry'. If row i has a leading 1, I will say the leading 1 is in position $j(i)$, so $a_{i,j(i)} = 1$ and $a_{i,k} = 0$ for $k < j(i)$.
- If row i has a leading 1, the other entries in the column $j(i)$ are all zero. That is, $a_{k,j(i)} = 0$ for $k \neq i$.
- If row i is zero, all later rows are zero.
- If there is a leading 1 in row i and row $i+1$, then $j(i) < j(i+1)$. That is, leading 1's move right as we go down the rows.

◊

"Echelon form" means the leading 1's (sometimes called "pivot elements") move down and to the right as you move down the matrix (look up the definition of the word 'echelon' and its relevance to military formations to get a good picture; you will see it matches with the examples below).

Example 73. $\begin{pmatrix} 1 & 0 & 3 & 0 & 1 \\ 0 & 1 & 2 & 0 & 3 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$ is an example of a matrix in reduced row echelon form. Notice that the first nonzero entry in each row is a 1; all other entries are zero in columns with a leading 1; all the rows with no leading 1's are at the end, and the leading 1's move to the right. Another example is

$$\begin{pmatrix} 0 & 0 & 1 & 3 & 4 & 5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The following matrices are not in reduced row echelon form. Why not?

$$\begin{pmatrix} 1 & 2 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

◊

Reduced row echelon form is nice because when M is in reduced row echelon form, it is straightforward to find the solutions to the system $A_M x = b$. First, let me show an example; then let me state the general process.

Example 74. Let $M = \begin{pmatrix} 1 & 0 & 3 & 0 & 1 \\ 0 & 1 & 2 & 0 & 3 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$. Then $A_M x = b$ translates to the five equations

$$\begin{aligned} x_1 + 3x_3 + x_5 &= b_1 \\ x_2 + 2x_3 + 3x_5 &= b_2 \\ x_4 &= b_3 \\ 0 &= b_4 \\ 0 &= b_5. \end{aligned}$$

The final two equations hold **if and only if** $b_4 = b_5 = 0$. So this sytem has no solutions if one of b_4, b_5 are nonzero. On the other hand, the first three equations can be rewritten as

$$\begin{aligned} x_1 &= b_1 - 3x_3 - x_5 \\ x_2 &= b_2 - 2x_3 - 3x_5 \\ x_4 &= b_3. \end{aligned}$$

Therefore this system has solutions if and only if $b_4 = b_5 = 0$. It has many solutions: we are allowed to choose the values of x_3, x_5 however we want (these are called 'independent variables'), while x_1, x_2, x_4 are determined by the b's and the choices of x_3, x_5. We can write the solutions as those vectors x of the form

$$x = \begin{pmatrix} b_1 - 3x_3 - x_5 \\ b_2 - 2x_3 - 3x_5 \\ x_3 \\ b_3 \\ x_5 \end{pmatrix}.$$

◊

The same thought process allows us to easily write the solutions to $A_M x = b$ whenever M is in reduced row echelon form.

Proposition 68. *Suppose M is an $m \times n$ matrix in reduced row echelon form. Suppose rows 1 through k have leading 1's, with $j(i)$ being the position of the leading 1 in row i. We say the variables $x_{j(1)}, \cdots, x_{j(k)}$ are "dependent variables", while all other x_r's are "independent variables".*

The equations $A_M x = b$ have a solution if and only if $b_{k+1} = \cdots = b_m = 0$ (that is, $b_i = 0$ if row i is zero / has no leading 1); when this is the case, we may freely choose all $x_1, \cdots, x_n$ other than *$x_{j(1)}, \cdots, x_{j(k)}$. After having chosen those values of x, there is a unique solution to this equation, with $x_{j(i)}$ given by*

$$x_{j(i)} = b_i - a_{i,j(i)+1}x_{j(i)+1} - \cdots - a_{i,n}x_n.$$

Notice that in the expression $b_i - a_{i,j(i)+1}x_{j(i)+1} - \cdots - a_{i,n}x_n$, no dependent variable appears, only the independent variables (which we chose to be whatever they wanted to be).

Remark 43. Applying this to the vector $b = 0$, this shows us a basis for $\ker(A_M)$: I have one basis vector for each independent variable x_r, corresponding to setting $x_r = 1$ and all other independent variables equal to zero. For example, in the case of $M = \begin{pmatrix} 1 & 0 & 3 & 0 & 1 \\ 0 & 1 & 2 & 0 & 3 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$, a basis for $\ker(A_M)$ is given by looking at the solution to our system of equations $A_M x = \vec{0}$ given by

$$\begin{aligned} x_1 &= -3x_3 - x_5 \\ x_2 &= -2x_3 - 3x_5 \\ x_4 &= 0 : \end{aligned}$$

setting $x_3 = 1$ and $x_5 = 0$ (and then setting $x_3 = 0$ and $x_5 = 1$) we see that the following two vectors form a basis for $\ker(A_M)$:

$$\begin{pmatrix} -3 \\ -2 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} -1 \\ -3 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

In particular, if M is in reduced row echelon form, $\text{null}(A_M)$ is the number of independent variables (while $\text{rank}(A_M)$ is the number of dependent variables). ◊

Thus a matrix M in reduced row echelon form allows us to easily compute its kernel, or even better, the set of solutions $A_M x = b$ for any vector $b \in \mathbb{F}^m$. There is a comparable notion which plays well with images, as well.

Definition 35. An $n \times m$ matrix M is said to be in **reduced column echelon form** if the following conditions hold:

- The first nonzero entry in each column of M (if there is one) is equal to 1. This is usually called a 'leading 1' or 'pivot entry'. If column j has a leading 1, I will say the leading 1 is in position $i(j)$, so $a_{i(j),j} = 1$ and $a_{k,j} = 0$ for $k < i(j)$.
- If column j has a leading 1, the other entries in the row $i(j)$ are all zero. That is, $a_{i(j),k} = 0$ for $k \neq j$.
- If column j is zero, all later columns are zero.
- If there is a leading 1 in column j and column $j+1$, then $i(j) < i(j+1)$. That is, leading 1's move down as we go right along the columns.

◊

Example 75. Here are two matrices in reduced column echelon form:

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

The matrix $\begin{pmatrix} 1 & 0 & 3 & 0 & 1 \\ 0 & 1 & 2 & 0 & 3 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$ is in reduced row echelon form, but it is not in reduced column echelon form (it fails almost every criterion). The two conditions of 'reduced row echelon form' and 'reduced column echelon form' are nearly opposites. ◊

This is almost (but not quite) the exact same set of conditions as I described in the Bonus exercise on HW4: if I hand you a matrix in reduced column echelon form, you automatically have a basis for its image!

Example 76. Consider the rcef matrix $M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 2 & 0 \\ 0 & 0 & 1 \\ 5 & 3 & 2 \end{pmatrix}$. For the associated linear map $A_M : \mathbb{F}^3 \to \mathbb{F}^5$, the output of $\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ is given by

$$A_M x = \begin{pmatrix} x_1 \\ x_2 \\ 3x_1 + 2x_2 \\ x_3 \\ 5x_1 + 3x_2 + 2x_3 \end{pmatrix}.$$

The nonzero columns of M provide a basis for the image, and it is now easy to check if a given vector b is in the image of A_M: set $x_1 = b_1, x_2 = b_2, x_3 = b_4$, and see if $b_3 = 3x_1 + 2x_2$ and $b_5 = 5x_1 + 3x_2 + 2x_3$. For instance, $b = \begin{pmatrix} 1 \\ 3 \\ 8 \\ -4 \\ 6 \end{pmatrix}$ **is** in the image, because if we set $x_1 = 1, x_2 = 3, x_3 = -4$, then we have

$$\begin{pmatrix} x_1 \\ x_2 \\ 3x_1 + 2x_2 \\ x_3 \\ 5x_1 + 3x_2 + 2x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 3(1) + 2(3) \\ -4 \\ 5(1) + 3(3) + 2(-4) \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 9 \\ -4 \\ 6 \end{pmatrix} = b,$$

so b is in the image of A_M.

On the other hand, $b = \begin{pmatrix} 0 \\ 2 \\ 0 \\ 3 \\ 0 \end{pmatrix}$ is not in the image of A_M, because if $A_M x = b$, we must have $x_1 = 0, x_2 = 2, x_3 = 3$; but

$$A_M \begin{pmatrix} 0 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ 3(0) + 2(2) \\ 3 \\ 5(0) + 3(2) + 2(3) \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ 4 \\ 3 \\ 12 \end{pmatrix} \neq b.$$

◊

The example above, stated generally, is given as follows.

Proposition 69. *Suppose M is an $m \times n$ matrix in reduced column echelon form, and that columns $1 \leqslant j \leqslant k$ are nonzero. Write $i(j)$ for the position of the leading 1 in each column. Then*

$$b \in im(A_M) \iff A_M \begin{pmatrix} b_{i(1)} \\ \cdots \\ b_{i(k)} \\ 0 \\ \cdots \\ 0 \end{pmatrix} = b.$$

Remark 44. Suppose M is a matrix in **both** reduced row echelon form and reduced column echelon form. Then M takes a very simple form, which is most easily expressed in the language of 'block matrices' (writing matrices in terms of smaller matrices). For instance,

$$M = \left(\begin{array}{cc|ccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

is both rref and rcef. In general, an $m \times n$ matrix of rank r which is both rref and rcef is necessarily of the form

$$M = \left(\begin{array}{c|c} I_{r\times r} & 0_{r\times(n-r)} \\ \hline 0_{(m-r)\times n} & 0_{(m-r)\times(n-r)} \end{array}\right),$$

where $I_{r\times r}$ refers to the $r \times r$ identity matrix, and $0_{i\times j}$ refers to the zero matrix with i rows and j columns. (In other words, M has 1's descending from the top left until eventually stopping, and all other entries are zero.)

There is exactly one rref and rcef matrix of a given rank. ◊

5.6.2 The Gauss–Jordan algorithm

The previous remarks are all well and good, but if I hand you a matrix M, it is usually not the case that M is in either rref or rcef. So it is not at all clear that the stuff we just discussed helps!

There are actually two (basically identical) Gauss–Jordan algorithms, one for row operations and one for column operations. I will go over the first in detail: the second discussion is almost identical. If you want to test your understanding, you can try to set up the theory of coluimn Gauss–Jordan, and see how the statements and arguments below change.

If you've ever attempted to solve systems of linear equations like

$$\begin{aligned} x_1 + 2x_2 + 3x_3 &= 0 \\ x_1 + 3x_3 &= 4 \\ x_2 - x_3 &= 5 \end{aligned}$$

before, you've proceeded in the following fashion:

- You added (a multiple of) one equation to another, which doesn't change the set of solutions.
- You scaled one equation by a nonzero quantity, which doesn't change the set of solutions.
- You might have swapped some equations to organize your work more easily, which certainly doesn't change the set of solutions.
- If you ever got an equation of the form $0 = 0$, it says nothing, so you discard it. If you ever get to an equation of the form $0 = b$ for $b \neq 0$, then you know your system has no solutions, because this is impossible.

You then proceeded until you found equations which were easily solvable. This is called 'row reduction', and the equations you found which were 'easily solvable' corresponded precisely to a matrix in reduced row echelon form.

Definition 36. Let $M = \begin{pmatrix} — & & — \\ & f_1 & \\ & \cdots & \\ & f_m & \\ — & & — \end{pmatrix}$ be an $m \times n$ matrix. Another $m \times n$ matrix M' is obtained from M by an **elementary row operation** if M' takes one of the following three forms:

(i) M' is obtained by swapping two rows i_1, i_2 of M. For instance,

$$M' = \begin{pmatrix} — & f_1 & — \\ — & f_4 & — \\ — & f_3 & — \\ — & f_2 & — \end{pmatrix}$$

is obtained by swapping rows 2 and 4 of M.

(ii) M' is obtained by scaling some single row of M **by a nonzero scalar**. For instance,

$$M' = \begin{pmatrix} — & f_1 & — \\ — & 4f_2 & — \\ — & f_3 & — \end{pmatrix}$$

is obtained by scaling the second row of M by $4 \neq 0$.

(iii) M' is obtained by adding a multiple of some row i_1 of M to some row i_2 of M. For instance,

$$M' = \begin{pmatrix} — & f_1 & — \\ — & f_2 & — \\ — & f_3 - 17f_4 & — \\ — & f_4 & — \end{pmatrix}$$

is obtained by adding -17 times row 4 to row 3.

Notice that these same operations can be applied to vectors ($m \times 1$ matrices). For instance, applying a row operation of type (ii) to $\begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}$ might give $\begin{pmatrix} 3 \\ 4 \\ 2 \end{pmatrix}$ (here we scaled the second row by 4). ◊

The discussion in my itemized list above (how solutions to the equations don't change when you perform these operations) is formalized as the following statement.

Proposition 70. *Suppose M is an $m \times n$ matrix and $b \in \mathbb{F}^m$ is a vector (an $m \times 1$ matrix). Suppose M' and b' are obtained by performing the same row operation on M and b. Then*

$$A_M x = b \iff A_{M'} x = b'.$$

For example, we have

$$\begin{pmatrix} 3 & 1 & 4 & 4 & 0 \\ 2 & 1 & 0 & 3 & 0 \\ 1 & 5 & 15 & 5 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \iff \begin{pmatrix} 1 & 0 & 4 & 1 & 0 \\ 2 & 1 & 0 & 3 & 0 \\ 1 & 5 & 15 & 5 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \\ 3 \end{pmatrix}.$$

Here I subtracted the second row from the first in both M and b.

Now let me show in example how to use row operations to **algorithmically** reduce a system of equations to one in reduced row echelon form, and thus extract all of its solutions easily. I will then state the Gauss–Jordan algorithm.

Example 77. Let's explicitly solve the system from before over $\mathbb{R}$ using row operations.

$$\begin{pmatrix}3&1&4&4&0\\2&1&0&3&0\\1&5&15&5&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\\x_5\end{pmatrix}=\begin{pmatrix}1\\2\\3\end{pmatrix}\iff\begin{pmatrix}1&1/3&4/3&4/3&0\\2&1&0&3&0\\1&5&15&5&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\\x_5\end{pmatrix}=\begin{pmatrix}1/3\\2\\3\end{pmatrix}$$

Subtract twice row 1 from row 2 $\iff\begin{pmatrix}1&1/3&4/3&4/3&0\\0&1/3&-8/3&1/3&0\\1&5&15&5&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\\x_5\end{pmatrix}=\begin{pmatrix}1/3\\4/3\\3\end{pmatrix}$

Subtract row 1 from row 3 $\iff\begin{pmatrix}1&1/3&4/3&4/3&0\\0&1/3&-8/3&1/3&0\\0&14/3&41/3&11/3&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\\x_5\end{pmatrix}=\begin{pmatrix}1/3\\4/3\\8/3\end{pmatrix}$

Scale row 2 by 3 $\iff\begin{pmatrix}1&1/3&4/3&4/3&0\\0&1&-8&1&0\\0&14/3&41/3&11/3&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\\x_5\end{pmatrix}=\begin{pmatrix}1/3\\4\\8/3\end{pmatrix}$

Subtract one-third row 2 from row 1 $\iff\begin{pmatrix}1&0&4&1&0\\0&1&-8&1&0\\0&14/3&41/3&11/3&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\\x_5\end{pmatrix}=\begin{pmatrix}-1\\4\\8/3\end{pmatrix}$

Subtract fourteen-thirds row 2 from row 3 $\iff\begin{pmatrix}1&0&4&1&0\\0&1&-8&1&0\\0&0&51&-1&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\\x_5\end{pmatrix}=\begin{pmatrix}-1\\4\\-16\end{pmatrix}$

Scale row 3 by $1/51$ $\iff\begin{pmatrix}1&0&4&1&0\\0&1&-8&1&0\\0&0&1&-1/51&1/51\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\\x_5\end{pmatrix}=\begin{pmatrix}-1\\4\\-16/51\end{pmatrix}$

Subtract four times row 3 from row 1 $\iff\begin{pmatrix}1&0&0&55/51&-4/51\\0&1&-8&1&0\\0&0&1&-1/51&1/51\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\\x_5\end{pmatrix}=\begin{pmatrix}13/51\\4\\-16/51\end{pmatrix}$

Add eight times row 3 to row 2 $\iff\begin{pmatrix}1&0&0&55/51&-4/51\\0&1&0&43/51&8/51\\0&0&1&-1/51&1/51\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\\x_5\end{pmatrix}=\begin{pmatrix}13/51\\76/51\\-16/51\end{pmatrix}.$

In the first step, I scaled the first row by $\frac{1}{3}$. In the next two steps, I subtracted off multiples of the first row from the next two (so there's only the single 1 in the first column). I then moved on to the next column, ignoring the first row, and scaling the first leading entry in the second row to be 1, then subtracting it off the correct multiple from the next row. I did the same thing in the last column.

In the end, we obtain a matrix in reduced-row-echelon form, with no all-zero rows. Therefore there is always a solution to the equations; in fact, there is a 2-dimensional family of solutions, with independent variables x_4, x_5 and dependent variables x_1, x_2, x_3. The equations now read

$$\begin{aligned} x_1 + \frac{55}{51}x_4 + \frac{-4}{51}x_5 &= \frac{13}{51} \\ x_2 + \frac{43}{51}x_4 + \frac{8}{51}x_5 &= \frac{76}{51} \\ x_3 + \frac{-1}{51}x_4 + \frac{1}{51}x_5 &= \frac{-16}{51}, \end{aligned}$$

which can be rewritten as

$$\begin{aligned} x_1 &= \frac{13 - 55x_4 + 4x_5}{51} \\ x_2 &= \frac{76 - 43x_4 - 8x_5}{51} \\ x_3 &= \frac{-16 + x_4 - x_5}{51}. \end{aligned}$$

Therefore, the solutions to this equation take precisely the form

$$\frac{1}{51}\begin{pmatrix} 13 - 55x_4 + 4x_5 \\ 76 - 43x_4 - 8x_5 \\ -16 + x_4 - x_5 \\ 51x_4 \\ 51x_5 \end{pmatrix}.$$

(I pulled out a scalar factor of 1/51 to make this expression cleaner.) As a sanity check, let's plug in some values and see if the original equation actually is satisfied. Taking $x_4 = x_5 = 0$, I find

$$\begin{pmatrix} 3 & 1 & 4 & 4 & 0 \\ 2 & 1 & 0 & 3 & 0 \\ 1 & 5 & 15 & 5 & 1 \end{pmatrix}\begin{pmatrix} 13/51 \\ 76/51 \\ -16/51 \\ 0 \\ 0 \end{pmatrix} = \frac{1}{51}\begin{pmatrix} 3\cdot 13 + 1\cdot 76 + 4\cdot(-16) \\ 2\cdot 13 + 1\cdot 76 + 0\cdot(-16) \\ 1\cdot 13 + 5\cdot 76 + 15\cdot(-16) \end{pmatrix} = \frac{1}{51}\begin{pmatrix} 39 + 76 - 64 \\ 26 + 76 \\ 13 + 380 - 240 \end{pmatrix} = \frac{1}{51}\begin{pmatrix} 51 \\ 102 \\ 153 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix},$$

as claimed.

(This sanity check is very valuable. I noticed a mistake in my original computation by carrying it out, and it's much much faster to check your work than it was to carry out the original computation.)

The only thing missing from the general algorithm in this example is the fact that when I moved onto the second column, there was no guarantee the first nonzero term was already in row 2. I might have had to swap some rows to guarantee this (notice we never used the row-swapping operation). This will be included in the phrasing in terms of the formal Guass–Jordan algorithm. ♢

Statement of the Guass–Jordan algorithm. Start with an $m \times n$ matrix M and a vector $b \in \mathbb{F}^m$. Apply the following operations (determined by the structure of M) to reduce it to a rref matrix, changing b with each operation as you go. The steps below are phrased as n steps for the n columns, with m sub-steps for each column.

- **Step C1.** Search to see if there exists a nonzero entry in column 1. If not, move on to step C2. If there is one, continue through the sub-steps below.
  - **Substep (swap).** If the first row with a nonzero entry in column 1 is row $i > 1$, swap row 1 with row i.
  - **Substep (scale).** Having done so, scale the first row by $1/a_{11}$, so that the first row starts with a leading 1.

  - **Substep (subtract).** Having done so, subtract a_{21} times the first row from the second row. Then subtract a_{31} times the first row from the third row, and so on, through subtracting a_{m1} times the first row from the m'th row. Having done so, move on to Step C2.

- $\cdots$ (continue down the columns...)

- **Step Cj.** Suppose rows 1 through i all have leading 1's, at this point. Search column j to see if there is a nonzero entry in a row lower than row i. If there is not, move on to Step C(j+1). If there is one, continue through the sub-steps below.
  - **Substep (swap).** If the first row after row i with a nonzero entry in column j is row $k > i+1$, swap row k with row $i+1$.
  - **Substep (scale).** Having done so, scale the $i+1$'th row by $1/a_{i+1,j}$, so that the row $i+1$ has a leading 1.
  - **Substep (subtract).** Having done so, subtract $a_{1,j}$ times the $i+1$'th row from the first row. Then subtract $a_{2,j}$ times the $i+1$'th row from the second row, and so on, skipping row $i+1$. After subtracting $a_{m,j}$ times the $i+1$'th row from the m'th row, move on to Step C(j+1).

- $\cdots$ (continue down the columns...)

- After completing Step Cn, the resulting matrix is in reduced row echelon form, and the algorithm is complete.

Each step requires no thought: you (or better, a computer) just follows the instruction. After each step C1, C2, ... we have made the first, second, etc column have the same behavior as the columns in an rref matrix. Ultimately, the matrix is actually rref, and we can read off the list of solutions. I suggest trying to see how the work we did in the previous example is the same as the work promised by this algorithm.

5.7 Bases and matrices

We have now developed a rather extensive understanding of matrices and how to work with them, as well as an algorithm for computing the kernel and image of a linear map $\mathbb{F}^n \to \mathbb{F}^m$. But most vector spaces are not of this form: for instance

$$V = \left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3 \quad \middle| \quad 3x - 2y + 5z = 0 \right\}$$

is 2-dimensional, but it is not literally $\mathbb{R}^2$, and I do not know how I would encode a linear map $V \to V$ as a matrix, much less compute its kernel or image.

In this section, I'll describe how to translate linear maps between general vector spaces into matrices. This depends on a choice of basis, and different choices will lead us to different matrices. (This is why I emphasized the distinction between linear maps and matrices!)

I want to warn that students often find this material very difficult. (I did when I learned it for the first time.) If you find the discussion here either hard to understand, or insufficient to develop understanding, I want to recommend some alternate references.

- I think that the discussion in Bretscher's book "Linear algebra with applications" in Section 3.4 is helpful, though the notation is rather heavy, and different from mine; I will try to explain the comparison. What he calls $[x]_\beta$ is what I would call $C_\beta^{-1}x$ below. If $T : V \to V$ is a linear transformation and β is a basis for V, he says something along the lines of: "Let B be the matrix with $B[x]_\beta = [Tx]_\beta$; we call this the matrix for T with respect for β." This is the matrix I would call $[T]_{\beta\to\beta}$ below.

- The discussion in Treil's "Linear Algebra Done Wrong", chapter 2.8, is also nice. His notation is closer to mine. Below I will refer to bases β and β', which he would call $\mathcal{A}$ and $\mathcal{B}$ respectively. The matrix I call $[T]_{\beta\to\beta'}$ is what he would call $[T]_{\mathcal{BA}}$. Notice the swapped order: he does this to remedy an irritation I will point out below.

In the end, you should learn to understand my notation and how to use it (since that's how I'm going to refer to this notion!) But for confusing topics, it's often useful to read other sources and see if one resonates with you. If you're only going to look at one of the above, I recommend Treil, since he does things closer to how I do them below.

5.7.1 Bases and coordinates

The first key ingredient is a map I have discussed in examples a number of times.

Definition 37. Let V be a finite-dimensional vector space, and let $\beta = (v_1, \cdots, v_n)$ be a basis for V. The **coordinate map** $C_\beta : \mathbb{F}^n \to V$ is the linear map

$$C_\beta \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} = a_1 v_1 + \cdots + a_n v_n.$$

◊

Let me record as a lemma something I've mentioned in examples before.

Lemma 71. *The map C_β defined above is invertible if and only if β is indeed a basis for V.*

Proof. The image of C_β is the set of vectors of the form $a_1 v_1 + \cdots + a_n v_n$; that is, $\mathrm{im}(C_\beta) = \mathrm{span}(\beta)$. So C_β is surjective if and only if β spans V.

The kernel of C_β is the set of vectors $\begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix}$ so that $a_1 v_1 + \cdots + a_n v_n = \vec{0}$. That is, it is the set of linear relations between the vectors of β. Because a linear map is injective if and only if its kernel is trivial, we see that C_β is injective if and only if β is linearly independent.

Combining these, we have proved the desired claim. □

From now on, if I ever write C_β, it is implicit that β is a basis. In particular, C_β is invertible. What is the map $C_\beta^{-1} : V \to \mathbb{F}^n$?

Remember that this should undo the map C_β. Because $C_\beta \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} = a_1 v_1 + \cdots + a_n v_n$, the fact that $C_\beta^{-1} C_\beta = I$ is the identity map means that

$$C_\beta^{-1} C_\beta \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} = \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix} \implies C_\beta^{-1}(a_1 v_1 + \cdots + a_n v_n) = \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix}.$$

In particular, $C_\beta(e_i) = v_i$ and $C_\beta^{-1}(v_i) = e_i$.

I think of this as follows. Every vector $v \in V$ can be written **in a unique way** as $a_1 v_1 + \cdots + a_n v_n$, because β is a basis. The map C_β^{-1} takes a vector v, finds out the unique way to write it as $v = a_1 v_1 + \cdots + a_n v_n$, and outputs

$$C_\beta^{-1}(v) = \begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix}.$$

I might call these the 'β-coordinates' of v, just as I might say that if $\vec{x} = \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} \in \mathbb{F}^n$ is a vector, then the x_i are the coordinates of $\vec{x}$.

The map C_β sends an element of $\mathbb{F}^n$ (a list of numbers) to the vector with those coordinates. On the other hand, the map C_β^{-1} takes a vector and *extracts* its β-coordinates.

Let me try to give some intuition with an example.

Example 78. Consider the subspace

$$V = \left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3 \quad \middle| \quad 3x - 2y + 5z = 0 \right\} \subset \mathbb{R}^3.$$

I'm going to choose some silly basis of this space; let's say

$$\beta = \left\{ \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}, \quad \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} \right\}.$$

Then the map $C_\beta : \mathbb{F}^2 \to V$ is defined by

$$C_\beta \begin{pmatrix} x \\ y \end{pmatrix} = x \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix} + y \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} = \begin{pmatrix} x + 2y \\ -x + 3y \\ -x \end{pmatrix}.$$

On the other hand, the map C_β^{-1} finds how to write a vector in β-coordinates and it sends it to those coordinates.

For instance, let's take the vector $\begin{pmatrix} 7 \\ 8 \\ -1 \end{pmatrix}$. Because $3 \cdot 7 - 2 \cdot 8 - 5 = 21 - 16 - 5 = 0$, this is indeed a vector in V. How do I write it as a linear combination of the basis vectors? If you want, you can run the Gauss–Jordan algorithm to solve this (good practice); I'll just tell you that

$$\begin{pmatrix} 7 \\ 8 \\ -1 \end{pmatrix} = 1 \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix} + 3 \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix}.$$

It follows that

$$C_\beta^{-1} \begin{pmatrix} 7 \\ 8 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix};$$

I send our vector to the coefficients in its representation as a linear combination of the basis vectors.

You might ask me for a formula for the linear map C_β^{-1}, but every answer I'm going to give you will be frustrating, I think. Part of the problem is that to give a formula for this map, you should first give me a formula for an arbitrary element of V. One of those is in terms of the basis β! For instance, a general form for a vector in V is $\begin{pmatrix} x + 2y \\ -x + 3y \\ -x \end{pmatrix}$, and indeed because

$$\begin{pmatrix} x + 2y \\ -x + 3y \\ -x \end{pmatrix} = x \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix} + y \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix},$$

we have $C_\beta^{-1} \begin{pmatrix} x + 2y \\ -x + 3y \\ -x \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}$. But all I've really told you here is that C_β^{-1} is the inverse of C_β.

$\Diamond$

Remark 45. For a more trivial example, suppose $V = \mathbb{F}^n$ and $\beta = (e_1, e_2, \cdots, e_n)$. Then the map $C_\beta : \mathbb{F}^n \to \mathbb{F}^n$ is the identity map. It sends a vector $\begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix}$ in $\mathbb{F}^n$ to the coefficients in the expression

$$\begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = x_1 e_1 + \cdots + x_n e_n;$$

that is, it sends $\begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix}$ to itself!

On the other hand, if I choose a different basis $\beta = (v_1, \cdots, v_n)$ for $\mathbb{F}^n$, then the map C_β will not be the identity; it will send $C_\beta e_j = v_j$. For instance,

$$V = \mathbb{F}^2, \quad \beta = \left(\begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right) \qquad \text{then} \qquad C_\beta = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

In this case one can actually compute C_β^{-1}, though usually it's not trivial to do so. In your homework, I present the fastest algorithm to do so in practice, as well as a fast formula for 2×2 matrices which in this case gives

$$C_\beta^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}.$$

There is a 'general formula' for an $n \times n$ matrix called Cramer's rule which I will present in an upcoming Curio, but this formula is useless in practice past the 3×3 case (and even that is stretching it). It does have some theoretical use. ◊

This is all well and good. The map C_β lets me trace out space, taking a sequence of numbers (the coordinates) to a particular point in V, and the map C_β^{-1} lets me take a point in V to the unique sequence of numbers which determines it (its coordinates). But it's not terribly interesting by itself.

5.7.2 The matrix associated to a pair of bases

We're going to use the language of the previous section to go from linear maps to matrices. Here is the crucial principle: **an $m \times n$ matrix comes from / corresponds to a map $\mathbb{F}^n \to \mathbb{F}^m$.** If V is a vector space and β is a basis for V of size n, we now have a way to go back and forth between $\mathbb{F}^n$ and V: sending a list of coordinates to the corresponding vector in V, and recovering the coordinates of a point on V:

$$C_\beta : \mathbb{F}^n \leftrightarrow V : C_\beta^{-1}$$

$$\mathbb{F}^n \underset{C_\beta^{-1}}{\overset{C_\beta}{\rightleftarrows}} V$$

If I follow one arrow and then the other (going back where I started), I get the identity map: these two maps are inverse.

Construction. Suppose $A : V \to W$ is a linear map. Choose a basis $\beta = (v_1, \cdots, v_n)$ for V and a basis $\beta' = (w_1, \cdots, w_m)$ for W. Consider the diagram of composable linear maps

$$\begin{array}{ccc} V & \xrightarrow{\quad A \quad} & W \\ {\scriptstyle C_\beta}\uparrow & & \downarrow{\scriptstyle C_{\beta'}^{-1}} \\ \mathbb{F}^n & & \mathbb{F}^m \end{array}$$

The composite $C_{\beta'}^{-1} A C_\beta$ defines a map $\mathbb{F}^n \to \mathbb{F}^m$, and corresponds to an $m \times n$ matrix.

Definition 38. Let $A : V \to W$ be a linear map, and choose a basis $\beta = (v_1, \cdots, v_n)$ for V, and a basis $\beta' = (w_1, \cdots, w_m)$ for W. Write $[A]_{\beta \to \beta'}$ for the $m \times n$ matrix which represents the linear map $C_{\beta'}^{-1} A C_\beta : \mathbb{F}^n \to \mathbb{F}^m$. ◊

To simplify notation, I will often simply write $[A]_{\beta \to \beta'}$ for either the $m \times n$ matrix or the linear map $C_{\beta'}^{-1} A C_\beta : \mathbb{F}^n \to \mathbb{F}^m$ it represents.

I want to try and explain what this means in a few ways. First off, let's try to understand $[A]_{\beta \to \beta'}$ by understanding what this composite does step-by-step:

- Given $\vec{x} = \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} \in \mathbb{F}^n$, the output $C_\beta \vec{x}$ sends it to $C_\beta \vec{x} = x_1 v_1 + \cdots + x_n v_n$, the vector in V with those coordinates.

- The map $A : V \to W$ takes vectors in V and produces vectors in W. At this stage, we plug $C_\beta \vec{x}$ into the map A, and $AC_\beta \vec{x}$ is a vector in W. Explicitly,

$$A(C_\beta \vec{x}) = A(x_1 v_1 + \cdots + x_n v_n) = x_1 A v_1 + \cdots + x_n A v_n.$$

- Now this vector can be expressed somehow as a linear combination of the β'-basis vectors $w_1, \cdots, w_m$; we have

$$w = (x_1 A v_1 + \cdots + x_n A v_n) = y_1 w_1 + \cdots + y_m w_m$$

for some $y_1, \cdots, y_m$, where $y_1, \cdots, y_m$ are what we call the β'-coordinates of this vector. Then we say $C_\beta^{-1} w = \begin{pmatrix} y_1 \\ \cdots \\ y_m \end{pmatrix}$.

More briefly: We send the list of numbers $\vec{x}$ to the vector in V with those β-coordinates. Then we plug that point in V to A, which produces a vector in W. Then we extract the β'-coordinates of this output. In the end, we have taken the list of n numbers $\vec{x} \in \mathbb{F}^n$ to a list of m numbers $[A]_{\beta \to \beta'} \vec{x} \in \mathbb{F}^m$.

I can also describe the columns of this matrix explicitly. The j'th column of $[A]_{\beta \to \beta'}$ is given by $C_{\beta'}^{-1} A C_\beta e_j = C_{\beta'}^{-1}(A v_j)$. Explicitly, $C_{\beta'}^{-1}$ takes a vector in W, writes $w = b_1 w_1 + \cdots + b_m w_m$ as a linear combination of the vectors in β', and $C_{\beta'}^{-1} w = \begin{pmatrix} b_1 \\ \cdots \\ b_m \end{pmatrix}$ extracts the coefficients of this linear combination (what we call the β'-coordinates of w). So the j'th column of $[A]_{\beta \to \beta'}$ is given by $\begin{pmatrix} b_{1j} \\ \cdots \\ b_{mj} \end{pmatrix}$, where $A v_j = b_{1j} w_1 + \cdots + b_{mj} w_m$.

Then

$$[A]_{\beta \to \beta'} \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_{11} x_1 + \cdots + b_{1n} x_n \\ \cdots \\ b_{m1} x_1 + \cdots + b_{mn} x_n \end{pmatrix}$$

means that

$$x_1 A v_1 + \cdots + x_n A v_n = \big(b_{11} x_1 + \cdots + b_{1n} x_n\big) w_1 + \cdots + \big(b_{m1} x_1 + \cdots + b_{mn} x_n\big) w_m.$$

To a large degree, this will be mostly useful for theoretical reasons, and in this class you will not need to compute the explicit matrices $[A]_{\beta \to \beta'}$. I still want to give an example or two as a way to get a sense for 'what is going on'.

Example 79. Let me take the space

$$V = \left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3 \quad \middle| \quad 3x - 2y + 5z = 0 \right\} \subset \mathbb{R}^3$$

from the preceding example, and the basis

$$\beta = \left(v_1 = \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}, \quad v_2 = \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} \right).$$

Now consider the map $A : \mathbb{F}^3 \to V$ given by

$$A\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2x + z \\ 3x + 5y - z \\ 2y - z \end{pmatrix};$$

you may check that the output is indeed an element of V. **Do not write A as a 3×3 matrix!** It is not a map from $\mathbb{F}^3$ to $\mathbb{F}^3$, but rather to a *subspace* of $\mathbb{F}^3$. It will be represented as a 3×2 matrix which depends on the basis we chose above!

Write std for the standard basis (e_1, e_2, e_3) of $\mathbb{F}^3$. Let's compute the matrix $[A]_{\mathsf{std}\to\beta}$. This means we should compute $C_\beta^{-1}Ae_j$, for $j = 1, 2, 3$. Even more explicitly, we should write each Ae_j as a linear combination of v_1 and v_2, and extract their coordinates.

We have

$$Ae_1 = A\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} = 0\begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix} + 1\begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} = 0v_1 + 1v_2$$

$$Ae_2 = A\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 5 \\ 2 \end{pmatrix} = -2\begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix} + 1\begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} = (-2)v_1 + 1v_2$$

$$Ae_3 = A\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix} = 1\begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix} + 0\begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} = 1v_1 + 0v_2.$$

From this we may read off the columns of

$$[A]_{\mathsf{std}\to\beta} = \begin{pmatrix} 0 & -2 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

♢

Remark 46. Choose a basis β for V. The identity map $1_V : V \to V$ always has $[1_V]_{\beta\to\beta} = I$ equal to the identity matrix. Proof: $[1_V]_{\beta\to\beta} = C_\beta^{-1}1_V C_\beta = C_\beta^{-1}C_\beta = I$. An exercise for you is to understand this explicitly in terms of the more concrete description outlined above. ♢

Remark 47. The matrix $[A]_{\beta\to\beta'}$ depends dramatically on the choice of bases β and β', and we will discuss the way it depends on them later. This is not a bad thing. In fact, even if we were only interested in studying concrete things in $\mathbb{F}^n$ (like in a concrete linear algebra class), it is still often useful to choose an *nonstandard* basis, which is better-adapted to understanding the object at hand. We'll see this in practice when we discuss diagonalization and the spectral theorem. ♢

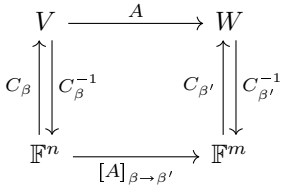
Remark 48. Suppose $A : \mathbb{F}^n \to \mathbb{F}^m$ is a linear map. Then what I have previously called 'the matrix representation for A' in this fancy language is the matrix $[A]_{\mathsf{std}\to\mathsf{std}}$: it is the linear map A when represented as a matrix in terms of the standard bases. ♢

It is sometimes also helpful to represent this diagramatically.

$$\begin{array}{ccc} V & \xrightarrow{\quad A \quad} & W \\ {\scriptstyle C_\beta}\uparrow & & \downarrow{\scriptstyle C_{\beta'}^{-1}} \\ \mathbb{F}^n & \xrightarrow[[A]_{\beta\to\beta'}]{} & \mathbb{F}^m \end{array}$$

In this picture, there are two ways to get from $\mathbb{F}^n$ to $\mathbb{F}^m$. The short path has a name written above it, $[A]_{\beta\to\beta'}$. This is just meant to indicate that it is what we get when we go around the 'long path': $[A]_{\beta\to\beta'} = C_{\beta'}^{-1}AC_\beta$. This is just a visual encoding of the definition of $[A]_{\beta\to\beta'}$, not a theorem.

A picture like this, where there are sometimes multiple ways to get from one point to another, is sometimes called a 'commutative diagram'. Using the fact that the vertical arrows used above are invertible, there's a slightly fancier way to write this diagram which contains more information:



Here the two adjacent vertical arrows are meant to indicate *you can go up or down, and going up and then down the same arrow undoes that arrow.* To say that this diagram 'commutes' means that any way I have from going from one point to another by following the arrows gives me the same answer. For instance, going from bottom-left to top-right, we see that

$$AC_\beta = C_{\beta'}[A]_{\beta\to\beta'},$$

or going from top-left to top-right, we see that

$$A = C_{\beta'}[A]_{\beta\to\beta'}C_\beta^{-1}.$$

Again, this is not a theorem, just a visual way to repackage the definition $[A]_{\beta\to\beta'} = C_{\beta'}^{-1}AC_\beta$ (and use facts like $C_\beta C_\beta^{-1} = 1_V$).

I'll conclude this discussion by showing that this representation of linear maps as matrices *with respect to chosen bases* plays well with composition / matrix multiplication — so long as you use the same basis on the vector space in the middle.

Proposition 72. *Suppose $B : V \to W$ is a linear map, and $A : W \to U$ is a linear map. Suppose we have chosen bases $\beta_V, \beta_W, \beta_U$ for V, W, U, respectively. Then we have*

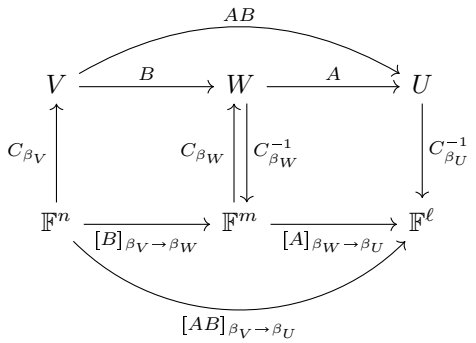
$$[AB]_{\beta_V\to\beta_U} = [A]_{\beta_W\to\beta_U}[B]_{\beta_V\to\beta_W}.$$

The weird ordering on the β's is not a mistake. It has to do with the fact that when you write composition in terms of arrows $V \xrightarrow{B} W \xrightarrow{A} U$, the composite seems to 'flip order' to $AB : V \to U$, and this relates to the fact that we write function application on the left: $(AB)v = A(Bv)$ means we apply B *first*, even though it comes second in the string "AB".

Proof. Let me first give a pure symbol-pushing formula argument, and then let me show a diagram which might help explain what's going on. By definition, we have

$$\begin{aligned}[A]_{\beta_W\to\beta_U}[B]_{\beta_V\to\beta_W} &= \left(C_{\beta_U}^{-1}AC_{\beta_W}\right)\left(C_{\beta_W}^{-1}BC_{\beta_V}\right) = C_{\beta_U}^{-1}AC_{\beta_W}C_{\beta_W}^{-1}BC_{\beta_V} \\ &= C_{\beta_U}^{-1}A1_WBC_{\beta_V} = C_{\beta_U}^{-1}ABC_{\beta_V} = [AB]_{\beta_V\to\beta_U}.\end{aligned}$$

All I used here is the definition of the expressions involved and the fact that $C_{\beta_W}^{-1}C_{\beta_W}$ is the identity map (these two maps undo each other). But I think a diagram might make this more clear.



The map $[A]_{\beta_W \to \beta_U}[B]_{\beta_V \to \beta_W}$ is what you get when you start at the bottom left, go up, right, down, up again, right, and down. When we went down and up again in the middle there, those two steps undo each other. This is indistinguishable from having gone up, right, right, and down — that is, apply C_{β_V}, then B, then A, then $C_{\beta_U}^{-1}$. But "apply B, then A" is what "apply AB'' means. So this reads: "Apply C_{β_V}, then AB, then $C_{\beta_U}^{-1}$." □

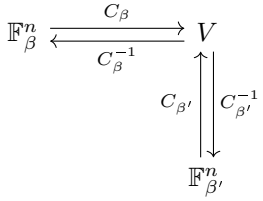
To a large degree, this tells us that *so long as you carefully keep track of bases*, you can go back and forth between the language of vector spaces, linear maps, and composition and the language of matrices and matrix multiplication.

Remark 49. If I chose different bases on W for the two different maps, I get nothing useful. There is no useful formula for the expression $[A]_{\beta'_W \to \beta_U}[B]_{\beta_V \to \beta_W}$. (Using some discussion from the next section, you can write an expression for this, but it's not something you will ever want to think about; it's basically useless.) ♢

5.7.3 Change of basis

When we have a linear map $A : V \to W$ and **chosen bases** β_V and β_W for V and W, we get an associated matrix $[A]_{\beta_V \to \beta_W}$. I had to make a choice in this construction! It's worth understanding how this choice affects the result: in principle (and in practice), we get many different matrices associated to the same linear map!

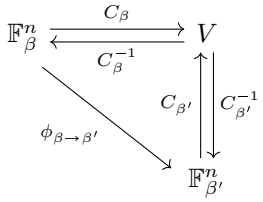
To understand this, let's go back to the earlier discussion of coordinates. Suppose I have two bases β and β' for V. We have two coordinate maps from $\mathbb{F}^n$ to V: how do they compare? In the diagram below, I write $\mathbb{F}^n_\beta$ to indicate that this is the home for the β-coordinates (it's still the same $\mathbb{F}^n$ as usual, the subscript is just to help keep track of what's going on).



Proposition 73. *Suppose V has two bases β and β'. There is a unique invertible linear map $\phi_{\beta \to \beta'} : \mathbb{F}^n \to \mathbb{F}^n$ which 'transitions' from one set of coordinates to another, in the sense that*

$$C_{\beta'}\phi_{\beta \to \beta'} = C_\beta.$$

This is depicted diagramatically below.



Proof. The map is simply given by the composite $\phi_{\beta \to \beta'} = C_{\beta'}^{-1}C_\beta$. Notice that this does indeed satisfy

$$C_{\beta'}\phi_{\beta \to \beta'} = C_{\beta'}C_{\beta'}^{-1}C_\beta = 1_W C_\beta = C_\beta,$$

as claimed.

This is the only such map because if $C_{\beta'}\phi_{\beta \to \beta'} = C_\beta$, then left-multiplying both sides by $C_{\beta'}^{-1}$ we see that

$$\phi_{\beta \to \beta'} = (C_{\beta'}^{-1}C_{\beta'})\phi_{\beta \to \beta'} = C_{\beta'}^{-1}(C_{\beta'}\phi_{\beta \to \beta'}) = C_{\beta'}^{-1}C_\beta.$$

□

Remark 50. In fact, $\phi_{\beta\to\beta'}$ is the matrix representative of the identity map $1_V : V \to V$ in terms of the bases β, β'; that is, $\phi_{\beta\to\beta'} = [1_V]_{\beta\to\beta'}$. This is the perspective preferred in Treil's book. ◊

These 'transition maps' behave as expected: the transition from coordinate system β to the same coordinate system is the identity; the transition from β' to β amounts to undoing the transition from β to β'; and the composite of the transition from β to β' with the transition from β' to β'' is precisely the transition from β to β''.

Lemma 74. *The transition maps described in Proposition 73 satisfy the following properties.*

a) For any basis β of V, we have $\phi_{\beta\to\beta} = I$ is the identity map $\mathbb{F}^n \to \mathbb{F}^n$.

b) For any two bases β, β' of V, we have $\phi^{-1}_{\beta\to\beta'} = \phi_{\beta'\to\beta}$.

c) For any three bases β, β', β'' we have $\phi_{\beta'\to\beta''}\phi_{\beta\to\beta'} = \phi_{\beta\to\beta''}$.

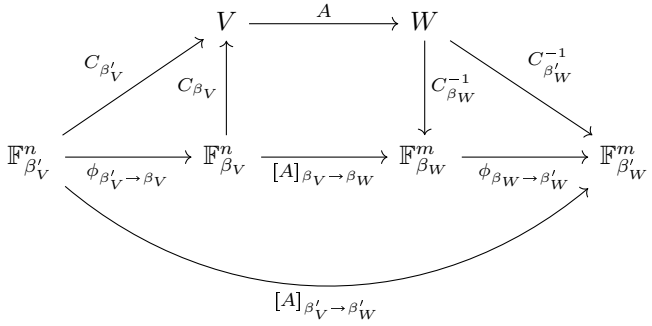
Proof. These are all definition-pushes using the fact that $\phi_{\beta\to\beta'} = C^{-1}_{\beta'}C_\beta$. For instance, the first follows because $C^{-1}_\beta C_\beta = I$. □

Now I can put this together to say how the matrices $[A]_{\beta\to\beta'}$ change as we chanage the basis on domain and codomain.

Proposition 75. *Suppose $A : V \to W$ is a linear map. Suppose that we have chosen two bases β_V, β'_V for V and two bases β_W, β'_W for W. Then the matrix representation of A with respect to these two bases are related by*

$$[A]_{\beta'_V\to\beta'_W} = \phi_{\beta_W\to\beta'_W}[A]_{\beta_V\to\beta_W}\phi_{\beta'_V\to\beta_V}.$$

The diagram below encodes the situation.



Proof. This is a symbol-pushing argument. We have

$$\begin{aligned}\phi_{\beta_W\to\beta'_W}[A]_{\beta_V\to\beta_W}\phi_{\beta'_V\to\beta_V} &= \left(C^{-1}_{\beta'_W}C_{\beta_W}\right)\left(C^{-1}_{\beta_W}AC_{\beta_V}\right)\left(C^{-1}_{\beta_V}C_{\beta'_V}\right)\\ &= C^{-1}_{\beta'_W}\left(C_{\beta_W}C^{-1}_{\beta_W}\right)A\left(C_{\beta_V}C^{-1}_{\beta_V}\right)C_{\beta'_V}\\ &= C^{-1}_{\beta'_W}AC_{\beta'_V} = [A]_{\beta'_V\to\beta'_W}.\end{aligned}$$

Try saying in words what this is supposed to mean, and try understanding this in terms of the diagram above (and what each map is supposed to 'do'). □

Allow me to summarize the results of this section.

If I have chosen a basis $\beta_V = (v_1, \cdots, v_n)$ for V and a basis $\beta_W = (w_1, \cdots, w_n)$ for W, I get a matrix $[A]_{\beta_V\to\beta_W} = M$. If I change the basis β_V to another basis β'_V, then I change this matrix by left-multiplication by an invertible matrix: $[A]_{\beta'_V\to\beta_W} = [A]_{\beta_V\to\beta_W}\phi_{\beta'_V\to\beta_V}$, or simply $M' = MT$ for the appropriate invertibile matrix T. This is called 'right-equivalence' on your homework, and you will show that two matrices are related this way ('changing perspective on the domain' / 'changing basis in the domain') if and only if the have the same image.

On the other hand, if I change the basis β_W to another basis β'_W (while leaving β_V as it is), I get a matrix $[A]_{\beta_V \to \beta'_W} = \phi_{\beta_W \to \beta'_W}[A]_{\beta_V \to \beta_W}$. That is, $M' = SM$ for an appropriate invertible matrix S; this is what I call left-equivalence on HW6#2, and you will show that two matrices are left-equivalent if and only if they have the same kernel.

Finally, if I change both bases, the matrix changes by replacing M with $M' = SMT$, for appropriate invertible matrices S, T. You will show on your homework that so long as M and M' have the same rank, it is always possible to find such invertible matrices S and T.

Exercise. Explain why this statement — HW6#2(c) — is equivalent to the statement 'For any map $A : V \to W$ of rank k, there exists a basis $(v_1, \cdots, v_n)$ of V and a basis $(w_1, \cdots, w_m)$ of W so that

$$Av_i = \begin{cases} w_i & 1 \leqslant i \leqslant k \\ \vec{0} & k < i \leqslant n \end{cases}$$

If you can explain this clearly, and you can show how to construct such bases, you're well on your way to giving a proof for HW6#2(c).

Chapter 6

Determinants and diagonalization

In this chapter, all vector spaces are finite-dimensional.

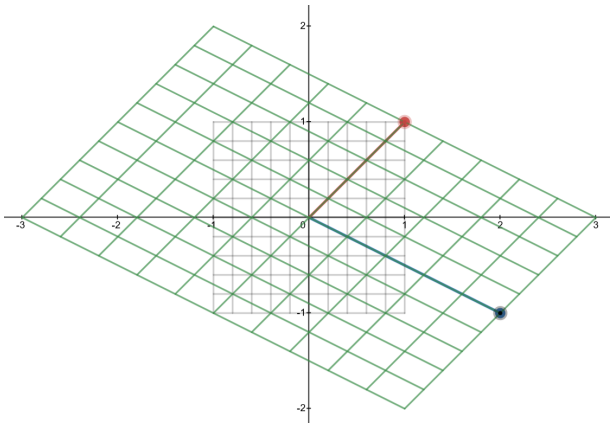
6.1 Determinants

In this section we'll explore the notion of *determinant* of an $n \times n$ matrix, a numerical quantity $\det(M) \in \mathbb{F}$ which *determines* whether or not the matrix is invertible. These notions were studied long, long before matrices or 'linear algebra' (more than two millenia ago by Chinese mathematicians, and rediscovered independently in Europe every century or so starting in the 1500s) because they *determine* whether or not the system of equations $Mx = b$ has a unique solution for any b.

The study of determinants is by necessity heavily algebraic, but the idea itself can be justified geometrically. I would like to give you a sense for how I *think about* determinants before I give you a sense for how I *work with* determinants: while I'd like some of my geometric thinking to translate into the algebraic manipulations I do, I have to admit there's less connection between them than I'd like.

6.1.1 Motivation

Suppose $A : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation. For instance, here's a picture of the linear transformation $A\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix}$ associated to the matrix $M = \begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix}$:



Thihs shows the before-and-after of applying the transformation A to the grid of squares depicted in grey, with the output of each grey square now being depicted as a green parallelogram. The first key point I want you to notice is that *all of the little unit squares are transformed in the same way*: they are all sent to different translates of a given parallelogram. If I look at the rectangle formed by two adjacent gray squares, the output is sent to the parallelogram formed by gluing two adjacent parallelograms.

The 'main green parallelogram' that everything is made out of is the parallelogram with sides

$$Ae_1,\ Ae_2 = \begin{pmatrix} 2 \\ -1 \end{pmatrix},\quad \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

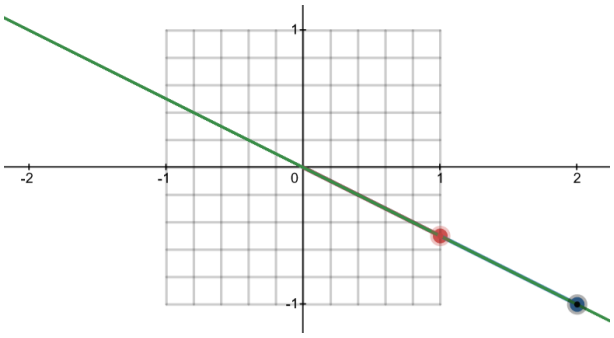
I happen to know that this parallelogram has area equal to 3, so all the area-one squares are sent to objects of area three. (I know this because I happen to know that the 'determinant' $ad - bc$ of a 2×2 matrix computes that area.) But when I glued two of these squares together to form an area-two rectangle, it was sent to the union of two area-three parallelograms: to something of area six.

What we're seeing is that *A scales area in a uniform way.* If I feed A a shape of area 3, it outputs a shape of area $3\text{Area}(S)$.

This fact is something we can formally prove as soon as we formally develop the notion of 'volume' — which we will not do in this course. For now, this fact will simply serve as motivation:

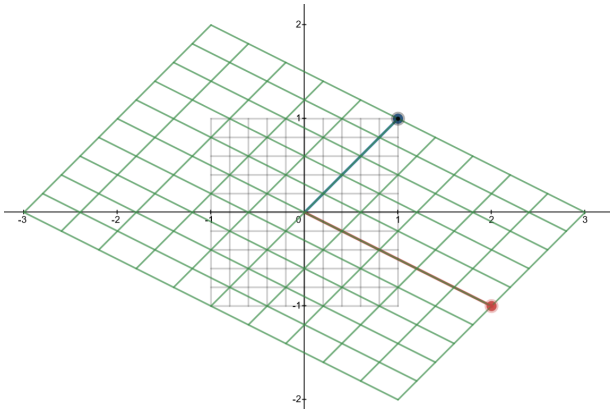
If $S \subset \mathbb{R}^n$ is any shape, and $A : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation, then
there is a constant $\det(A) \in \mathbb{R}$ so that $\text{vol}(AS) = |\det(A)|\text{vol}(S)$.

This quantity has some nice properties. First off, if $\det(A)$ is zero, then A must 'collapse' all of $\mathbb{R}^n$ to a smaller-dimensional space (we have crushed the cube, of positive volume, to something with no volume — somehow we must have collapsed all of $\mathbb{R}^n$ to something of strictly smaller dimension, as in the following picture:



On the other hand, if $\det(A)$ is nonzero, then the unit cube is sent to something of positive volume, which must take up a whole region in n-dimensional space, hence must have image of dimension n. Because $A : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation between vector spaces of the same dimension, and $\dim \text{im}(A) = n$, we must have that A is surjective and by the invertible map theorem that A is invertible. (Again, this is a heuristic argument!)

Now let me return to the formula $\text{vol}(AS) = |\det(A)|\text{vol}(S)$. It is certainly not true that the quantity $ad - bc$ is always positive, so something is lost when I write that absolute value. What does the *sign* of $\det(A)$ encode? In the example above where $M = \begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix}$, the determinant was $3 > 0$. If I had instead swapped the first and second columns, the picture would almost look the same...



Except: compare this to the first picture I gave. The red and blue vectors have swapped positions! Before, the red vector was positioned *counter-clockwise* from the blue vector; in the new picture, the red vector is positioned *clockwise* from the blue vector. The blue vector represents Ae_1 and the red vector represents Ae_2. Before applying A, the vector e_2 lies counter-clockwise to e_1. After appyling A, the vector Ae_2 lies counter-clockwise from Ae_1 for the transformation associated to $\begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix}$, but the opposite is true for the transformation associated to $\begin{pmatrix} 1 & 2 \\ 1 & -1 \end{pmatrix}$.

When I swapped the two columns of A, I swapped their roles, and thus negated the effect of A on the way this arrangement was oriented (for the first, it sends counter-clockwise to counter-clockwise; for the second, it sends counter-clockwise to clockwise).

In general, the sign of $\det(A) \in \mathbb{R}\backslash\{0\}$ determines whether or not a linear transformation preserves or reverses the 'orientation' in $\mathbb{R}^n$. What orientation means is hard to describe in a non-circular fashion, but I can give some examples. In one dimension, a line is either oriented forward or backwards. In two dimensions, the plane is oriented either clockwise or counter-clockwise. In three dimensions, space is oriented in either a 'left-hand' fashion or a 'right-hand' fashion notice that there is no way to rotate your left hand to make it resemble your right hand: you would have to reflect it in a mirror.

So we expect to find a scalar value $\det(A) \in \mathbb{R}$ with the properties above, and with some pleasant algebraic properties that make it amenable to calculation. Let me summarize the properties we expect.

Theorem 76. *There is a function of $n \times n$ matrices called the determinant $\det(M) \in \mathbb{F}$ which satisfies the following properties.*

(i) We have $\det(MN) = \det(M)\det(N)$.

(ii) We have $\det(M) \neq 0$ if and only if M is invertible.

(iii) When $\mathbb{F} = \mathbb{R}$, we may understand $\det(M)$ as being the constant so that applying M scales volume by $|\det(M)|$, and preserves or reverses orientation depending on whether $\det(M) > 0$ or $\det(M) < 0$.

(iv) There are multiple ways to compute $\det(M)$, some of which are useful for explicit computations, and some of which are useful for theoretical investigation.

(Three of these claims will be proved below; you should cite the more specific claims, not this theorem. This statement is merely to help summarize what the 'point' of the next two sections are.)

To actually define and investigate this geometric quantity, I'm going to get very, very algebraic.

6.1.2 The defining property

If one thinks of the determinant as encoding *signed volume* of an n-dimensional box in the n-dimensional vector space V, in some sense, then one is led to a few expected algebraic properties. First, this quantity

is encoded algebraically as a function $D : V^n \to \mathbb{F}$, which takes as input a list of n vectors $(v_1, \cdots, v_n)$ and produces the volume $D(v_1, \cdots, v_n)$ of the parallelepiped with these sides.

If you scale **one** side of a box by c, the area of that box is scaled by c. A legitimate n-dimensional box with nonzero volume cannot have parallel edges (or else the box would have smaller dimension than n), so $D(v_1, \cdots, v_n)$ vanishes if one of its entries is repeated. And lastly, if you shear a box, the volume does not change (this is called Cavalieri's principle in geometry; to demonstrate it, take a deck of cards and push it so that it's tilted. Certainly you will agree the resulting object contains the same volume of paper.)

Instead of searching for a function with precisely these properties (the shearing invariance one in particular is kind of irritating to work with), I'm going to investigate functions satisfying a property that appears stronger at first glance.

Definition 39. Let V be a vector space over $\mathbb{F}$. A **multilinear function** $D : V^m \to \mathbb{F}$ of m variables is a function which takes as input a list $(v_1, \cdots, v_m)$ of vectors in V and produces as output an element $D(v_1, \cdots, v_m) \in \mathbb{F}$, which satisfies the following two additional properties.

(M1) D respects addition in one coordinate at a time:

$$D(v_1, \cdots, v_{i-1}, w+u, v_{i+1}, \cdots, v_n) = D(v_1, \cdots, v_{i-1}, w, v_{i+1}, \cdots, v_n) + D(v_1, \cdots, v_{i-1}, u, v_{i+1}, \cdots, v_n).$$

(M2) D respects scaling in one coordinate at a time:

$$D(v_1, \cdots, v_{i-1}, cv_i, v_{i+1}, \cdots, v_n) = cD(v_1, \cdots, v_{i-1}, v_i, v_{i+1}, \cdots, v_n). \quad \Diamond$$

Said another way, $D : V^m \to \mathbb{F}$ is multilinear if, for all $1 \leqslant i \leqslant n$ and all lists $v_1, \cdots, v_{i-1}, v_{i+1}, \cdots, v_n$, the function

$$w \mapsto D(v_1, \cdots, v_{i-1}, w, v_{i+1}, \cdots, v_n)$$

is a linear function of w. We say D is 'linear in each input'.

Example 80. The most famous example of a bilinear function (multilinear function of two variables) is the dot product $\cdot : \mathbb{F}^n \times \mathbb{F}^n \to \mathbb{F}$, defined as

$$\begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ \cdots \\ y_n \end{pmatrix} = x_1 y_1 + \cdots + x_n y_n.$$

This is linear in each of the two inputs, because for instance

$$\begin{pmatrix} x_1 + x_1' \\ \cdots \\ x_n + x_n' \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ \cdots \\ y_n \end{pmatrix} = (x_1+x_1')y_1 + \cdots + (x_n+x_n')y_n = (x_1y_1 + \cdots + x_ny_n) + (x_1'y_1 + \cdots + x_n'y_n) = \left(\begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} + \begin{pmatrix} x_1' \\ \cdots \\ x_n' \end{pmatrix} \right) \cdot \begin{pmatrix} y_1 \\ \cdots \\ y_n \end{pmatrix}.$$

This shows $(M1)$ for addition in *the first coordinate*, but the same argument applies for the second coordinate. In general, we find that

$$(v + v') \cdot w = v \cdot w + v' \cdot w, \quad (cv) \cdot w = c(v \cdot w), \qquad v \cdot (w + w') = v \cdot w + v \cdot w', \quad v \cdot (cw) = c(v \cdot w).$$

These facts say that the dot product is linear *in each variable separately*; that is, that it is bilinear. $\Diamond$

The example of the dot product has $v \cdot w = w \cdot v$.

Definition 40. Let $D : V^m \to \mathbb{F}$ be a multilinear function of m variables.

- We say D is **symmetric** if the outputs do not depend on the order of its inputs. That is, D is symmetric if
$$D(v_1, \cdots, v_i, \cdots, v_j, \cdots, v_m) = D(v_1, \cdots, v_j, \cdots, v_i, \cdots, v_m)$$
is unchanged after swapping two of its inputs.
- We say D is **alternating** if, whenever $v_i = v_j$ for $i \neq j$, we have $D(v_1, \cdots, v_n) = 0$. $\Diamond$

Lemma 77. *If $D : V^m \to \mathbb{F}$ is alternating and I swap two of its inputs, the output is negated. That is,*

$$D(v_1, \cdots, v_i, \cdots, v_j, \cdots, v_n) = -D(v_1, \cdots, v_j, \cdots, v_i, \cdots, v_n).$$

Further, if I add a multiple of one input to another input, the output is unchanged;

$$D(v_1, \cdots, v_i, \cdots, v_j + cv_i, \cdots, v_n).$$

Proof. Notice that

$$D(v_1, \cdots, v_i + v_j, \cdots, v_i + v_j, \cdots, v_n) = 0$$

because D is alternating, but using multilinearity we see that

$$D(v_1, \cdots, v_i+v_j, \cdots, v_i+v_j, \cdots, v_n) = D(v_1, \cdots, v_i, \cdots, v_i+v_j, \cdots, v_n)+D(v_1, \cdots, v_j, \cdots, v_i+v_j, \cdots, v_n),$$

which simplifies further to

$$D(v_1, \cdots, v_i, \cdots, v_i, \cdots, v_n)+D(v_1, \cdots, v_j, \cdots, v_i, \cdots, v_n)+D(v_1, \cdots, v_j, \cdots, v_i, \cdots, v_n)+D(v_1, \cdots, v_j, \cdots, v_j, \cdots, v_n).$$

Because D is alternating, the first and last terms also vanish Ithey have repeated inputs). Therefore

$$D(v_1, \cdots, v_i, \cdots, v_j, \cdots, v_n) + D(v_1, \cdots, v_j, \cdots, v_i, \cdots, v_n) = 0,$$

so that the first is -1 times the second.

Similarly,

$$D(v_1, \cdots, v_i, \cdots, v_j + cv_i, \cdots, v_n) = D(v_1, \cdots, v_n) + cD(v_1, \cdots, v_i, \cdots, v_i, \cdots, v_n) = D(v_1, \cdots, v_n);$$

in the first equality I used multilinearity on the j'th input, and in the second equality I used that D is alternating and the input v_i is repeated. □

The first part means that if I reorder the inputs of D, the output changes by ± 1 depending on how many times I performed a "swap". For instance, if D is an alternating function of four variables, we have

$$D(v_4, v_1, v_2, v_3) = -D(v_1, v_4, v_2, v_3) = (-1)^2 D(v_1, v_2, v_4, v_3) = (-1)^3 D(v_1, v_2, v_3, v_4),$$

where at each stage I swapped two variables (introducing a minus sign). If you've ever heard of the 'cross product', you've seen this phenomenon before. If not, the first place it will be really meaningful is for the determinant.

Let me quickly record this idea of 'reordering the variables' as a general definition.

Definition 41. Suppose $\sigma : \{1, \cdots, n\} \to \{1, \cdots, n\}$ is a bijection. Such a bijection is often described by listing out the elements of $\{1, \cdots, n\}$ in the order $(\sigma(1), \cdots, \sigma(n))$; this is a list of the numbers from 1 to n in some strange order, such as $(3, 4, 1, 5, 2)$.[1]

We say the **sign** of this bijection is

$$\epsilon(\sigma) = (-1)^{\#\text{ swaps needed to reorder the list to be the standard order}}. \quad \Diamond$$

Such bijections are often called **permutations** of $\{1, \cdots, n\}$, and there are $n! = n(n-1)(n-2)\cdots 2 \cdot 1$ of them.

For instance, we showed above that $\epsilon(4123) = (-1)^3 = -1$, whereas

$$\epsilon(34152) = -\epsilon(14352) = (-1)^2\epsilon(12354) = (-1)^3\epsilon(12345) = (-1)^3.$$

What this discussion amounts to showing is that *if D is an alternating function and you rearrange its inputs, the output changes by ± 1, depending on how many swaps it takes to reorder them back to normal.* That is,

$$D(v_{\sigma(1)}, \cdots, v_{\sigma(n)}) = \epsilon(\sigma) D(v_1, \cdots, v_n).$$

[1] This notation is **not the same as** cycle notation for permutations, if you've ever heard of that. If not, don't worry.

It is not obvious that the number of swaps needed to reorder this list back to standard order is either always odd or always even, so that this sign is well-defined / doesn't depend on the choice of swaps, but this is true. It's not worth our time to prove it[2], so I'm going to move on to the upshot of all of this.

The point is that these properties are *almost* enough to uniquely define the determinant.

Theorem 78. *There exists a unique alternating multilinear function of n variables $D_c : (\mathbb{F}^n)^n \to \mathbb{F}$ for which $D_c(e_1, \cdots, e_n) = c$.*

That is, there is almost exactly one alternating multilinear function of n variables; we have a 1-parameter family of them. If we further specify that $D(e_1, \cdots, e_n)$, then we get a unique such function. Notice that condition (M2) showed up in our discussion of the idea of determinants above, while condition (M1) is closely related to the discussion of Cavalieri's principle and the fact that shearing does not change volume, and the alternating condition is related to our discussion of orientation.

I'll give a proof of this theorem at the end of this section. First, I want to identify some upshots. The first is that only one of these functions actually matters:

Corollary 79. *For any $c \in \mathbb{F}$, we have $D_c(v_1, \cdots, v_n) = cD_1(v_1, \cdots, v_n)$. In particular, $D_0(v_1, \cdots, v_n) = 0$.*

Proof. The function $D(v_1, \cdots, v_n) = cD_1(v_1, \cdots, v_n)$ is an alternating multilinear function because D_1 is and a scalar multiple of a multilinear function is still multilinear, and similarly a scalar multiple of an alternating function is alternating.

Because $D(e_1, \cdots, e_n) = cD_1(e_1, \cdots, e_n) = c$, and there is a unique alternating multilinear function for which $D_c(e_1, \cdots, e_n) = c$, we must have $D = D_c$. Therefore

$$D_c(v_1, \cdots, v_n) = cD_1(v_1, \cdots, v_n)$$

for all $v_1, \cdots, v_n$. □

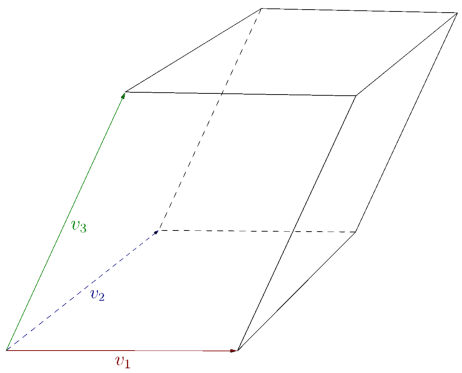
Definition 42. We call the function $D_1 : (\mathbb{F}^n)^n \to \mathbb{F}$ the **determinant**, and write it $\det(v_1, \cdots, v_n)$.

Let $M = \begin{pmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{pmatrix}$ be an $n \times n$ matrix. Its **determinant** is

$$\det(M) = \det(v_1, \cdots, v_n) = \det(Me_1, \cdots, Me_n);$$

that is, take the determinant of the list of the columns of M. ◊

Remark 51. In keeping with the discussion that opened this section, when working over $\mathbb{R}$ you should understand the quantity $\det(v_1, \cdots, v_n)$ as the n-dimensional *signed volume* of the n-dimensional parallelepiped (sheared box) whose sides are $v_1, \cdots, v_n$. The sign encodes the orientation of this box. Notice that if this box has two repeated sides, it is necessarily of dimension $\leqslant (n-1)$, so that this quantity should be zero (the alternating property).



[2]If you want to prove it, let $N(\sigma)$ be the number of pairs $i < j$ so that $\sigma(i) > \sigma(j)$. Then show that if σ' is obtained from σ by a swap, then $N(\sigma) - N(\sigma')$ is odd. Probably not a good use of your time.

(The quantity $\det(v_1, v_2, v_3)$ will be the signed volume of the box above. In this case, the orientation of (v_1, v_2, v_3) is positive, so the signed volume is just the volume.)

Then if M is a matrix, $\det(M)$ is the constant so that M scales volume by $|\det(M)|$ and changes the orientation by the sign of $\det(M)$. This is encoded into the idea above because the unit box (with sides $e_1, \cdots, e_n$) has volume 1, and it is sent to the parallelepiped with sides $Me_1, \cdots, Me_n$, which has volume $\det(Me_1, \cdots, Me_n)$.

Because M should scale the volume of every quantity by the same constant $\det(M)$, this constant must be equal to the volume $\det(Me_1, \cdots, Me_n)$.

I will not be able to justify the claim that the algebraic quantity 'the determinant' is signed volume in any rigorous sense until later, if only because I have failed to give a precise definition of volume. ◊

To rephrase our preceding discussion in terms of matrices, the determinant is the unique function on matrices with the following properties.

- If M' is obtained from M by swapping two of its columns, I have $\det(M') = -\det(M)$. Further, if M has a repeated column, then $\det(M) = 0$.
- If M' is obtained from M by scaling one of its columns by c, then $\det(M') = c\det(M)$.
- Suppose

$$M = \begin{pmatrix} | & & | & & | \\ v_1 & \cdots & w+u & \cdots & v_n \\ | & & | & & | \end{pmatrix} \text{ while } M_w = \begin{pmatrix} | & & | & & | \\ v_1 & \cdots & w & \cdots & v_n \\ | & & | & & | \end{pmatrix} \text{ and } M_u = \begin{pmatrix} | & & | & & | \\ v_1 & \cdots & u & \cdots & v_n \\ | & & | & & | \end{pmatrix}.$$

  Then $\det(M) = \det(M_w) + \det(M_u)$.
- If I is the $n \times n$ identity matrix, we have $\det(I) = 1$.

The first condition is the alternating condition, the next two multilinearity, and the last the condition that $D_1(e_1, \cdots, e_n) = 1$.

Before moving on to actually constructing the determinant, let me point out that these properties show already that the determinant is well-behaved under matrix mutliplication. The following was stated earlier as Theorem 76(i) and half of (ii).

Theorem 80. *If M and N are $n \times n$ matrices, we have $\det(MN) = \det(M)\det(N)$. As a corollary, if M is invertible, we have $\det(M) \neq 0$, and in fact $\det(M^{-1}) = \det(M)^{-1}$.*

Proof. This is kind of a trick using Theorem 78. My intuition is that $\det(MN) = \det(MNe_1, \cdots, MNe_n)$ should be how MN scales volume, so it should be "how M scales volume" times "how N scales volume". I'll try to make this intuition into a proof by thinking of the operation 'determinant of a list I've applied M to', and showing that this is 'determinant of that list times determinant of M'.

Consider the function $D : (\mathbb{F}^n)^n \to \mathbb{F}$ given by

$$D(v_1, \cdots, v_n) = \det(Mv_1, \cdots, Mv_n).$$

Notice that this is multilinear: it is linear in each coordinate, as

$$\begin{aligned} \det(Mv_1, \cdots, M(aw+bu), \cdots, Mv_n) &= \det(Mv_1, \cdots, aMw + bMu, \cdots, Mv_n) \\ &= a\det(Mv_1, \cdots, Mw, \cdots, Mv_n) + b\det(Mv_1, \cdots, Mu, \cdots, Mv_n), \end{aligned}$$

the first equality because multiplication by M is linear and the last equality because determinant is multilinear, and notice that D is alternating, as if $v_i = v_j$, then $Mv_i = Mv_j$ and thus $\det(Mv_1, \cdots, Mv_n) = 0$ as the determinant is alternating.

Thus by the combination of Theorem 78 and Corollary 79 we have

$$D(v_1, \cdots, v_n) = c\det(v_1, \cdots, v_n) \quad \text{where} \quad c = D(e_1, \cdots, e_n) = \det(Me_1, \cdots, Me_n) = \det(M).$$

This gives a formalization of the idea that "Applying M scales the volume of every parallelepiped by the same quantity $\det(M)$." If $(v_1, \cdots, v_n)$ are the sides of a parallelepiped, then $(Mv_1, \cdots, Mv_n)$ are the sides

of the parallelepiped obtained by applying M to my first parallelepiped, and the formula above shows that its volume is $\det(M)$ times the volume of the original.

To conclude, $\det(MN)$ is defined to be $\det(MNe_1, \cdots, MNe_n)$, and we've now established that

$$\det(MNe_1, \cdots, MNe_n) = D(Ne_1, \cdots, Ne_n) = \det(M)\det(Ne_1, \cdots, Ne_n) = \det(M)\det(N),$$

where in the last equality I just recalled that $\det(Ne_1, \cdots, Ne_n)$ is the definition of $\det(N)$.

As for the corollary, suppose M is invertible and set $N = M^{-1}$. Then $MN = I$ is the identity, so $\det(M)\det(M^{-1}) = \det(I) = 1$. Thus $\det(M)$ and $\det(N)$ are both nonzero. Dividing both sides by $\det(M)$ we see that $\det(M^{-1}) = \det(M)^{-1}$. □

All this said and done, let's actually prove that the determinant exists and is uniquely determined by these properties.

6.1.3 Proof of Theorem 78

I will argue first that an alternating multilinear function $D_c : (\mathbb{F}^n)^n \to \mathbb{F}$ must take a very particular form, at which point we will see precisely what form it will take (giving a formula for the determinant).

Let's write the output of some arbitrary list of vectors as

$$D_c\left(\begin{pmatrix} a_{11} \\ \cdots \\ a_{n1} \end{pmatrix}, \cdots, \begin{pmatrix} a_{1n} \\ \cdots \\ a_{nn} \end{pmatrix}\right) = D_c(a_{11}e_1 + \cdots + a_{n1}e_n, \cdots, a_{1n}e_1 + \cdots + a_{nn}e_n).$$

Notice that using multilinearity *once*, we can pull out all the scaling and addition from the first coordinate:

$$D_c(a_{11}e_1 + \cdots + a_{n1}e_n, \cdots, a_{1n}e_1 + \cdots + a_{nn}e_n) = \sum_{i_1=1}^{n} a_{i_1,1} D_c(e_{i_1}, \cdots, a_{1n}e_1 + \cdots + a_{nn}e_n).$$

One may do the same in the second coordinate, and so on through the n'th, so that

$$D_c(a_{11}e_1 + \cdots + a_{n1}e_n, \cdots, a_{1n}e_1 + \cdots + a_{nn}e_n) = \sum_{i_1=1}^{n} \cdots \sum_{i_n=1}^{n} a_{i_1,1} \cdots a_{i_n,n} D_c(e_{i_1}, \cdots, e_{i_n}).$$

That is, we can separate all of these additions and pull all of the scalars out front (so we're scaling each $D_c(e_{i_1}, \cdots, e_{i_n})$ a total of n times).

Now if $(i_1, \cdots, i_n)$ has any repeated entries, we know $D_c(e_{i_1}, \cdots, e_{i_n}) = 0$ by the assumption that D_c is alternating. Otherwise, $(i_1, \cdots, i_n) = \sigma$ is some permutation of $\{1, \cdots, n\}$, and by the alternating property we know that

$$D_c(e_{i_1}, \cdots, e_{i_n}) = \epsilon(\sigma) D_c(e_1, \cdots, e_n) = \epsilon(\sigma)c,$$

swapping terms repeatedly until we end up back the usual ordering, where we already know that $D_c(e_1, \cdots, e_n) = c$ by hypothesis. In this case, the big product may be rewritten as

$$a_{i_1,1} \cdots a_{i_n,n} = a_{\sigma(1),1} \cdots a_{\sigma(n),n}.$$

Altogether, what this shows us is that the function D_c **must be**

$$D_c(a_{11}e_1 + \cdots + a_{n1}e_n, \cdots, a_{1n}e_1 + \cdots + a_{nn}e_n) = c \sum_{\sigma \text{ a permutation of } \{1,\cdots,n\}} \epsilon(\sigma) a_{\sigma(1),1} \cdots a_{\sigma(n),n}.$$

On the other hand, it's straightfoward to verify that this function is indeed multilinear, alternating, and has $D_c(e_1, \cdots, e_n) = c$. This proves both existence and uniqueness, and I'll record the resulting formula for the determinant.

Corollary 81. *The function* $\det(M)$ *can be defined explicitly as follows. If* $M = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$, *then*

$$\det(M) = \sum_{\sigma \text{ a permutation of } \{1,\cdots,n\}} \epsilon(\sigma) a_{\sigma(1),1} \cdots a_{\sigma(n),n}.$$

Visual interpretation. Choosing a permutation σ corresponds to choosing one element $a_{\sigma(j),j}$ from each column so that no column has more than one element. For instance,

$$\begin{pmatrix} & & & * & \\ * & & & & \\ & & * & & \\ & * & & & \\ & & & & * \end{pmatrix}$$

has dots in entries $a_{21}, a_{42}, a_{33}, a_{14}, a_{55}$, so corresponds to the permutation $\sigma = (24315)$. The sign of this permutation is

$$\epsilon(24315) = -\epsilon(14325) = (-1)^2\epsilon(12345) = +1.$$

Whenever we have such an arrangements of dots in our matrix, we take the product of all of those entries, with a sign (+1 if I can swap rows to get to the usual diagonal arrangement with an even number of swaps, −1 if it takes an odd number of swaps). Then we add up over all such arrangements of dots.

This definition is very computable for small matrices, but completely intractable for large-dimensional matrices. (It requires you multiple n things a total of $n!$ times and then add them all up.) But still, let's write down what it says in the small cases.

Example 81. Let M be a 2×2 matrix, $M = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$.

There are two ways to order (12): there is the ordering (12) itself and there is the reverse ordering (21), obtained by a single swap. The determinant of M is

$$\det(M) = \epsilon(12)a_{11}a_{22} + \epsilon(21)a_{12}a_{21} = a_{11}a_{22} - a_{12}a_{21}.$$

If I rename these variables as $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, this expression simplifies to the more familiar $\det(M) = ad - bc$ that you have now worked with many times. ◊

Example 82. Let $M = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$ be a 3×3 matrix. Now there are six ways to permute (123): the ones which require an even number of swaps are (123), (231), (312), corresponding to the dot arrangements

$$\begin{pmatrix} * & & \\ & * & \\ & & * \end{pmatrix}, \quad \begin{pmatrix} & & * \\ * & & \\ & * & \end{pmatrix}, \quad \begin{pmatrix} & * & \\ & & * \\ * & & \end{pmatrix},$$

whereas the ones which require an odd number of swaps are (132), (213), (321), corresponding to dot arrangements

$$\begin{pmatrix} * & & \\ & & * \\ & * & \end{pmatrix}, \quad \begin{pmatrix} & * & \\ * & & \\ & & * \end{pmatrix}, \quad \begin{pmatrix} & & * \\ & * & \\ * & & \end{pmatrix}.$$

Correspondingly, we find

$$\begin{aligned} \det(M) = {} & a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ & - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}. \end{aligned}$$

For instance,

$$\det\begin{pmatrix}1 & 3 & 8\\ 0 & 2 & 4\\ 3 & -1 & -3\end{pmatrix} = 1\cdot 2\cdot(-3) + 3\cdot 4\cdot 3 + 8\cdot 0\cdot(-1) \quad - 1\cdot 4\cdot(-1) - 3\cdot 0\cdot(-3) - 8\cdot 2\cdot 3,$$

which simplifies to $-6 + 36 + 0 + 4 + 9 - 48 = -5$. ◊

Past 3×3 matrices this very quickly becomes a useless formula. In particular, it's not super helpful for computing determinants of arbitrary-dimensional $n \times n$ matrices. Our next goal is to find more useful ways of thinking about and computing this quantity.

Remark 52. It may be helpful to understand the argument in the previous argument in a special case. First, let me run through all of the details for $n = 2$; then let me run through some of the details for $n = 3$.

For $n = 2$, we are trying to determine what $D_c(a_{11}e_1 + a_{21}e_2, a_{12}e_1 + a_{22}e_2)$ is. Expanding using multilinearity, this simplifies to

$$\begin{aligned} D_c(a_{11}e_1 + a_{21}e_2, a_{12}e_1 + a_{22}e_2) &= a_{11}D_c(e_1, a_{12}e_1 + a_{22}e_2) + a_{21}D_c(e_2, a_{12}e_1 + a_{22}e_2)\\ &= a_{11}a_{12}D_c(e_1, e_1) + a_{11}a_{22}D_c(e_1, e_2) + a_{21}a_{12}D_c(e_2, e_1) + a_{21}a_{22}D_c(e_2, e_2). \end{aligned}$$

We then observe that because D_c is supposed to be alternating, we have $D_c(e_1, e_1) = D_c(e_2, e_2) = 0$. So the above expression simplifies to

$$a_{11}a_{22}D_c(e_1, e_2) + a_{12}a_{21}D_c(e_2, e_1).$$

Next, we use the alternating property again to swap e_2 and e_1 to obtain $D_c(e_2, e_1) = -D_c(e_1, e_2)$, so that the whole expression simplifies to $(a_{11}a_{22} - a_{12}a_{21})D_c(e_1, e_2)$, which we can now simplify to $c(a_{11}a_{22} - a_{12}a_{21})$. The quantiy in parentheses is precisely the determinant

$$\det\begin{pmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{pmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

Now in the 3×3 case, suppose we've already gone through the steps of expanding using multilinearity repeatedly, so we've seen that

$$D_c\left(\begin{pmatrix}a_{11}\\ a_{21}\\ a_{31}\end{pmatrix}, \begin{pmatrix}a_{12}\\ a_{22}\\ a_{32}\end{pmatrix}, \begin{pmatrix}a_{13}\\ a_{23}\\ a_{33}\end{pmatrix}\right) = \sum_{i_1=1}^{3}\sum_{i_2=1}^{3}\sum_{i_3=1}^{3} a_{i_1,1}a_{i_2,2}a_{i_3,3}D_c(e_{i_1}, e_{i_2}, e_{i_3}).$$

This is a sum over $27 = 3^3$ terms. But most of them vanish; for instance, $D(e_1, e_1, e_3) = 0$, as it has a repeated term. The only possible choices of (i_1, i_2, i_3) for which $D_c(e_{i_1}, e_{i_2}, e_{i_3})$ is nonzero is when (i_1, i_2, i_3) are all distinct. There are only six such options:

$$(i_1, i_2, i_3) \text{ is one of } (1,2,3),\ (1,3,2),\ (2,1,3),\ (2,3,1),\ (3,1,2),\ (3,2,1).$$

Since we can throw out all of the other terms, our expression simplfies to

$$\begin{aligned} &a_{11}a_{22}a_{33}D_c(e_1, e_2, e_3) + a_{11}a_{32}a_{22}D_c(e_1, e_3, e_2)\\ &+a_{21}a_{32}a_{13}D_c(e_2, e_3, e_1) + a_{21}a_{12}a_{33}D_c(e_2, e_1, e_3)\\ &+a_{31}a_{12}a_{23}D_c(e_3, e_1, e_2) + a_{31}a_{22}a_{13}D_c(e_3, e_2, e_1). \end{aligned}$$

Now notice that

$$D_c(e_1, e_3, e_2) \text{ and } D_c(e_3, e_2, e_1) \text{ and } D_c(e_2, e_1, e_3) \text{ are all equal to } -D_c(e_1, e_2, e_3) = -c,$$

because I can swap two of their entries to get the standard list. On the other hand,

$$D_c(e_2, e_3, e_1) \text{ and } D_c(e_3, e_1, e_2) \text{ are both equal to } (-1)^2 D_c(e_1, e_2, e_3) = +c,$$

as it takes *two* swaps to reorder this list to the standard order. Try visualizing the dot-diagrams I mentioned above and seeing that this is true in that picture, as well.

This finally gives (with some rearranging) that our sum is equal to

$$c(a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{32}a_{21} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}).$$

Setting $c = 1$, this gives the formula for the determinant of a 3×3 matrix. ◊

6.2 Computing with determinants

In the previous section, we investigated a notion called the 'determinant'. This is a function inspired by geometry: if $v_1, \cdots, v_n \in \mathbb{F}^n$, then $\det(v_1, \cdots, v_n)$ is meant to measure the 'oriented volume' of the parallelepiped with sides $v_1, \cdots, v_n$. If M is an $n \times n$ matrix, we defined $\det(M) = \det(Me_1, \cdots, Me_n)$ to be the volume of the parallelepiped spanned by its columns (equivalently, $\det(M)$ is the quantity so that applying M scales volume of some object by $\det(M)$.)

We saw that this notion can be axiomatized: it is linear in each variable (multilinear) and alternating (if an input repeats, the output is zero; if one swaps two inputs, the output negates), and satisfies $\det(e_1, \cdots, e_n) = 1$, or in terms of matrices, $\det(I) = 1$: the determinant of the identity matrix is 1.

We saw that it's multiplicative, in the sence that $\det(MN) = \det(M)\det(N)$. I concluded by giving a formula for this, but one that is awful to use in practice:

$$\det(M) = \sum_{\sigma} \epsilon(\sigma) a_{\sigma(1),1} \cdots a_{\sigma(n),n},$$

where the sum is over all rearrangements $(\sigma(1), \cdots, \sigma(n))$ of the string $(1, 2, \cdots, n)$, and

$$\epsilon(\sigma) = (-1)^{\#\text{ swaps needed to change to the original order}}.$$

Visually, this corresponds to summing over all ways to write a dot in every row of M so that no column has more than one dot, multipling all the entries, and either adding or subtracting (depending on whether or not it takes an even number of row/column swaps to move this dot arrangement to the usual one, with dots down the main diagonal).

In this section I want to introduce two more ways to actually compute determinants in practice, which are often just as (if not more) effective than the definition in terms of a sum over permutations.

6.2.1 Laplace expansion: an inductive definition

The next interpretation of the determinant allows for some much simpler computations. It's not much more than a rephrasing of the definition above, but one that's usually a lot easier to think about. It involves reducing an $n \times n$ determinant to a sum of $(n-1) \times (n-1)$ determinants.

Definition 43. Let M be an $n \times n$ matrix. The "(i,j)'th minor of M" $M_{i,j}$ for the $(n-1) \times (n-1)$ matrix obtained by *removing* the i'th row and j'th column from M:

$$M_{i,j} = \begin{pmatrix} a_{11} & \cdots & a_{1,j-1} & a_{1,j+1} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{i-1,1} & \cdots & a_{i-1,j-1} & a_{i-1,j+1} & \cdots & a_{i-1,n} \\ a_{i+1,1} & \cdots & a_{i+1,j-1} & a_{i+1,j+1} & \cdots & a_{i+1,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{n,j-1} & a_{n,j+1} & \cdots & a_{nn} \end{pmatrix}.$$

◊

Sometimes the (i,j)'th minor is visualized as

$$M_{i,j} = \begin{pmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cancel{a_{i1}} & \cdots & \cancel{a_{ij}} & \cdots & \cancel{a_{in}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nj} & \cdots & a_{nn} \end{pmatrix}.$$

For instance, if $M = \begin{pmatrix} 4 & 3 & 7 & 1 \\ 3 & 1 & 2 & 3 \\ 2 & 2 & 2 & 103 \\ 0 & 3 & 0 & 4 \end{pmatrix}$, its minor $M_{2,4}$ is given by

$$M_{2,4} = \begin{pmatrix} 4 & 3 & 7 & 1 \\ \cancel{3} & \cancel{1} & \cancel{2} & \cancel{3} \\ 2 & 2 & 2 & 103 \\ 0 & 3 & 0 & 4 \end{pmatrix} = \begin{pmatrix} 4 & 3 & 7 \\ 2 & 2 & 2 \\ 0 & 3 & 0 \end{pmatrix}.$$

Here is an observation. The determinant $\det(M) = \sum_\sigma \epsilon(\sigma) a_{\sigma(1),1} \cdots a_{\sigma(n),n}$ is obtained as a sum over all permutations $(\sigma(1), \cdots, \sigma(n))$ of $(1, \cdots, n)$ (or equivalently, is a sum over all ways of putting a dot in each column of M so that no two dots are in the same row). Let's focus on a single column. There are exactly n possibilities for where we put the dot in the j'th column (we place it in row $\sigma(j)$).

For instance, there are exactly four spots we could place a dot in the final column of

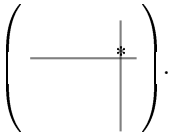
$$M = \begin{pmatrix} 4 & 3 & 7 & 1 \\ 3 & 1 & 2 & 3 \\ 2 & 2 & 2 & 103 \\ 0 & 3 & 0 & 4 \end{pmatrix}.$$

I can break up the big sum defining $\det(M)$ into smaller sums, depending on what integer $1 \leqslant i \leqslant n$ the value of $\sigma(j)$ is:

$$\det(M) = \sum_{\text{permutations } \sigma} \epsilon(\sigma) a_{\sigma(1),1} \cdots a_{\sigma(n),n} = \sum_{i=1}^{n} a_{ij} \sum_{\substack{\text{permutations } \sigma \\ \text{for which } \sigma(j)=i}} \epsilon(\sigma) a_{\sigma(1),1} \cdots a_{\sigma(j-1),j-1} a_{\sigma(j+1),j+1} \cdots a_{\sigma(n),n}.$$

Here I pulled the product term $a_{\sigma(j),j} = a_{ij}$ out to the front of the sum, as it appears in every term there.

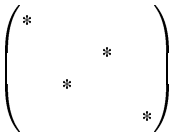
Let's try to understand the smaller sum. In the case of the previous 4×4 matrix, the first sum ranges over $1 \leqslant i \leqslant n$, choosing which entry the dot in the fourth column lies in. If I look at $i = 2$ (so that the dot is in row 2, column 4), all remaining dots must be in the region not crossed off in the following picture:



That is, the dots must be arranged to have one in each row and column of $M_{2,4}$. More generally, if we investigate the dot-arrangements which contain a dot in entry (i, j) (so $\sigma(j) = i$), the remaining dots lie in the smaller matrix $M_{i,j}$ whose entries lie off of row i and column j.

The smaller sum above is almost precisely the same as $\det(M_{i,j})$! It's a sum over ways to arrange dots in each column of $M_{i,j}$ so that each row has exactly one dot, and the sum is over products of the values at each of those dots, together with a sign.

There's only one difference, which is the sign $\epsilon(\sigma)$. This is a rather irritating point, but the issue is (to give an example) in comparing the number of row/column swaps needed to put



into diagonal form (one), and the number of swaps needed to put

$$M_{2,3} = \begin{pmatrix} * & & & \\ & \cancel{*} & & \\ & * & & \\ & & * & \end{pmatrix} = \begin{pmatrix} * & & \\ & * & \\ & & * \end{pmatrix}$$

into diagonal form (zero) — the results are not always the same!

As it turns out, the dots in $M_{i,j}$ can be put into standard form with $i+j$ more swaps than those in the original matrix; if $i+j$ is even, the sign is the same, whereas if $i+j$ is odd, the sign differs by a factor of -1. The argument for this fact isn't important (we will never use the argument again), but I will include it in the following remark, which I **strongly** encourage you skip unless you're exceptionally interested.

Remark 53. Suppose $(\sigma(1),\cdots,\sigma(n))$ is a permutation of $(1,\cdots,n)$ with $\sigma(j)=i$. There is a corresponding permutation $(\sigma'(1),\cdots,\sigma'(n-1))$ of $(1,\cdots,n-1)$, obtained by 'deleting $\sigma(j)$': we set

$$\sigma'(t)=\begin{cases}\sigma(t) & t<j \text{ and } \sigma(t)<i\\ \sigma(t-1) & t>j \text{ and } \sigma(t)<i\\ \sigma(t)-1 & t<j \text{ and } \sigma(t)>i\\ \sigma(t-1)-1 & t>j \text{ and } \sigma(t)>i\end{cases}$$

For instance, if σ is the permutation $(4,5,3,1,2)$, and we are deleting $\sigma(3)=3$, the resulting permutation σ' is $(3,4,1,2)$ (every value larger than 3 is shifted down); if $\sigma=(251364)$ and we are deleting $\sigma(4)=3$, the resulting permutation is $\sigma'=(24153)$: every value larger than 3 gets shifted down.

Now I claim $\epsilon(\sigma')=(-1)^{i+j}\epsilon(\sigma)$. To see this, swap $\sigma(j)=i$ to the back of the permutation, which requires a total of $n-j$ swaps (one for each $j<t\leqslant n$): set

$$\sigma_0=(\sigma(1),\cdots,\sigma(j-1),\sigma(j+1),\cdots,\sigma(n),i),\quad \text{so}\quad \epsilon(\sigma_0)=(-1)^{n-j}\epsilon(\sigma).$$

For instance, for $\sigma=(251364)$, swapping 3 to the back takes two swaps, and $\sigma_0=(251643)$, which has the same sign.

Now σ' is obtained from σ_0 by deleting the last term and decreasing every value larger than $\sigma(j)$. If we perform a swap among all but the last terms in σ_0, this corresponds to the same swap for σ'; and if we perform swaps on σ_0 so that the result is ordered correctly (except for the last term), the same will be true for σ'.

That is, if we write σ_1 for the permutation obtained by ordering all but the last term of σ_0, we have

$$\epsilon(\sigma')=\#\text{ swaps needed to go from }\sigma_0\text{ to }\sigma_1=\epsilon(\sigma_0)\epsilon(\sigma_1).$$

For instance, for $\sigma=(251364)$ where we delete $\sigma(4)=3$, the resulting terms are

$$\sigma'=(24153),\quad \sigma_0=(251643),\quad \sigma_1=(124563).$$

The process of going from σ_0 to σ_1 takes four swaps:

$$\sigma_0=(251643)\to(152643)\to(125643)\to(124653)\to(124563)=\sigma_1.$$

This parallels the corresponding process of ordering σ':

$$\sigma'=(24153)\to(14253)\to(12453)\to(12354)\to(12345).$$

What remains is to determine $\epsilon(\sigma_1)$. Because this is ordered correctly except for the last term $\sigma(j)=i$, it takes $n-i$ swaps to move it into its proper place (past all of the terms $i+1\leqslant t\leqslant n$). For instance, it takes $\sigma_1=(124563)$ a total of three swaps:

$$(124563)\to(124536)\to(124356)\to(123456).$$

Therefore $\epsilon(\sigma_1)=(-1)^{n-i}$. Combining all of this, we see that

$$\epsilon(\sigma')=(-1)^{n-i}\epsilon(\sigma_0)=(-1)^{n-i}(-1)^{n-j}\epsilon(\sigma)=(-1)^{i+j}\epsilon(\sigma).\quad\diamondsuit$$

What we have argued is the following way to reduce an $n\times n$ determinant to the calculation of $(n-1)\times(n-1)$ determinants. I only went through the analysis for columns, but it holds true for rows as well.

Theorem 82 (Laplace expansion)**.** *Let M be an $n \times n$ matrix. Fix a column (say, the j'th column). Then we have*

$$\det(M) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det(M_{i,j}),$$

the signed sum of determinants of the (i,j) minors — the minors corresponding to deleting the j'th column and some other row.

Similarly, if one fixes a row (say, the i'th row), we have

$$\det(M) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(M_{i,j}),$$

Let me apply this in an example.

Example 83. Let's use this discussion to compute $\det(M)$ for $M = \begin{pmatrix} 3 & 0 & 0 & 7 \\ 0 & 1 & 2 & 3 \\ 0 & 1 & 0 & -1 \\ 2 & 0 & 0 & 0 \end{pmatrix}$. I see that the bottom row has almost all zeroes, so it seems productive for Laplace expansion. The Laplace expansion formula gives me

$$\begin{aligned} \det(M) &= (-1)^{4+1} 2 \det(M_{4,1}) + (-1)^{4+2} 0 \det(M_{4,2}) + (-1)^{4+3} 0 \det(M_{4,3}) + (-1)^{4+4} 0 \det(M_{4,4}) \\ &= -2 \det \begin{pmatrix} 3 & 0 & 0 & 7 \\ 0 & 1 & 2 & 3 \\ 0 & 1 & 0 & -1 \\ 2 & 0 & 0 & 0 \end{pmatrix} = -2 \det \begin{pmatrix} 0 & 0 & 7 \\ 1 & 2 & 3 \\ 1 & 0 & -1 \end{pmatrix}. \end{aligned}$$

Now that we're at a 3×3 matrix, I could try doing this by hand; there are only six terms in the relevant sum. But let me try Laplace expansion down the second row (again, almost all zeroes!) to simplify this expression further. I get

$$-2 \det \begin{pmatrix} 0 & 0 & 7 \\ 1 & 2 & 3 \\ 1 & 0 & -1 \end{pmatrix} = -2\big((-1)^{2+1} 0 \det(M_{1,2}) + (-1)^{2+2} 2 \det(M_{2,2}) + (-1)^{2+3} 0 \det(M_{3,2})\big),$$

which simplifies to

$$-4 \det \begin{pmatrix} 0 & 7 \\ 1 & -1 \end{pmatrix} = (-4)(0 \cdot (-1) - 7 \cdot 1) = 28.$$

◊

Example 84. This lets us quickly recover our previous formula for the determinants of 3×3 matrices, in a way which is maybe easier to remember. We have

$$\det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = a_{11} \det \begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} - a_{12} \det \begin{pmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{pmatrix} + a_{13} \begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix},$$

remembering the minus sign that appears in the second term. Expanding this out gives the formula for $\det(M)$ which is a sum over six terms. ◊

In fact, this computation technique can be used to very quickly analyze the determinants of upper-triangular matrices.

Proposition 83. *Let*

$$T = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix}$$

be an upper-triangular matrix. Then $\det(T) = a_{11} a_{22} \cdots a_{nn}$ is the product of the diagonal entries of T.

Proof. I will prove the claim by induction on the size n of the matrix. For 1×1 upper-triangular matrices, the answer is especially boring: $\det\begin{pmatrix} a_{11} \end{pmatrix} = a_{11}$, as promised.

Now suppose (as per my inductive hypothesis) that the stated formula gives the determinant of an $n \times n$ upper-triangular matrix. Using the Laplace expansion (Theorem 82), let's prove that the formula also holds for $(n+1) \times (n+1)$ upper-triangular matrices. To do so, let's expand along the leftmost column (as I notice the only nonzero term there is the top-left entry): I have

$$\det \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1,n+1} \\ 0 & a_{22} & \cdots & a_{2,n+1} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_{n+1,n+1} \end{pmatrix} = \sum_{i=1}^{n+1} (-1)^{i+1} a_{i1} \det(T_{i,1}).$$

Now a_{i1} is only nonzero for $i = 1$, where it is a_{11}, so this simplifies to

$$\det(T) = (-1)^{1+1} a_{11} \det(T_{1,1}) = a_{11} \det(T_{1,1}).$$

Now

$$T_{1,1} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1,n+1} \\ 0 & a_{22} & \cdots & a_{2,n+1} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_{n+1,n+1} \end{pmatrix}, = \begin{pmatrix} a_{22} & \cdots & a_{2,n+1} \\ \cdots & \cdots & \cdots \\ 0 & \cdots & a_{n+1,n+1} \end{pmatrix},$$

and in particular $T_{1,1}$ is an $n \times n$ upper-triangular matrix. By our inductive hypothesis, we know that its determinant is the product of its diagonal entries: $\det(T_{1,1}) = a_{22} \cdots a_{n+1,n+1}$. Combining this with the result of Laplace expansion, we see that

$$\det(T) = a_{11} \det(T_{1,1}) = a_{11} a_{22} \cdots a_{n+1,n+1}.$$

Said another way, we recursively use the Laplace expansion down the first column of each successive matrix. Each time, the determinant is the top-left entry times the determinant of the bottom-right block (because all but one term is zero). □

We can use this to prove a technical result that will be useful momentarily.

Corollary 84. *Let A be a matrix in reduced-row echelon form. Then*

$$\det A = \begin{cases} 1 & A = I \text{ is the identity matrix} \\ 0 & \text{otherwise} \end{cases}.$$

Proof. If A is in reduced-row echelon form, the first nonzero entry in row i must occur in column i or further to the right (prove this by induction, using the fact that leading 1's move right as we descend down the rows). Thus the first j for which a_{ij} is nonzero has $j \geqslant i$. This means precisely that A vanishes below the main diagonal, so A is upper-triangular.

Suppose $\det(A) \neq 0$. This means the entries on the main diagonal of A are all nonzero. As explained above, the first j for which a_{ij} could possibly be nonzero is a_{ii}, in which case it's a leading 1 for that row; by assumption that $a_{ii} \neq 0$, we assume $a_{ii} = 1$ is a leading 1 in row i.

Because $a_{ii} \neq 0$ for all i, there is a leading 1 in every row, in position i. Therefore the i'th column is equal to e_i (if a_{ij} is a leading one in row i, column j, then by definition of reduced row echelon form the j'th column is e_i). Therefore $A = I$ and $\det(A) = 1$.

Thus if A is a reduced-row-echelon-form matrix other than the identity, $\det(A) = 0$. □

6.2.2 Row and column operations

In the previous section we recast the definition of determinant inductively, 'expanding along rows and columns'. In this section I want to explain that it's also amenable to computation using the Gauss–Jordan algorithm[3], and in fact this is the most efficient way (by far!) to compute determinants.

[3]Actually, because we have already computed determinants of upper-triangular matrices, it suffices to use a weaker version of the algorithm where we do not 'cancel out' the terms above leading 1s. This gives a computational speedup by a factor of about 2

Theorem 85. *Suppose M' is an $n \times n$ matrix obtained from the matrix M by an elementary row or column operation. Then the determinant changes in the following ways:*

(i) If M' is obtained by swapping two rows of M or two columns of M, then $\det(M') = -\det(M)$.

(ii) If M' is obtained by scaling a column or row of M by the scalar c, then $\det(M') = c\det(M)$.

(iii) If M' is obtained by adding a multiple of one column to another, or one row to another, then $\det(M') = \det(M)$.

For *columns*, these are contained in my axiomatization of the determinant: it is the unique alternating multilinear function with $\det(e_1, \cdots, e_n) = 1$. The second fact follows immediately from multilinearity:

$$\det(M') = \det(v_1, \cdots, cv_i, \cdots, v_n) = c\det(v_1, \cdots, v_n) = c\det(M).$$

The first and third follow from the fact that det is alternating and from Lemma 77.

To get the same results for *row operations*, there is an algebra trick we can use, which allows us to interchange thinking about row operations and column operations.

Definition 44. Suppose

$$M = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

is an $m \times n$ matrix. Its **transpose** is the $n \times m$ matrix given by interchanging the roles of the columns and rows, or 'flipping it along the top-left to bottom-right diagonal". More formally, its transpose is

$$M^T = \begin{pmatrix} a_{11} & \cdots & a_{m1} \\ a_{12} & \cdots & a_{m2} \\ \cdots & \cdots & \cdots \\ a_{1n} & \cdots & a_{mn} \end{pmatrix}.$$

◊

Notice that a row operation on M^T corresponds to a column operation on M, and vice versa, because the rows of M^T correspond to the columns of M. This matrix operation can be understood in terms of linear maps between vector spaces, but not in an obvious way: see the first part of Curio 4.

Here, the reason M^T is relevant is to prove the theorem above for row operations, by using column operations on the transpose:

Lemma 86. *If M is an $n \times n$ matrix, we have $\det(M^T) = \det(M)$.*

Proof. Write $M = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$. The transpose matrix has $M^T = \begin{pmatrix} a_{11} & \cdots & a_{n1} \\ \cdots & \cdots & \cdots \\ a_{1n} & \cdots & a_{nn} \end{pmatrix}$. To clarify notation in the following computation, it will be convenient to write the entry of M^T in row i and column j by the name b_{ij}, so that $b_{ij} = a_{ji}$.

We have

$$\det(M) = \sum_{\text{permutations } \sigma} \epsilon(\sigma) a_{\sigma(1),1} \cdots a_{\sigma(n),n},$$

while

$$\det(M^T) = \sum_{\text{permutations } \sigma} \epsilon(\sigma) b_{\sigma(1),1} \cdots b_{\sigma(n),n} = \sum_{\text{permutations } \sigma} \epsilon(\sigma) a_{1,\sigma(1)} \cdots a_{n,\sigma(n)}.$$

I claim these two sums are equal. If σ is a permutation, there is a unique *inverse permutation* σ^{-1} so that $\sigma(j) = i \iff \sigma^{-1}(i) = j$. In particular, if $\sigma(j) = i$, then

$$a_{\sigma(j),j} = a_{i,j} = a_{i,\sigma^{-1}(i)}.$$

Because every permutation σ corresponds to exactly one permutation σ^{-1}, I can therefore rewrite the second sum as

$$\det(M^T) = \sum_{\text{permutations } \sigma} \epsilon(\sigma^{-1}) a_{1,\sigma^{-1}(1)} \cdots a_{n,\sigma^{-1}(n)} = \sum_{\text{permutations } \sigma} \epsilon(\sigma^{-1}) a_{\sigma(1),1} \cdots a_{\sigma(n),n}.$$

This is almost the same as the sum defining $\det(M)$, except that one has $\epsilon(\sigma)$ and the other $\epsilon(\sigma^{-1})$. This is rather irritating. The easiest way I have to argue this runs as follows. Consider σ as a bijection $\{1, \cdots, n\} \to \{1, \cdots, n\}$. Then σ^{-1} is the inverse of this bijection, $\sigma^{-1}\sigma = \sigma\sigma^{-1} =$ Identity map. The simplest possible such bijections are swaps σ_{ij}, defined by $\sigma_{ij}(i) = j$ and $\sigma_{ij}(j) = i$ and $\sigma_{ij}(k) = k$ otherwise. Notice that $\sigma_{ij}^{-1} = \sigma_{ij}$.

When we talk about swapping elements of σ until it's in the standard order, this amounts to saying 'write σ as a composition of swaps σ_{ij} and record the number of swaps we compose'. Now if $\sigma = \sigma_{i_1,j_1} \cdots \sigma_{i_k,j_k}$ is the composition of k swaps, then

$$\sigma^{-1} = \sigma_{i_k,j_k}^{-1} \cdots \sigma_{i_1,j_1}^{-1} = \sigma_{i_k,j_k} \cdots \sigma_{i_1,j_1}$$

is also a composition of k swaps, and thus $\epsilon(\sigma^{-1}) = (-1)^k = \epsilon(\sigma)$. This completes the argument. □

Remark 54. The 'transpose' operation will return when we discuss inner products later. It is interesting to observe that $(AB)^T = B^T A^T$. You can either prove this by hand or use the perspective in Curio 4. ◊

This in hand, the theorem follows: a row operation on M is the same as a column operation on M^T; we know these transform correctly under column operations; apply that $\det(M) = \det(M^T)$.

Let me explain how to actually *compute* using these.

Example 85. Take $M = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{pmatrix}$. It would be tremendously painful to compute $\det(M)$ using either the definition or using Laplace expansion. However, I can rapidly compute $\det(M)$ by performing row operations and seeing how the determinant changes as I perform them. First I would carry out some row subtractions,

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{pmatrix} \to \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \end{pmatrix} \to \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Adding or subtracting multiples of one row to another will not change the determinant. Because this matrix has an all-zero row, its determinant must be zero (use Laplace expansion along the bottom row, or use the fact that determinant is linear in each row, as $\det(M) = \det(M^T)$ and the determinant is already known to be linear in each column.)

On the other hand, take $M = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 256 \end{pmatrix}$. Again, you absolutely do not want to try to compute this using the definition! Let's try row-reduction again. A few row-subtractions gives me

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 256 \end{pmatrix} \to \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 3 & 8 & 15 \\ 0 & 7 & 26 & 255 \end{pmatrix} \to \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 2 & 6 \\ 0 & 0 & 12 & 234 \end{pmatrix} \to \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 2 & 6 \\ 0 & 0 & 0 & 198 \end{pmatrix}.$$

This list of row-subtractions (which did not change the determinant!) has reduced is to an upper-triangular matrix, and thus I can immediately read off that its determinant is $\det(M) = 1 \cdot 1 \cdot 2 \cdot 198 = 396$. In particular, M is not invertible (over, say, $\mathbb{R}$). ◊

This allows us to prove the crucial property of matrices.

Corollary 87. *An $n \times n$ matrix M has $\det(M) \neq 0$ if and only if M is invertible.*

Proof. Notice that elementary row operations do not change whether $\det(M)$ is zero or not: swap two rows and you negate $\det(M)$ (which doesn't change whether or not it's zero); scale a row by $c \neq 0$ and the determinant becomes $c\det(M)$, which is zero if and only if $\det(M)$ is zero; and add or subtract a multiple of a row, and the determinant does not change (so the property of whether or not it is zero certainly does not change).

The Gauss–Jordan algorithm shows that I can transform M into a matrix $\operatorname{rref}(M)$ in reduced-row-echelon form with finitely many row operations, so by the preceding discussion $\det M \neq 0 \iff \det \operatorname{rref}(M) \neq 0$. But we saw in Corollary 84 that $\det \operatorname{rref}(M) \neq 0 \iff \operatorname{rref}(M) = I$, so $\det M \neq 0 \iff \operatorname{rref}(M) = I$. But you proved in your homework that $\operatorname{rref}(M) = I$ if and only if M is invertible, so this completes the proof. □

6.3 Diagonal matrices and eigen(things)

It is not infrequent in applications that one has a linear map $A : V \to V$ from a vector space **back to itself** and one wants to *iterate it*, so that we can examine the behavior of $A^n : V \to V$ (the n-fold composition of A with itself) as n gets very large. The most common application of this is where A represents the way some system evolves over time: for instance, see predator-prey models in biology (where an element of $V = \mathbb{R}^3$ measures the population of three different species in a given month, and Av is meant to encode the population of these three species after a month has passed). If we study the behavior of A^n as $n \to \infty$, then we understand the 'limiting state' of these populations: does one go extinct? Do they reach an equilibrium?

Similar applications arise in the study of 'Markov chains' (see also the Google PageRank algorithm, or rather, the publicly available version of this algorithm).

So how do we investigate such a thing? If I choose a basis $\beta = (v_1, \cdots, v_n)$ for V, you showed on your homework that the matrix $[A]_{\beta\to\beta}$ has

$$[A^n]_{\beta\to\beta} = [A]^n_{\beta\to\beta}.$$

So the naive answer is: 'choose a basis for V, then take powers of the corresponding matrix and see what happens as $n \to \infty$." But, as a simple example which shows how awful this is in general, set

$$M = \begin{pmatrix} 3/2 & 0 & -3/2 \\ -1/2 & 2 & 9/2 \\ 0 & -1 & -2 \end{pmatrix}.$$

I encourage you to compute $M^2 = MM$ and $M^3 = M(M^2)$ and see that these computations are quite painful. How are we expected to understand the behavior of M^n when n is large? In fact, I can promise you that (working over the reals so limits make sense) we in fact have

$$\lim_{n\to\infty} M^n = \begin{pmatrix} 3 & 3 & 3 \\ -3 & -3 & -3 \\ 1 & 1 & 1 \end{pmatrix},$$

but how can we possibly see that when the computations are so unpleasant?

The key point is to choose a **smart basis.** Suppose I can find a basis β in which

$$[A]_{\beta\to\beta} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

is diagonal. Finding such a basis is called *diagonalizing* the linear transformation $A : V \to V$. If we can do this, then

$$[A^m]_{\beta\to\beta} = [A]^m_{\beta\to\beta} = \begin{pmatrix} \lambda_1^m & 0 & \cdots & 0 \\ 0 & \lambda_2^m & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n^m \end{pmatrix}.$$

This is much more comprehensible, and very easily computable!

Let's unpack exactly what it is I'm trying to establish. First, recall how to interpret $[A]_{\beta\to\beta}$. The j'th column of this matrix $\begin{pmatrix} a_{1j} \\ \cdots \\ a_{nj} \end{pmatrix}$ is obtained by taking the j'th basis vector v_j and writing

$$Av_j = a_{1j}v_1 + \cdots + a_{nj}v_n$$

as a linear combination of the same basis vectors. If $[A]_{\beta\to\beta}$ has j'th column equal to $\begin{pmatrix} 0 \\ \cdots \\ \lambda_j \\ \cdots \\ 0 \end{pmatrix}$, this means precisely that

$$Av_j = 0v_1 + \cdots + \lambda_j v_j + \cdots + 0v_n = \lambda_j v_j.$$

That is, A merely scales this vector, by the quantity λ_j. A vector with this property (and the scalar quantity it is scaled by) is an important notion, and I want to record it as a definition. Before doing so, let me point out that $A\vec{0} = \lambda\vec{0}$ for *any scalar* λ *whatsoever*, so in the following definition I will need to take care to exclude the zero vector (less this definition have no content).

Definition 45. Let $A : V \to V$ be a linear map from a vector space (over $\mathbb{F}$) to itself. Suppose $v \neq \vec{0}$ is a **nonzero** vector and $\lambda \in \mathbb{F}$ is a scalar such that $Av = \lambda v$, then we say v is an **eigenvector** of A with **eigenvalue** λ.

We say the matrix A is **diagonalizable** if there exists a basis $\beta = (v_1, \cdots, v_n)$ of eigenvectors $Av_i = \lambda_i v_i$, in which case we have that

$$[A]_{\beta\to\beta} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \lambda_n \end{pmatrix}$$

is a diagonal matrix. ♢

Remark 55. The word 'eigenvalue' is a weird mish-mash of German and English. The prefix 'eigen' means "characteristic, own", as in, a something which belongs to A or is characteristic of A. A popular folk etymology says that the term is used because $Av = \lambda v$ says that the vector is merely scaled, so it belongs to its 'own line'. This is not the original usage of the term. A better word might be 'characteristic value' and 'characteristic vector', as these values and vectors tell you an immense amount about the behavior of A. Unfortunately, the Germglish has stuck. ♢

Example 86. If $M = \begin{pmatrix} 3 & -1 \\ 1 & 1 \end{pmatrix}$, I happen to know that the vector $v = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ is an eigenvector: we have

$$A_M v = \begin{pmatrix} 3 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} = 2v.$$

Therefore v is an eigenvector of A_M (or, by an abuse of notation, I will often say 'an eigenvector of M') with associated eigenvalue $\lambda = 2$.

In fact, it turns out that the only eigenvalue of M is $\lambda = 2$, and all eigenvectors are of the form $\begin{pmatrix} a \\ a \end{pmatrix}$ for $a \neq 0$. It may come as some surprise that there is only one eigenvalue, but in general, there are only finitely many — and for an $n \times n$ matrix, no more than n. ♢

The appearance of the vector v in the preceding example came like magic. It is (presumably!) not at all clear how I would have found this vector, so for the rest of this section I'd like to explore how to compute the collection of eigenvalues and eigenvectors of a linear transformation $A : V \to V$.

6.3.1 Finding eigenvalues and eigenvectors

It is usually not so easy to find a particular eigenvector. But if we recast the problem in terms of finding the *entire space of eigenvectors*, then it becomes susceptible to our existing tools.

Definition 46. Let $A : V \to V$ be a linear map and $\lambda \in \mathbb{F}$ be any scalar whatsoever. The **λ-eigenspace** is the set

$$E_\lambda = \{v \in V \mid Av = \lambda v\}.$$

$\Diamond$

Exercise. Check that $E_\lambda \subset V$ is a linear subspace. Then observe that E_λ is precisely the set of λ-eigenvectors *together with the zero vector*.

The reason I think that this subspace is worth focusing on is that it can be quickly rephrased as the *kernel* of a certain linear map. Recall that we have the identity map $1_V : V \to V$ (which I will henceforth write as I for convenience of notation); its scalar multiple $\lambda I : V \to V$ is defined by

$$(\lambda I)(v) = \lambda v.$$

Then the eigenvalue equation can be rewritten and rearranged as

$$Av = \lambda v \quad \iff \quad Av = (\lambda I)v \quad \iff \quad (\lambda I - A)v = \vec{0}.$$

Here the linear map $\lambda I - A$ is defined by $(\lambda I - A)v = \lambda v - Av$ (and you can quickly check that this is indeed linear; in fact, any sum or scalar multiple of linear maps is linear). Further, one may quickly check that if the corresponding matrix is

$$[A]_{\beta\to\beta} = M = \begin{pmatrix} a_{11} & a_{12}\cdots & a_{1n} & \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix},$$

then

$$[\lambda I - A]_{\beta\to\beta} = \lambda I - M = \begin{pmatrix} \lambda - a_{11} & -a_{12} & \cdots & -a_{1n} \\ -a_{21} & \lambda - a_{22} & \cdots & -a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ -a_{n1} & -a_{n2} & \cdots & \lambda - a_{nn} \end{pmatrix}.$$

Thus, what we have stated is

$$E_\lambda = \ker(\lambda I - A),$$

and we have determined the corresponding description at the level of matrices. This is fantastic, for two reasons!

- Using the Gauss–Jordan algorithm, we can quickly compute E_λ for any fixed scalar λ. However, this is still not quite good enough: if we're working over, say, $\mathbb{R}$, there are uncountably many scalars. We can't just compute all of these one by one.
- We now have a great way of determining whether or not E_λ is non-trivial. Because $\lambda I - A$ is an $n \times n$ matrix, it has a determinant, and $\det(\lambda I - A) = 0$ if and only if $\lambda I - A$ is non-invertible. By the invertible map theorem, the map $\lambda I - A$ is non-invertible if and only if $\ker(\lambda I - A) \neq \{\vec{0}\}$, so that $\det(\lambda I - A) = 0$ **if and only if** A has an eigenvector with associated eigenvalue λ.

Let's record the quantity occurring in the second bullet point as a definition; it will be an important character in what follows.

Definition 47. Let M be an $n \times n$ matrix over the field $\mathbb{F}$. Its **characteristic polynomial** is the polynomial $p_M(\lambda) = \det(\lambda I - M)$.

More generally, let $A : V \to V$ be a linear map. Choose a basis $\beta = (v_1, \cdots, v_n)$ for V, and let $M = [A]_{\beta\to\beta}$ be the corresponding $n \times n$ matrix. Then the **characteristic polynomial** of A is the polynomial

$$p_A(\lambda) = \det(\lambda I - M).$$

$\Diamond$

The discussion in the preceding bullet point amounts to a proof of the following fact.

Proposition 88. *Let $A : V \to V$ be a linear map. Then $\lambda_0 \in \mathbb{F}$ is an eigenvalue of A if and only if $p_A(\lambda_0) = 0$, and when this is the case, $E_{\lambda_0} = \ker(\lambda_0 I - A)$.*

This can shed some light on the particular examples I mentioned before:

Example 87. Let $M = \begin{pmatrix} 3 & -1 \\ 1 & 1 \end{pmatrix}$. Then $\lambda I - M = \begin{pmatrix} \lambda - 3 & 1 \\ -1 & \lambda - 1 \end{pmatrix}$, so that

$$p_M(\lambda) = (\lambda - 3)(\lambda - 1) - (-1) = \lambda^2 - 4\lambda + 4.$$

I can factor this as $p_M(\lambda) = (\lambda - 2)^2$, so that the only root of $p_M(\lambda)$ is $\lambda = 2$. It thus follows that the only eigenvalue of M is 2. Let's compute the corresponding eigenspace.

We have

$$E_2 = \ker(2I - M) = \ker \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}.$$

You can either see by inspection that the kernel of this linear map is span $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, or you can run the Gauss–Jordan algorithm to reduce to the rref matrix $\begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix}$ (which has the same kernel, as row operations preserve the kernel of a matrix), where you can read off that the relevant equation is $x_1 - x_2 = 0$.

This matrix is not diagonalizable. I cannot find a basis of eigenvectors, because there is only one eigenvalue, and the eigenspace is 1-dimensional. It is therefore impossible to find two linearly independent eigenvectors. ◊

Now I defined the characteristic polynomial above for general linear maps by *making a choice of basis*. When I make a choice, I have to verify that the resulting quantity didn't depend on that choice — otherwise it does not depend on A, but rather the pair (A, β) of A and a choice of basis. Fortunately, this is not the case.

To check this, I want to use the relation between 'changing basis' and 'passing from M to SMS^{-1}', which you established on your homework: if we choose two bases β and β' for V, then $[A]_{\beta' \to \beta'} = \phi_{\beta \to \beta'} [A]_{\beta \to \beta} \phi^{-1}_{\beta \to \beta'}$, so $M' = SMS^{-1}$ for M' the matrix associated to β', whereas M is the matrix associated to β, and S is the matrix $\phi_{\beta \to \beta'}$.

I will see the latter idea many times. I want to give it a name.

Definition 48. Suppose M and M' are two $n \times n$ matrices. We say that M and M' are **similar** (or **conjugate**) if there exists an invertible $n \times n$ matrix S so that $M' = SMS^{-1}$. ◊

Lemma 89. *Suppose M' and M are conjugate matrices. Then $\det(M') = \det(M)$.*

Proof. If M' is conjugate to M, then $M' = SMS^{-1}$ for some invertible S. Then

$$\det(M') = \det(SMS^{-1}) = \det(S)\det(M)\det(S^{-1}) = \det(S)\det(M)\det(S)^{-1} = \det(S)\det(S)^{-1}\det(M) = \det(M).$$

The second equality uses that determinants are multiplicative, Theorem 80, while the next uses that $\det(S^{-1}) = \det(S)^{-1}$ (contained in that same argument). Next, I used the fact that the determinant is an element of $\mathbb{F}$, and unlike matrix multiplication, products in $\mathbb{F}$ are commutative; thus I can move $\det(S)^{-1}$ 'past' $\det(M)$. The two determinants of S and S^{-1} cancel out, and thus $\det(M') = \det(M)$. □

Proposition 90. *The characteristic polynomial $p_A(\lambda)$ does not depend on the choice of basis β. Equivalently, if M is an $n \times n$ matrix and S is an invertible $n \times n$ matrix, we have $p_M(\lambda) = p_{SMS^{-1}}(\lambda)$.*

Proof. The fact for matrices follows quickly from the previous discussion. We have

$$p_{SMS^{-1}}(\lambda) = \det(\lambda I - SMS^{-1}).$$

The trick is to observe that $\lambda I - M$ and $\lambda I - SMS^{-1}$ are conjugate. In fact,

$$S(\lambda I - M)S^{-1} = S(\lambda I)S^{-1} - SMS^{-1} = \lambda SIS^{-1} - SMS^{-1} = \lambda SS^{-1} - SMS^{-1} = \lambda I - SMS^{-1}.$$

Here the crucial point is that $S(\lambda I)S^{-1}$ is simply λI: scaling by λ commutes with all other linear maps by the definition of linearity:

$$[A(\lambda I)]v = A(\lambda v) = \lambda Av = [(\lambda I)A]v.$$

Thus

$$p_{SMS^{-1}}(\lambda) = \det(\lambda I - SMS^{-1}) = \det(\lambda I - M) = p_M(\lambda).$$

□

To summarize, associated to a linear map A we have an associated polynomial $p_A(\lambda)$. It is defined in terms of matrices, but it doesn't depend on the choice of basis used to turn A into a matrix $[A]_{\beta\to\beta}$. The roots of this polynomial, the 'characteristic polynomial', are precisely the eigenvalues of A. Once we compute the list of eigenvalues $\lambda_1, \cdots, \lambda_k$, we can then use the Gauss–Jordan algorithm to compute the eigenspaces E_{λ_i}. If we're lucky, these are large enough to produce a basis of eigenvectors for A, at which point we've finished our work. We are not always so lucky.

I would like to conclude this subsection with a few examples of diagonalizing matrices. For the first, we will be able to diagonalize the matrix; for the latter two, we will not be able to.

The first example will take a little while, partly because I want to go through at least one somewhat intricate example in full detail.

Example 88. Let the matrix $M = \begin{pmatrix} 3/2 & 0 & -3/2 \\ -1/2 & 2 & 9/2 \\ 0 & -1 & -2 \end{pmatrix}$ be the matrix from the beginning of this section. I claim M is diagonalizable; let's diagonalize it.

First, let's compute the characteristic polynomial. I have

$$\lambda I - M = \begin{pmatrix} \lambda - 3/2 & 0 & 3/2 \\ 1/2 & \lambda - 2 & -9/2 \\ 0 & 1 & \lambda + 2 \end{pmatrix}.$$

Using Laplace expansion along the bottom row, I find its determinant is

$$\det(\lambda I - M) = 0 \det \begin{pmatrix} 0 & 3/2 \\ \lambda - 2 & -9/2 \end{pmatrix} - 1 \det \begin{pmatrix} \lambda - 3/2 & 3/2 \\ 1/2 & -9/2 \end{pmatrix} + (\lambda + 2) \det \begin{pmatrix} \lambda - 3/2 & 0 \\ 1/2 & \lambda - 2 \end{pmatrix},$$

(check the statement of Laplace expansion and confirm I wrote this correctly, including the signs!) which simplifies to

$$(\lambda + 2)(\lambda - 3/2)(\lambda - 2) - (\lambda - 3/2)(-9/2) + (3/2)(1/2) = \lambda^3 - \frac{3}{2}\lambda^2 + \frac{1}{2}\lambda.$$

I can factor this completely as

$$p_M(\lambda) = \lambda(\lambda - 1)(\lambda - 1/2),$$

so there are three distinct eigenvalues, $\{0, 1/2, 1\}$.

Let's compute the eigenspaces. Using row reduction, I find

$$E_0 = \ker(-M) = \ker(M) = \ker \begin{pmatrix} 3/2 & 0 & -3/2 \\ -1/2 & 2 & 9/2 \\ 0 & -1 & -2 \end{pmatrix} = \ker \begin{pmatrix} 1 & 0 & -1 \\ -1/2 & 2 & 9/2 \\ 0 & -1 & -2 \end{pmatrix}$$
$$= \ker \begin{pmatrix} 1 & 0 & -1 \\ 0 & 2 & 4 \\ 0 & -1 & -2 \end{pmatrix} = \ker \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & -1 & -2 \end{pmatrix} = \ker \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

This matrix is in reduced row echelon form, and I can quickly read off that

$$\ker \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix} = \operatorname{span} \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

This is a 0-eigenvector.

Next, I find

$$E_{1/2} = \ker(\frac{1}{2}I - M) = \ker\begin{pmatrix} -1 & 0 & 3/2 \\ 1/2 & -3/2 & -9/2 \\ 0 & 1 & 5/2 \end{pmatrix} = \ker\begin{pmatrix} 1 & 0 & -3/2 \\ 1/2 & -3/2 & -9/2 \\ 0 & 1 & 5/2 \end{pmatrix}$$
$$= \ker\begin{pmatrix} 1 & 0 & -3/2 \\ 0 & -3/2 & -15/4 \\ 0 & 1 & 5/2 \end{pmatrix} = \ker\begin{pmatrix} 1 & 0 & -3/2 \\ 0 & 1 & 5/2 \\ 0 & 1 & 5/2 \end{pmatrix} = \ker\begin{pmatrix} 1 & 0 & -3/2 \\ 0 & 1 & 5/2 \\ 0 & 0 & 0 \end{pmatrix}.$$

I can now read off from this rref matrix that

$$E_{1/2} = \ker\begin{pmatrix} 1 & 0 & -3/2 \\ 0 & 1 & 5/2 \\ 0 & 0 & 0 \end{pmatrix} = \text{span}\begin{pmatrix} 3/2 \\ -5/2 \\ 1 \end{pmatrix} = \text{span}\begin{pmatrix} 3 \\ -5 \\ 2 \end{pmatrix}.$$

Lastly, I find

$$E_1 = \ker(I - M) = \ker\begin{pmatrix} -1/2 & 0 & 3/2 \\ 1/2 & -1 & -9/2 \\ 0 & 1 & 3 \end{pmatrix} = \ker\begin{pmatrix} 1 & 0 & -3 \\ 1/2 & -1 & -9/2 \\ 0 & 1 & 3 \end{pmatrix}$$
$$= \ker\begin{pmatrix} 1 & 0 & -3 \\ 0 & -1 & -3 \\ 0 & 1 & 3 \end{pmatrix} = \ker\begin{pmatrix} 1 & 0 & -3 \\ 0 & 1 & 3 \\ 0 & 1 & 3 \end{pmatrix} = \ker\begin{pmatrix} 1 & 0 & -3 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{pmatrix}.$$

From here I can read off that

$$E_1 = \ker\begin{pmatrix} 1 & 0 & -3 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{pmatrix} = \text{span}\begin{pmatrix} 3 \\ -3 \\ 1 \end{pmatrix}.$$

Now you can check that the list

$$\beta = \left(\begin{pmatrix} 3 \\ -3 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 3 \\ -5 \\ 2 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \right)$$

is a basis for $\mathbb{F}^3$, and these are eigenvectors of M. In the basis β, we have

$$[M]_{\beta\to\beta} = D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and $D = \phi_{\text{std}\to\beta} M \phi_{\beta\to\text{std}}$. By definition $\phi_{\beta\to\beta'} = C_{\beta'}^{-1} C_\beta$, so $\phi_{\beta\to\text{std}} = C_\beta$ (as C_{std} is the identity map). Therefore $D = S^{-1}MS$ where $S = \begin{pmatrix} 3 & 3 & 1 \\ -3 & -5 & -2 \\ 1 & 2 & 1 \end{pmatrix}$, and you compute (using, say, the algorithm on Homework 6 #5) that

$$S^{-1} = \begin{pmatrix} 1 & 1 & 1 \\ -1 & -2 & -3 \\ 1 & 3 & 6 \end{pmatrix}.$$

Switching terms around, we find $M = SDS^{-1}$ for this matrix S, or written out,

$$\begin{pmatrix} 3 & 3 & 1 \\ -3 & -5 & -2 \\ 1 & 2 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 \\ -1 & -2 & -3 \\ 1 & 3 & 6 \end{pmatrix}.$$

Earlier I mentioned above that M^n behaves well as $n \to \infty$. I can see this from the perspective of this problem. Notice that

$$M^n = (SDS^{-1})^n = SDS^{-1}SDS^{-1}\cdots SDS^{-1};$$

all the pairs $S^{-1}S$ cancel out to give the identity, and this reduces to $M^n = SD^nS^{-1}$. (This is the content of HW6 #7.) Now observe that

$$M^n = \begin{pmatrix} 3 & 3 & 1 \\ -3 & -5 & -2 \\ 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2^n & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ -1 & -2 & -3 \\ 1 & 3 & 6 \end{pmatrix}$$

clearly has a limit as $n \to \infty$: it limits to

$$\begin{pmatrix} 3 & 3 & 1 \\ -3 & -5 & -2 \\ 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ -1 & -2 & -3 \\ 1 & 3 & 6 \end{pmatrix} = \begin{pmatrix} 3 & 3 & 3 \\ -3 & -3 & -3 \\ 1 & 1 & 1 \end{pmatrix},$$

as promised. ◊

Now let me do some much shorter non-examples.

Example 89. Earlier I mentioned that the matrix $M = \begin{pmatrix} 3 & -1 \\ 1 & 1 \end{pmatrix}$ was not diagonalizable. We did find at least one eigenvector: $v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ was an eigenvector with eigenvalue 2. I can't complete this to a basis of eigenvectors, as discussed earlier. However, let's see what happens when I choose v_2 arbitrarily (let's say $v_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ to keep things simple). Then $Mv_1 = 2v_1$, whereas we can express Mv_2 in terms of this basis as

$$Mv_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix} = -1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + 2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = -v_1 + 2v_2,$$

so that if $\beta = (v_1, v_2)$, we have

$$[M]_{\beta \to \beta} = \begin{pmatrix} 2 & -1 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 3 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

This is not a diagonal matrix, but it is at least upper-triangular, and relatively easy to compute powers of. (If you want, you can compute inductively a formula for T^n, where T is the matrix $[M]_{\beta \to \beta}$.) Just like in the previous problem, I determined the matrix on the right as $\phi_{\beta \to \text{std}} = C_\beta$ as the matrix whose columns are the vectors of β.

We will see later that while not every matrix can be diagonalized, **over the complex numbers** every matrix can be made upper-triangular. ◊

Example 90. On the other hand, over the reals, it's possible that matrices do not even have a single eigenvalue. Consider, for instance, $M = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. This matrix represents $\text{rot}_{\pi/2}$, rotation of the plane counter-clockwise by angle $\pi/2$.

You should compute that the characteristic polynomial is $p_M(\lambda) = \lambda^2 + 1$, which does not have a single real root, as for real numbers λ we have $\lambda^2 + 1 \geqslant 0$. This matrix does not have a single real eigenvalue.

This corresponds to the visual fact that if I rotate a nonzero vector, it points in a different direction than it started. In fact, rot_θ never has any real eigenvalues for any $0 < \theta < \pi$; the argument is similar.

On the other hand, if I consider this a matrix **over** $\mathbb{C}$, so that M corresponds to a linear map $A_M : \mathbb{C}^2 \to \mathbb{C}^2$, then it *is* diagonalizable: we have two roots $\lambda = i, -i$, for which

$$\ker(iI - M) = \ker \begin{pmatrix} i & 1 \\ -1 & i \end{pmatrix} = \ker \begin{pmatrix} 1 & -i \\ -1 & i \end{pmatrix} = \ker \begin{pmatrix} 1 & -i \\ 0 & 0 \end{pmatrix} = \text{span} \begin{pmatrix} i \\ 1 \end{pmatrix},$$

whereas

$$\ker(-iI - M) = \ker \begin{pmatrix} -i & 1 \\ -1 & -i \end{pmatrix} = \ker \begin{pmatrix} 1 & i \\ -1 & -i \end{pmatrix} = \ker \begin{pmatrix} 1 & i \\ 0 & 0 \end{pmatrix} = \text{span} \begin{pmatrix} -i \\ 1 \end{pmatrix}.$$

Therefore M *is* diagonalizable over the complex numbers in the basis given by the two eigenvectors identified above; over the complex numbers its diagonalization is given by

$$D = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}.$$

The relationship between this and M is given by

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} i & -i \\ 1 & 1 \end{pmatrix} \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \begin{pmatrix} i & -i \\ 1 & 1 \end{pmatrix}^{-1}.$$

This is algebraically useful, but geometrically very difficult to parse! This refers to the behavior of this linear map on $\mathbb{C}^2$ (which is too large for our feeble 3-dimensional brains to visualize), whereas the corresponding linear transformation of $\mathbb{R}^2$ has no eigenvalues whatsoever. ◊

6.3.2 Additional properties of the characteristic polynomial

Before concluding this section, I would like to discuss some important but simple facts about the characteristic polynomial.

Lemma 91. *Let $A : V \to V$ be a linear map on a vector space of* $\dim V = n$*. The characteristic polynomial is a polynomial of degree n, with*

$$p_A(\lambda) = \lambda^n + a_1\lambda^{n-1} + \cdots + a_n.$$

Proof. Choose an $n \times n$ matrix M representing $[A]_{\beta\to\beta}$ with respect to some basis β. We have

$$p_A(\lambda) = \det \begin{pmatrix} \lambda - a_{11} & -a_{12} & \cdots & -a_{1n} \\ -a_{21} & \lambda - a_{22} & \cdots & -a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ -a_{n1} & -a_{n2} & \cdots & \lambda - a_{nn} \end{pmatrix},$$

where the a_{ij} are the coefficients in the matrix M.

The determinant is a sum of products of entries of $\lambda I - M$, where the sum is over ways to choose entries in each column so that there is exactly one in each row.

Here, we are multiplying n terms which either look like $\lambda - a_{ii}$ or which look like $-a_{ij}$ for $i \neq j$. The only way for this product of n terms to produce a power of λ^n is if all n terms are of the form $\lambda - a_{ii}$. This is the product corresponding to the 'diagonal' arrangement of dots $\begin{pmatrix} * & & & \\ & * & & \\ & & \cdots & \\ & & & * \end{pmatrix}$, which introduces no sign; this contributes

$$(\lambda - a_{11}) \cdots (\lambda - a_{nn}) = \lambda^n + \cdots$$

to the sum, with exactly one λ^n term and all terms of lower λ-degree. □

One interesting fact is that we can identify the coefficients of the characteristic polynomial. Two are particularly well-known.

Lemma 92. *The coefficient a_n of the characteristic polynomial of an $n\times n$ matrix M is given by $(-1)^n \det(M)$. The coefficient a_1 is a quantity called $-tr(M)$, where*

$$tr \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} = a_{11} + \cdots + a_{nn}$$

is the sum of the diagonal entries, called the 'trace'.

Proof. As $p_M(\lambda) = \lambda^n + \cdots + a_{n-1}\lambda + a_n$, we see that $p_M(0) = a_n$. Now

$$p_M(0) = \det(-M) = \det \begin{pmatrix} -a_{11} & \cdots & -a_{1n} \\ \cdots & \cdots & \cdots \\ -a_{n1} & \cdots & -a_{nn} \end{pmatrix}.$$

To compare this to the determinant of M, I want to pull out a scalar of -1 from each column. There are a total of n columns, and each time I pull out a scalar of -1, the determinant scales by -1; doing so a total of n times, I find $\det(-M) = (-1)^n \det(M)$.

Alternatively, observe that $-M = (-I)M$, and that $\det(-I) = \det \begin{pmatrix} -1 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & -1 \end{pmatrix} = (-1)^n$.

As for the statement about the coefficient of λ^{n-1}, this is a bit more subtle, and requires going back to the 'dot-diagram' argument above. Let me recall the determinant

$$\det \begin{pmatrix} \lambda - a_{11} & -a_{12} & \cdots & -a_{1n} \\ -a_{21} & \lambda - a_{22} & \cdots & -a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ -a_{n1} & -a_{n2} & \cdots & \lambda - a_{nn} \end{pmatrix}$$

in terms of the sum over certain products of its entries.

Above, I discussed one important input to this:

$$(\lambda - a_{11}) \cdots (\lambda - a_{nn}) = \lambda^n - (a_{11} + \cdots + a_{nn})\lambda^{n-1} + \cdots$$

the rest involving terms of the form λ^{n-2} or lower. There's our trace! What I need to establish is that no other products contribute to the λ^{n-1} term.

Suppose I take a different dot-diagram, with at least one entry off the diagonal (say one of the terms I take a product of lies in entry (i, j) with $i \neq j$). *Then at least two entries are off of the diagonal:* for instance, $\begin{pmatrix} * & & & & \\ & & * & & \\ & * & & & \\ & & & * & \\ & & & & * \end{pmatrix}$. Because there is only one entry in row i, there cannot be a dot in entry (i, i); because there is only one entry in column j, there cannot be a dot in entry (j, j). In the example, $(i, j) = (2, 3)$; there is no dot in entry $(2, 2)$ or $(3, 3)$.

Thus such a dot-diagram contributes a product of at most $n - 2$ terms of the form $(\lambda - a_{ii})$ and at least 2 terms of the form $-a_{ij}$. As a result, this can only contribute terms with λ-degree at most $n - 2$; they do not contribute to the coefficient of λ^{n-1}. □

There is a similar formula for the coefficients a_k in terms of a sum over 'determinants of $k \times k$ minors'. It is not used very often and I will not discuss it further.

Remark 56. The trace is a remarkable and very useful quantity. In some sense, it is the *derivative* of the determinant. Maybe we will make this precise next term. Maybe not. One interesting fact about the trace is that for any two $n \times n$ matrices M and N, we have $\mathrm{tr}(MN) = \mathrm{tr}(NM)$. (There is no product formula for the trace of a product of two matrices, so you will not be able to use any such formula to prove this fact.)

This can be proved by writing down explicitly what these two quantities are and verifying that they are the same. Interestingly, you can use this to show that $\mathrm{tr}(SMS^{-1}) = \mathrm{tr}(M)$, though this is not obvious from the definition of trace.

As a fun little exercise, you can try using the trace to prove that there over $\mathbb{R}$ there are no matrices A and B for which $AB - BA = I$ is the identity. The argument fails over $\mathbb{F}_2$, and it might be fun to find 2×2 matrices A, B over $\mathbb{F}_2$ for which $AB - BA = I$ actually holds. ◊

There is one more thing worth saying. We can usually compute the determinant and trace in terms of the *eigenvalues* of a matrix.

Definition 49. Suppose $p(\lambda) = \lambda^n + a_1\lambda^{n-1} + \cdots + a_n$ is a polynomial over $\mathbb{F}$. We say $p(\lambda)$ **splits into linear factors** if there exist $\lambda_1, \cdots, \lambda_k \in \mathbb{F}$ and integers $m_1, \cdots, m_k \geqslant 1$ for which

$$p(\lambda) = (\lambda - \lambda_1)^{m_1} \cdots (\lambda - \lambda_k)^{m_k}.$$

Then we say that $p(\lambda)$ has roots $\lambda_1, \cdots, \lambda_k$ with **multiplicity** $m_1, \cdots, m_k$. ◊

Remark 57. Notice that $n = m_1 + \cdots + m_k$. ◊

Example 91. For instance, the polynomial $p(\lambda) = \lambda^2(\lambda-1)^2(\lambda-2)$ has three roots, $0, 1, 2$. The roots 0 and 1 appear with multiplicy two, while the root 2 appears with multiplicity 1. This polynomial splits into linear factors.

However, the polynomial $p(\lambda) = \lambda^2 + 1$, considered as a polynomial over $\mathbb{R}$, has no roots. Therefore, it does not split into linear factors. On the other hand, it does split into linear factors over the complex numbers: $p(t\lambda = (\lambda - i)(\lambda + i)$, so $p(\lambda)$ has roots $i, -i$, each with multiplicity 1. ◊

The following proposition helps me really get a sense for what determinants do in terms of 'scaling volume'.

Proposition 93. *Suppose M is an $n \times n$ matrix over $\mathbb{F}$ for which $p_A(\lambda)$ splits into linear factors $\lambda_1, \cdots, \lambda_k$ with multiplicity $m_1, \cdots, m_k$. (We say the eigenvalue λ_i has algebraic multiplicity m_i.) Then*

$$\det(M) = \lambda_1^{m_1} \cdots \lambda_k^{m_k} \quad \textit{and} \quad \textit{tr}(M) = m_1\lambda_1 + \cdots + m_k\lambda_k.$$

Proof. We have

$$p_A(\lambda) = (\lambda - \lambda_1)^{m_1} \cdots (\lambda - \lambda_k)^{m_k}.$$

It is perhaps helpful to write this as the product

$$(\lambda - c_1) \cdots (\lambda - c_n),$$

where $c_1, \cdots, c_{m_1}$ are all equal to λ_1, while $c_{n-m_k+1}, \cdots, c_n$ are all equal to λ_k. This can be understood as a sum over 2^n products, where for each term in the sum I either multiple a 'λ' factor or a '$(-c_i)$' factor.

To get a term with coefficient λ^0, I must take a product over all of the $(-c_i)$ terms and no λ terms, ultimately giving

$$a_n = (-1)^{m_1}\lambda_1^{m_1} \cdots (-1)^{m_k}\lambda_k^{m_k} = (-1)^n\lambda_1^{m_1} \cdots \lambda_k^{m_k}.$$

Because the last coefficient of the characteristic polynomial is $a_n = (-1)^n \det(M)$, this gives the claimed result.

On the other hand, to get a term with coefficient λ^{n-1}, I must take a product over all but one λ term and only one $-c_i$ term. This gives me

$$(-c_1 - \cdots - c_n)\lambda^{n-1} = (-m_1\lambda_1 - \cdots - m_k\lambda_k)\lambda^{n-1},$$

as each λ_i appears m_i times among the c's. Because this coefficient of the characteristic polynomial is $a_1 = -\mathrm{tr}(M)$, this gives the claimed result.

Interestingly, all other coefficients a_k of the characteristic polynomial can also be described in this way but by a more careful argument: the coefficient is $(-1)^k$ times the 'kth symmetric polynomial' of the eigenvalues $\lambda_1, \cdots, \lambda_1, \cdots, \lambda_k, \cdots, \lambda_k$, where I list each eigenvalue λ_i a total of m_i times. □

Example 92. For the matrix $M = \begin{pmatrix} 3 & -1 \\ 1 & 1 \end{pmatrix}$, the characteristic polynomial was $(\lambda - 2)^2$, so 2 appears as an eigenvalue with multiplicity two. The preceding formula gives us $\det(M) = 2^2 = 4$ (correct) and $\mathrm{tr}(M) = 2 \cdot 2$ (correct).

For the matrix $M = \begin{pmatrix} 3/2 & 0 & -3/2 \\ -1/2 & 2 & 9/2 \\ 0 & -1 & -2 \end{pmatrix}$, we found that the characteristic polynomial is

$$p_M(\lambda) = \lambda^3 - \frac{3}{2}\lambda^2 + \frac{1}{2}\lambda = \lambda(\lambda - 1)(\lambda - 1/2).$$

The sum of the eigenvalues is $3/2$, which is indeed equal to $\mathrm{tr}(M) = 3/2 + 2 - 2 = 3/2$; the product of the eigenvalues is zero, which is indeed the determinant of M (as M is not invertible). ◊

Suppose M is diagonalizable. Then the corresponding diagonal matrix is

$$D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_k & 0 \\ 0 & 0 & \cdots & 0 & \lambda_k \end{pmatrix},$$

where λ_1 appears m_1 times, and so on, through λ_k appearing m_k times. This matrix scales m_i different coordinates by λ_i; each of these scales volume by λ_i. Thus each eigenvalue represents a 'stretch factor' of a different direction. Overall, we have scaled m_i different directions by λ_i, which overall scales volume by the product $\lambda_1^{m_1} \cdots \lambda_k^{m_k}$ appearing in the theorem above.

The geometric interpretation of trace is substantially more subtle, so I will not try.

6.4 Diagonalizing a matrix

Let's recall our big goal. We have a linear map $A : V \to V$ from a finite-dimensional vector space (over $\mathbb{F}$) to itself. We want to find a basis for V consisting entirely of eigenvectors, in which case $[A]_{\beta\to\beta}$ will be represented by a diagonal matrix, with diagonal entries λ_i the eigenvalue associated to the i'th basis vector.

So far, we have defined a polynomial $p_A(\lambda)$, the characteristic polynomial of A If $p_A(\lambda)$ is the characteristic polynomial, the eigenvalues $\lambda_1, \cdots, \lambda_k$ are its roots. The λ_i-eigenvectors are precisely the nonzero vectors in the eigenspace $E_{\lambda_i} = \ker(\lambda_i I - A)$.

It suffices to determine whether or not the eigenvectors *span* V, because if so, we can take a spanning set of eigenvectors and trim it to a basis by the basis reduction lemma. The span of the eigenvectors of V is precisely

$$\text{span of all eigenvectors } = E_{\lambda_1} + \cdots + E_{\lambda_k} = \left\{ v \in V \mid v = a_1 v_1 + \cdots + a_k v_k, \quad v_i \in E_{\lambda_i} \right\}.$$

If $E_{\lambda_1} + \cdots + E_{\lambda_k} = V$, we win! The span is everything; we can trim a spanning set of eigenvectors into a basis. If $E_{\lambda_1} + \cdots + E_{\lambda_k} \subsetneq V$ is a proper subspace, we lose: the eigenvectors do not even span V, so we certainly cannot find a basis of eigenvectors.

Goal: If $A : V \to V$ is a linear map with eigenvalues $\lambda_1, \cdots, \lambda_k$, we want to find a way to compute $\dim(E_{\lambda_1} + \cdots + E_{\lambda_k})$. If this is equal to $\dim V$, then $E_{\lambda_1} + \cdots + E_{\lambda_k} = V$ by Theorem 41; if not, they cannot possibly be equal.

Let me start with some observations about the way the different eigenspaces interact, before discussing the individual eigenspaces. The eigenspaces satisfy a useful property. (The following property does not have a standard name in the literature, so I chose one that should feel natural, given the result to follow.)

Definition 50. Let $W_1, \cdots, W_k \subset V$ be linear subspaces of V. We say that these subspaces are **independent** if, for all $1 < i \leqslant k$, we have

$$W_i \cap \left(W_1 + \cdots + W_{i-1} \right) = \{\vec{0}\}.$$

$\Diamond$

I mention these for two reasons. First, the E_λ's satisfy this property:

Lemma 94. *If $\lambda_1, \cdots, \lambda_k$ are the eigenvalues of $A : V \to V$, then the subspaces $E_{\lambda_1}, \cdots, E_{\lambda_k}$ are independent. That is, for all $1 < i \leqslant k$, we have*

$$E_{\lambda_i} \cap \left(E_{\lambda_1} + \cdots + E_{\lambda_{i-1}} \right) = \{\vec{0}\}.$$

Proof. You'll verify this on your homework, but try writing a proof right now! It is a good exercise in understanding the definitions. $\square$

Further, this property helps me compute dimensions.

Lemma 95. *If $W_1, \cdots, W_k$ are independent subspaces of V, then* $\dim(W_1 + \cdots + W_k) = \dim(W_1) + \cdots + \dim(W_k)$. *In fact, if $(v_{1,1}, \cdots, v_{1,\dim W_1}), \cdots, (v_{k,1}, \cdots, v_{k,\dim W_k})$ are bases for $W_1, \cdots, W_k$, then*

$$(v_{1,1}, \cdots, v_{1,\dim W_1}, \cdots, v_{k,1}, \cdots, v_{k,\dim W_k})$$

is a basis for $W_1 + \cdots + W_k$.

Proof. This is proved by induction. The base case $k = 2$ was #6 on the recent midterm; there, you further assumed that $V = W_1 + W_2$, but that was not necessary to prove $\dim(W_1 + W_2) = \dim(W_1) + \dim(W_2)$. Inductively, suppose the claim about bases is true for a sum of k independent subspaces; we will prove it true for $k + 1$ independent subspaces. But if I abbreviate $W = W_1 + \cdots + W_k$, then

$$W_1 + \cdots + W_k + W_{k+1} = W + W_{k+1}.$$

The definition of 'independent subspaces' means precisely that $W_{k+1} \cap W = \{\vec{0}\}$. Applying the $k = 2$ case once more, we see that

$$\dim(W_1 + \cdots + W_{k+1}) = \dim(W + W_{k+1}) = \dim(W) + \dim(W_{k+1}) = \dim(W_1) + \cdots + \dim(W_{k+1}),$$

and similarly one can extract the more precise claim about bases. □

Therefore

$$\dim\left(E_{\lambda_1} + \cdots + E_{\lambda_k}\right) = \dim\left(E_{\lambda_1}\right) + \cdots + \dim\left(E_{\lambda_k}\right).$$

I now need to get an understanding of each one of these terms.

Recall from earlier the notion of *algebraic multiplicity* of an eigenvalue. Let me restate the definition slightly more generally than I used it earlier. If λ_i is a root of $p_A(\lambda)$, we say that its **algebraic multipicity** is the largest m_i so that $p_A(\lambda)$ can be factored as

$$p_A(\lambda) = (\lambda - \lambda_i)^{m_i} q(\lambda),$$

where $q(\lambda)$ is another polynomial, which necessarily has $q(\lambda_i) \neq 0$, or else we could pull out one more $(\lambda - \lambda_i)$ factor.

This definition allows for the possibility that p_A cannot be fully split into linear factors. For instance, in the polynomial $p_A(\lambda) = \lambda^4 - 2\lambda^3 + 2\lambda^2$ considered over the reals, we can factor

$$p_A(\lambda) = (\lambda^2 + 1)(\lambda^2 - 2\lambda + 1) = (\lambda^2 + 1)(\lambda - 1)^2,$$

so that $p_A(\lambda)$ has one real root ($\lambda_1 = 1$) with multiplicity $m_1 = 2$. If I work over $\mathbb{C}$, this can be factored further as

$$p_A(\lambda) = (\lambda - 1)^2(\lambda - i)(\lambda + i), \quad \text{so that over } \mathbb{C}, \text{ we have } \lambda_1 = 1,\ \lambda_2 = i,\ \lambda_3 = -i,$$

with multiplicities $m_1 = 2,\ m_2 = 1,\ m_3 = 1$.

The following propositions shows that this algebraic multiplicity 'controls' the dimension of E_λ (or rather, bounds it).

Lemma 96. *Let $A : V \to V$ be a linear map. Suppose λ_0 is an eigenvalue of A, with algebraic multiplicity m. Then $1 \leqslant \dim E_{\lambda_0} \leqslant m$.*

Proof. The inequality $1 \leqslant \dim E_{\lambda_0}$ is straightforward: to say λ_0 is an eigenvector of A means that there exists some nonzero vector $v \in E_{\lambda_0}$, hence $\{v\}$ is a linearly independent subset of E_{λ_0} with at least one element.

Ultimately, I have to refer back to both bases and determinants somehow: the notion of 'algebraic multiplicity' refers to the characteristiic polynomial $p_A(\lambda)$, which is defined as $\det(\lambda I - M)$, where $M = [A]_{\beta \to \beta}$ is a matrix representation of A in an appropriate basis for A.

Let's choose our basis so that it knows about the eigenspace E_{λ_0}. Begin by picking a basis $(v_1, \cdots, v_d)$ for E_{λ_0}; by the basis extension lemma, we may extend this to a basis $\beta = (v_1, \cdots, v_n)$ for V.

By the assumption that $v_j \in E_{\lambda_0}$, we know $Av_j = \lambda_0 v_j$ for all $1 \leqslant j \leqslant d$. Thus the j'th column of $[A]_{\beta\to\beta}$ is given by $\lambda_0 e_j$ for $1 \leqslant j \leqslant d$. We have absolutely no information about Av_j for $j > d$, and thus the $d+1$ column and onward could be absolutely anything.

With respect to the basis β, we can write

$$[A]_{\beta\to\beta} = \left(\begin{array}{ccc|ccc} \lambda_0 & \cdots & 0 & a_{1,d+1} & \cdots & a_{1,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \lambda_0 & a_{d,d+1} & \cdots & a_{d,n} \\ \hline 0 & \cdots & 0 & a_{d+1,d+1} & \cdots & a_{d+1,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & a_{n,d+1} & \cdots & a_{n,n} \end{array}\right) = \left(\begin{array}{c|c} \lambda_0 I_d & M_{12} \\ \hline 0_{(n-d)\times d} & M_{22} \end{array}\right),$$

where M_{12} is some $d \times (n-d)$ matrix, and M_{22} is some $(n-d)\times(n-d)$ matrix, and I_d is the $d \times d$ identity matrix.

Thus

$$\lambda I - M = \left(\begin{array}{ccc|ccc} \lambda-\lambda_0 & \cdots & 0 & -a_{1,d+1} & \cdots & -a_{1,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \lambda-\lambda_0 & -a_{d,d+1} & \cdots & -a_{d,n} \\ \hline 0 & \cdots & 0 & \lambda-a_{d+1,d+1} & \cdots & -a_{d+1,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & -a_{n,d+1} & \cdots & \lambda-a_{n,n} \end{array}\right) = \left(\begin{array}{c|c} \lambda I_d - \lambda_0 I_d & -M_{12} \\ \hline 0_{(n-d)\times d} & \lambda I_{(n-d)} - M_{22} \end{array}\right).$$

I can take the determinant of this matrix by Laplace expanding down the first d columns one by one; inductively, I obtain

$$p_A(\lambda) = \det(\lambda I - M) = (\lambda-\lambda_0)^d \det(\lambda I - M_{22}) = (\lambda-\lambda_0)^d p_{M_{22}}(\lambda).$$

Therefore I can factor at least $d = \dim E_{\lambda_0}$ factors of $(\lambda-\lambda_0)$ out of this polynomial. Because m is the largest number of such factors I can extract, we have $d \leqslant m$, as desired. □

This is enough to completely determine when a map is diagonalizable.

Theorem 97. *A linear map $A : V \to V$ is diagonalizable if and only if its characteristic polynomial $p_A(\lambda)$ splits into linear factors, and for every eigenvalue λ_i, we have $\dim E_{\lambda_i}$ = algebraic multiplicity of λ_i.*

Proof. Suppose A has eigenvalues $\lambda_1, \cdots, \lambda_k$ with algebraic multiplicity $m_1, \cdots, m_k$. This means that we may factor

$$p_A(\lambda) = (\lambda-\lambda_1)^{m_1} \cdots (\lambda-\lambda_k)^{m_k} q(\lambda),$$

where $q(\lambda)$ has no roots over $\mathbb{F}$ (such as $q(\lambda) = \lambda^2 + 1$ over $\mathbb{F} = \mathbb{R}$). In particular, we have

$$\dim V = \deg(p_A) = m_1 + \cdots + m_k + \deg(q).$$

The first equality comes from Lemma 91; the next from the fact that degree of polynomials is additive under multiplication of polynomials. Notice that $\dim V = m_1 + \cdots + m_k$ if and only if q is a constant (that is, if and only if p_A can be fully split into linear factors).

Now we have

$$\dim(E_{\lambda_1} + \cdots + E_{\lambda_k}) = \dim(E_{\lambda_1}) + \cdots + \dim(E_{\lambda_k}) \leqslant m_1 + \cdots + m_k \leqslant \dim V.$$

The first equality follows from the combination of Lemmas 94 and 95, while the first inequality is the content of Lemma 96. Finally, the second inequality is contained in the discussion of the previous paragraph.

The first inequality is an equality if and only if $\dim(E_{\lambda_i}) = m_i$ for all i. The second inequality is an equality if and only if $p_A(\lambda)$ can be fully factored into linear parts.

Now A is diagonalizable if and only if $\dim(E_{\lambda_1} + \cdots + E_{\lambda_k}) = \dim V$, and the previous discussion establishes that this is true if and only if $p_A(\lambda)$ can be fully factored (over $\mathbb{F}$) and $\dim E_{\lambda_i} = \text{algmult}(\lambda_i)$ for all i. □

Here is a useful corollary, frequently stated in linear algebra texts as the main result of the diagonalization theory.

Corollary 98. *Suppose $A : V \to V$ is a linear map which has $n = \dim V$ distinct eigenvalues $\lambda_1, \cdots, \lambda_n \in \mathbb{F}$, then A is diagonalizable.*

Proof. To say that we have n distinct eigenvalues implies $p_A(\lambda)$ has n roots, hence can be written as $p_A(\lambda) = (\lambda - \lambda_1) \cdots (\lambda - \lambda_n)$. Thus p_A can be fully factored.

On the other hand, $1 \leqslant \dim E_{\lambda_i} \leqslant \text{algmult}(\lambda_i) = 1$ for all $1 \leqslant i \leqslant n$, so that each $\dim E_{\lambda_i}$ is equal to the algebraic multiplicity of λ_i. Thus both hypotheses of Theorem 97 are satisfied. □

6.4.1 Triangulization

Not every linear map can be diagonalized, for two reasons:

- The characteristic polynomial may not split into linear factors.
- Even if it does, the eigenvalues λ_i may have $\dim E_{\lambda_i} < \text{algmult}(\lambda_i)$.

In some algebraic contexts, the first issue goes away entirely.

Definition 51. A field $\mathbb{F}$ is called **algebraically closed** if every polynomial $p(\lambda) = \lambda^n + a_1\lambda^{n-1} + \cdots + a_n$ with $a_1, \cdots, a_n \in \mathbb{F}$ splits into linear factors (aka, can be fully factored into linear parts. ◊

Remark 58. Equivalently, a field is algebraically closed if every polynomial has a root. (Proof: induct on degree; because $p(\lambda)$ has a root, it can be factored as $p(\lambda) = (\lambda - t_0)q(\lambda)$ where $q(\lambda)$ has degree one less.) ◊

You have seen an example of an algebraically closed field before.

Theorem 99 (The fundamental theorem of algebra)**.** *The complex numbers $\mathbb{C}$ are an algebraically closed field.*

Every proof of this fact is difficult, and requires some amount of analysis (though you may be able to keep that amount to a minimum). Therefore, the proof is beyond the scope of this Honors Math A. If we're lucky, it will be covered after we discuss line integrals in Honors Math B. (Yes, line integrals can be used to prove this algebraic fact.)

In fact, much as $\mathbb{R}$ sits inside the algebraically closed field $\mathbb{C}$, every field $\mathbb{F}$ can be extended to a *larger field* $\overline{\mathbb{F}}$ which is algebraically closed and is the smallest algebraically closed field containing $\mathbb{F}$. This is called its algebraic closure, and you might learn about it in the Modern Algebra sequence, depending on the taste of the instructor.

Corollary 100. *Let $A : V \to V$ be a linear map from V to itself, where V is a finite-dimensional vector space over the **algebraically closed** field $\mathbb{F}$. Then A has an eigenvector: there exists some $v \neq \vec{0}$ and $\lambda \in \mathbb{F}$ so that $Av = \lambda v$.*

Proof. This is merely the statement that $p_A(\lambda)$ has a root; if λ_0 is such a root, then E_{λ_0} is nontrivial, and one may take v to be any nonzero element in this vector space. □

Still, even over $\mathbb{C}$ (or any other algebraically closed field), the second issue remains. It is simply not true, for instance, that $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ is diagonalizable over any field whatsoever. Its characteristic polynomial is λ^2, so it has only the eigenvalue 0, but $E_0 = \ker M = \text{span}(e_1)$ is 1-dimensional. No dice.

However, we can still get something computationally useful.

Theorem 101. *Suppose $A : V \to V$ is a linear map from V to itself, where V is a finite-dimensional vector space over the **algebraically closed** field $\mathbb{F}$. Then there exists a basis β for V so that the matrix $[A]_{\beta\to\beta}$ is upper-triangular:*

$$[A]_{\beta\to\beta} = \begin{pmatrix} \lambda_1 & a_{12} & \cdots & a_{1n} \\ 0 & \lambda_2 & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}.$$

Proof. This can be proved by induction on the dimension of V (remember that in this chapter everything is finite-dimensional; abandon all hope, ye who enter infinite-dimensional vector spaces).

In the base case $\dim V = 1$ this is very nearly tautological. Pick any nonzero vector $v \in V$, which serves as a basis $\beta = (v)$ for V. With respect to this basis, A is given by a 1×1 matrix $M = \begin{pmatrix} a_{11} \end{pmatrix}$, which is tautologically upper-triangular. This has nothing to do with algebraic closedness; every 1×1 matrix is upper-triangular.

Suppose inductively that the claim holds for every linear map $A : W \to W$ where $\dim W = n$. We will prove the claim for linear maps $A : V \to V$ where $\dim V = n + 1$.

Apply Corollary 100 to see that there exists an eigenvector $v_0 \in V$ for which $Av_0 = \lambda_0 v_0$. Because v_0 is nonzero (by definition of eigenvector), the list (v_0) is linearly independent. It can therefore be extended to a basis $\beta = (v_0, \cdots, v_n)$ for V. Unfortunately, with respect to this basis, all we know is that $[A]_{\beta \to \beta}$ takes the form

$$[A]_{\beta \to \beta} = \left(\begin{array}{c|ccc} \lambda_0 & a_{01} & \cdots & a_{0n} \\ \hline 0 & a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & a_{n1} & \cdots & a_{nn} \end{array}\right) = \left(\begin{array}{c|c} \lambda_0 & M_{01} \\ \hline 0_{n \times 1} & M_{11} \end{array}\right),$$

where M_{01} is a $1 \times n$ matrix and M_{11} is an $n \times n$ matrix, and these matrices can be absolutely anything whatsoever.

What I notice is that M_{11} is an $n \times n$ matrix. If I write $W = \operatorname{span}(v_1, \cdots, v_n)$ and $\beta_W = (v_1, \cdots, v_n)$, then M_{11} defines a linear map $B : W \to W$, where $\dim W = n$. By the inductive hypothesis, I can *rechoose* this basis $\beta'_W = (w_1, \cdots, w_n)$ for W so that with respect to this basis, we have

$$[B]_{\beta'' \to \beta''} = \begin{pmatrix} \lambda_1 & b_{12} & \cdots & b_{1n} \\ 0 & \lambda_2 & \cdots & b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}.$$

If I set $\beta' = (v_0, w_1, \cdots, w_n)$ to be the new basis for V, then we have

$$[A]_{\beta' \to \beta'} = \begin{pmatrix} \lambda_0 & b_{01} & \cdots & b_{0n} \\ 0 & \lambda_1 & \cdots & b_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}.$$

This completes the inductive step, and this the result holds for all linear maps $A : V \to V$ between finite-dimensional vector spaces over an algebraically closed field. □

This is not the final word in the story. There is a still-better description, called the *Jordan normal form* of a linear map. For any linear map $A : V \to V$ (where V is defined over an algebraically closed field $\mathbb{F}$) one may choose a basis β for which $[A]_{\beta \to \beta}$ is a *particularly simple* upper-triangular matrix, called its Jordan normal form. Stated correctly, the Jordan normal form actually completely determines the conjugacy class of two matrices. If M and M' are $n \times n$ matrices over an algebraically closed field $\mathbb{F}$, then there exists an invertible $n \times n$ matrix S so that $SMS^{-1} = M'$ if and only if M and M' have the same Jordan normal form (where 'the same' should be interpreted carefully).

Thus while not every matrix can be diagonalized, there is still something we can say, and it is almost as good as a diagonalization. This will probably be discussed in Curio 5.

Chapter 7

Inner products, orthogonality, and the spectral theorem

7.1 Inner product spaces

Before giving a general definition, I want to give you two motivating examples.

Example 93. Let x, y be vectors in $\mathbb{R}^n$. Their **dot product** is

$$\begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ \cdots \\ y_n \end{pmatrix} = x_1 y_1 + \cdots + x_n y_n.$$

Notice a few things: first, that this is a bilinear operation (linear on each of the two inputs seperately); second, that it is symmetric; third, that $x \cdot y = y \cdot x$ because multiplication and addition in $\mathbb{R}$ are commutative); finally, that the operation $x \cdot x$ has the special property that

$$x \cdot x = x_1^2 + \cdots + x_n^2 = \|x\|^2$$

is the **length-squared** of x. In particular,

$$x \cdot x \geqslant 0 \quad \text{and} \quad x \cdot x = 0 \iff x = \vec{0}.$$

$\Diamond$

Example 94. The same idea does not naively work for the complex numbers, because if $z = x + iy$, there is no reason to believe $z^2 = (x^2 - y^2) + i(2xy)$ is a non-negative real number (and there is no useful notion of "positive complex number" in general). There is a trick to repair this, however.

Definition 52. If $z = x + iy$ is a complex number, its **complex conjugate** is $\overline{z} = x - iy$. Its **absolute value** is $|z| = \sqrt{x^2 + y^2}$. Its **real and imaginary parts** are $\mathrm{Re}(z) = x$ and $\mathrm{Im}(z) = y$. $\Diamond$

The complex conjugate has some useful properties, which you can verify by simple by-hands computation:

- We have $\overline{z + w} = \overline{z} + \overline{w}$, and $\overline{zw} = \overline{z}\,\overline{w}$.
- We have $\overline{z} = z$ if and only if $z = x + i0$ is a real number, whereas $\overline{z} = -z$ if and only if $z = 0 + iy$ is a purely imaginary number (its real part is zero).
- We have
$$z\overline{z} = (x + iy)(x - iy) = x^2 + y^2 + i(yx - xy) = x^2 + y^2 = |z|^2$$
is a non-negative real number, for any z, and in fact $z\overline{z} = 0$ if and only if $z = 0$.

Now let me define a 'dot product' over $\mathbb{C}^n$. If $z \in \mathbb{C}^n$ and $w \in \mathbb{C}^n$, we set their **complex dot product** to be

$$\begin{pmatrix} z_1 \\ \cdots \\ z_n \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ \cdots \\ w_n \end{pmatrix} = z_1\overline{w_1} + \cdots + z_n\overline{w_n}.$$

As a concrete example,

$$\begin{pmatrix} 1+i \\ 3-i \end{pmatrix} \cdot \begin{pmatrix} 2-i \\ 3+i \end{pmatrix} = (1+i)(\overline{2-i}) + (3-i)(\overline{3+i}) = (1+i)(2+i) + (3-i)(3-i) = (1+3i) + (8-6i) = 9-3i.$$

This new operation satisfies the property that $z \cdot z = |z_1|^2 + \cdots + |z_n|^2 \geqslant 0$, with equality if and only if $z = \vec{0}$, mimicking the corresponding property in the real case.

This is additive in both coordinates:

$$(z + z') \cdot w = z \cdot w + z' \cdot w; \quad z \cdot (w + w') = z \cdot w + z \cdot w'$$

However, one crucial property is different. While the dot product respects scaling in the *first* coordinate:

$$(cz) \cdot w = \begin{pmatrix} cz_1 \\ \cdots \\ cz_n \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ \cdots \\ w_n \end{pmatrix} = (cz_1)\overline{w_1} + \cdots + (cz_n)\overline{w_n} = c(z_1\overline{w_1} + \cdots + z_n\overline{w_n}) = c(z \cdot w).$$

However, it does not respect scaling in the second coordinate! Rather, we have

$$z \cdot (cw) = z_1(\overline{cw_1}) + \cdots + z_n(\overline{cw_n}) = \overline{c}(z_1\overline{w_1} + \cdots + z_n\overline{w_n}) = \overline{c}(z \cdot w).$$

When we pull out a scalar c from the second coordinate, it scales the dot product by the **complex conjugate** of that scalar.

Similarly, we have

$$z \cdot w = \overline{w \cdot z};$$

if we swap the terms, the dot product is changed by a complex conjugation.

Because the property $z \cdot z \geqslant 0$ (and the relation of $z \cdot z$ to length) is so important to the theory to come, we have to just accept these irritating and unexpected difficulties. They will be included in the definition of inner product space. $\Diamond$

In practice, the two examples above are far and away the most important. Even so, we should define the general notion of inner product space, *if only so that we can pass to subspaces as necessary.* The infinite-dimensional generalizations to 'Hilbert spaces' are also important, and in that context the more abstract phrasing is important (where many Hilbert spaces appear in naturte whose inner products do not take the form above).

Definition 53. An **inner product space** over $\mathbb{R}$ or $\mathbb{C}$ consists of the following data, satisfying the following properties.

(D1) A vector space V over the field $\mathbb{F}$, which is either $\mathbb{R}$ or $\mathbb{C}$.

(D2) A function $\langle -, - \rangle : V \times V \to \mathbb{F}$ which takes two input vectors and returns an element of the underlying field.

We demand this data satisfy the following four axioms.

(I1) **(Positive.)** The quantity $\langle v, v \rangle$ is a non-negative real number, which we write as $\langle v, v \rangle \geqslant 0$. This quantity is called the 'norm-squared', and we say that the 'norm' of v is the real number $\|v\| = \sqrt{\langle v, v \rangle}$.

(I2) **(Definite.)** The quantity $\langle v, v \rangle$ is zero if and only if $v = \vec{0}$.

(I3) **(Linearity in the first input.)**For any $a, b \in \mathbb{F}$ and any $v, w, u \in V$, we have

$$\langle av + bw, u \rangle = a\langle v, u \rangle + b\langle w, u \rangle.$$

(I4) **(Conjugate symmetry.)** For any $v, w \in V$, we have

$$\langle w, v \rangle = \overline{\langle v, w \rangle}.$$

If $\mathbb{F} = \mathbb{R}$, this should be interpreted as $\langle w, v \rangle = \langle v, w \rangle$, as there is no concept of 'complex conjugation' of real numbers.

$\Diamond$

Notice that in (I3) I have packaged both 'respects addition' (set $a = b = 1$) and 'respects scaling' (set $w = \vec{0}$) into the same formula. Further, notice that (I3) + (I4) imply that we have

$$\begin{aligned} \langle v, aw + bu \rangle &= \overline{\langle aw + bu, v \rangle} = \overline{a\langle w, v \rangle + b\langle u, v \rangle} \\ &= \overline{a}\overline{\langle w, v \rangle} + \overline{b}\overline{\langle u, v \rangle} \\ &= \overline{a}\langle v, w \rangle + \overline{b}\langle v, u \rangle, \end{aligned}$$

so that $\langle v, w \rangle$ is 'conjugate-linear' (or sometimes 'antilinear') in the second input: it is additive, and scalars pull out as their complex-conjugates.

Example 95. The most important examples of inner products are the dot products on $\mathbb{R}^n$ and $\mathbb{C}^n$ discussed above. $\Diamond$

Example 96. Suppose V is an inner product space, and $W \subset V$ is a subspace. Then restricting the inner product to W gives W the structure of an inner product space as well. $\Diamond$

I mentioned above that in the case of the dot product on $\mathbb{R}^n$ or $\mathbb{C}^n$, the quantity $\sqrt{v \cdot v}$ measures the length of a vector. (I will discuss this in more detail in the next section.) This is worth writing down in general.

Definition 54. If V is an inner product space, the **norm** (sometimes 'magnitude') is a function $\| \cdot \| : V \to [0, \infty)$ defined by

$$\sqrt{\langle v, v \rangle} \in [0, \infty).$$

$\Diamond$

Lemma 102. *The norm on an inner product space satisfies the following properties.*

(N1) We have $\|v\| \geqslant 0$*, with* $\|v\| = 0$ *if and only if* $v = \vec{0}$*.*

(N2) We have $\|cv\| = |c|\|v\|$*.*

(N3) We have

$$\|v + w\|^2 = \|v\|^2 + \|w\|^2 + 2Re\,\langle v, w\rangle.$$

In particular, we have $\|v\|^2 + \|w\|^2 = \|v + w\|^2$ *if and only if* $\langle v, w\rangle$ *has zero real part.*

Proof. The first item is a restatement of property (I1) and (I2) of inner product spaces. The next follows from linearity in the first input and conjugate-linearity in the second input:

$$\|cv\|^2 = \langle cv, cv\rangle = c\langle v, cv\rangle = c\bar{c}\langle v, v\rangle = |c|^2\|v\|^2.$$

Taking square roots of these non-negative numbers gives the desired claim. The final claim follows because

$$\begin{aligned}\|v + w\|^2 &= \langle v + w, v + w\rangle = \langle v, v + w\rangle + \langle w, v + w\rangle \\ &= \langle v, v\rangle + \langle v, w\rangle + \langle w, v\rangle + \langle w, w\rangle \\ &= \|v\|^2 + \|w\|^2 + \langle v, w\rangle + \langle w, v\rangle.\end{aligned}$$

Now if $\langle v, w\rangle$ is the complex number $x + iy$, we have $\langle w, v\rangle = \overline{\langle v, w\rangle} = x - iy$, so

$$\langle v, w\rangle + \langle w, v\rangle = 2x = 2\text{Re}\,\langle v, w\rangle.$$

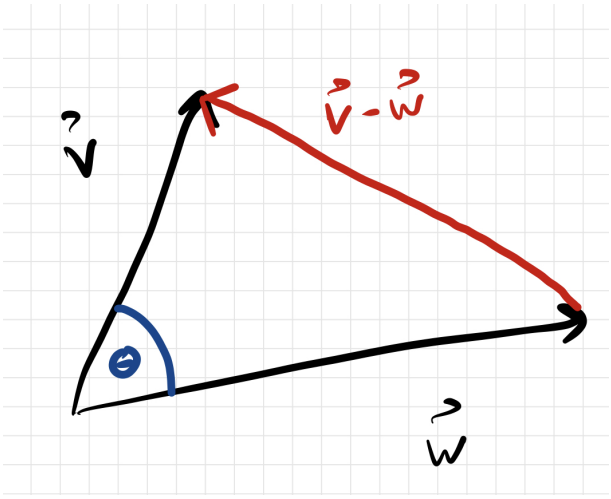
This gives the desired claim. □

The statement above looks quite a lot like the Pythagorean theorem. In fact, we use it as inspiration for the following definition.

Definition 55. Let V be an inner product space. We say that two vectors $v, w \in V$ are **orthogonal** if $\langle v, w\rangle = 0$. ◊

The notion of orthogonality will be crucial in what follows. Notice that no vector (other than the zero vector) is orthogonal to itself, as $\langle v, v\rangle = \|v\|^2$, which is only zero for the zero vector.

As an aside, suppose $v, w \in \mathbb{R}^n$. One may form a triangle with sides $v, w, v - w$, as follows:



The rule of cosines from trigonometry (which I will not prove) asserts a relation between the length of these three sides, which generalizes the Pythagorean theorem:

$$\|v - w\|^2 = \|v\|^2 + \|w\|^2 - 2\|v\|\|w\|\cos(\theta).$$

On the other hand, our inner product formulas imply that

$$\|v-w\|^2 = \langle v-w, v-w\rangle = \langle v,v\rangle - \langle w,v\rangle - \langle v,w\rangle + \langle w,w\rangle = \|v\|^2 + \|w\|^2 - (\langle v,w\rangle + \langle w,v\rangle).$$

Because we are working over the *reals*, we have $\langle w,v\rangle = \langle v,w\rangle$, so that we have just proved

$$\|v\|^2 + \|w\|^2 - 2\|v\|\|w\|\cos(\theta) = \|v\|^2 + \|w\|^2 - 2\langle v,w\rangle.$$

Canceling out like terms and dividing by -2, we have established a geometric formula for an algebraic quantity, the standard inner product on $\mathbb{R}^n$.

Proposition 103. *Let $v, w \in \mathbb{R}^n$ be vectors with angle $0 \leqslant \theta < \pi$ between them, where $\mathbb{R}^n$ is equipped with the standard inner product (the dot product). Then we have*

$$\langle v,w\rangle = \|v\|\|w\|\cos(\theta).$$

Note, in particular, that this implies $\langle v,w\rangle$ is zero if and only if one of the two vectors v, w is zero or the angle between them is $\theta = \pi/2$ (so they are perpendicular in the sense you're used to.) Further, if $\langle v,w\rangle$ is positive, the angle between these two vectors is acute; if $\langle v,w\rangle$ is negative, the angle is obtuse.

7.1.1 Orthogonal complements

In a finite-dimensional inner product space, there is a canonical complementary subspace to a given subspace $W \subset V$, the set of vectors *perpendicular to this subspace*. Imagine a plane in 3D space; the line perpendicular through the origin perpendicular to it is complementary. We encode this idea in the following definition.

Definition 56. Let V be an inner product space, and let $W \subset V$ be a subspace. Its **orthogonal complement** is the set $W^\perp \subset V$ defined by

$$W^\perp = \{v \in V \mid \forall_{w\in W} \langle v,w\rangle = 0\}.$$

That is, $W^\perp$ consists of those vectors which are orthogonal to every vector in W. ◊

It will take us some work to actually prove this space is complementary. In general, it at least intersects W trivially.

Proposition 104. *The orthogonal complement $W^\perp$ is a subspace of V, and $W^\perp \cap W = \{\vec{0}\}$.*

Proof. For (S1), suppose $v, v' \in W^\perp$. Then for all $w \in W$, we have

$$\langle v+v', w\rangle = \langle v,w\rangle + \langle v',w\rangle = 0 + 0 = 0,$$

where those inner products vanish because $v, v' \in W^\perp$. Therefore $v + v' \in W^\perp$. A similar argument applies for (S2): for all $w \in W$ we have

$$\langle cv, w\rangle = c\langle v,w\rangle = c(0) = 0.$$

Finally, for (S3), observe that for all $w \in W$, we have $\langle 0, w\rangle = 0$ because the function $v \mapsto \langle v,w\rangle$ is a linear function, and linear functions send zero to zero. (In fact, $\langle 0,v\rangle$ is zero for all vectors $v \in V$.)

For the last claim, notice that $0 \in W^\perp \cap W$ as the latter is a subspace. For the reverse inclusion, suppose $w \in W^\perp \cap W$. Then by definition w is orthogonal to every vector in W, and in particular w is perpendicular to itself, so

$$0 = \langle w,w\rangle = \|w\|^2.$$

By axiom (I2), definiteness — rephrased later as Lemma [**?**] (N1) — we thus have $w = \vec{0}$, as claimed. □

In fact, in *finite dimensions*, we have a better result. (Something comparable is true in infinite dimensions, but you'll need to learn something about limits.)

Proposition 105. *Suppose V is a finite-dimensional inner product space. If $W \subset V$ is a subspace, then $W^\perp$ is a complementary subspace to W: in addition to the property that $W^\perp \cap W = \{\vec{0}\}$, we also have that $W^\perp + W = V$ (every vector in V can be written as a sum of a vector in W and a vector in $W^\perp$).*

I am going to postpone the proof of this result somewhat, because it requires some new ideas (which I could obscure to give the proof faster, but what's the fun in that?). To see where we're going, remember that the way we discussed complementary subspaces on the last question of the Midterm before we introduced inner products was to introduce bases. The key point is to incorporate the inner product into our study and use of bases.

Definition 57. A **orthonormal list of vectors** in a finite-dimensional inner product space V is a list of vectors $(v_1, \cdots, v_k)$ for which

$$\|v_i\|^2 = 1 \quad \text{and} \quad v_i \cdot v_j = 0 \text{ for all } i \neq j.$$

That is, it is a list of length-one vectors which are mutually orthogonal. ◊

Lemma 106. *If $(v_1, \cdots, v_k)$ is an orthonormal list, it is necessarily linearly independent.*

Proof. Suppose $a_1v_1 + \cdots + a_kv_k = \vec{0}$ is a linear relation. Take the inner product with v_j: we have

$$\begin{aligned} 0 = \langle \vec{0}, v_j \rangle &= \langle a_1v_1 + \cdots + a_kv_k, v_j \rangle \\ &= a_1\langle v_1, v_j \rangle + \cdots + a_k\langle v_k, v_j \rangle = a_j, \end{aligned}$$

as $\langle v_i, v_j \rangle = 0$ for $i \neq j$ and $\langle v_j, v_j \rangle = 1$. It follows that $a_j = 0$ for all j, so the given linear relation is the trivial linear relation. Thus $(v_1, \cdots, v_k)$ is linearly independent. □

Proposition 107. *Let V be a finite-dimensional inner product space. Every list of orthonormal vectors $(v_1, \cdots, v_k)$ in V can be extended to an orthonormal **basis** for V.*

Proof. The algorithm to produce such a basis is called the 'Gram-Schmidt algorithm'. We will prove that each list $(v_1, \cdots, v_k)$ of orthonormal vectors which does not span V can be extended to a list $(v_1, \cdots, v_{k+1})$ of orthonormal vectors; inducting on k, we see that eventually we have produced an orthonormal list $(v_1, \cdots, v_n)$ where $n = \dim V$; because this list is linearly independent, by Corollary 42 we see that it spans V, hence is a basis. (Finite-dimensionality is essential here!)

Because $(v_1, \cdots, v_k)$ is a linearly independent list which does not span V, there exists some vector $v \notin \operatorname{span}(v_1, \cdots, v_k)$. Pick any such v. I am going to use v to construct a new vector not in $\operatorname{span}(v_1, \cdots, v_k)$ which is also perpendicular to all of the previous vectors; then I will scale it to be a unit-length vector. This will provide the desired extension to an orthonormal list $(v_1, \cdots, v_{k+1})$ of length $k+1$, and as discussed above, iterating this procedure leads us to a basis for V.

First, for $1 \leqslant i \leqslant k$, set $a_i = \langle v, v_i \rangle$. Then set

$$v' = v - a_1v_1 - \cdots - a_kv_k.$$

If $v' \in \operatorname{span}(v_1, \cdots, v_k)$, then v is as well; conversely because v is not in $\operatorname{span}(v_1, \cdots, v_k)$, neither is v'. Furthermore, observe that

$$\langle v', v_j \rangle = \langle v - a_1v_1 - \cdots - a_kv_k, v_j \rangle = \langle v, v_j \rangle - a_1\langle v_1, v_j \rangle - \cdots - a_k\langle v_k, v_j \rangle.$$

Now because $(v_1, \cdots, v_k)$ is orthonormal, $\langle v_i, v_j \rangle = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$, so this expression simplifies to

$$\langle v, v_j \rangle - a_j = \langle v, v_j \rangle - \langle v, v_j \rangle = 0.$$

Therefore $\langle v', v_j \rangle = 0$ for all v_j, and v' is not in $\operatorname{span}(v_1, \cdots, v_k)$. Lastly, renormalize: set

$$v_{k+1} = \frac{1}{\|v'\|}v'.$$

By linearity, again v_{k+1} is orthogonal to all of $v_1, \cdots, v_j$, and it remains nonzero; but

$$\|v_{k+1}\| = \frac{1}{\|v'\|}\|v'\| = 1,$$

by Lemma 102 (N3).

Thus we have extended the orthonormal list $(v_1, \cdots, v_k)$ to an orthonormal list $(v_1, \cdots, v_{k+1})$. □

Starting from the empty list, we obtain the following.

Corollary 108. *Every finite-dimensional inner product space has an orthonormal basis.*

We are now ready to prove

Proof of Proposition 105. Let $W \subset V$ be a subspace. Choose an orthonormal basis $(v_1, \cdots, v_k)$ for this subspace using Corollary 108, and extend it to an orthonormal basis $(v_1, \cdots, v_n)$ for V using Proposition 107. I claim that $W^\perp = \operatorname{span}(v_{k+1}, \cdots, v_n)$, at which point we see immediately that $W^\perp$ is complementary to W. To see the reverse containment, choose $v \in \operatorname{span}(v_{k+1}, \cdots, v_n)$. Observe that if $w = a_1v_1 + \cdots + a_kv_k \in W$, then

$$\langle v, w\rangle = \langle a_{k+1}v_{k+1} + \cdots + a_nv_n, a_1v_1 + \cdots + a_kv_k\rangle = \sum_{i=1}^{k}\sum_{j=k+1}^{n} a_j\overline{a_i}\langle v_j, v_i\rangle,$$

which is zero as $\langle v_j, v_i\rangle = 0$ for $j \neq i$ (and here $j \geqslant k+1$ while $i \leqslant k$). Therefore $\langle v, w\rangle = 0$ for all $w \in W$, so that $v \in W^\perp$.

For the forward containment, observe that if $w \in W^\perp$, we may express it as $w = a_1v_1 + \cdots + a_nv_n$ for some $a_i \in \mathbb{F}$. But because $v_j \in W$ for $1 \leqslant j \leqslant k$, by definition of $W^\perp$ see that

$$0 = \langle w, v_j\rangle = \langle a_1v_1 + \cdots + a_nv_n, v_j\rangle = \sum_{i=1}^{n} a_i\langle v_i, v_j\rangle = a_j,$$

once again by orthonormality of the basis. Therefore $a_j = 0$ for $1 \leqslant j \leqslant k$ and so

$$w = a_{k+1}v_{k+1} + \cdots + a_nv_n \in \operatorname{span}(v_{k+1}, \cdots, v_n),$$

as desired. □

This gives the following rather intuitive fact: the orthogonal complement to an orthogonal complement is the original space.

Corollary 109. *Let $W \subset V$ be a subspace of a finite-dimensional inner product space. Then $(W^\perp)^\perp = W$.*

Proof. It is always true that $W \subset (W^\perp)^\perp$: if $w \in W$ then $\langle w, v\rangle = 0$ for all $v \in W^\perp$. Now in the finite-dimensional case, because W and $W^\perp$ are complementary, we have $\dim W^\perp = \dim V - \dim W$, so that $\dim(W^\perp)^\perp = \dim V - (\dim V - \dim W) = \dim W$. Thus $W \subset (W^\perp)^\perp$ is a subspace of a finite-dimensional vector space of the same dimension (Theorem 41). □

Remark 59. The fact that this fails in general is related to the fact that $W + W^\perp$ need not always be the whole space. Examples are necessarily infinite-dimensional, so perhaps a bit aggravating, but here is the simplest one I can give, which uses the notion of convergent series. The space denoted ℓ^2 consists of (let's say real) sequences $(a_1, a_2, \cdots)$ for which $\sum_{i=1}^{\infty} |a_i|^2 < \infty$. There is an inner product defined by

$$\langle (a_i), (b_i)\rangle = \sum_{i=1}^{\infty} a_ib_i;$$

that this sum is convergent follows from the fact that $a_ib_i \leqslant \frac{a_i^2+b_i^2}{2}$ (because $(a_i + b_i)^2 \geqslant 0$). ◊

7.2 The transpose and the dot product

We previously introduced the *transpose* of a matrix in Definition 44 as a tool to let us pass between row operations and column operations. The transpose finds its real significance in discussions of the standard inner product on $\mathbb{R}^n$ (though this is not unrelated to passing betwen row and column operations, as perhaps or perhaps not elucidated by the discussion opening Curio 4). We will see this below; while discussing the same for $\mathbb{C}^n$, we need a slightly more complicated definition.

Definition 58. Let $M = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$ be an $m \times n$ matrix over $\mathbb{C}$. Its **conjugate transpose** is the $n \times m$ complex matrix given by

$$M^* = \begin{pmatrix} \overline{a_{11}} & \cdots & \overline{a_{m1}} \\ \overline{a_{12}} & \cdots & \overline{a_{m2}} \\ \cdots & \cdots & \cdots \\ \overline{a_{1n}} & \cdots & \overline{a_{mn}} \end{pmatrix}.$$

That is, we take the transpose of M and then the complex conjugate of each of its entries. ◊

Example 97. The conjugate transpose of $M = \begin{pmatrix} 3 & 1-i \\ 2i & \pi + 3i \\ 7-2i & 3 \end{pmatrix}$ is $M^* = \begin{pmatrix} 3 & -2i & 7+2i \\ 1+i & \pi - 3i & 3 \end{pmatrix}$. ◊

Remark 60. Notice that $(A^T)^T = A$ and $(A^*)^* = A$: transpose twice and you end up back where you started. ◊

Before moving on to its relevance to the standard inner product, let me observe a basic fact about theh transpose (or conjugate transpose) of a product.

Lemma 110. *If M is an $\ell \times m$ matrix and N an $m \times n$ matrix over the field $\mathbb{F}$, then the transpose of the product satisfies $(MN)^T = N^T M^T$. If M and N are defined over $\mathbb{C}$ (so that the conjugate transpose is defined), we have $(MN)^* = N^* M^*$.*

Proof. The proof is by direct computation (though the ideas leading Curio 4 can be used to give a conceptual proof).

Write $M = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \cdots & \cdots & \cdots \\ a_{\ell 1} & \cdots & a_{\ell m} \end{pmatrix}$ and $N = \begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \cdots & \cdots & \cdots \\ b_{m1} & \cdots & b_{mn} \end{pmatrix}$. Then the entry $(MN)_{ij}$ in row i, column j, is given by

$$(MN)_{ij} = a_{i1}b_{1j} + \cdots + a_{im}b_{mj}, \quad \text{so} \quad (MN)^T_{ij} = (MN)_{ji} = a_{j1}b_{1i} + \cdots + a_{jm}b_{mi}.$$

On the other hand, one has

$$(N^T M^T)_{ij} = \begin{pmatrix} b_{1i} & \cdots & b_{mi} \end{pmatrix} \begin{pmatrix} a_{j1} \\ \cdots \\ a_{jm} \end{pmatrix} = b_{1i}a_{j1} + \cdots + b_{mi}a_{jm} = a_{j1}b_{1i} + \cdots + a_{jm}b_{mi},$$

the same result. Here we use that multipilication in $\mathbb{F}$ is commutative.

The extension to conjugate transposes follows by the same argument, with the addition point that $\overline{zw} = \overline{z}\,\overline{w}$: complex conjugation respects multiplication. □

The reason these are relevant is that I can write an *formula* for the dot product (the standard inner product on $\mathbb{R}^n$ and $\mathbb{C}^n$) using the transpose (or, over $\mathbb{C}$, the conjugate tranpose).

If $v \in \mathbb{R}^n$ or $v \in \mathbb{C}^n$, then we think of v as a column vector $\begin{pmatrix} v_1 \\ \cdots \\ v_n \end{pmatrix}$, an $n \times 1$ matrix. Using the discussion above, I can also define a $1 \times n$ matrix: in the real case I will be interested in $v^T = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix}$, whereas in the complex case we will be interested in the conjugate transpose $v^* = \begin{pmatrix} \overline{v}_1 & \cdots & \overline{v}_n \end{pmatrix}$. Furthermore, because the product of a $1 \times n$ matrix and an $n \times 1$ matrix is a 1×1 matrix (a single number!), the expression $w^T v$ in the first case and $w^* v$ in the second case gives us an element of $\mathbb{R}$ (resp. $\mathbb{C}$). In fact, these are precisely the inner products.

Lemma 111. *Let $v, w \in \mathbb{R}^n$. Then $\langle v, w \rangle = w^T v$. Similarly, if $v, w \in \mathbb{C}^n$, we have $\langle v, w \rangle = w^* v$.*

Proof. Let me write the complex case (the argument in the real case is the same): we have

$$w^* v = \begin{pmatrix} \overline{w}_1 & \cdots & \overline{w}_n \end{pmatrix} \begin{pmatrix} v_1 \\ \cdots \\ v_n \end{pmatrix} = \overline{w}_1 v_1 + \cdots + \overline{w}_n v_n = v_1 \overline{w}_1 + \cdots + v_n \overline{w}_n = \langle v, w \rangle.$$

□

Remark 61. One may take this perspective from the start, and derive the basic properties of the inner product using this formula. For instance, conjugate symmetry comes from the fact that

$$\overline{\langle v, w\rangle} = (w^* v)^* = v^* w^{**} = v^* w = \langle w, v\rangle. \quad \Diamond$$

The following corollary is the reason I've discussed all of this. It is immensely useful, and if you take a different perspective, the formulas below are so important they can be used to *define* the transpose (and conjugate transpose) matrix. This is not the same as but not unrelated to the perspective of Curio 4.

Corollary 112 (The transpose and inner products)**.** *Let $v \in \mathbb{R}^n$ and $w \in \mathbb{R}^m$. If $A : \mathbb{R}^n \to \mathbb{R}^m$ is a linear map, then we have*

$$\langle Av, w\rangle = \langle v, A^T w\rangle.$$

Similarly for the complex inner product: we would have

$$\langle Av, w\rangle = \langle v, A^* w\rangle.$$

Proof. I will write the proof in the complex case; the real case is the same argument. We have

$$\langle v, A^* w\rangle = (A^* w)^* v = w^* A^{**} v = w^* Av = \langle Av, w\rangle.$$

Here I used that the conjugate transpose flips the order of a product, and that taking the conjugate transpose twice returns you to the original matrix. □

Remark 62. When $A : \mathbb{F}^n \to \mathbb{F}^n$ runs from the same space to itself, so that $\langle v, Aw\rangle$ makes sense, observe that we have the same formula here by (conjugate)-symmetry:

$$\langle v, Aw\rangle = \overline{\langle Aw, v\rangle} = \overline{\langle w, A^* v\rangle} = \langle A^* v, w\rangle. \quad \Diamond$$

Remark 63. If $A : V \to W$ is a linear map between finite-dimensional inner product spaces, the formula above defines a map $A^* : W \to V$ of complex inner product spaces: it is the unique linear map so that

$$\langle v, A^* w\rangle = \langle Av, w\rangle$$

for all $v \in V$ and $w \in W$. (One must argue that such a linear map exists and is unique.) The same argument applies for A^T in the real case. From the perspective of Curio 4, the point is that if you have a linear map $V \to W$, there is a canonical dual map $W^* \to V^*$ between the dual vector spaces. Then the fact that V and W are inner-product spaces gives an isomorphism $V \cong V^*$ by sending v to the functional $\varphi_v(w) = \langle w, v\rangle$; this is an isomorphism because it is an injective inear map between two finite-dimensional inner product spaces of the same dimension (this isomorphism is called the Riesz representation). Then the composite map $W \to W^* \to V^* \to V$ is precisely the map A^* defined by the formula above (A^T in the real case). $\Diamond$

This fact is incredibly useful, and you should keep it written down somewhere circled — you will use it again. Let me attempt to give a canonical example of its use. It also demonstrates a very standard trick (to show that a vector is zero, show that its norm-squared is zero).

Proposition 113. *Let $A : \mathbb{F}^n \to \mathbb{F}^m$ be a linear map (where $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$). For convenience, write A^* for the conjugate transpose or the usual transpose, depending on $\mathbb{F}$. Then we have*

$$\ker(A^*) = \mathrm{im}(A)^\perp, \qquad \mathrm{im}(A^*) = \ker(A)^\perp.$$

Proof. I will write the proof over $\mathbb{C}$ (the only change over $\mathbb{R}$ is writing the transpose instead of the conjugate transpose). This is a double containment argument. Suppose $v \in \ker(A^*)$. Let's show $\langle v, w\rangle = 0$ for all $w \in \mathrm{im}(A)$, or equivalently, $\langle v, Au\rangle = 0$ for all $u \in \mathbb{F}^n$. We have

$$\langle v, Au\rangle = \langle A^* v, u\rangle = \langle \vec{0}, u\rangle = 0,$$

as claimed. This proves the forward containment.

Conversely, suppose $v \in \mathrm{im}(A)^\perp$, so that $\langle v, Au\rangle = 0$ for all $u \in \mathbb{C}^n$. We want to show that $A^*v = 0$. **Now for an infinitely-useful magic trick.** Observe that

$$\|A^*v\|^2 = \langle A^*v, A^*v\rangle = \langle v, AA^*v\rangle = 0.$$

The second equality follows because we may move A^* across the inner product at the cost of adding a conjugate transpose (and $(A^*)^* = A$). Because $\|A^*v\|^2 = 0$ and the norm of a vector is zero if and only if that vector is zero, it follows that $A^*v = 0$, so that $v \in \ker(A^*)$. We have completed the double-containment; this proves the first stated equality.

The second fact follows by applying the first, and the fact that $(W^\perp)^\perp = W$: we have

$$\mathrm{im}(A^*) = \left(\mathrm{im}(A^*)^\perp\right)^\perp = (\ker((A^*)^*))^\perp = \ker(A)^\perp,$$

where the last equality uses that $(A^*)^* = A$. □

The theorem above will be essential when we later discuss the 'spectral theorem': for certain matrices, we will be able to say that their eigenspaces are orthogonal. (As for why this matters, you'll have to wait and see.) I can give an immediate application, though.

Corollary 114. *Let $A : \mathbb{F}^n \to \mathbb{F}^m$ be a linear map. Then* $\mathrm{rank}(A^*) = \mathrm{rank}(A)$.

Remark 64. That is, the maximal number of columns of A which are linearly independent is equal to the maximal number of *rows* of A which are linearly independent, as these are the (complex conjugates of) the columns of A^*. This is often called the 'row rank = column rank' theorem in elementary linear algebra texts. ◊

Proof. We have

$$\mathrm{rank}(A^*) = \dim \mathrm{im}(A^*) = \dim \ker(A)^\perp = n - \dim \ker(A) = \dim \mathrm{im}(A) = \mathrm{rank}(A).$$

The first and last equalities are the definition of rank; the second equality is given by the preceding Proposition; the third equality is Midterm #6 together with Proposition 105 that the orthogonal complement is a complementary subspace; the fourth equality is the rank-nullity theorem. □

7.3 Special classes of linear maps

There are privileged classes of linear maps between inner product spaces — most often from an inner product space to itself, and this is the case we will take up below. (One may slightly generalize this discussion to allow for maps between different inner product spaces called 'isometric embeddings', but we won't gain much from this generalization, so I won't bother.)

Definition 59. Let V be an inner product space. An **isometry** is an invertible linear map $A : V \to V$ with the additional property that for all $v_1, v_2 \in V$, we have

$$\langle Av_1, Av_2\rangle = \langle v_1, v_2\rangle.$$

When V is a real inner product space, these transformations are called **orthogonal transformations**. When V is complex, they are called **unitary transformations**. If $V = \mathbb{R}^n$ the corresponding $n \times n$ matrices are called **orthogonal matrices**, and if $V = \mathbb{C}^n$ the corresponding $n \times n$ matrices are called **unitary matrices**. The set of orthogonal $n \times n$ matrices is often denoted $O(n)$, whereas the set of unitary $n \times n$ matrices is often denoted $U(n)$. ◊

The significance of the given equation is that A *preserves the inner-product structure*: you can compute the inner product of two vectors either before or after applying A. This immediately implies that A preserves the length (norm) of vectors:

$$\|Av\| = \sqrt{\langle Av, Av\rangle} = \sqrt{\langle v, v\rangle} = \|v\|.$$

In fact, the converse is true as well; one can think of isometries as length-preserving linear maps (or distance-preserving linear maps, if you like). I include this below in a longer list of equivalent conditions, some of which are easily checkable.

Proposition 115. *The following are all equivalent conditions on a linear map $A : V \to V$ of finite-dimensional inner product spaces. (Here I write A^* to mean the conjugate transpose in the complex case, and the usual transpose in the real case.)*

(a) A is an isometry.

(b) A preserves norms: for all $v \in V$, we have $\|Av\| = \|v\|$.

(c) A satisfies the equation $A^*A = I$.

(d) If $V = \mathbb{F}^n$ so that A may be identified with an $n \times n$ matrix, the columns of A form an orthonormal basis for $\mathbb{F}^n$.

Proof. We first prove that (a) and (b) are equivalent, and then move on to showing that $(a) \implies (c) \implies (d) \implies (a)$, which shows those three conditons are all equivalent.

We saw that $(a) \implies (b)$ above. As for the reverse direction (which I will prove in the complex case), observe that we established in Lemma 102(N3) that

$$\frac{\|v+w\|^2 - \|v\|^2 - \|w\|^2}{2} = \operatorname{Re}\langle v, w\rangle.$$

It follows that if A preserves norms, then

$$\operatorname{Re}\langle Av, Aw\rangle = \frac{\|A(v+w)\|^2 - \|Av\|^2 - \|Aw\|^2}{2} = \frac{\|v+w\|^2 - \|v\|^2 - \|w\|^2}{2} = \operatorname{Re}\langle v, w\rangle.$$

In the first equality I used that A is linear to combine $Av + Aw = A(v+w)$ and in the second that A is assumed to preserve norms. So at least A preserves the *real part* of the inner product, which establishes $(b) \implies (a)$ in the real case; for the complex case, a trick shows that the same argument handles the imaginary part: observe that if $\langle v, w\rangle = x + iy$ so that

$$\langle v, iw\rangle = -i\langle v, w\rangle = -i(x+iy) = y - ix,$$

we have

$$\operatorname{Re}\langle v, iw\rangle = \operatorname{Im}\langle v, w\rangle.$$

Because A is complex-linear, and we have established that A preserves the real part of the inner product, it follows that A preserves the imaginary part as well; so $\langle Av, Aw\rangle = \langle v, w\rangle$.

Let's move on to the remaining two conditions. Suppose A is an isometry. I claim that $A^*A = I$. To see this, observe that for all $v, w \in V$, we have

$$\langle A^*Av, w\rangle = \langle Av, Aw\rangle = \langle v, w\rangle,$$

the first equality using the formula relating the conjugate transpose to the inner product, and the next the assumption that A is an isometry.

I claim that this implies $A^*Av = v$ for all v. To see this, set $w = A^*Av - v$, and observe that

$$\|A^*Av - v\|^2 = \langle A^*Av - v, A^*Av - v\rangle = \langle A^*Av, A^*Av\rangle - \langle v, A^*Av\rangle - \langle A^*Av, v\rangle + \langle v, v\rangle;$$

applying the previous equality repeatedly, we see that $\langle A^*Av, A^*Av\rangle = \langle A^*Av, v\rangle = \langle v, v\rangle$, and similarly for the other terms, so that this simplifies to

$$\|A^*Av - v\|^2 = \|v\|^2 - \|v\|^2 - \|v\|^2 + \|v\|^2 = 0.$$

Therefore $A^*Av - v = \vec{0}$, so $A^*Av = v$ for all $v \in V$. Thus $A^*A = I$. This gives $(a) \implies (c)$. Conversely, if $A^*A = I$, then using the equation relating the inner product to the conjugate transpose, we have for all $v, w \in V$ that

$$\langle Av, Aw\rangle = \langle A^*Av, w\rangle = \langle v, w\rangle,$$

so A is an isometry.

Lastly I will show that when $V = \mathbb{F}^n$ we have $(c) \iff (d)$. Write the corresponding matrix as $M = \begin{pmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{pmatrix}$. Then

$$M^*M = \begin{pmatrix} — & v_1^* & — \\ & \cdots & \\ — & v_n^* & — \end{pmatrix} \begin{pmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{pmatrix} = \begin{pmatrix} v_1^*v_1 & \cdots & v_1^*v_n \\ \cdots & \cdots & \cdots \\ v_n^*v_1 & \cdots & v_n^*v_n \end{pmatrix} = \begin{pmatrix} \langle v_1, v_1\rangle & \cdots & \langle v_1, v_n\rangle \\ \cdots & \cdots & \cdots \\ \langle v_n, v_1\rangle & \cdots & \langle v_n, v_n\rangle \end{pmatrix}.$$

To say that this is the identity matrix means precisely

$$\langle v_i, v_j\rangle = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

which is precisely what it means to say that the list of columns $(v_1, \cdots, v_n)$ is orthonormal; because this is an orthornormal list (so linearly independent) of the same dimension as V, it spans V by Corollary 42, so is an orthonormal basis. □

7.3.1 Examples of orthogonal and unitary matrices

The standard examples of orthogonal matrices are rotations and reflections. In fact, every 2×2 orthogonal matrix takes one of the following forms:

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}.$$

(Notice that the first column represents some arbitrary unit-length vector in $\mathbb{R}^2$, and the second column the two possibilities for a unit vector perpendicular to it.) The first of these is rotation by an angle θ, and the second reflection across a line which lies $\theta/2$ radians counter-clockwise from the x-axis.

We can generalize this to arbitrary dimensions, as follows.

Definition 60. Let V be a real inner product space of dimension n, and let $W \subset V$ be a subspace of dimension 2. Choose an orthonormal basis (v_1, v_2) for W and an orthonormal basis $(v_1, \cdots, v_n)$ of V extending it. Then **Rotation along W by angle θ** is the map $R_{W,\theta} : V \to V$ defined with respect to this basis as

$$R_{W,\theta}(v_i) = \begin{cases} \cos(\theta)v_1 + \sin(\theta)v_2 & i = 1 \\ -\sin(\theta)v_1 + \cos(\theta)v_2 & i = 2 \\ v_i & i > 2 \end{cases}.$$

That is, it rotates the plane W by an angle of θ and keeps the orthogonal complement $W^\perp = \text{span}(v_3, \cdots, v_n)$ fixed. ♢

It is a straightforward computation to verify that $R_{W,\theta}$ sends the orthonormal basis $(v_1, \cdots, v_n)$ to another orthonormal basis; this implies $R_{W,\theta}$ is an orthogonal transformation:

$$\begin{aligned} \langle R_{W,\theta}v, R_{W,\theta}w\rangle &= \left\langle R_{W,\theta}(\sum_{i=1}^n a_i v_i), R_{W,\theta}\sum_{j=1}^n b_j v_j \right\rangle = \sum_{i=1}^n \sum_{j=1}^n a_i b_j \langle R_{W,\theta}v_i, R_{W,\theta}v_j\rangle \\ &= \sum_{i=1}^n \sum_{j=1}^n a_i b_j \langle v_i, v_j\rangle = \langle \sum_{i=1}^n a_i v_i, \sum_{j=1}^n b_j v_j\rangle = \langle v, w\rangle \end{aligned}$$

with the equality between the first and second row arising because

$$\langle R_{W,\theta}v_i, R_{W,\theta}v_j\rangle = \langle v_i, v_j\rangle = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

and $\langle av, bw\rangle = ab\langle v, w\rangle$ because we are working in a *real* inner product space, for which the inner product is linear in both terms. (Were we working in a complex inner product space, this expression would instead read $\langle av, bw\rangle = a\bar{b}\rangle v, w\rangle$.)

Here is a description of what's going on in 3D. Pick a line in 3D space, and rotate 'around' that line by some angle θ. This is what is called above the rotation along the plane perpendicular to that line.

I will not prove the following theorem, but it is valuable for getting intuition for what an orthogonal matrix 'does'.

Theorem 116. *Every orthogonal transformation $A : V \to V$ with determinant $\det A > 0$ can be written as a composite of rotations.*

Even more precisely, there exists a sequence $W_1, \cdots, W_m$ of orthogonal planes inside V (orthogonal meaning: for all $w_i \in W_i$ and $w_j \in W_j$ with $i \neq j$ we have $\langle w_i, w_j\rangle = 0$; the subspaces are orthogonal to one another, though not orthogonal complements to one another), so that A is given by rotation by some $\theta_1, \cdots, \theta_m$ along each of these planes. Thus an orthogonal transformation with positive determinant should be understood as 'rotating around some collection of orthogonal planes'.

The case of $\det A < 0$ should be understood in terms of *reflections*.

Definition 61. Let V be a finite-dimensional inner product space and let $W \subset V$ be a non-trivial proper subspace. The **reflection along** W is the unique linear map $\mathrm{ref}_W : V \to V$ which is the identity on W and acts as -1 on $W^\perp$. If $(v_1, \cdots, v_k)$ is an orthonormal basis for W and this is extended to an orthonormal basis $(v_1, \cdots, v_n)$ for V, then in terms of this basis we have

$$\mathrm{ref}_W(v_i) = \begin{cases} v_i & 1 \leqslant i \leqslant k \\ -v_i & k < i \leqslant n \end{cases} \quad \lozenge$$

Once again, you can check that this is an orthogonal transformation because it sends an orthonormal basis to an orthonormal basis. Interestingly, it is also a symmetric transformation in the sense of the next section (prove by hand that $\langle \mathrm{ref}_W v, w\rangle = \langle v, \mathrm{ref}_W w\rangle$ for all $v, w \in V$).

The determinant of ref_W can be computed in the basis $(v_1, \cdots, v_n)$ above to be $(-1)^{n-k}$. If $k = n-1$, so that W is one dimension smaller than the ambient space it lives in, this is called 'reflection across a hyperplane'; it's what happens when you look in a mirror in 3D space. In this case, $\det(\mathrm{ref}_W) = -1$. The analogue of Theorem 116 in this context is that an orthogonal matrix with $\det(A) < 0$ can be written as a composite of an orthogonal transformation with $\det(A) > 0$ and a reflection along across a hyperplane (so a composite of a bunch of rotations and one reflection).

The unitary case is easier.

Let z be a complex number with $|z| = 1$. Because $|x + iy| = \sqrt{x^2 + y^2}$, such numbers take the form $\cos(\theta) + i\sin(\theta)$. These are often written as $e^{i\theta}$ because of a relation between the Taylor series of these two functions, which we will discuss next term. Every complex number can be written as $re^{i\theta}$ where $r \geqslant 0$ is real and θ is some angle; multiplication of complex numbers in this form satisfies $(re^{i\theta})(se^{i\psi}) = rse^{i(\theta+\psi)}$. Multiplying by $re^{i\theta}$ has the visual effect of scaling the complex plane by the factor r and rotating the complex plane by the angle θ.

Unitary matrices allow us to compress these 'rotations' into a single entry. The most common unitary matrix is a diagonal unitary matrix:

$$D = \begin{pmatrix} e^{i\theta_1} & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & e^{i\theta_n} \end{pmatrix},$$

which has the effect of rotating each complex plane factor in $\mathbb{C}^n$ by a factor of θ_j. In fact, we will actually prove the following as a consequence of the 'spectral theorem'.

Theorem 117. *Let $A : V \to V$ be a unitary transformation of a finite-dimensional complex inner product space. Then there exists a **orthonormal** basis of eigenvectors $(v_1, \cdots, v_n)$. With respect to this basis, we have*

$$[A]_{\beta\to\beta} = \begin{pmatrix} e^{i\theta_1} & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & e^{i\theta_n} \end{pmatrix}$$

for appropriate angles $\theta_1, \cdots, \theta_n$, *where* $Av_j = e^{i\theta_j}$.

7.3.2 (skew)-Symmetric and (skew)-Hermitian matrices

The next class of matrices are worth introducing partly for their relationship to quadratic functions (and we will conclude the term with a discussion of quadratic functions), but they also appear rather often. (I often see them in the study of partial differential equations, where the relation to the inner product is especially important.)

Definition 62. Let M be an $n \times n$ matrix with real entries. If $M = M^T$ we say that M is **symmetric**, whereas if $M^T = -M$ we say that M is **skew-symmetric**.

Let M be an $n \times n$ matrix with complex entries. If $M = M^*$ we say that M is **Hermitian**[1], whereas if $M^* = -M$ we say that M is **skew-Hermitian**. ◊

Example 98. Let me discuss both the general form of such matrices (which we can easily write down, unlike the case of orthogonal matrices; the key is that the set of symmetric matrices forms a vector space, as do the other three classes above, whereas orthogonal matrices firmly do not — they form a 'matrix group' or in modern terminology a 'Lie group', named after the mathematician Sophus Lie).

$$M \text{ symmetric} \implies M = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{12} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{pmatrix}$$

is a matrix which 'looks the same' when you flip it across the diagonal; $a_{ij} = a_{ji}$. For instance,

$$M = \begin{pmatrix} 4 & 3 & 2 & 17 \\ 3 & \pi & 6 & -1 \\ 3 & 6 & 0 & -2 \\ 17 & -1 & -2 & e^2 \end{pmatrix}$$

is a 4×4 symmetric matrix. On the other hand, a skew-symmetric matrix has $a_{ji} = -a_{ij}$; when $i = j$ this reads $a_{ii} = -a_{ii}$, so that $2a_{ii} = 0$ and hence $a_{ii} = 0$, so the general form is

$$M \text{ skew-symmetric} \implies M = \begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ -a_{12} & 0 & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ -a_{1n} & -a_{2n} & \cdots & 0 \end{pmatrix}.$$

The diagonal entries are always zero, and the other entries negate when we flip the matrix. For instance, an example of a skew-symmetric matrix is

$$M = \begin{pmatrix} 0 & 3 & -2 \\ -3 & 0 & -1 \\ 2 & 1 & 0 \end{pmatrix}.$$

If M is Hermitian with entries $M_{k\ell} = a_{k\ell} + ib_{k\ell}$, the relevant formula is $\overline{M}_{k\ell} = M_{\ell k}$. Along the diagonal, this gives $\overline{M}_{kk} = M_{kk}$, so the diagonal entries are **real numbers**; off the diagonal the entries change by complex conjugation when you flip the matrix. The general form is

$$M \text{ Hermitian} \implies M = \begin{pmatrix} a_{11} & a_{12} + ib_{12} & \cdots & a_{1n} + ib_{1n} \\ a_{12} - ib_{12} & a_{22} & \cdots & a_{2n} + ib_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1n} - ib_{1n} & a_{2n} - ib_{2n} & \cdots & a_{nn} \end{pmatrix}.$$

For instance, $M = \begin{pmatrix} 3 & 2-i & i \\ 2+i & 4 & 1 \\ -i & 4 & 0 \end{pmatrix}$ is Hermitian.

[1] Named after mathematician Charles Hermite

On the other hand, if M is skew-Hermitian, we have $\overline{M}_{k\ell} = -M_{\ell k}$. On the diagonal, we thus have $\overline{M}_{kk} = -M_{kk}$; that is, $M_{kk} = ib_{kk}$ is purely imaginary. The general form of a skew-Hermitian matrix is

$$M \text{ skew-Hermitian} \implies M = \begin{pmatrix} ib_{11} & a_{12}+ib_{12} & \cdots & a_{1n}+ib_{1n} \\ -a_{12}+ib_{12} & ib_{22} & \cdots & a_{2n}+ib_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ -a_{1n}+ib_{1n} & -a_{2n}+ib_{2n} & \cdots & ib_{nn} \end{pmatrix}.$$

An example of a skew-Hermitian matrix is $M = \begin{pmatrix} 2i & 3+i & 4 \\ -3+i & -i & 12-5i \\ -4 & -12-5i & 0 \end{pmatrix}$. Notice that $0 = 0 + 0i$ is still a purely imaginary number, as it has zero real part. ◊

The first crucial property of these matrices is their relation to the inner product.

Lemma 118. *A real matrix is symmetric (and a complex matrix is Hermitian) if and only if, for all $v, w \in V$, we have $\langle Mv, w\rangle = \langle v, Mw\rangle$. A real matrix is skew-symmetric (and a complex matrix is skew-Hermitian) if and only if, for all $v, w \in V$, we have $\langle Mv, w\rangle = -\langle v, Mw\rangle$.*

Proof. I will handle the Hermitian case. All other cases are similar. If M is Hermitian we have

$$\langle Mv, w\rangle = \langle v, M^*w\rangle = \langle v, Mw\rangle.$$

Conversely, if $\langle Mv, w\rangle = \langle v, Mw\rangle$, observe that this implies $\langle v, Mw\rangle = \langle v, M^*w\rangle$ for all v, w. In particular, $\langle v, (M - M^*)w\rangle = 0$ for all v, w. Applying this to $w = e_i$ and $v = (M - M^*)e_i$ we see that $\|(M - M^*)e_i\| = 0$ for all i, so that $Me_i - M^*e_i = \vec{0}$, or rather $Me_i = M^*e_i$. We have proved these two matrices have the same columns, hence are the same matrix. □

The second is that their eigenvalues are much more constrained than those of general matrices. The first of these facts

Proposition 119. *The eigenvalues of a symmetric or Hermitian matrix are all real (and its characteristic polynomial splits into linear factors over $\mathbb{R}$). The eigenvalues of a skew-symmetric or skew-Hermitian matrix are all purely imaginary.*

Proof. If M is a symmetric matrix (so an $n \times n$ matrix with real entries and $M^T = M$) we may view it as a complex matrix whose entries all happen to be real (that is, have no imaginary part); from this perspective M is a Hermitian matrix. Similarly for skew-symmetric and skew-Hermitian matrices. It suffices to argue the claim in the complex case. This is one place where the complex perspective is actually very valuable, even if you only care about real matrices!

Suppose M is Hermitian and $v \in \mathbb{C}^n$ is an eigenvector of M with eigenvalue λ. That is, $Mv = \lambda v$. Consider the quantity $\langle Mv, v\rangle$. (This is a good idea because of two reasons: "Mv" shows up in the definition of eigenvalue, and the statement that M is Hermitian means something about its relation to inner-products.)

We can compute this in two ways:

$$\lambda\|v\|^2 = \langle \lambda v, v\rangle = \langle Mv, v\rangle = \langle v, Mv\rangle = \langle v, \lambda v\rangle = \overline{\lambda}\|v\|^2,$$

where in the second step I used that M is Hermitian and in the last step I used that $\langle v, cw\rangle = \bar{c}\langle v, w\rangle$. Thus $\lambda\|v\|^2 = \overline{\lambda}\|v\|^2$. Beacuse $\|v\|^2 \neq 0$ (to say that v is an eigenvector means, in particular, it is nonzero), we may divide it from this equation to see that $\lambda = \overline{\lambda}$ — that is, the eigenvalue is real.

A similar argument applies in the skew-Hermitian setting:

$$\lambda\|v\|^2 = \langle \lambda v, v\rangle = \langle Mv, v\rangle = -\langle v, Mv\rangle = -\langle v, \lambda v\rangle = -\overline{\lambda}\|v\|^2;$$

proceeding as above we see that $\lambda = -\overline{\lambda}$, so that the eigenvalue λ is purely imaginary. □

Remark 65. It follows that if M is skew-symmetric and invertible, it has no real eigenvectors whatsoever! So it is certainly not diagonalizable over the reals. There are versions of diagonalization that do hold (but I will not endeavor to prove them here). ◊

The last class of operators we will introduce will be precisely those for which the 'spectral theorem' is applicable. Notice that the definition below subsumes *all of* unitary, Hermitian, and skew-Hermitian operators.

Definition 63. We say that a linear map $A : \mathbb{C}^n \to \mathbb{C}^n$ is **normal** if we have $A^*A = AA^*$. ◊

HW. Prove that if A is normal, then $\mathrm{im}(A) = \ker(A)^\perp$.

Notice that this contains both the statement of Proposition applied to either a Hermitian or skew-Hermitian matrix, and the fact that for A unitary we have $\mathrm{im}(A) = \mathbb{C}^n$ and $\ker(A) = \{\vec{0}\}$.

7.4 The spectral theorem

In this section, we will prove the following theorem, a good capstone to the course.

Theorem 120 (The spectral theorem over $\mathbb{C}$)**.** *Suppose V is a finite-dimensional inner-product space over $\mathbb{C}$ (say $\dim V = n$), and $A : V \to V$ is a linear map. Then the following three claims are equivalent:*

(a) There exists an orthonormal basis β for V consisting of eigenvectors of A.

(b) There exists an orthonormal basis β for V so that $[A]_{\beta\to\beta}$ is diagonal.

(c) There exists a unitary map $U : \mathbb{C}^n \to V$ so that $U^{-1}AU$ is diagonal.

(d) The linear map A is normal: $A^*A = AA^*$.

We say that A is unitarily diagonalizable *if any of the first three conditions hold.*

The equivalence (a) $\iff$ (b) $\iff$ (c) are straightforward: if β is such a basis, then $U = C_\beta$ is such a unitary map; if we have such a unitary map U, take the basis to be $\beta = (Ue_1, \cdots, Ue_n)$. The interesting claim is that (a-c) are equivalent to (d). You will prove the claim (c) $\implies$ (d) on your homework; we will focus on the harder direction (d) $\implies$ (a).

My first goal is to convince you that this is an interesting and useful result. My second goal is to prove it over $\mathbb{C}$. In the section after that, we will prove the corresponding result over $\mathbb{R}$; this is what we'll use in Calculus.

7.4.1 Motivation for the spectral theorem

Before talking about orthogonal diagonalization, let me discuss how we intuit what *diagonalization* means to begin with. Consider the map $A : \mathbb{R}^2 \to \mathbb{R}^2$ with associated matrix $M = \begin{pmatrix} 3 & -2 \\ 1 & 0 \end{pmatrix}$. You can compute that this matrix has two eigenvalues — $\lambda = 1, 2$ — and has associated eigenspaces

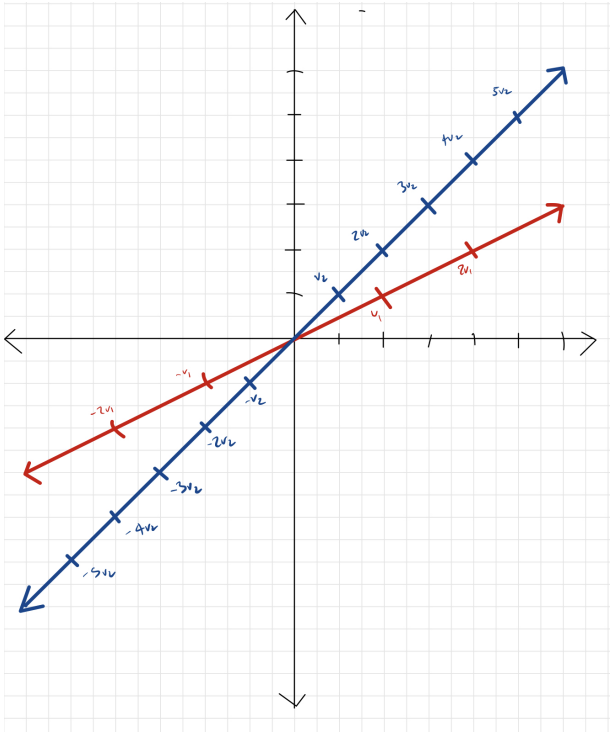
$$E_1 = \mathrm{span}\begin{pmatrix} 2 \\ 1 \end{pmatrix}, \qquad E_2 = \mathrm{span}\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

In the basis

$$\beta = \left(v_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad v_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right),$$

we have that $[M]_{\beta\to\beta} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$.

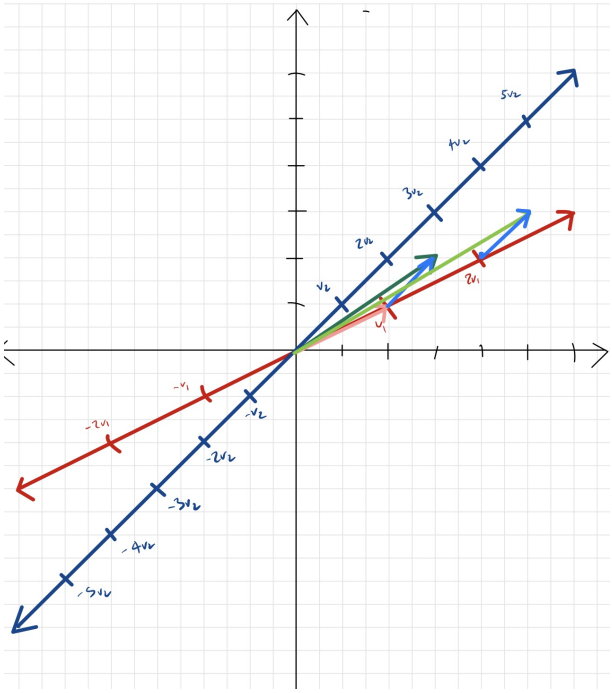
Visually, what this means is that if I draw the plane with a *different set of coordinate axes and tickmarks* — one axis being $\mathrm{span}(v_1)$, the other being $\mathrm{span}(v_2)$, and the tickmarks representing multiples of v_1 and v_2.



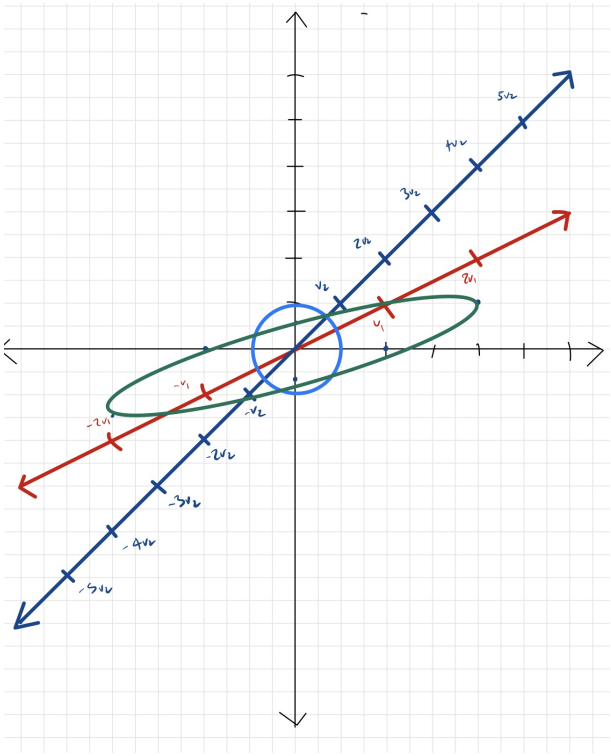
The fact that $Av_1 = 2v_1$ and $Av_2 = v_2$ tells us that *in terms of these coordinate axes*, we can see what A does rather easily. For instance,

$$A\begin{pmatrix}3\\2\end{pmatrix} = A(v_1 + v_2) = 2v_1 + v_2 = \begin{pmatrix}5\\3\end{pmatrix},$$

visualized as follows. Notice that the second vector has twice as much 'red part', and the same amount of 'blue part'.



This is very algebraically useful, but very hard to work with geometrically... I can very vaguely see that the 'red part' of the dark-green vector got longer when I drew the light-green vector, but very vaguely. One thing that's even harder to visualize is what A does to *shapes in the plane*, as opposed to individual vectors.. As an example, let's take this picture showing what A does to the unit circle; the unit circle is drawn in light-blue, while its image under A is drawn in dark-blue.



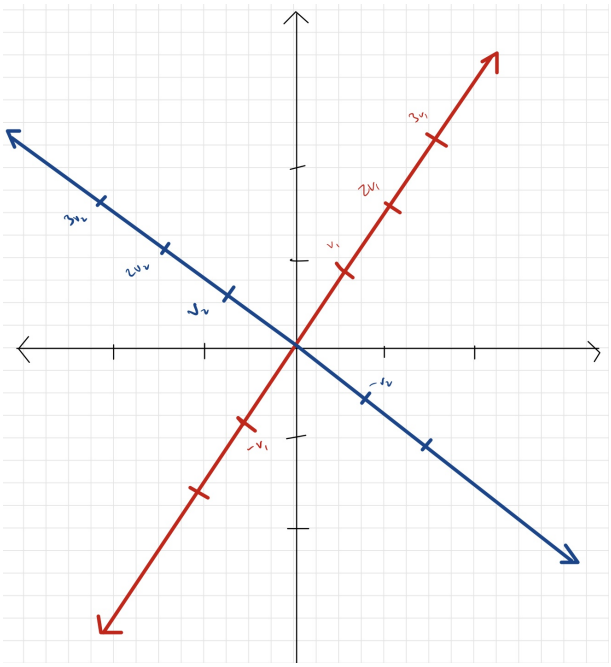
I find it basically impossible to see what happens to the circle from the coordinate axis picture! Part of the issue is that *the coordinate system itself is hard to visualize.*

On the other hand, consider the matrix $M = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$. Some of you showed on your homework that this is diagonalizable, with eigenvalues $\phi = \frac{1+\sqrt{5}}{2}$ and $-\phi^{-1} = \frac{1-\sqrt{5}}{2}$. I can compute the eigenspaces in a particularly suggestive way: they're given by

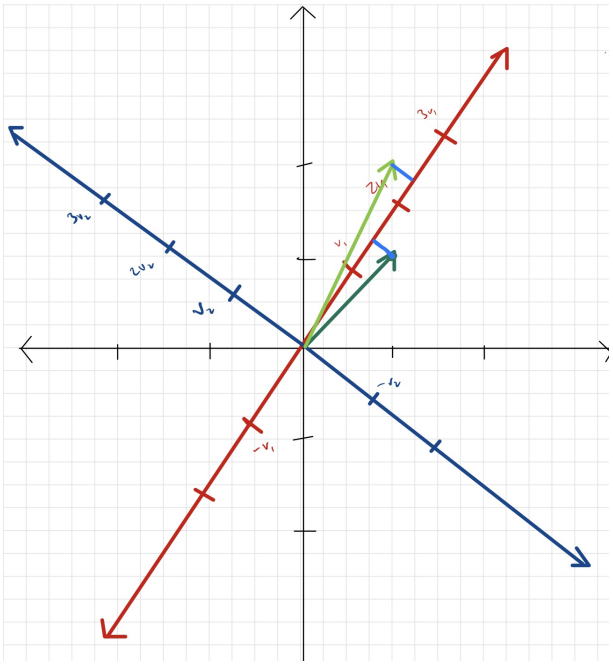
$$E_\phi = \text{span}\begin{pmatrix} 1/(\sqrt{\phi^2+1}) \\ \phi/\sqrt{\phi^2+1} \end{pmatrix}, \quad E_{-\phi^{-1}} = \begin{pmatrix} -\phi/\sqrt{1+\phi^2} \\ 1/\sqrt{1+\phi^2} \end{pmatrix}.$$

I divide by those factors because the two eigenvectors listed here form an *orthonormal basis*: they are perpendicular and have length 1! Approximate decimal values are $v_1 \approx \begin{pmatrix} 0.526 \\ 0.857 \end{pmatrix}$ and $v_2 \approx \begin{pmatrix} -0.857 \\ 0.526 \end{pmatrix}$.

Here is a picture of the corresponding coordinate axes.

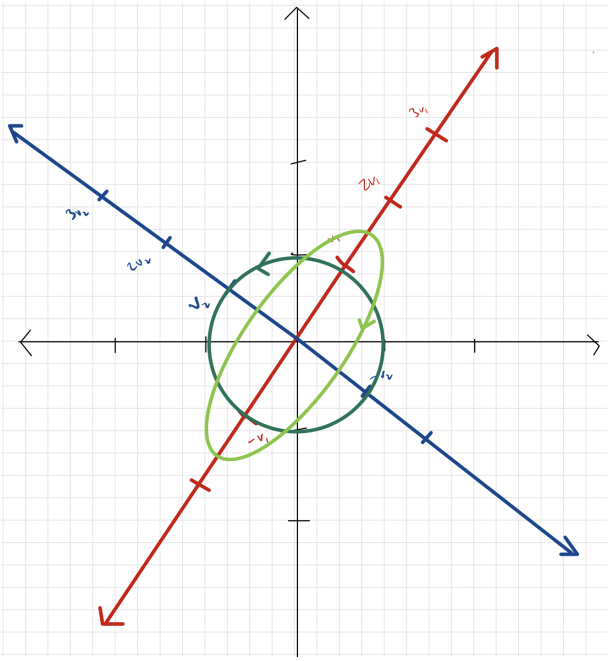


M stretches the red axis by a factor of $\phi \approx 1.62$ — so it stretches it by a factor of around 60% — and it scales the blue axis by a factor of $-\phi^{-1} \approx -0.62$ (so it flips backwards, and shrinks by a factor of around 40%). For instance, here's a picture visualizing $M \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ in terms of this coordinate system.



See how the 'red part' of the vector gets longer, while the 'blue part' negates (it should also be smaller; the fact that it's not is a small visual error.)

Now let me draw what happens to the unit circle in *this* picture.



I can actually see what's going on in terms of the coordinate axes here, and I can predict the shape the resulting circle is sent to! The point is that this coordinate system looks a lot like the standard coordinate system — in fact, it's obtained from the standard coordinate system by a rotation (a type of orthogonal transformation) and a reflection (the counterclockwise-oriented circle is sent to a clockwise-oriented ellipse). This is precisely what it means to be orthogonally diagonalizable: there is a coordinate sytem (related to the usual one by an orthogonal transformation, which you may think of as a composite of rotations). We'll actually use this in the next part.

The unitary case is the same idea, but harder to visualize!

7.4.2 Proof of the spectral theorem over $\mathbb{C}$

The proof will be given by induction on $\dim V$. When $\dim V = 1$, the only linear maps $A : V \to V$ are given by scaling; $A(v) = cv$ for some $c \in \mathbb{C}$. Any unit vector defines an orthonormal basis for V consisting of eigenvectors of A. So the interesting step is the inductive step.

The idea is as follows.

- Because A is a complex matrix, it has *some* eigenvector v_1, which may be rescaled to be a unit vector. Call $W = \text{span}(v_1)^\perp$.
- Because A is normal, we can argue that A maps W into itself, and therefore we may restrict its domain and codomain to a map $A_W : W \to W$. We then check that A_W is normal and $\dim W = \dim V - 1$. For convenience, say $\dim V = n$.
- We can now apply the inductive hypothesis to find an orthonormal basis $(v_2, \cdots, v_n)$ for W consisting of eigenvectors of A_W. Notice that $(v_1, \cdots, v_n)$ defines an orthonormal basis for V. Further, because v_1 is an eigenvalue of A, and $v_2, \cdots, v_n$ are eigenvalues of A_W (which is just the restriction of A to W!) Therefore $(v_1, \cdots, v_n)$ is an orthonormal basis for V consisting of eigenvectors of A, as desired.

Let me give more details for each of these three steps.

Step 1. Because $\mathbb{C}$ is algebraically closed by the fundamental theorem of algebra, Corollary 100 guarantees that there exists an eigenvector of A. Call this v. By definition of eigenvector, $v \neq \vec{0}$, so by definiteness of the inner-product $\|v\| \neq 0$. Set

$$v_1 = \frac{1}{\|v\|} v; \quad \text{we have} \quad \|v_1\| = \left\| \frac{1}{\|v\|} v \right\| = \left| \frac{1}{\|v\|} \right| \|v\| = \frac{\|v\|}{\|v\|} = 1.$$

Here I used that $1/\|v\|$ is a positive real number, so its absolute value is itself. Because v_1 is a scalar multiple of v, it is again an eigenvector of A. (This is teh step that will be most subtle when we move to the real case: such an eigenvector is not so simply guaranteed.)

Step 2. Consider the subspace $W = \text{span}(v_1)^\perp$. It follows from Midterm #6 and Proposition 105 that have $\dim W = \dim V - 1$ (because v_1 is nonzero, (v_1) is a linearly independent set, hence forms a basis for $\text{span}(v_1)$, so that $\dim \text{span}(v_1) = 1$.

My first claim is that $A(W) \subset W$. That is, if w has the property that $\langle w, v_1\rangle = 0$, then $\langle Aw, v_1\rangle = 0$ too. (Notice that if w is orthogonal to v_1, it is also orthogonal to cv_1 for all $c \in \mathbb{C}$; so $\langle w, v_1\rangle = 0$ is equivalent to the claim $w \in \text{span}(v_1)^\perp$.) This is the hardest part of the proof by a long shot. The first fact we need is a relationship between the norms of Av and A^*v.

Lemma 121. *Let V be a finite-dimensional inner product space and $A : V \to V$ a normal operator. Then $\|Av\| = \|A^*v\|$ for all $v \in V$.*

Proof. If A is normal, then we have

$$\begin{aligned}\|Av\|^2 &= \langle Av, Av\rangle = \langle v, A^*Av\rangle = \langle v, AA^*v\rangle \\ &= \overline{\langle AA^*v, v\rangle} = \overline{\langle A^*v, A^*v\rangle} = \langle A^*v, A^*v\rangle = \|A^*v\|^2.\end{aligned}$$

□

The reason this is relevant is a simple formula for computing the lengths in terms of an orthonormal basis.

Lemma 122. *Suppose $(v_1, \cdots, v_n)$ is an orthonormal basis for the inner product space V. Then*

$$\|a_1v_1 + \cdots + a_nv_n\|^2 = |a_1|^2 + \cdots + |a_n|^2.$$

Proof. We have

$$\begin{aligned}\|a_1v_1 + \cdots + a_nv_n\|^2 &= \langle \sum_{i=1}^n a_iv_i, \sum_{j=1}^n a_jv_j\rangle = \sum_{i=1}^n \sum_{j=1}^n a_i\overline{a}_j\langle v_i, v_j\rangle \\ &= \sum_{i=1}^n a_i\overline{a}_i = \sum_{i=1}^n |a_i|^2,\end{aligned}$$

as desired. □

We can combine these to prove the crucial lemma.

Lemma 123. *Let $A : V \to V$ be a normal operator, and let $v_1 \in V$ be an eigenvector of A. Then $\langle w, v_1\rangle = 0$ implies $\langle Aw, v_1\rangle = 0$.*

Proof. Let me try to give some conceptual insight here into what normality buys us. It's easiest to describe this in terms of matrices, so extend v_1 to an orthonormal basis $(v_1, \cdots, v_n)$ for V. In terms of this orthonormal basis, because $Av_1 = a_{11}v_1$ for some scalar a_{11}, we have

$$[A]_{\beta\to\beta} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & a_{n2} & \cdots & a_{nn}\end{pmatrix}.$$

Notice that the entries are $a_{ij} = \langle Av_i, v_j\rangle$. In particular,

$$\overline{a}_{ji} = \overline{\langle Av_j, v_i\rangle} = \langle v_i, Av_j\rangle = \langle A^*v_i, v_j\rangle,$$

so that

$$[A^*]_{\beta\to\beta} = [A]^*_{\beta\to\beta} = \begin{pmatrix} \overline{a}_{11} & 0 & \cdots & 0 \\ \overline{a}_{12} & \overline{a}_{22} & \cdots & \overline{a}_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ \overline{a}_{1n} & \overline{a}_{2n} & \cdots & \overline{a}_{nn}\end{pmatrix}.$$

The key point is that the previous two lemmas show us that *because A is normal, the norm of each row is the same as the norm of each column.* More precisely, $Av_j = a_{1j}v_1 + \cdots + a_{nj}v_n$, whereas $A^*v_j = \overline{a}_{j1}v_1 + \cdots + \overline{a}_{jn}v_n$. It follows that

$$|a_{1j}|^2 + \cdots + |a_{nj}|^2 = \|Av_j\|^2 = \|A^*v_j\|^2 = |\overline{a}_{j1}|^2 + \cdots + |\overline{a}_{jn}|^2 = |a_{j1}|^2 + \cdots + |a_{jn}|^2.$$

Applying this to the first column we see that

$$|a_{11}|^2 = |a_{11}|^2 + |a_{12}|^2 + \cdots + |a_{1n}|^2 \implies |a_{12}|^2 + \cdots + |a_{1n}|^2 = 0.$$

It follows that $a_{12} = \cdots = a_{1n} = 0$, and

$$[A]_{\beta\to\beta} = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & a_{n2} & \cdots & a_{nn} \end{pmatrix}.$$

It follows that for all $j > 1$, we have

$$\langle Av_j, v_1\rangle = \langle a_{2j}v_2 + \cdots + a_{nj}v_n, v_1\rangle = \sum_{i=2}^{n} a_{ij}\langle v_i, v_1\rangle = 0.$$

Therefore, if $w \in \operatorname{span}(v_1)^\perp$ — so that $w = b_2v_2 + \cdots + b_nv_n$ — we have

$$\langle Aw, v_1\rangle = \langle \sum_{j=2}^{n} b_jAv_j, v_1\rangle = \sum_{j=2}^{n} b_j\langle Av_j, v_1\rangle = 0,$$

as desired. □

Recall that we are writing $W = \operatorname{span}(v_1)^\perp$. We have proved that A restricts to a linear map $A_W : W \to W$ (as if $w \in W$, then $Aw \in W$ as well). We also verified that $\dim W = \dim V - 1$. Notice that A_W is still normal. This is easiest to see at the level of matrices. In the orthonormal basis $(v_1, \cdots, v_n)$, writing $\beta' = (v_2, \cdots, v_n)$, we saw above that $[A]_{\beta\to\beta}$ takes the block-matrix form

$$[A]_{\beta\to\beta} = \left(\begin{array}{c|c} a_{11} & 0 \\ \hline 0 & [A_W]_{\beta'\to\beta'} \end{array}\right),$$

so that

$$[A^*A]_{\beta\to\beta} = [A]_{\beta\to\beta} = \left(\begin{array}{c|c} \overline{a}_{11}a_{11} & 0 \\ \hline 0 & [A_W^*A_W]_{\beta'\to\beta'} \end{array}\right), \quad \text{while} \quad [AA^*]_{\beta\to\beta} = \left(\begin{array}{c|c} a_{11}\overline{a}_{11} & 0 \\ \hline 0 & [A_WA_W^*]_{\beta'\to\beta'} \end{array}\right).$$

Because $A^*A = AA^*$, it follows that $A_W^*A_W = A_WA_W^*$ as well.

Step 3. We have now establishes that $A_W : W \to W$ is a normal operator on an inner product space of dimension $\dim W = \dim V - 1$. By inductive hypothesis, there exists an orthonormal basis $(v_2, \cdots, v_n)$ for W which consists of eigenvectors of W. Because $\operatorname{span}(v_1)$ and $\operatorname{span}(v_1)^\perp$ are complementary subspaces — and all of these vectors are perpendicular to the length-1 vector v_1 — it follows that $(v_1, \cdots, v_n)$ is an orthonormal basis for V consisting of eigenvectors. This completes the proof.

7.4.3 The real case

The real spectral theorem works for a smaller class of operators, but they're exactly the operators that show up in calculus when trying to understand second derivatives. The first part of the argument is easier

Theorem 124 (The spectral theorem over $\mathbb{R}$)**.** *Suppose V is a finite-dimensional inner-product space over $\mathbb{R}$ (say $\dim V = n$), and $A : V \to V$ is a linear map. Then the following three claims are equivalent:*

(a) There exists an orthonormal basis β for V consisting of eigenvectors of A.

(b) There exists an orthonormal basis β for V so that $[A]_{\beta\to\beta}$ is diagonal.

(c) There exists an orthogonal transformation $O : \mathbb{R}^n \to V$ so that $O^{-1}AO$ is diagonal.

(d) The linear map A is symmetric: $A^T = A$.

The equivalence (a) $\iff$ (b) $\iff$ (c) is exactly as above, and the implication (b) $\implies$ (d) is similar to (but simpler than) the complex case. Again, I'll focus on (d) $\implies$ (a).

Again, I'll prove this by induction on $\dim V$, and the 1×1 case is tautological (every 1×1 matrix is symmetric and diagonal). Our steps are the same as before:

- Prove that there exists an eigenvector of A. Rescaling it, write v_1 for a length-1 eigenvector of A.
- Let $W = \text{span}(v_1)^\perp$. Prove that $A(W) \subset W$, so restricting domain and codomain defines a linear map $A_W : W \to W$. Argue that A_W remains symmetric.
- Choosing an orthonormal basis $(v_2, \cdots, v_n)$ for W consisting of eigenvectors of A_W, we find that $(v_1, \cdots, v_n)$ is an orthonormal basis of V consisting of eigenvectors, as desired.

This time, the first step is harder, and the second step is easier. In fact, for the first step, we have little choice but to think about complex operators.

Lemma 125. *Let $A : V \to V$ be a symmetric map on a finite-dimensional real inner product space, meaning $\langle Av, w\rangle = \langle v, Aw\rangle$ for all $v, w \in V$. Then the characteristic polynomial $p_A(\lambda)$ has only real roots, and therefore splits into linear factors over the real numbers.*

Proof. Choose an orthonormal basis $\beta = (v_1, \cdots, v_n)$ for V. With respect to this basis, $[A]_{\beta\to\beta}$ is a symmetric matrix: as discussed earlier, the entries a_{ij} are $\langle Av_i, v_j\rangle$, and therefore

$$a_{ji} = \langle Av_j, v_i\rangle = \langle v_j, Av_i\rangle = \langle Av_i, v_j\rangle = a_{ij},$$

in the second-to-last step using that real inner products are symmetric.

We showed in Proposition 119 that a symmetric matrix has only real eigenvalues, which is the stated claim. Notice that in that proof we had to think about complex operators! □

Because $p_A(\lambda)$ has only real roots, it has *some* real root, and hence $A : V \to V$ has some eigenvalue λ — that is, there exists $\lambda \in \mathbb{R}$ so that E_λ is non-trivial. Then any nonzero vector in E_λ is an eigenvector of A. Normalizing it to a length-1 vector, this finishes step one.

For the next step, we have an easier version of the lemma about normal operators:

Lemma 126. *Let $A : V \to V$ be a symmetric map on a finite-dimensional real inner product space. If $v_1 \in V$ is an eigenvector, and $w \in V$ is orthogonal to v_1, then Aw is as well.*

Proof. We have

$$\langle Aw, v_1\rangle = \langle w, Av_1\rangle = \langle w, \lambda v_1\rangle = \lambda\langle w, v_1\rangle = 0.$$

□

The rest of the proof now goes through as before.

7.5 Quadratic functions on $\mathbb{R}^n$ and the spectral theorem

I want to conclude with one application that appears, at first glance, to be outside the realm of linear algebra: *quadratic* functions!

Definition 64. A **quadratic function** on $\mathbb{R}^n$ is a function of the form

$$q(x_1, \cdots, x_n) = \sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{i-1} a_{ij} x_i x_j.$$

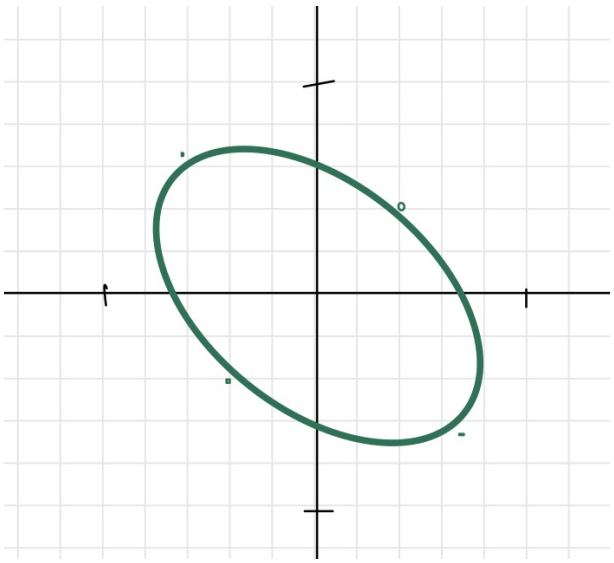
$\Diamond$

That is, the general form of a quadratic function of two real variables is $ax^2 + bxy + cy^2$, while the general form of a quadratic function of three real variables is

$$q(x, y, z) = ax^2 + by^2 + cz^2 + dxy + exz + fyz,$$

where $a, b, c, d, e, f \in \mathbb{R}$ are fixed real numbers (here, e does not denote the exponential constant, but rather an arbitrary real number).

Our big goal is as follows. These functions can be hard to understand in general; for instance, can you visualize what $q(x, y) = 3x^2 + 4xy + 3y^2$ 'looks like'? For instance, can you visualize what the shape $3x^2 + 4xy + 3y^2 = 1$ should look like? (I can't immediately.) But the actual picture is rather simple:



It's an ellipse with major axis the line $y = -x$ and minor axis the line $y = x$. If I rotate it by $\pi/4$ radians counter-clockwise, it becomes a standard ellipse, stretched along the x- and y-axes.

But how can I see this from looking at the function itself, without drawing a graph? This turns out to be closely related to our theory of orthogonal diagonalization!

The first observation is that these quadratic functions can be encoded in terms of *bilinear* functions, inspired by the fact that xy is a bilinear function of x and y. Before explaining this equivalence, let me recall the definition of symmetric bilinear functions, and then give a useful way to think about these in terms of matrices.

Definition 65. If V is a real vector space, an *symmetric bilinear function* on V is a function $B : V \times V \to \mathbb{R}$ so that $B(v, w)$ is linear in each coordinate separately and so that $B(v, w) = B(w, v)$. $\Diamond$

Lemma 127. *Let B be a symmetric bilinear function on $\mathbb{R}^n$. Then there exists a unique $n \times n$ symmetric matrix M for which*

$$B(v, w) = \langle Mv, w \rangle = w^T M v.$$

Proof. First, observe that for a symmetric matrix M, the expression $B(v,w) = \langle Mv, w\rangle$ does indeed define a symmetric bilinear form. It is bilinear because M is linear and the real inner product is bilinear; it is symmetric because

$$B(w,v) = \langle Mw, v\rangle = \langle v, Mw\rangle = \langle Mv, w\rangle = B(v,w),$$

the second-to-last equality by the assumption that M is symmetric.

Explicitly, suppose

$$M = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{pmatrix}, \quad \text{and we write} v = \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix} = \sum_{i=1}^n x_i e_i \quad \text{and} \quad w = \begin{pmatrix} y_1 \\ \cdots \\ y_n \end{pmatrix} = \sum_{j=1}^n y_j e_j.$$

Here, because M is symmetric, $b_{ij} = b_{ji}$ for all i, j. Then the associated bilinear form is

$$\begin{aligned} B(v,w) = \langle Mv, w\rangle &= \left\langle \begin{pmatrix} b_{11}x_1 + \cdots + b_{1n}x_n \\ \cdots \\ b_{n1}x_1 + \cdots + b_{nn}x_n \end{pmatrix}, \begin{pmatrix} y_1 \\ \cdots \\ y_n \end{pmatrix} \right\rangle \\ &= (b_{11}x_1 + \cdots + b_{1n}x_n)y_1 + \cdots + (b_{n1}x_1 + \cdots + b_{nn}x_n)y_n \\ &= \sum_{i=1}^n \sum_{j=1}^n b_{ij}x_i y_j. \end{aligned}$$

Remembering that $b_{ij} = b_{ji}$, when $i \neq j$ we can collect the expressions $x_i y_j$ and $x_j y_i$ into the same term; writing this as a sum over the terms where $i = j$ and the terms where $i < j$, we thus have

$$B(v,w) = \sum_{i=1}^n b_{ii}x_i y_i + \sum_{1 \leqslant i < j \leqslant n} b_{ij}(x_i y_j + x_j y_i).$$

Showing that every bilinear form arises in this way, from a unique matrix M, is a bit like the proof that the determinant must take the form it does. Suppose B is an arbitrary bilinear form. Then Then

$$B(v,w) = B\left(\sum_{i=1}^n x_i e_i, \sum_{j=1}^n y_j e_j\right) = \sum_{i=1}^n \sum_{j=1}^n x_i y_j B(e_i, e_j).$$

Because $B(e_i, e_j) = B(e_j, e_i)$, we may rewrite this as

$$\sum_{i=1}^n B(e_i, e_i)x_i y_i + \sum_{1 \leqslant i < j \leqslant n} B(e_i, e_j)(x_i y_j + x_j y_i).$$

By comparing these two expressions, we see that there is a unique symmetric matrix M so that $B(v,w) = \langle Mv, w\rangle$: set the entries of M to have $b_{ij} = B(e_i, e_j)$. □

Now that we've compared the study of symmetric bilinear forms on $\mathbb{R}^n$ and symmetric $n \times n$ matrices, let's bring quadratic functions into the mix.

Lemma 128. *Let $B : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a symmetric bilinear function, meaning $B(v,w)$ is linear in each input and $B(v,w) = B(w,v)$. Then the function $B(v,v)$ is a quadratic function, and in fact, the assignment $\mathsf{SymBilinear}(\mathbb{R}^n) \to \mathsf{Quadratic}(\mathbb{R}^n)$ is a bijection.*

Proof. Suppose $B(v,w) = \langle Mv, w\rangle$, where $M = \begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \cdots & \cdots & \cdots \\ b_{1n} & \cdots & b_{nn} \end{pmatrix}$ is a symmetric matrix. Writing $v = \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix}$, by the formula discussed above we have

$$B(v,v) = \sum_{i=1}^n b_{ii}x_i^2 + \sum_{1 \leqslant i < j \leqslant n} 2b_{ij}x_i x_j.$$

This is exactly the form of a quadratic function defined above, where $a_{ii} = b_{ii}$ and $a_{ij} = 2b_{ij}$.

If I have a quadratic function

$$q(x_1, \cdots, x_n) = \sum_{i=1}^{n} a_{ii}x_i^2 + \sum_{1 \leqslant i < j \leqslant n} a_{ij}x_i x_j,$$

then it arises as $\langle Mv, v\rangle$ where M is the symmetric matrix

$$M = \begin{pmatrix} a_{11} & a_{12}/2 & \cdots & a_{1n}/2 \\ a_{12}/2 & a_{22} & \cdots & a_{2n}/2 \\ \cdots & \cdots & \cdots & \cdots \\ a_{1n}/2 & a_{2n}/2 & \cdots & a_{nn} \end{pmatrix}.$$

□

In the discussion before getting into the linear algebra, I talked about 'rotating the ellipse $\pi/4$ radians clockwise'. In terms of the expression $q(v) = 1$, the rotated ellipse corresponds to the equation $q(Av) = 1$, where A is the linear map which rotates the plane $\pi/4$ radians counter-clockwise. This suggests that the function sending v to $q(Av)$ is worth naming and studying.

Definition 66. Suppose $A : \mathbb{R}^n \to \mathbb{R}^n$ is a linear map. If $q : \mathbb{R}^n \to \mathbb{R}$ is a quadratic function, then the *pullback along* A is the function$A^*q : \mathbb{R}^n \to \mathbb{R}$ defined by $(A^*q)(v) = q(Av)$.

Similarly, if $B : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a symmetric bilinear function, its *pullback along* A is $(A^*B)(v, w) = B(Av, Aw)$. ◊

I've really stated the same definition twice, as if $q(v) = B(v, v)$, then $(A^*q)(v) = q(Av) = B(Av, Av) = (A^*B)(v, v)$. Note that the pulled-back quadratic function is still quadratic. I'll record this statement as a lemma, but the fact for bilinear forms is immediate from the fact that A is linear (and this implies the fact for quadratic functions by the discussion above).

Lemma 129. *If q is a quadratic function on $\mathbb{R}^n$ (or B is a symmetric bilinear function), then the pullback A^*q is again quadratic (and A^*B again symmetric bilinear).*

More important is that we can explicitly compute the matrix associated to A^*q.

Lemma 130. *Suppose q is a quadratic function on $\mathbb{R}^n$ with associated symmetric matrix M. Then A^*q has associated symmetric matrix A^TMA.*

Proof. Write $q(v) = \langle Mv, v\rangle$. Then

$$(A^*q)(v) = q(Av) = \langle MAv, Av\rangle = (Av)^T(MAv) = v^TA^TMAv = \langle (A^TMA)v, v\rangle. \quad \square$$

This is *almost exactly* the expression we see when we write a linear map in terms of a different basis! The only difference is the appearance of a transpose instead of an inverse; usually the expression would be $A^{-1}MA$.

However, for orthogonal transformations, we have $A^T = A^{-1}$. This is the content of Proposition 115(a) = (c) in the real case: $A^TA = I$, and as A is a square matrix, Our theory of orthogonal diagonalization will let us transform any quadratic function into an easily-visualizable one.

Theorem 131. *Let $q : \mathbb{R}^n \to \mathbb{R}$ be a quadratic function with associated symmetric matrix M. Then there exists an orthonormal basis $\beta = (v_1, \cdots, v_n)$ of $\mathbb{R}^n$ for which*

$$q(a_1v_1 + \cdots + a_nv_n) = \lambda_1a_1^2 + \cdots + \lambda_na_n^2.$$

That is, there exists an orthogonal transformation $O : \mathbb{R}^n \to \mathbb{R}^n$ so that $(O^*q)(x_1, \cdots, x_n) = \lambda_1x_1^2 + \cdots + \lambda_nx_n^2$. (Here, O is the transformation C_β, whose associated matrix has columns the vectors $v_1, \cdots, v_n$.) The scalars $\lambda_1, \cdots, \lambda_n$ are the eigenvalues of the matrix M.

Proof. By the spectral theorem, there exists an orthogonal matrix O for which

$$O^{-1}MO = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \text{ is diagonal.}$$

The matrix $O^{-1}MO = [M]_{\beta\to\beta}$ is the matrix M written in the basis $\beta = (Oe_1, \cdots, Oe_n)$, as this reads

$$\phi_{std\to\beta}[M]_{std\to std}\phi_{\beta\to std}, \quad \text{where} \quad \phi_{\beta\to std} = C_{std}^{-1}C_\beta = C_\beta.$$

In particular, $M(Oe_i) = \lambda_i Oe_i$, so the diagonal entries are the eigenvalues of M.

Because $O^{-1} = O^T$, we see that the matrix associated to O^*q is $O^TMO = O^{-1}MO = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \lambda_n \end{pmatrix}$.

Thus

$$(O^*q)(x_1, \cdots, x_n) = \lambda_1 x_1^2 + \cdots + \lambda_n x_n^2.$$

□

Often the eigenvalues are arranged so that $\lambda_1 > \cdots > \lambda_n$.

If you think of O as being a composite of rotations (and maybe one reflection, if its determinant is negative), the picture is that we can make q into a standard quadratic function $\lambda_1 x_1^2 + \cdots + \lambda_n x_n^2$ by rotating the coordinate axes appropriately. In particular, we can understand the set of vectors $q(v) = 1$ very easily and visually.

Remark 66. If you allow non-orthogonal transformations so that I can stretch the axes in addition to rotating them, you can get every quadratic function into an even simpler form. Suppose $\lambda_1, \cdots, \lambda_k > 0$, while $\lambda_{k+1}, \cdots, \lambda_{k+\ell} = 0$, and $\lambda_{k+\ell+1}, \cdots, \lambda_n < 0$. We say that the *signature* of the quadratic function is $(n_+, n_0, n_-) = (k, \ell, n-k-\ell)$; these are the number of positive, zero, and negative eigenvalues. Then if

$$q(x_1, \cdots, x_n) = \lambda_1 x_1^2 + \cdots + \lambda_n x_n^2,$$

then if I write

$$a_i = \begin{cases} 1/\sqrt{\lambda_i} & \lambda_i > 0 \\ 1 & \lambda_i = 0 \\ 1/\sqrt{-\lambda_i} & \lambda_i < 0 \end{cases},$$

I find that

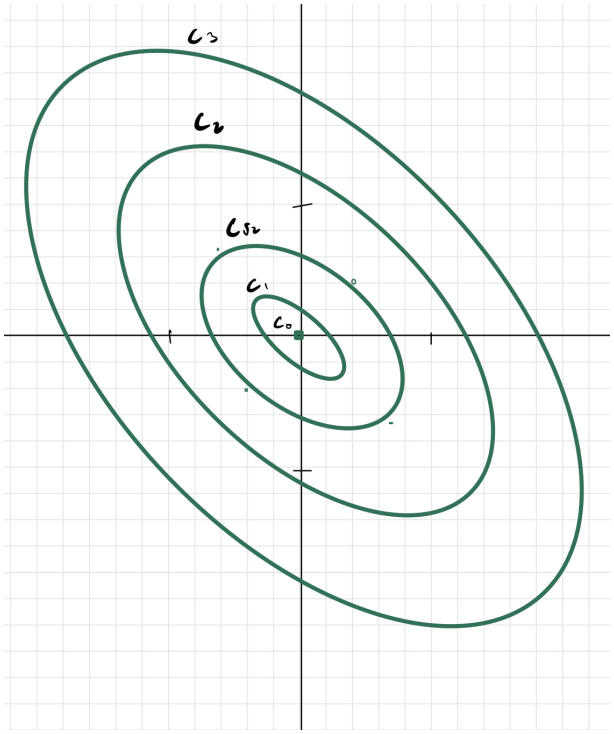
$$q(a_1x_1, \cdots, a_nx_n) = x_1^2 + \cdots + x_k^2 - x_{k+\ell+1}^2 - \cdots - x_n^2.$$

That is, applying the appropriate transformation A, the function A^*q squares the coordinates, adds some, and subtracts some.

If q and q' are quadratic functions, there exists an invertible transformation $A : \mathbb{R}^n \to \mathbb{R}^n$ so that $A^*q = q'$ if and only if the signatures (n_+, n_0, n_-) of q and q' agree. This is called Sylvester's inertia theorem, for some reason or another. ◊

7.5.1 The 2×2 case

Consider a function $q(x, y) = ax^2 + bxy + cy^2$, where not all of the coefficients are zero (as then the function would simply be the zero function); the associated symmetric matrix is $M = \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix}$. I want to understand what this function 'looks like', and a good first step at that is understanding its level curves $C_{q,k} = \{(x, y) \mid q(x, y) = k\}$ as k varies.



I claim that there are four possibilities for the shape of these sets.

I The quadratic function could be 'elliptic', in which case the sets $C_{q,k}$ are either empty, the origin, or an ellipse centered at the origin. The function q is either non-negative or non-positive; this case splits into two sub-cases, depending on whether $q \geqslant 0$ or $q \leqslant 0$.

   (a) If $q \geqslant 0$, then $C_{q,k}$ is empty for $k < 0$, while $C_{q,0} = \{(0,0)\}$, and $C_{q,k}$ is an ellipse for $k > 0$. Think $q(x,y) = x^2 + y^2$.

   (b) If $q \leqslant 0$, then $C_{q,k}$ is an ellipse for $k < 0$, while $C_{q,0} = \{(0,0)\}$, and $C_{q,k}$ is empty for $k > 0$. Think $q(x,y) = -x^2 - y^2$.

II The quadratic function could be 'hyperbolic', in which case $C_{q,k}$ is a hyperbola for all $k \neq 0$. The set $C_{q,0}$ is the union of two lines through the origin. Think $q(x,y) = x^2 - y^2$.

III The quadratic form could be 'degenerate', in which case the sets $C_{q,k}$ is always either a line, a union of two parallel lines, or empty. Think $q(x,y) = x^2$. There $x^2 = 0$ is the vertical axis and $x^2 = 1$ is the union of two vertical lines.

In case I(a), the origin is the minimum of the function: if $(x,y) \neq (0,0)$, then $q(x,y) > 0$. In case I(b), the origin is the maximum of the function: if $(x,y) \neq (0,0)$ then $q(x,y) < 0$. In case II the origin is neither a maximum nor a minimum; q is positive in some places and negative in others. (In the last case, the origin is again either a local maximum or a local minimum depending on the sign of q, but it is a 'degenerate local max/min': there are other points nearby with the same value (a whole line of them).

The punchline is that we can use linear algebra to determine immediately which case we're in! (The ideas here are certainly older than linear algebra, but I'm very fond of the way it allows us to package them together.) This statement is often used in multivariable calculus classes to state a 'second derivative test' for functions of two real variables.

Theorem 132. *The quadratic function $q(x,y) = ax^2 + bxy + cy^2$ is elliptic if and only if $b^2 < 4ac$, hyperbolic if and only if $b^2 > 4ac$, and degenerate if and only if $b^2 = 4ac$.*

Suppose we are in the first case, and $q(x,y) = ax^2 + bxy + cy^2$ is elliptic. Then a and c have the same nonzero sign, and $a, c > 0$ if and only if $q \geqslant 0$, while $a, c < 0$ if and only if $q \leqslant 0$.

Proof. By the discussion in the preceding section, there is an orthogonal matrix O for which $(O^*q)(x,y) = \lambda_1 x^2 + \lambda_2 y^2$; we have

$$C_{O*q,k} = \{(x,y) \mid q(O(x,y)) = k\} = \{O^{-1}(x,y) \mid q(x,y) = k\} = O^{-1}C_{q,k}.$$

The transformation $O(x,y)$ is either a rotation by some angle or a reflection (and in fact one can assume O is a rotation, though I won't discuss this in detail), and any rotation or reflection of an ellipse, hyperbola, line, or union of lines is another shape of the same type.

So first let's discuss the case $\lambda_1 x^2 + \lambda_2 y^2$. If both λ_1, λ_2 are positive, then $\lambda_1 x^2 + \lambda_2 y^2 \geqslant 0$, and it is equal to zero if and only if $(x,y) = (0,0)$; if $k > 0$ the set $\lambda_1 x^2 + \lambda_2 y^2 = k$ is precisely an ellipse, a stretched version of the standard circle $x^2 + y^2 = 1$, and we are in case I(a). One may write this ellipse as the set of (x,y) for which $(\sqrt{\lambda_1/k}x)^2 + (\sqrt{\lambda_2/k}y)^2 = 1$.

If $\lambda_1, \lambda_2 < 0$, then $\lambda_1 x^2 + \lambda_2 y^2 \leqslant 0$, with equality if and only if $(x,y) = (0,0)$; in this case for $k < 0$ the set $\lambda_1 x^2 + \lambda_2 y^2 = k$ is again an ellipse, a stretched version of $-x^2 - y^2 = -1$ (the standard unit circle), and we are in case I(b). One may write this ellipse as the set of (x,y) for which $(\sqrt{-\lambda_1/k}x)^2 + (\sqrt{-\lambda_2/k}y)^2 = 1$.

If $\lambda_1 > 0$ and $\lambda_2 < 0$ have opposite sign, then $\lambda_1 x^2 + \lambda_2 y^2 = 0$ defines a union of lines; this can be rewritten as $(\sqrt{\lambda_1}x)^2 = (\sqrt{-\lambda_2}y)^2$, which gives as solutions the two lines $y = \pm\sqrt{-\lambda_1/\lambda_2}x$. For $k \neq 0$, it is instead a hyperbola, stretched from the standard hyperbola $x^2 - y^2 = 1$:

$$(\sqrt{\lambda_1/k}x)^2 - (\sqrt{-\lambda_2/k}y)^2 = 1.$$

There we are in case II.

Finally, if one of the λ's is zero — say, $\lambda_2 = 0$ — then the equations are $\lambda_1 x^2 = k$. When $k = 0$ this is a single line, when k is nonzero and it has the same sign as λ_1 this is the two lines $x = \pm\sqrt{k/\lambda_1}$, when they have opposite sign it is empty. This is case III.

(It is not possible for $\lambda_1 = \lambda_2 = 0$, as this would imply $q(x,y) = 0$, and I assumed the function is not zero.)

So the cases are determined by the eigenvalues λ_1, λ_2 of the matrix $M = \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix}$, and in particular what their signs are. Here's the fun part. *The determinant of a matrix is the product of its eigenvalues* (counted with multiplicity, so if M has characteristic polynomial $(\lambda - 2)^2$, its determinant is 4); so here

$$ac - b^2/4 = \det M = \lambda_1\lambda_2, \quad \text{or rephrased} \quad 4ac - b^2 = 4\lambda_1\lambda_2.$$

If one of the eigenvalues is zero — the degenerate case — then $\det M = \lambda_1\lambda_2 = 0$. This means $4ac - b^2 = 0$, or $4ac = b^2$.

If the eigenvalues have opposite sign (the hyperbolic case), then $\det M = \lambda_1\lambda_2 < 0$, so that $4ac - b^2 < 0$, or $b^2 > 4ac$. Finally, if the eigenvalues have the same sign (positive or negative!) we have $\det M = \lambda_1\lambda_2 > 0$. This gives us case (I).

As for identifying between cases I(a) and I(b), notice that in case I(a) $q(x,y) > 0$ away from the origin and in case I(b) $q(x,y) < 0$ away from the origin. As we have $a = q(1,0)$ and $c = q(0,1)$, the signs of either of these determine which of the two sub-cases we are in. □

This can be rephrased in an amusing but useless way in terms of the characteristic polynomial.

Corollary 133. *The point $(0,0)$ is an isolated local maximum of the function $q(x,y) = ax^2 + bxy + cy^2$ if and only if $M = \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix}$ has characteristic polynomial $p_M(t) = t^2 + dt + e$ where $d, e > 0$.*

Proof. The only case of the four in which $(0,0)$ is an isolated local maximum is case I(b), where $\lambda_1, \lambda_2 < 0$. We have

$$p_M(t) = (\lambda - \lambda_1)(\lambda - \lambda_2) = \lambda^2 - (\lambda_1 + \lambda_2)\lambda + \lambda_1\lambda_2.$$

The quantity $e = \lambda_1\lambda_2$ is the product of the eigenvalues — the determinant of that matrix — and $d = -\lambda_1 - \lambda_2$ is the negative of their sum. If $e = \lambda_1\lambda_2 > 0$, then both λ_1 and λ_2 have the same sign; if $d > 0$ then thaht sign must be negative. So under the given assumptiom, both eigenvalues are negative — so we are in case I(b) above. □

Remark 67. Similar logic as the preceding discussion implies that if $q(\vec{x})$ is a quadratic functon with associated symmetric matrix M, the origin $\vec{0}$ is an isolated local maximum of q if and only if M has all negative eigenvalues (and $\vec{0}$ is an isolated local minimum if and only if M has all positive eigenvalues).

I believe it is the case that all of the values $\lambda_1, \cdots, \lambda_n$ are negative if and only if the coefficients of $p_M(t)$ are all positive; equivalently, that

$$\lambda_1, \cdots, \lambda_n < 0 \quad \iff \quad \text{for all } k, \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant n} \prod_{j=1}^{n} \lambda_{i_j} < 0.$$

The expression on the right is called the *k'th symmetric polynomial*. For $n = 2$, these are $\lambda_1\lambda_2$ and $\lambda_1 + \lambda_2$. For $n = 3$, they are

$$\lambda_1\lambda_2\lambda_3, \quad \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3, \quad \lambda_1 + \lambda_2 + \lambda_3.$$

This would imply that $\vec{0}$ is an isolated local minimum of q if and only if $p_M(t)$ has all positive coefficients. An amusing, but computationally useless, criterion. (In real practice, you would probably just determine the eigenvalues in a more computationally stable way.) ♢

See you next term, where we'll actually use this to discuss the second-derivative test in multivariable calculus.

Remark 68. One can understand the principal axes of the ellipse, too, by determining the precise eigenspaces of the associated matrix $M = \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix}$. For the example opening this section, $M = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}$. The minor axis is parallel to $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, while the major axis is parallel to $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$. With respect to this basis the formula for the ellipse is $5y_1^2 + y_2^2 = 1$, or $(\sqrt{5}y_1)^2 + y_2^2 = 1$. There is a stretch factor of $1/\sqrt{5}$ in the direction of the minor axis. ♢