Linear Algebra Done Half-Right: A Semi-Practical Introduction

Connor Harris

Draft as of September 24, 2024

Preface

Linear algebra began as a set of *ad hoc* matrix algorithms for solving systems of linear equations, and most students of science and engineering still learn it in the same way. Teaching linear algebra purely as matrix manipulation avoids the need to introduce mathematical abstractions, such as axiomatic definitions of vector spaces and linear transformations. But there are tradeoffs: many simple notions get obscured behind matrix notation, and applying matrix algebra to fields such as quantum mechanics and real analysis, in which vector spaces and operators are often infinite-dimensional or don't have obvious matrix representations, is harder than applying a more abstract theory that treats vector spaces and operators by themselves.

A few other books—perhaps the most popular is *Linear Algebra Done Right* by Sheldon Axler, an inspiration for much of this book—instead deemphasize matrices and work with abstract, axiomatically defined vector spaces and linear maps. These books are aimed at students of pure mathematics, and though they illustrate the conceptual unity of the subject well, they're less helpful in teaching practical applications of matrix algebra. Axler, for instance, favors theoretical elegance to the point of defining many concepts with vital applications—most notably determinants, relegated to one short section almost at the end of *Linear Algebra Done Right*—in terms of abstractions that most scientists and engineers won't need.

In this book, I've tried to fill in the gap between these two approaches, and create something that could help two groups of readers:

1. Students who are taking a standard applied course in linear algebra, focused on matrix algorithms, who may not be used at first to dealing with axiomatically defined mathematical objects but who think a more abstract view on the same material will help solidify their understanding and make the algorithms more intuitive.

2. Mathematicians who are used to working at a higher level of abstraction who would like to see how the more abstract concepts they are likely already used to translate into matrix manipulations.

For the first group, I've tried to make the introduction to thinking at a higher level of abstraction as gentle as I could. A short introductory Chapter 0, intended more for reference than for reading straight through, is a quick introduction to these concepts for readers that need it. (If you're looking for a fuller introduction, I recommend *Tools of the Trade: Introduction to Advanced Mathematics*, by Paul J. Sally.) While writing, I've had these objectives constantly in mind:

1. Make frequent connections between the two languages of matrix algebra and axiomatic vector spaces, but keep them conceptually distinct. In particular, I usually don't treat matrices and column vectors as linear maps and vectors in their

own right: instead, matrices are introduced as representations of more abstract objects.

This aim has sometimes required a bit of nonstandard notation. For instance, while many books use $\mathbb{R}^3$ to denote the set of column vectors with three real entries (that is, real-valued matrices with three rows and one column), I call this space $\mathrm{Col}_3(\mathbb{R})$ and reserve $\mathbb{R}^3$ for the set of ordered triples of real numbers without matrix structure: the column vector $\begin{bmatrix} a \\ b \\ c \end{bmatrix} \in \mathrm{Col}_3(\mathbb{R})$ is the "column vector representation relative to the standard basis" of the ordered triple $(a, b, c) \in \mathbb{R}^3$.

Keeping vector spaces and maps conceptually distinct from their matrix representations may seem a bit pedantic, but I think this separation will clarify many topics and prevent some conceptual confusions from ever arising: the idea that a matrix and its diagonalization, for instance, produce two representations of the same underlying operator is much clearer when you're used to treating matrices and linear operators as different objects with separate terminologies. This approach also generalizes more easily to the theory of infinite-dimensional vector spaces required in many other fields.

2. Illustrate abstract results, as often as possible, in simple concrete settings. Many statements and results about general vector spaces, for instance, are illustrated with an easy-to-understand (and, when possible, easy-to-visualize) example in a simple vector space such as $\mathbb{R}^2$ or $\mathbb{R}^3$.

3. Provide full proofs for most results, but present them in a form that I hope is clearer for newcomers to higher mathematics. For instance, when a section involves a proof of a complicated theorem, I've tried to explain in advance why the theorem will eventually prove useful in order to help readers stay oriented. I've tried to make the proof language easier to understand even at the cost of some concision, including providing redundant but easier-to-understand definitions and breaking up complicated deductions into explicit numbered lists of key steps.

I haven't yet provided a full set of exercises for every section: I anticipate that most readers will use this book as a supplement to another book or course rather than as a source in its own right. But many sections (eventually, all) have an introductory list of easy "key questions" that ask about that section's core concepts and definitions. These questions can serve as a quick check of comprehension and as a guide to review: if you can answer a section's key questions without too much difficulty, you probably remember the section's key concepts well. If you use flashcards to study (or a spaced repetition software such as Anki), you may want to make the key questions into flashcards. Key questions that you should be able to answer off the top of your head are unmarked, questions that might take a moment of thought are marked with $(\star)$, and questions that take more thought (and possibly work with pencil and paper) are marked with $(\star\star)$.

This book is essentially complete in scope, but it is still a work in progress, and I would appreciate any corrections or suggestions for improvement. I can be reached by Twitter DMs (handle `@cmhrrs`) or email (`connorh94 at-sign gmail dot com`).

Contents

Chapter 0

Introduction to set and function concepts

This book uses notation and vocabulary for sets and functions that may be new to you if you haven't taken higher-level mathematics courses before. This section is a quick introduction. You don't need to memorize everything right away: you can learn the details as you go along. If most of the material in this section looks familiar to you after a quick skim, feel free to jump into the main material in Chapter 1.

0.1 Sets

A *set* is a collection of items, called "elements." Sets can't contain duplicate elements: any object is contained in a set either zero times or one time. The order of elements in a set also doesn't matter. (There are whole textbooks in set theory devoted to making these ideas as precise as possible, but these common-sense notions are good enough for our purposes.)

To say that some object x is in the set S, we write $x \in S$ or (more rarely) $S \ni x$. To write that x is not an element of S, write $x \notin S$.

To write out a set, list its elements separated by commas within curly braces: $\{1, 2, 4\}$ is a set that contains the elements 1, 2, and 4. (Another way of writing the same set is $\{2, 4, 1\}$, because elements in a set don't have an order.) The set with no elements is called the empty set. We can write it as $\{\}$ or as $\emptyset$.

If S and T are two sets, then the set of elements that are in both S and T is called the *intersection*, and denoted $S \cap T$. Two sets that don't have any elements in common are *disjoint*. The set of elements in either S or T (or both) is called the *union*, and denoted $S \cup T$. The set of elements in S but not T is called the *set difference* and denoted $S \setminus T$. For instance, if $S = \{1, 4, 5, 8\}$ and $T = \{4, 5, 9, 10, 12\}$, then:

- The intersection $S \cap T$ is $\{4, 5\}$.
- The union $S \cup T$ is $\{1, 4, 5, 8, 9, 10, 12\}$.
- The set difference $S \setminus T$ is $\{1, 8\}$.
- The set difference $T \setminus S$ is $\{9, 10, 12\}$.

The number of elements in a set is denoted with the sign $|\cdot|$. For instance, with S and T as above, $|S| = 4$, $|T| = 5$, and $|S \cup T| = 7$.

If every element of S is also in T, then S is called a *subset* of T, and T is a *superset* of S. These statements are written $S \subseteq T$ and $T \supseteq S$. If T also has at least one element that S doesn't have (that is, S and T aren't identical), then S is called a *strict subset* of T and T is a *strict superset* of S. This is written $S \subset T$ and $T \supset S$, or sometimes $S \subsetneq T$ and $T \supsetneq S$ (with a slash denoting negation through the bottom bar) for absolute clarity. The easiest way to prove that two sets S and T equal each other is often to give a proof that $S \subseteq T$ (that is, if $x \in S$, then $x \in T$) and a separate proof that $T \subseteq S$.

A few common sets have special symbols:

- $\mathbb{N}$ is the set of positive (or "natural") integers: $\{1, 2, 3, \ldots\}$.
- $\mathbb{N}_0$ is the set of non-negative integers: $\{0, 1, 2, 3, \ldots\}$. (A minority of books denote this set by $\mathbb{N}$ instead, but we'll always use $\mathbb{N}$ to exclude zero.)
- $\mathbb{Z}$ is the set of all integers: $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$. (The letter Z comes from the German word *Zahlen*, which means "numbers.")
- $\mathbb{Q}$ is the set of rational numbers. (The Q comes from *quotient*.)
- $\mathbb{R}$ is the set of real numbers.
- $\mathbb{C}$ is the set of complex numbers.

If X is any set, then X^2 is the set of ordered pairs of elements in X. Likewise, X^3 is the set of ordered triples, X^4 the set of ordered quadruples, and so on. If X and Y are possibly different sets, then $X \times Y$ is the set of ordered pairs with the first element from X and the second element from Y.

0.2 Set-builder notation

Set-builder notation denotes a set by giving a common formula for its elements. This notation has two variants, both of which use a formula with two parts separated by a colon or a vertical bar. In one variant, the left-hand side of the formula specifies a variable standing for elements of the set, and the right-hand side gives conditions For example:

- $\{n \in \mathbb{Z} : |n| \leq 2\}$ is $\{-2, -1, 0, 1, 2\}$, the set of integers with absolute value at most 2.
- $\{(a, b) \in \mathbb{N}^2 : a + b = 5, a < b\}$ is the set of pairs of positive integers, the first integer less than the second, whose sum is 5. This set is $\{(1, 4), (2, 3)\}$, with two elements.

In the second variant, we put a formula on the left and the allowable values of the formula's variables on the right. Any value of the formula is an element of the set. For example:

- $\{n^3 : n \in \mathbb{N}\}$ is the set of all positive cubes: $\{1, 8, 27, 64, 125, \ldots\}$.
- $\{(n, \sqrt{n}) : n \in \mathbb{N}, n \leq 5\}$ is the set of ordered pairs of the first five natural numbers and their square roots: $\{(1, 1), (2, 1.414), (3, 1.732), (4, 2), (5, 2.236)\}$ (rounding the square roots).

0.3 Functions

0.3.1 Core definitions; domain and range

A *function* f from one set X to another set Y, denoted $f : X \to Y$, is a pairing of every element of X to another element in Y. The element of Y that is matched with some element $x \in X$ is denoted $f(x)$. (If you've done computer programming in languages with static types, it may help to think of the notation $f : X \to Y$ as a type signature.)

Every element of X needs exactly one partner in Y, but not the other way around: one element of Y could be paired with multiple values of X, or none at all. X is called the *domain* of f, and Y is the *codomain*. The set of elements in Y that f actually uses (i.e. every y for which there's some $x \in X$ such that $f(x) = y$) is called the *image* or *range* of f and can be written $f(X)$ or $\operatorname{im} f$.

You're probably used to seeing functions defined by a formula, such as $f(x) = x^2$ or $g(x) = \cos \log |x + 1|$. But the definition of a function in set theory doesn't require the function to have a nice formula.

For example, suppose $X = \{1, 2, 3\}$, and $Y = \{a, b, c, d\}$. You might define a function $f : X \to Y$ based on the pairs $(1, a)$, $(2, c)$, $(3, a)$, in which case the values of f are $f(1) = a$, $f(2) = c$, $f(3) = a$. The image of f is $\{a, c\}$.

0.3.2 Injective, surjective, bijective

A function f is called:

- *injective*, if different elements of its domain always have different values (that is, if $f(x_1) \neq f(X_{RM})$ whenever $x_1 \neq X_{RM}$);
- *surjective*, if the range of f is all of Y;
- *bijective*, if it is both injective and surjective.

If $f : X \to Y$ is any function and $S \subseteq X$, we can write $f|_S : S \to Y$ for the *restriction* of f to S. That is, $f|_S(x) = f(x)$ if $x \in S$, and $f|_S(x)$ is undefined if $x \in X \setminus S$.

0.3.3 Function composition

The *composition* of two functions $X \to Y$ and $g : Y \to Z$, denoted $g \circ f$, is the result of applying f and then g. (To reiterate: composition runs *right to left*. This is crucial to remember, and as you may come to appreciate later, it's arguably a defect of standard mathematical notation.) If $f(x) = y$ and $g(y) = z$, then $(g \circ f)(y) = z$. Function composition is associative. Suppose we have three functions $e : W \to X$, $f : X \to Y$, and $g : Y \to Z$ and elements $w \in W, x \in X, y \in Y, z \in Z$ such that $e(w) = x$, $f(x) = y$, and $g(y) = z$. Then $(f \circ e)(w) = y$ and thus $g \circ (f \circ e)(w) = g(y) = z$. Similarly, $((g \circ f) \circ e)(w) = (g \circ f)(x) = z$. So $(g \circ f) \circ e = g \circ (f \circ e)$.

0.3.4 Identity and inverse functions

The *identity function* from a set X to itself, sometimes denoted $\mathrm{id}_X : X \to X$, is the function that leaves every element unchanged: $\mathrm{id}_X(x) = x$ for every element $x \in X$.

If $f : X \to Y$ is bijective, then it has an *inverse function* $f^{-1} : Y \to X$ that reverses the pairing of elements defined by f: if $f(x) = y$, then $f^{-1}(y) = x$, and vice versa. If g is the inverse of f, then f is also the inverse of g, as you might like to convince yourself.

Even if $f : X \to Y$ isn't bijective and so doesn't have an inverse, we can still define the *preimage* $f^{-1}(S)$ for any subset $S \subseteq Y$. This is the set $\{x \in X : f(x) \in S\}$ of every element that f maps into S. For instance, if we define the (non-injective) function $f : \mathbb{R} \to \mathbb{R}$ as $f(x) = x^2$, then the preimage of the set $S = \{-1, 0, 4\}$ is $f^{-1}(S) = \{-2, 0, 2\}$ (because $f(0) = 0$, $f(-2) = f(2) = 4$, and there's no real number x such that $f(x) = -1$).

If $f : X \to Y$ is a bijective function and $g : Y \to X$ is its inverse, then $g \circ f$ is the identity function on X and $f \circ g$ is the identity function on Y. It's impossible to have $g \circ f = \mathrm{id}_X$ without $f \circ g = \mathrm{id}_Y$, or vice versa, as long as f and g are both bijective.[1] (Proof: suppose f, g are bijective and $g \circ f = \mathrm{id}_X$. For every $y \in Y$, take $x \in X$ to be the necessarily unique element in Y such that $f(x) = y$. Then since $g \circ f = \mathrm{id}_X$, so $g(y) = (g \circ f)(x) = x$ and so $(f \circ g)y = f(x) = y$ for every element $y \in Y$, so $f \circ g = \mathrm{id}_Y$. The proof that $f \circ g = \mathrm{id}_Y$ implies $g \circ f = \mathrm{id}_X$ is symmetrical.)

Finally, functions can be written in a shorthand form with the symbol $\mapsto$. For example, $x \mapsto x^2$ is shorthand for the function f given by the formula $f(x) = x^2$.

0.4 Quantifiers

0.4.1 Definitions

The symbols $\exists$ and $\forall$ are called "quantifiers" and they mean, respectively, "there exists" and "for all." For example, $(\exists x \in S)x > 0$ means "at least one element in S is positive," while $(\forall x \in S)x > 0$ means "every element in S is greater than zero."

0.4.2 $\forall$ and $\exists$ can't be reversed

These quantifiers can be combined in the same expression, but be warned that changing the order of terms with $\exists$ and $\forall$ changes the meaning of a statement! For example, suppose X and Y are two sets. Then $(\forall x \in X)(\exists y \in Y)x + y \in \mathbb{Z}$ means "for every element x in X, there's some other element $y \in Y$, which may depend on x, such that $x + y$ is an integer. But $(\exists y \in Y)(\forall x \in X)x + y \in \mathbb{Z}$ means "there's some specific element of Y that, when added to every element of X, produces an integer." This is a much stronger claim. For instance, the pair of sets $X = \{0.4, 0.7\}$ and $Y = \{0.3, 1.6\}$ satisfies the first statement (if you take 0.4 from X, then you can choose 1.6 from Y; and if you take 0.7 from X, then you can take 0.3 from Y) but not the second (there's no element $y \in Y$—in fact, no real number at all—such that $y + 0.4$ and $y + 0.7$ are both integers).

As a more concrete example that also illustrates why quantifiers can disambiguate thoughts that would be ambiguous if expressed in English, [2], let S be the set of all humans in the world, and let $L(x, y)$ be a predicate[3] representing the sentence "x loves

[1] If one of f or g isn't bijective (so it doesn't have a real inverse), then it's possible for $f \circ g$ but not $g \circ f$ to be an identity. Consider, for instance, the functions $f, g : \mathbb{N} \to \mathbb{N}$ on the set of positive integers defined as $f(n) = \max(n - 1, 1)$ and $g(n) = n + 1$.

[2] Borrowed, I believe, from *A Modern Formal Logic Primer* by Paul Teller.

[3] *Predicate* is a logicians' term for a function that takes the values *true* and *false*.

y": that is, it's a function Then the sentence "Everybody loves somebody" could be interpreted in two ways:

- $(\forall x \in S)(\exists y \in S)L(x, y)$: that is, "Everybody loves at least one person."
- $(\exists y \in S)(\forall x \in S)L(x, y)$: that is, "There is some specific person, y, who is loved by everyone in the world."

These are quite clearly not equivalent statements!

A big portion of adjusting to higher mathematics is getting comfortable with working with nested quantifiers. Even in earlier mathematics, you've probably seen and (likely) been initially confused by a few definitions that used multiple layers of quantifiers, though probably not with our logicians' notation here. Perhaps the best example is the epsilon–delta criterion for a function f on the real numbers to be continuous at some point x_0:

$$(\forall \epsilon \in \mathbb{R}^+)(\exists \delta \in \mathbb{R}^+)(\forall x \in \mathbb{R})(|x - x_0| \geq \delta \text{ or } |f(x) - f(x_0)| < \epsilon).$$

0.4.3 Negation of quantifier statements

If $P(x)$ is some predicate (which you can think of as a function that maps possible values of x to the set $\{\text{true}, \text{false}\}$) and $\overline{P}(x)$ is the negation of P, then the negation of $(\forall x)P(x)$ is $(\exists x)\overline{P}(x)$ and the negation of $(\exists x)P(x)$ is $(\forall x)\overline{P}(x)$. This should be relatively intuitive: if the statement "$P(x)$ is true for all $x \in S$" is false, then there must be some element $x \in S$ for which $P(x)$ is false. This extends to multiple quantifiers: for instance, the negation of $(\forall x)(\exists y)(\forall z)P(x, y, z)$ is $(\exists x)(\forall y)(\exists z)\overline{P}(x, y, z)$.

As a more concrete example, the negation of the statement $(\forall x \in S)x < 0$ (that is, S contains only negative numbers) is $(\exists x \in S)x \geq 0$ (S contains at least one positive number).

I once took a formal logic course that used the symbol $\neg$ to denote negation, and the instructor referred to these equivalences somewhat playfully as the "swimmy-past" rule: in $\neg(\forall x)P(x)$, the $\neg$ statement can "swim past" the $\forall$ term, in the process changing it to $\exists$, producing the result $(\exists x)\neg P(x)$. Symmetrically, if you start with $(\exists x)\neg P(x)$, then the $\neg$ symbol can swim to the left and flip $\exists$ back to $\forall$. Similarly, $\neg(\exists x)P(x)$ would turn into $(\forall x)\neg P(x)$. It's also clearer how this rule can be applied one quantifier at a time: for instance, $\neg(\forall x)(\exists y)(\forall z)P(x, y, z)$ turns into $(\exists x)\neg(\exists y)(\exists z)P(x, y, z)$ by the first application, and then $(\exists x)(\forall y)\neg(\exists z)P(x, y, z)$ on the second and finally $(\exists x)(\forall y)(\exists z)P(x, y, z)$ on the third.

These negation formulas underlie the most commonly used proof technique in this book: proof by contradiction. Suppose, for example, that we're studying a class of objects called *frobnicators*, each of which have associated real numbers with some special property called *transmogrifying values*.[4] We may have to prove a statement like this:

Proposition. *Every frobnicator has at least one positive transmogrifying value.*

This statement, spelled out in symbols, would be:

$(\forall F \in \text{set of frobnicators})(\exists x \in \mathbb{R})(x > 0 \text{ and } x \text{ is a transmogrifying value of } F.$

[4]These obviously aren't real concepts; the point of choosing fake ones is to focus your general attention on the abstract form of the argument, not on any particular mathematical object.

To prove this, we can disprove the negation, namely:

$$(\exists F \in \text{set of frobnicators})(\forall x \in \mathbb{R})(x \leq 0 \text{ or } x \text{ is not a transmogrifying value of } F).$$

(The negation of "P and Q" is "not-P or not-Q," and vice versa.) With this in mind, we can start writing a proof by assuming the existence of a counterexample, following the negation:

Proof. Suppose that F is a frobnicator that does not have a positive transmogrifying value. By the Transmogrifiability Lemma, F must have at least one transmogrifying value; call it x. Necessarily, $x \leq 0$. Therefore, ... [and try to prove that F or x must have some impossible property, such as contradicting another known result on frobnicators] □

0.4.4 Quantification over empty sets

Any statement involving $\forall$ is true if the set in the $\forall$ statement is empty, and any statement involving $\exists$ is false. For instance, if $S = \emptyset$, then $(\forall x \in S)x > 0$ and $(\forall x \in S)x < 0$ are both true.

This may seem a bit paradoxical, but the convention for $\exists$ should at least make sense: it's hard to say that a sentence of the form "there exists some element of S such that ..." could be true if there are no elements of S to begin with. And to keep the common-sense rule that "there is at least one x such that P" and "for all x, not-P" should be opposite, the rule for $\forall$ follows. This will save us from having to insert a lot of special cases into theorem definitions.

0.5 The colon-equals definition symbol

Sometimes we'll use the symbol $\coloneqq$ to mean equality by definition: a new name on the left refers to the known object given on the right. $=:$ is symmetrical: the known object is on the left and the new name is on the right. This lets us reserve the symbol $=$ for expressing an equality between two preexisting objects, not giving one known object a new name.

For example, we might express a result on two sets X and Y as:

> Therefore, $X \cap Y$ is the set $\{n^3 : n \in \mathbb{N}, n \leq 10\}$ of the first ten positive cubes. Call this new set S.

Or we can use the new symbol to abbreviate this as:

> Therefore, $S \coloneqq X \cap Y = \{n^3 : n \in \mathbb{N}, n \leq 10\}$.

If this sounds confusing, don't worry: it will be clearer in practice. And in any case, most writers will use $=$ to mean definitions when it's clear from the context that a new object is being defined, and use $\coloneqq$ only when needed for disambiguation.

Chapter 1

Vector spaces

Overview. The core concept in linear algebra is a *vector space*, a collection of objects that can be added to each other ("vector addition") or multiplied by single numbers ("scalar multiplication"). A prototypical example is the set of all ordered pairs of real numbers, commonly denoted $\mathbb{R}^2$: we can add two ordered pairs by adding their components (for instance, $(1, 2) + (5, -3) = (6, -1)$, or multiply an ordered pair by another real number by another real number by multiplying the individual components (for instance, $4(1, 2) = (4, 8)$). Expounding this intuition further is the job of Section 1.1.

Most of the vector spaces that you'll deal with in linear algebra are simple lists of numbers like these, but the concept of a vector space is even more abstract: for instance, the set of differentiable functions defined on a certain interval of the real number line can be considered as a vector space, and this fact lets us apply concepts from linear algebra to solve differential equations. In particular, a vector space as a set of objects with two arithmetic operations that must satisfy a specific list of properties modeled after addition and multiplication, which ultimately amount to "you can manipulate algebraic expressions on arbitrary vector spaces, without knowing anything about the structure of the vectors themselves, in the same way as vectors in more ordinary spaces such as $\mathbb{R}^2$.

The axiomatic definition of a vector space also relies on two more axiomatic concepts: an *abelian group* (roughly: a collection of objects that can be added or subtracted, such as the set of integers) and a *field* (an abelian group in which elements can also be multiplied and divided, such as the set of real numbers). Sections 1.2 through 1.4 present these axioms. If you're not used to working with axiomatically defined structures, then these sections will be a useful introduction; but by the same token, however, if you find them confusing, you can (at least for the moment) skim over them and continue with Section 1.5, confident that the upshot is "algebra on vector spaces works the way you would expect it to."

Sections 1.5 through 1.8 present the core concepts of *linear combination*, *span*, *subspace*, *linear independence*, and *basis*, each of which is immediately tied into the two core vector space operations of addition and mlultiplication. A linear combination of a set of vectors is an expression that you can make by multiplying some of the starting vectors by scalars and adding them together; the set of values of all such combinations is the span of the starting set. For instance, if S is the set of vectors $\{(1, 4), (4, 2)\}$ in $\mathbb{R}^2$, then $2(1, 4) + 3(1, 2)$ is a linear combination in S, and its value $(5, 14)$ is an element of the span of S.

A subspace is a set that is its own span: the sums and scalar multiples of all the

elements of the subspace are also in the subspace: in essence, it's a smaller vector space contained in a larger one. Another key concept is that of *subspace sum*: the smallest subspace that contains the union of two or more smaller subspaces. This notion is crucial for some fundamental theorems presented in later chapters, which rest on the decomposition of a vector space into a sum of subspaces with special properties.

The trickiest notion in this set is *linear independence*: essentially, a linearly independent set is as small as a set with its span could possibly be—removing any element from the set means shrinking its span. Finally, a *basis* for a vector space (or subspace) is a linearly independent set whose span is the entire space or subspace.

These notions are basic and a bit boring, but important to understand thoroughly: a large number of practical problems in applied linear algebra come down to determining whether a set is linearly independent.

Finally, 1.9 introduces the notion of an "affine space" or "coset" of a subspace, and gives a few equivalent definitions. An affine space is the result of adding a fixed vector to every element of a subspace; intuitively speaking, it's a space parallel to but offset from its corresponding vector subspace.

There will be more information presented later on the basic theory of vector spaces as objects in their own right, particularly in Chapter 5. Chapter 1 is limited to the basic knowledge necessary for the basic theory of linear maps presented in Chapter 2, which in turn is the minimum required to make intuitive the core matrix algorithms such as Gauss–Jordan elimination in Chapter 3.

1.1 Motivation; vectors in physics

Key questions.

1. What is a vector in physics? What does it mean to break a vector into components?

2. What two properties of vectors in physics are the basis for our more generalized idea of vector spaces?

3. What does it mean for a function to be "linear"?

To oversimplify, linear algebra is the study of functions called *linear transformations*. When you took algebra in middle or high school, you probably learned that "linear functions" are functions of the form $f(x) = ax + b$, and their graphs are straight lines. The definition of linearity in linear algebra, though, is a bit different, and only some of the "linear" functions from high school algebra are "linear" in the linear algebra definition—namely, the function $f(x) = ax$, whose graphs are lines through the origin. (This is a good opportunity for a general warning: different fields of mathematics often use the same word for slightly different concepts!)[1]

Linear functions such as $f(x) = ax$ have two important properties. First, if you multiply the input to f by some factor k, you cause the output to be multiplied by the same amount: $f(kx) = kf(x)$. Mathematicians will say as a shorthand that a function f with this property "respects multiplication."

[1]Functions like $f(x) = ax + b$ for $b \neq 0$ are "affine" in the vocabulary of linear algebra, but we won't talk about affine functions much.

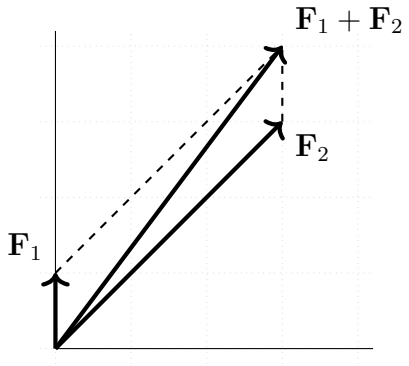
Second, if you add two inputs before giving them to f, you get the same result as if you give them to f separately and then add the outputs: $f(x+y) = f(x) + f(y)$. Mathematicians would say that f "respects addition." To summarize: f is linear if, when you add two inputs to f together, or when you scale one input up, the output acts the same way.

Of course, f is simple: it takes one input and gives one output. But you can find more complicated functions with multiple inputs or outputs that fit the same principle. Let's consider functions of the class $f(x,y) = ax+by$, where a and b are some constants. Then f is a function with two inputs and one output, but it also, in its own way, respects addition and multiplication. If we define addition on pairs of numbers component-by-component as $(x_1, y_1) + (x_2, y_2) = (x_1 + X_2, y_1 + y_2)$, then f respects addition: $f(x_1 + x_2, y_1 + y_2) = f(x_1, y_1) + f(x_2, y_2) = ax_1 + ax_2 + by_1 + by_2$. If we define multiplication of a single number k by a pair of numbers (x, y) component-by-component, as $k \times (x, y) = (kx, ky)$, then f also respects multiplication: $f(kx, ky) = kax + kby = k(ax + by) = kf(x, y)$.

Linear algebra, in essence, is the classification of linear functions like these. In linear algebra, functions are defined on abstract mathematical structures called *vector spaces*, which are sets of elements called *vectors*. A linear function sends each element of one vector space to an element of a possibly different vector space. The function $f(x, y) = kx + ky$, for instance, maps ordered pairs of real numbers to single real numbers, and the set of real numbers and the set of real number pairs are both vector spaces. In mathematical symbols, we denote the vector space of real numbers by $\mathbb{R}$, and the vector space of pairs of real numbers by $\mathbb{R}^2$.

But what is a vector space, and why are $\mathbb{R}$ and $\mathbb{R}^2$ examples? It will help to look at a field with a more concrete concept called "vector": in physics, a vector is a quantity with both magnitude and direction. Velocity, force, momentum, and electric field, for example, are "vector" quantities: an object's velocity might be four meters per second northwest, a force might be 10 newtons upward, an electric field may be 20 volts per meter southward at an angle 25 degrees below horizontal. And vectors can be broken into components showing their extent along different axes. For example, a velocity of 4 m/s northwest can be broken into components in the cardinal directions as $2\sqrt{2} \approx 2.83$ m/s west, and $2\sqrt{2} \approx 2.83$ m/s north. You could represent this vector as an ordered pair relative to a coordinate axis, such as $(-2.83, 2.83)$ with an east-pointing x-axis and a north-pointing y-axis (or, in three-dimensional space, $(-2.83, 2.83, 0)$ with a vertical z-axis). Quantities such as mass, energy, electric potential, and electric charge, on the other hand, are "scalar": they're just magnitudes that don't point anywhere.

Finally, you can do two important arithmetic operations on vectors in physics, and these operations motivate the more general definition of "vector" in pure mathematics. First, you can add vectors component-by-component. If one force $\mathbf{F}_1$ of 1 newton is pulling an object north, say, and another force $\mathbf{F}_2$ of $3\sqrt{2}$ newtons is pulling it northeast—broken into components, that's $(0, 1)$ and $(3, 3)$—then the resulting force can be calculated comes from adding the components together to get $(3, 4)$. So the total force $\mathbf{F}_1 + \mathbf{F}_2$ is 5 newtons, pulling at a compass heading of $\arctan(3/4) = 36.9°$, a little bit north of northeast. Geometrically, $\mathbf{F}_1 + \mathbf{F}_2$ looks like the endpoint of $\mathbf{F}_2$ if you moved it to start at $\mathbf{F}_1$'s end, or vice versa.



Second, you can multiply vectors by scalars, producing another vector that points in the same (or the opposite) direction. If $\mathbf{E}$ is an electric field of 10 newtons per coulomb pointing to the north, then $3\mathbf{E}$ is a field of 30 N/C pointing north. Likewise, $\mathbf{E}/2$ is 5 N/C pointing north, and $-2\mathbf{E}$ is 20 N/C pointing south. If q is a charge of 4 coulombs (a scalar quantity), then $q\mathbf{E}$ is a force vector of 40 newtons pointing north—that is, the vector of force that the electric field $\mathbf{E}$ would exert on an object with charge q. (For the most part, in pure mathematics, we won't deal with quantities that have physical units attached, but all the concepts transfer into physics quite directly.)

A vector space in linear algebra is just a set of objects that we'll call "vectors" and that have these two properties. You can add vectors to other vectors, and you can scale individual vectors up and down, to get results in the same vector space. And a vector space is a set of vectors together with definitions of addition and scaling. The elements of a vector space could be a wide variety of objects that have lots of additional properties beyond just these two operations, and we'll see important subcategories of vector spaces in which the vectors have other operations defined on them as well. But to qualify as a vector space, those two operations, as long as the way they're defined satisfies a few important properties, are all that a set needs. We'll see examples soon enough.

Answers to key questions.

1. A vector in physics is a quantity that has both a magnitude and a direction. Breaking a vector into components means taking a set of coordinate axes and writing it as a list of projected lengths onto each axis.

2. The two properties of vectors in physics that give rise to our more general notion of vectors in linear algebra are that vectors can be added to each other and also scaled up or down.

3. A function is linear if vector operations on its inputs create the same effect on its output: adding two inputs also adds their outputs together, and scaling an input scales the resulting output.

1.2 Abelian groups

Key questions.

1. What is an abelian group? What four properties must the operation on an abelian group satisfy?

2. How many identities can an abelian group have? How many inverses can any element of an abelian group have?

3. ($\star$) Is the set of integers with the operation of addition an abelian group? What about the set of even integers? What about the set of real numbers? What about the set of positive real numbers?

4. ($\star$) Why isn't the set of complex numbers with the operation of multiplication an abelian group? Can you make it an abelian group by removing any elements?

5. ($\star\star$) Consider the set of nonnegative real numbers with the operation $a \star b = |a - b|$ (the absolute value of $a - b$). Which abelian group axioms does this structure satisfy?

6. ($\star\star$) Consider the set of real numbers $\mathbb{R}$ with the operation $a \star b = a + b + ab$. Prove that if you remove one real number x, then $\mathbb{R} \setminus \{x\}$ is an abelian group with the same operation $\star$. What is x?

To define vector spaces formally, we need a more basic concept: a *field*. A field is a set of numbers that has addition, subtraction, multiplication, and division by nonzero elements.[2] Some examples of fields that you might already be familiar with are the rational numbers $\mathbb{Q}$, the real numbers $\mathbb{R}$, and the complex numbers $\mathbb{C}$. One set of numbers that isn't a field is the integers $\mathbb{Z}$, because the quotient of two integers isn't always an integer—while on the other hand, the quotient of two (nonzero) real numbers is always a real number.

The best way to define fields is to use an even more basic concept: an *abelian group*. Intuitively, an abelian group is a set of objects with a concept of addition or subtraction, but not multiplication. The prototypical abelian group is $\mathbb{Z}$, the set of integers: integers can be added or subtracted, but not divided, as the quotient of two integers may not be an integer itself. (There is a notion of multiplication in the integers, but multiplication is an extra operation not contained in the abelian group axioms.)

More precisely: an abelian group is a set of elements G with an binary operation $\star$ "defined" on them. "Binary" means that the operation takes two inputs and produces one output. You can imagine constructing an operation by going through every ordered pair of elements $(a, b) \in G$ (including the pairs that include the same element twice) and choosing some element $c \in G$ to be the value of $a \star b$. (There's nothing stopping us from choosing $a \star b$ to be one of a or b.)

But for an operation to make a set into an abelian group, the operation can't be completely arbitrary. It has to satisfy a few properties, which are modeled on the properties of addition of ordinary numbers:

1. *Associativity*: For all triples of elements $a, b, c \in G$, $(a \star b) \star c = a \star (b \star c)$. That is, if $x = a \star b$ and $y = b \star c$, then $x \star c = a \star y$. This axiom also has to be true when two or three of the elements a, b, c are the same. (In general, whenever mathematicians make a statement of the form "for all $x, y \in S$ for some set $S \ldots$," this includes if x and y are the same element of S.

[2] In physics or multivariable calculus, you may have learned about "vector fields": regions of space with a vector defined at each point. Vector fields are a completely different concept that we won't ever discuss in this book; don't get confused.

This axiom implies its own generalization to expressions with four or more terms: all possible ways to parenthesize these have the same value. For instance, $w \star x \star y \star z$ can be parenthesized in five ways: as $((w \star x) \star y) \star z$, as $(w \star x) \star (y \star z)$, as $w \star (x \star (y \star z))$, as $(w \star (x \star y)) \star z$, or as $w \star ((x \star y) \star z)$. If the operation $\star$ is associative, though, then all of these operations must have the same value. For instance, $((w \star x) \star y) \star z = (w \star x) \star (y \star z)$ is just the axiom $(a \star b) \star c = a \star (b \star c)$ with $a = w \star x, b = y, c = z$. (You may want to prove for yourself that the other ways of parenthesizing $w \star x \star y \star z$ are all equivalent as well, perhaps by showing that in any parenthesization, the parentheses can always be shifted "all the way to the left" to form $(((w \star x) \star y) \star z)$, and then try writing a general proof that works for five or more items.)

Since associativity works for any number of items, we can and will write expressions like $w \star x \star y \star z$ without providing parentheses to clarify the order in which the $\star$ operators should be evaluated.

2. *Commutativity*:[3] $a \star b = b \star a$ for all pairs $a, b \in S$.

   This axiom also implies its own generalization to expressions with three or more terms: we can reorder such expressions in any way that we like. For instance, $a \star b \star c = b \star a \star c = b \star c \star a$.

3. *Identity*: There's some fixed element, conventionally called e, such that, for every element $a \in G$, $a \star e = e \star a = a$. (To reemphasize, the same e has to work for *every* element a.)

   This axiom also implies that there can't be more than one identity. Suppose e_1 and e_2 are identities of the same abelian group G. Then $e_1 \star e_2 = e_1$ (because e_2 is an identity, so $a \star e_2 = a$ for any $a \in G$). But also $e_1 \star e_2 = e_2$, because e_1 is an identity and so $e_1 \star a = a$. And since $e_1 \star e_2 = e_2$ and $e_1 \star e_2 = e_1$, then $e_1 = e_2$.

4. *Inverses*: For every element $a \in G$, there's some element $b \in G$ such that $a \star b = e$. (We could have $a = b$; in particular, the identity element is always its own inverse.)

   As with the identity of an abelian group, the inverse of any particular element is also always unique. Suppose that a is an element of G with two inverses b_1, b_2. Consider the expression $b_1 \star a \star b_2$. We can parenthesize this as $(b_1 \star a) \star b_2 = e \star b_2 = b_2$, or as $b_1 \star (a \star b_2) = b_1 \star e = b_1$. But since the operation $\star$ is associative, these two expressions have to be equal, so $b_1 = b_2$.

These are important enough definitions that we'll give them conspicuous definition paragraphs.

Definition. *A **binary operation** on a set S is a function that takes an ordered pair of elements from S and returns an element in S.*

Definition. *An **abelian group** is a set with a binary operation that satisfies the four axioms of associativity, commutativity, identity, and inverses.*

[3] A group whose binary operation satisfies the other axioms in this list but not commutativity is just called a "group," not an "abelian group." The theory of general groups is much more complicated than the theory of abelian groups, but we won't need it for linear algebra.

Here are some examples of sets (with an associated operation) that are abelian groups, and a few that aren't.

1. The integers (commonly written with the symbol $\mathbb{Z}$) with the operation of addition are an abelian group. Associativity and commutativity are true because $(a+b)+c = a+(b+c)$ and $a+b = b+a$ for all integers $a, b, c \in \mathbb{Z}$. The identity element e is 0 (because $a+0 = 0+a = a$), and the inverse of a is the negative $-a$ (because $a+(-a) = 0$).

2. The set of *even* integers (commonly written $2\mathbb{Z}$) with the operation of addititon is also an abelian group.

3. The set of *odd* integers (which we'll denote $2\mathbb{Z}+1$) with the operation of addition is *not* an abelian group, because the sum of two odd integers is not an odd integer.

4. The set of integers under *subtraction* is not an abelian group, because subtraction is not associative and commutative: in general, $a-b \neq b-a$ and $(a-b)-c \neq a-(b-c)$.

5. The set of integers under *multiplication* is almost but not quite an abelian group. Multiplication is associative and commutative ($(ab)c = a(bc)$ and $ab = ba$), and 1 is a multiplicative identity ($1 \times a = a \times 1 = a$ for every integer a). But most integers do not have integer inverses: the solution x to $ax = 1$ (that is, $x = 1/a$) is generally not an integer when a is an integer.

6. The set of nonzero real numbers under multiplication, however, does satisfy the axiom of inverses: the multiplicative inverse of any real number a is $1/a$. We have to leave zero out because anything times zero is zero, so there's no solution to $0x = 1$. But multiplication is still defined as an operation on $\mathbb{R} \setminus \{\mathbf{0}\}$: if a and b are two nonzero real numbers, then ab is a nonzero real number as well. This means that removing 0 from $\mathbb{R}$ doesn't mean removing the result of the group operation on elements that are left in the group.

7. The real numbers (including zero) under addition are also an abelian group. Zero is the identity and negatives are inverses, just like with $\mathbb{Z}$.

Answers to key questions.

1. An abelian group is a set of elements G, together with a binary operation $\star$ that takes ordered pairs of elements in G and produces single elements in G. This operation must obey the axioms of associativity ($a \star (b \star c) = (a \star b) \star c$ for all $a, b, c \in G$), commutativity ($a \star b = b \star a$ for all $a, b \in G$), identity (there's some element $e \in G$ such that for every other $a \in G$, $a \star e = e \star a = e$), and inverses (for every element $a \in G$ there's some $b \in G$ such that $a \star b = e$).

2. Only one, for both questions.

3. The set of integers, set of even integers, and set of real numbers are all abelian groups with the operation of addition. The set of nonnegative real numbers isn't because it fails the axiom of identity: there's no positive real number e such that $a+e = a$ for every other positive real number a. (The only additive identity in the real numbers is zero, and of course zero isn't positive.)

4. The set of complex numbers with the operation of multiplication fails the axiom of inverses: the identity for multiplication is 1, but 0 doesn't have an inverse because there's no real number b such that $0b = 1$. If we take 0 out of the real numbers, then every element of $\mathbb{R} \setminus \{0\}$ has a multiplicative inverse (specifically, the inverse of a is $1/a$), and $\mathbb{R} \setminus \{0\}$ satisfies the other abelian group axioms.

5. This structure doesn't satisfy associativity: in general, $a \star (b \star c) = |a - |b - c||$ doesn't equal $(a \star b) \star c = ||a - b| - c|$. One simple example: $4 \star (3 \star 2) = 4 \star 1 = 3$ but $(4 \star 3) \star 2 = 1 \star 2 = 1$.

   It does satisfy all other abelian group axioms: commutativity ($a - b = -(b - a)$ so $|a - b| = |b - a|$), identity ($|a - 0| = |0 - a| = a$ for all nonnegative real numbers a, so 0 is an identity), and inverses (every element is its own inverse because $|a - a| = 0$).

6. Let's start by checking which abelian group axioms are satisfied by $\mathbb{R}$ with the operation $a \star b = a + b + ab$. This structure satisfies the axiom of associativity: $a \star (b \star c) = a + (b \star c) + a(b \star c) = a + b + c + bc + ab + ac + abc$ and $(a \star b) \star c = (a \star b) + c + (a \star b)c = a + b + ab + c + ac + bc + abc$, and these expressions equal each other (you just have to rearrange the terms to make the expressions identical). It also satisfies the axiom of commutativity ($b \star a = b + a + ba$, which of course equals $a + b + ab$ because addition and multiplication of real numbers are also commutative). It also satisfies the axiom of identity, with 0 as the identity element ($a \star 0 = 0 \star a = a$).

   It does not, however, satisfy the axiom of inverses: in particular, -1 doesn't have an inverse. Since 0 is the identity element, the inverse of any element a is the solution b to $a \star b = 0$; that is, $a + b + ab = 0$. The solution is $b = \frac{a}{a+1}$, but this isn't defined when $a = -1$.

   But if we remove -1, then $\mathbb{R} \setminus \{-1\}$ is an abelian group with the operation $\star$: the only solutions to $a \star b = -1$ are when $a = -1$ or $b = -1$ (you can factor $a \star b = a + b + ab = -1$ as $(a + 1)(b + 1) = 0$), so taking -1 out of the set doesn't remove any values of $\star$ on inputs other than -1, so $\star$ is still defined as an operation on $\mathbb{R} \setminus \{-1\}$.

1.3 Fields

Key questions.

1. How many operations are defined on fields? Which operation makes the whole field into an abelian group? Which operation makes the field with one element removed an abelian group, and which element do you have to remove?

2. Which of the three field axioms determines how the two operations defined on a field interact with each other?

3. ($\star\star$) Let S be a set containing two elements $\{s, t\}$. Define two operations on S: addition as $s + s = t + t = s$ and $s + t = t + s = s$, and multiplication as $ss = st = ts = s$ and $tt = t$. Prove that S with these two operations is a field.

1.3.1 Operations and axioms

A *field* is a special kind of abelian group. In particular, a field is a set $\mathbb{F}$ with two binary operations: addition and multiplication. Just like with normal algebra, we'll write addition with the $+$ sign and multiplication with no sign at all: ab means a times b. We need to define these operations to satisfy a few axioms, modeled after how addition and multiplication work in familiar number systems such as the real numbers and rational numbers.

1. $\mathbb{F}$ is an *abelian group with addition as the binary operation*: addition is associative ($(a+b)+c = a+(b+c)$ for all triples $a, b, c \in \mathbb{F}$) and commutative ($a+b = b+a$ for all pairs $a, b \in \mathbb{F}$). There's also an additive identity, which we'll denote 0, such that $a + 0 = 0 + a$, and every element in a has an additive inverse (which we'll write $-a$) such that $a + (-a) = (-a) + a = 0$.

   We can define subtraction of an element as addition of its additive inverse: that is, we'll write $a - b$ to mean $a + (-b)$.

2. The set $\mathbb{F} \setminus \{0\}$—that is, $\mathbb{F}$ without the additive identity—is an *abelian group with multiplication as the binary operation*, with a multiplicative identity 1. We'll denote multiplication of field elements by writing them together without an operator sign. Every element a must have a multiplicative inverse (we'll write it a^{-1}) such that $aa^{-1} = 1$. Division by an element is just multiplication by its multiplicative inverse; that is, we can write a/b for ab^{-1}.

3. Multiplication and addition obey the *distributive property* $a(b + c) = ab + ac$. (Just like with normal algebra, in expressions involving fields, multiplication has higher precidence than multiplication, so $ab + ac$ means $(ab) + (ac)$.)

   One immediate consequence is that anything times 0 is itself: we know that $0 + 0 = 0$, so for for an arbitrary element x, we have $0x = (0 + 0)x = 0x + 0x$, and adding $-0x$ to each side of the equation $0x = 0x + 0x$ gives us $0 = 0x$. A further consequence is that -1 (that is, the negative multiplicative identity) times any element is its additive inverse: $0 = 0x = (-1 + 1)x = (-1)x + x$.

4. The additive and multiplicative identities are different: $0 \neq 1$. This axiom rules out the so-called "field with one element": any structure that satisfies the other field axioms but has $0 = 1$ must have only one element (because $x = 1x = 0x = 0$ for any possible other element x).

Again, let's put this in a short summary paragraph with conspicuous typography.

Definition. *A **field** is an abelian group (whose group operation is called "addition") with an additional operation called "multiplication," defined such that the group with the additive identity removed is an abelian group under multiplication, and multiplication and addition obey the distributive property $a(b + c) = ab + ac$.*

You may have noticed that all of these axioms are satisfied by a few familiar number systems: for instance, the rational numbers $\mathbb{Q}$, the real numbers $\mathbb{R}$, and the complex numbers $\mathbb{C}$. These axioms let you manipulate algebraic expressions for elements of fields just like how you're used to doing in regular algebra. Introductory linear algebra mostly deals with $\mathbb{R}$ and $\mathbb{C}$. But it's worth knowing that fields have an axiomatic

definition, because if we can prove that a structure satisfies the field axioms, then we know that every result that we can prove just from the field axioms—that is, without using any special property of an individual field such as $\mathbb{R}$ or $\mathbb{C}$—will apply to that structure automatically, no matter how weird its definition is. And the majority of results in this book do apply to arbitrary fields.

1.3.2 Basic properties

Most algebraic manipulations and simple deductions on familiar fields such as $\mathbb{R}$ also apply to any axiomatically defined field. We outlined a few of these as we were presenting the field axioms. We'll just present one additional result on axiomatic fields that we will frequently rely on implicitly.

Proposition. *The product ab of two arbitrary elements a, b of any field equals zero if and only if either $a = 0$ or $b = 0$.*

Proof. A "P if and only if Q" statement is really two statements: P implies Q, and Q implies P. We often have to prove these separately. We'll do that here:

1. *If $a = 0$ or $b = 0$, then $ab = 0$.* First, we'll prove that $0b = 0$ for all field elements b. We start with the equation $0+0 = 0$ (because 0 is an additive identity, so $x+0 = x$ for all x), and multiply this equation on the right by a to get $0b = (0+0)a = 0a+0a$. If we add $-(0b)$, the additive inverse of $0b$, to each side of this equation, then we get $0 = 0b$.

   The proof that $a0 = 0$ for all field elements a is symmetrical.

2. *If $ab = 0$, then either $a = 0$ or $b = 0$.* Suppose $ab = 0$ but $a \neq 0$. Then we can multiply by a^{-1} on the left to get $a^{-1}(ab) = a^{-1}0$. But $a^{-1}(ab) = (a^{-1}a)b = b$ (axiom of associativity of multiplication), and $a^{-1}0 = 0$ (because we just proved that anything times 0 is 0). That is, $ab = 0$ and $a \neq 0$ together imply $b = 0$, so $ab = 0$ implies either $a = 0$ or $b = 0$.

□

The proof of statement 2 crucially relies on the existence of multiplicative inverses. There is a more general structure than a field in abstract algebra, called a *ring*, in which addition and multiplication exist and follow all of the field axioms except that multiplication doesn't have to have inverses and may not be commutative. There are examples of rings in which two nonzero elements have a product that equals zero. (On the other hand, even in a ring, the product of anything with zero, on the left or on the right, is also zero: our proof of statement 1 didn't use the axioms of multiplicative commutativity or multiplicative inverses.) The general theory of rings isn't relevant to linear algebra, though, so we'll leave it off here.

One more useful result:

Proposition. *The additive inverse of any field element a is $(-1)a$ (that is, the additive inverse of the multiplicative identity, times a).*

Proof. We know that $0a = 0$ for all a, and further that $0 = 1 + (-1)$ by definition of additive inverses. So by the distributive property, $0 = (1 + (-1))a = 1a + (-1)a$, so $1a$ and $(-1)a$ are additive inverses.

□

1.3.3 Fields of characteristic 2

Though the field axioms let us prove that several familiar results from $\mathbb{R}$ and $\mathbb{C}$ also apply in arbitrary fields, there are a few that don't. Most relevant for our purposes, there are fields in which the crazy-seeming equations $1 = -1$ and $1 + 1 = 0$ are true: the multiplicative identity is its own additive inverse.

One example is the field defined in key question 3 of this section. In this field, s is the additive identity 0 and t is the multiplicative identity 1, with addition defined as $0 + 0 = 1 + 1 = 0$ and $0 + 1 = 1 + 0 = 0$, and multiplication as $0 \times 0 = 0 \times 1 = 1 \times 0 = 1$ and $1 \times 1 = 1$. (If you're familiar with modular arithmetic on integers, then you'll note that arithmetic operations on this set are just integer operations modulo 2.)

There's a special term for fields with this property:

Definition. *A field in which $1 + 1 = 0$ has **characteristic 2**.*

The "characteristic" of a field in general is a slightly more abstract notion: a field is said to have characteristic n if n is the smallest positive integer such that $\underbrace{1 + \cdots + 1}_{n \text{ times}} = 0$, and zero if no such integer exists. We won't cover the general theory of field characteristics much more in this book, but it's worth flagging that a few results in this book don't apply (or need special proofs) to fields of characteristic 2. We'll flag those as they come up.

If you only need linear algebra for engineering or physics, then you're likely only to ever use $\mathbb{R}$ and $\mathbb{C}$ as underlying fields, so you can safely ignore all discussion of characteristic-2 fields. If you are going to study more pure mathematics, though, you'll need to know the caveats.

One small result:

Proposition. *In a field of characteristic 2, every element, not just the multiplicative identity, is its own additive inverse.*

Proof. Let a be an arbitrary element of the field. We've already shown that in any field regardless of characteristic, $1a = a$ (because 1 is defined as the element that has this property) and that $(-1)a$ is the additive inverse of a. But in characteristic 2, $1 = -1$, so $(-1)a = 1a = a$.

□

Answers to key questions.

1. Fields have two operations: addition and multiplication. Addition makes the whole field into an abelian group (with the identity denoted 0). A field is only an abelian group under multiplication of 0 is removed, because anything times 0 in a field equals 0.

2. The axiom of distributivity $a(b+c) = ab+ac$ governs how field operations interact.

3. To prove that S is a field, we have to check that it obeys three field axioms: first, that it is an abelian group under addition; second, that it is an abelian group under multiplication once the additive identity is taken out; and third, that multiplication and addition follow the distributive property.

   The first step, proving that S is an abelian group under addition, requires checking that it satisfies the four abelian group axioms:

- *Associativity* is the most tedious axiom to check: you have to check that $a + (b + c) = (a + b) + c$ for each of the $2^3 = 8$ possible assignments of the variables a, b, c to values in $\{s, t\}$. Checking each one is straightforward, though. For instance, for the assignment $a = t, b = t, c = s$, you can compute $a+(b+c) = t+(t+s) = t+t = s$ and similarly $(a+b)+c = (t+t)+s = s+s = s$, so $a + (b + c) = (a + b) + c$ is true for this assignment of variables.
- *Commutativity* is shorter: you have to check $a + b = b + a$ for the four assignments of a and b to values in $\{s, t\}$. This is obviously true if $a = b$, and it's true for the assignments $(a, b) = (s, t)$ and $(a, b) = (t, s)$ because addition is defined as $s + t = t + s = s$.
- The *identity* for addition is s, as $s+s = s$ and $s+t = t+s = s$ by the definition of addition.
- Each element of s is its own *inverse* under addition: $s + s = s$ and $t + t = t$.

We next have to check that S minus its additive identity s is an abelian group under multiplication. This is straightforward: $S \setminus \{s\}$ contains only one element, namely t, and multiplication is defined as an operation on $S\setminus\{s\}$ because the only product of elements in $S\setminus\{s\}$, namely the product of t with itself, is also in $S\setminus\{s\}$. You can quickly check that associative and commutative axioms $(ab)c = a(bc)$ and $ab = ba$ are true for the assignment $a = b = c = t$. Finally, $S \setminus \{s\}$ has a multiplicative identity (namely t), and its only element (namely t) has an inverse (namely t).

Finally, we have to check the distributive identity $a(b + c) = ab + ac$ for all eight assignments of variables a, b, c to elements of S. We can simplify our work a lot though, by going back to the original definitions of addition and multiplication to note that $tx = x$ and $sx = s$ for both $x = s$ and $s = t$. So if $a = s$, then $s(b + c) = s$ (both if $b + c = x$ or if $b + c = y$) and similarly $sb + sc = s + s = s$, so $s(b + c) = sb + sc$ regardless of the values of b and c. This proves the distributive identity for four of the eight possible variable assignments. For the other four assignments, namely those with $a = y$, we have $t(b+c) = b+c$ and $tb+tc = b+c$, again regardless of the values of $a + c$.

This field, incidentally, is just the integers with addition and multiplication defined modulo 2, with $s = 0$ and $t = 1$. In other books, you may see it denoted as $\mathbb{Z}/2\mathbb{Z}$ or as $\mathbb{F}_2$. We won't cover fields with finite numbers of elements much in this book, and they're not that useful for most applications of linear algebra, but they appear a lot elsewhere in pure mathematics.

1.4 Vector spaces

Key questions.

1. How many operations are defined between two elements of a vector space? How many are defined between an element of a vector space and an element of its base field?

2. List the five vector space axioms.

3. What vector space is denoted by $\mathbb{R}^2$? How are addition and multiplication defined in this space? What is the additive identity?

4. ($\star$) Suppose that we defined scalar multiplication in $\mathbb{R}^2$ as simply $k(x, y) = (x, y)$; that is, multiplication of a vector by any scalar simply returns the vector unchanged. Vector addition is unchanged. The resulting structure satisfies every vector space axiom except one. Which one does it break?

5. ($\star$) Prove that in any vector space, if $k \neq 0$ and $\mathbf{v} \neq \mathbf{0}$, then $k\mathbf{0} \neq \mathbf{0}$. (Hint: if $k \neq 0$ and $k\mathbf{v} = 0$, then what is $(k^{-1}k)\mathbf{v}$?)

6. ($\star\star$) Define a *wacktor space* as a set of elements W with operations of vector addition and scalar multiplication by elements of a field $\mathbb{F}$ that satisfy all vector space axioms except the axiom of multiplicative identity. That is, W is an abelian group under vector addition, scalar multiplication distributes over vector and field addition, and scalar multiplication is pseudo-associative with field multiplication, but there may be an element $\mathbf{w}$ such that $1\mathbf{w} \neq \mathbf{w}$. (One example of a wacktor space that is not a vector space is $\mathbb{R}^2$ with multiplication defined as $k(x, y) = (0, ky)$ and addition defined like normal, as $(x_1, y_1) + (x_2, y_2) = (x_1 + y_1, x_2 + y_2)$.)

   Suppose W is a wacktor space that is not a vector space. Prove that there is some element $\mathbf{w} \in W \setminus \{\mathbf{0}\}$ such that $1\mathbf{w} = \mathbf{0}$. (Hint: you know there's some element $\mathbf{w}'$ such that $1\mathbf{w}' \neq \mathbf{w}'$. What can you say about $\mathbf{w}' - 1\mathbf{w}'$?) Prove that this element $\mathbf{w}$ satisfies $k\mathbf{w} = \mathbf{0}$ for all scalars $k \in \mathbb{F}$, not just 1.

1.4.1 Definition and axioms

We're finally ready to define vector spaces. A vector space V over a field $\mathbb{F}$—specifying the field is crucial: vector spaces always have an associated field—is a set of elements with two associated operations. One operation, *vector addition*, takes pairs of elements of V and produces other elements of V. The other operation, *scalar multiplication*, takes one input in V and the other input (important!), called a *scalar*, from the base field $\mathbb{F}$, and produces another input in V. The geometric intuition behind scalar multiplication is scaling up or down: $2\mathbf{v}$ is twice the size of $\mathbf{v}$, and $\frac{1}{3}\mathbf{v}$ is one-third the size of $\mathbf{v}$. And $-\mathbf{v}$ (that is, -1 times $\mathbf{v}$) is $\mathbf{v}$ with the direction reversed.

As usual, we'll denote vector addition with the sign $+$ and multiplication without any written sign, with parentheses to clarify order of operations when necessary. These two operations must satisfy the following axioms:

1. V with addition is an *abelian group*. This implies four sub-axioms, the familiar properties of an abelian group:

   (a) Associativity: $\mathbf{u}+(\mathbf{v}+\mathbf{w}) = (\mathbf{u}+\mathbf{v})+\mathbf{w}$ for all triples of elements $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$. (In this book, we'll use boldface letters to denote elements of vector spaces. In handwriting, you'll see vectors marked with arrows over the top: $\vec{u}, \vec{v}, \vec{w}$.)

   (b) Commutativity: $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ for all pairs of elements $\mathbf{u}, \mathbf{v} \in V$.

   (c) Identity: there's some fixed element $\mathbf{0} \in V$ such that $\mathbf{v} + \mathbf{0} = \mathbf{v}$ for every $\mathbf{v} \in V$.

   (d) Inverses: every vector $\mathbf{v}$ has some additive inverse $\mathbf{w}$ such that $\mathbf{v} + \mathbf{w} = \mathbf{0}$. We'll denote the additive inverse with the minus sign ($\mathbf{w} = -\mathbf{v}$).

2. Multiplication *distributes over vector addition*: that is, $k(\mathbf{u}+\mathbf{v}) = k\mathbf{u}+k\mathbf{v}$ for all $k \in \mathbb{F}$ and $\mathbf{u}, \mathbf{v} \in V$.

3. Multiplication *distributes over field addition*: that is, $(a+b)\mathbf{v} = a\mathbf{v}+b\mathbf{v}$ for all $a, b \in \mathbb{F}$ and $\mathbf{v} \in V$.

4. Multiplication in $\mathbb{F}$ and scalar multiplication in V follow the *pseudo-associativity* (sometimes also called *compatibility*) property $(ab)\mathbf{v} = a(b\mathbf{v})$. That is, the scalar ab times $\mathbf{v}$ equals the scalar a times the vector $b\mathbf{v}$. (We call this *pseudo*-associativity because $(ab)\mathbf{v}$ and $a(b\mathbf{v})$ involve operations of different kinds: $(ab)\mathbf{v}$ involves one multiplication of two scalars and a multiplication of the resultant scalar by a vector, while $a(b\mathbf{v})$ involves two scalar-by-vector multiplications.)

5. *Multiplicative identity*: scalar multiplication by the field multiplicative identity 1 leaves vectors unchanged; that is, $1\mathbf{v} = \mathbf{v}$.

Again, to summarize in

Definition. *A **vector space over a field** $\mathbb{F}$ is an abelian group V (whose operation is called **vector addition**) with an additional **scalar multiplication** operation that takes an input from $\mathbb{F}$. These operations must obey the following additional axioms: scalar multiplication distributes over field addition, scalar multiplication distributes over vector addition, and field multiplication is pseudo-associative (a.k.a. compatible) with vector multiplication. $\mathbb{F}$ is called the **base field** of V.*

1.4.2 Basic consequences of vector space axioms

From these axioms, we can prove a few basic facts regarding multiplication by the zero vector, or by the scalars 0 and -1, quite similar to the analogous results for field multiplication. The proofs look quite similar, as well. We will often rely on these results implicitly: citing them for every relevant small algebraic manipulation would get tedious. Throughout, V is a vector space and $\mathbb{F}$ is its base field.

1. $0\mathbf{v} = \mathbf{0}$ for any $\mathbf{v} \in V$. Proof: $0 = 0+0$ because 0 is an additive identity, so $0\mathbf{v} = (0+0)\mathbf{v} = 0\mathbf{v}+0\mathbf{v}$ by the distributive property. Subtracting (that is: adding the additive inverse of) $0\mathbf{v}$ from each side of the equation $0\mathbf{v} = 0\mathbf{v}+0\mathbf{v}$ leaves $\mathbf{0} = 0\mathbf{v}$.

2. $k\mathbf{0} = \mathbf{0}$ for any $k \in \mathbb{F}$. Proof: $k\mathbf{0} = k(\mathbf{0}+\mathbf{0}) = k\mathbf{0}+k\mathbf{0}$, and subtracting $k\mathbf{0}$ from each side leaves $\mathbf{0} = k\mathbf{0}$.

3. If $k\mathbf{v} = \mathbf{0}$, then either $k = 0$ or $\mathbf{v} = \mathbf{0}$. Proof: suppose $k\mathbf{v} = \mathbf{0}$ but $k \neq 0$. Then k^{-1} exists, and $k^{-1}(k\mathbf{v}) = k^{-1}\mathbf{0}$. But $k^{-1}\mathbf{0} = \mathbf{0}$ (because, as we just proved, anything times $\mathbf{0}$ is $\mathbf{0}$), and $k^{-1}(k\mathbf{v}) = (k^{-1}k)\mathbf{v} = 1\mathbf{v} = \mathbf{v}$ (by the pseudo-associativity and multiplicative identity axioms). So $k\mathbf{v} = \mathbf{0}$ implies $k = 0$ or $\mathbf{v} = \mathbf{0}$.

4. -1 times any vector is its additive inverse. Proof: for any vector $\mathbf{v}$ we have $\mathbf{0} = 0\mathbf{v} = (-1+1)\mathbf{v} = (-1)\mathbf{v}+1\mathbf{v}$ by field axiom 3, and $1\mathbf{v} = \mathbf{v}$ by vector space axiom 7, so $\mathbf{v}$ and $-1\mathbf{v}$ add to $\mathbf{0}$.

In most vector spaces, there's a natural way to define addition and multiplication that obviously fits all of the axioms. We chose this set of axioms because they're specific enough to describe the vector spaces that we care most about—the sets of ordered pairs, triples, and so on of real or complex numbers—while also general enough that we can use them to prove results that apply to other spaces as well.

1.4.3 Additive inverses in characteristic 2 and otherwise

There's one more result that will be useful further down the line:

Proposition. *Suppose $\mathbb{F}$ is a field that doesn't have characteristic 2, and V is a vector space over $\mathbb{F}$. Then:*

1. *If $\mathbb{F}$ has characteristic 2, then every element of V is its own additive inverse.*
2. *If $\mathbb{F}$ does not have characteristic 2, then the only element of V that is its own additive inverse is $\mathbf{0}$.*

Proof. We've already proved that for $k \in \mathbb{F}$ and $\mathbf{v} \in V$ arbitrary (and regardless of the characteristic of $\mathbb{F}$), $k\mathbf{v} = \mathbf{0}$ if and only if $k = 0$ or $\mathbf{v} = \mathbf{0}$.

Let $\mathbf{v}$ be an element in V. that is its own additive inverse: that is, such that $\mathbf{v}+\mathbf{v} = \mathbf{0}$. Then $\mathbf{v}+\mathbf{v} = 1\mathbf{v}+1\mathbf{v} = (1+1)\mathbf{v}$ by the axioms of multiplicative identity and distribution over field addition, so $\mathbf{v}$ is its own additive inverse if and only if $1 + 1 = 0$ (that is, $\mathbb{F}$ has characteristic 2) or $\mathbf{v} = \mathbf{0}$.

□

1.4.4 Examples

So, what are some examples of vector spaces? The most common vector spaces that we'll investigate in this book—and, indeed, the *only* vector spaces that some other books on linear algebra ever work with—are the spaces $\mathbb{F}^n$ for some field $\mathbb{F}$ and positive integer n. This is the set of all ordered lists of n elements of the field $\mathbb{F}$. For example, $\mathbb{R}^3$ is the set of all triples of real numbers; some elements of $\mathbb{R}^3$ are $(2, 2, \pi)$ and $(-1, 0, \sqrt{2}+e^{420.69})$. For another example, $\mathbb{C}^2$ is the set of pairs of complex numbers, and has elements such as $(1 + i, -3 + \pi^2 i)$.

The base field for $\mathbb{F}^n$ is just $\mathbb{F}$, and scalar multiplication and addition work component by component. For example, take two elements $\mathbf{u} = (2, 5)$ and $\mathbf{v} = (0.3, -2)$ of $\mathbb{R}^2$. Then $2\mathbf{u} = (4, 10)$: each component of $2\mathbf{u}$ is twice the component of $\mathbf{u}$ in the same position. And $\mathbf{u} - \mathbf{v} = (1.7, 7)$: each component of $\mathbf{u} - \mathbf{v}$ is the corresponding component of $\mathbf{u}$ minus the corresponding component of $\mathbf{v}$.

$\mathbb{R}^2$ and $\mathbb{R}^3$ are especially important because they can be represented geometrically. $\mathbb{R}^2$ is the Cartesian plane; any element (x, y) of $\mathbb{R}^2$ can be represented as as a line from the origin to the point (x, y). Likewise, $\mathbb{R}^3$ is three-dimensional Cartesian space. Results about the algebraic structure of $\mathbb{R}^2$ and $\mathbb{R}^3$ therefore have important consequences for geometry—and, of course, for many applied fields that rely on geometry, such as computer graphics.

Finally, n can be 1. Every field can act as a vector space over itself, with its own elements serving as both vectors and scalars: "vector addition" and "scalar multiplication" are just regular addition and multiplication in the field.

There are a few other spaces that we'll encounter as well:

1. $\mathbb{F}^{\mathbb{N}}$ is the set of all infinite sequences $(a_1, a_2, a_3, \ldots)$ of elements in $\mathbb{F}$. Multiplication and addition in $\mathbb{F}^{\mathbb{N}}$, again, work component by component.

2. One important subset of $\mathbb{F}^{\mathbb{N}}$, denoted $\mathbb{F}^{\infty}$, is the set of sequences with only a finite number of nonzero elements. This is also a vector space. (You should convince yourself that it's closed under multiplication and addition: that is, if you take a sequence with only a finite number of nonzero elements and multiply every entry by a scalar, or if you take two such sequences and add corresponding entries, then the result also has a finite number of nonzero elements.)

3. The set of *functions* defined on the real line or the complex plane is a vector space. Addition and multiplication are defined point-by-point: for example, if $f(x) = x^2$ and $g(x) = \sin x$, then $f + g$ is the function $x \mapsto x^2 + \sin x$, and $-2g$ is the function $x \mapsto -2 \sin x$. The base field of this vector space is the field from which the *values* of the functions are drawn. For instance, the set of complex-valued functions on the real line is a vector space over $\mathbb{C}$.

4. Various subspaces of function spaces are also vector spaces. For example, the set of *continuous* functions on a certain interval is a vector space, because sums and scalar products of continuous spaces are also continuous. The set of functions that are differentiable n times is also a vector space. (This observation lets us use the theory of linear algebra to solve several large classes of differential equations. We'll see one example later.)

Answers to key questions.

1. Vector spaces have one operation defined on pairs of vectors (namely vector addition) and one operation defined on pairs of a vector and a field element (namely scalar multiplication).

2. The five vector space axioms are: (1) the vector space is an abelian group with the operation of addition, (2) scalar multiplication distributes over vector addition, (3) scalar multiplication distributes over field addition, (4) scalar multiplication follows a pseudo-associative property with field multiplication, and (5) 1 is a multiplicative identity for scalar multiplication.

3. $\mathbb{R}^2$ is the set of ordered pairs of real numbers. Addition and multiplication are defined component by component as $(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2)$ and $k(x, y) = (kx, ky)$, and the additive identity is $(0, 0)$.

4. The axiom that doesn't hold is distribuitivity of scalar multiplication over field addition. For instance, if $\mathbf{v} = (1, 0)$, then $1\mathbf{v}$ and $2\mathbf{v}$ would both equal $(1, 0)$ with our new definition of scalar multiplication, but distributivity over field addition would require $2\mathbf{v} = (1 + 1)\mathbf{v} = 1\mathbf{v} + 1\mathbf{v} = (1, 0) + (1, 0) = (2, 0)$.

5. If $k\mathbf{v} = 0$ and $k \neq 0$, then by the multiplicative identity axiom $1\mathbf{v} = \mathbf{v}$, but by the pseudo-associativity axiom $1\mathbf{v} = (k^{-1}k)\mathbf{v} = k^{-1}(k\mathbf{v}) = k^{-1}\mathbf{0} = \mathbf{0}$ (because the vector space axioms imply that any scalar times $\mathbf{0}$ equals $\mathbf{0}$), so $\mathbf{v} = 0$. So there can't be elements $k \in \mathbb{F} \setminus \{0\}, \mathbf{v} \in V \setminus \{\mathbf{0}\}$ such that $k\mathbf{v} = 0$.

6. Let $\mathbf{w}'$ be an element in W such that $\mathbf{w}' \neq 1\mathbf{w}'$, and define $\mathbf{w} = \mathbf{w}' - 1\mathbf{w}'$. Then

$$\begin{aligned} 1\mathbf{w} &= 1(\mathbf{w}' - 1\mathbf{w}') && \text{(by definition of } \mathbf{w}') \\ &= 1\mathbf{w}' - 1(1\mathbf{w}') && \text{(scalar multiplication distributes over vector addition)} \\ &= 1\mathbf{w}' - (11)\mathbf{w}' && \text{(pseudo-associativity)} \\ &= 1\mathbf{w}' - 1\mathbf{w}' = \mathbf{0}. \end{aligned}$$

So $1\mathbf{w} = \mathbf{0}$. By immediate consequence, $k\mathbf{w} = (k1)\mathbf{0} = k\mathbf{0} = \mathbf{0}$ for any scalar k (our proof that anything times $\mathbf{0}$ equals $\mathbf{0}$ did not rely on the axiom of multiplicative identity, so it's still valid in wacktor spaces).

1.5 Linear combinations and span

Key questions.

1. What is a "linear combination" of a set of vectors?
2. What does it mean for a linear combination to be "trivial"?
3. What's the relationship between linear combinations and span? Why does the span of a set of vectors always include $\mathbf{0}$?
4. ($\star$) Prove that the sets $\{(1,2,3),(1,1,1)\}$ and $\{(100,200,300),(105,205,305)\}$ have the same span in $\mathbb{R}^3$.

In the last section, we defined a vector space V over a field $\mathbb{F}$ as a set of elements called "vectors." Vector spaces allow two operations: adding two vectors together, and scaling a single vector through multiplication with an element in $\mathbb{F}$.

1.5.1 Definitions

A *linear combination* of a set S of vectors is a sum of a finite number of scalar multiples of the elements of S. (Generic vector spaces don't have a concept of infinite sums.[4]) For example, some linear combinations of two vectors $\mathbf{u}, \mathbf{v}$ are $\frac{2}{3}\mathbf{u} - \mathbf{v}$ (in a vector space over $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$), or $-\mathbf{u} + \pi\mathbf{v}$ (in a vector space over $\mathbb{R}$ or $\mathbb{C}$), or $(1+i)\mathbf{u} + (\sqrt{2} - 3i)\mathbf{v}$ (in a vector space over $\mathbb{C}$)—or just $-4\mathbf{v}$, giving $\mathbf{u}$ a coefficient of zero.

More generally:

Definition. *A **linear combination** of a set of vectors S from a vector space V is either:*

[4]To define the sum of an infinite series, we need a concept of a *limit* of the sequence of partial sums of this series, and limits of sequences require a concept of *distance between vectors* (or, equivalently, *size* of the difference between two vectors) so that terms in a sequence can get arbitrarily close to the sequence limit. The standard vector space axioms don't have a concept of distance. Some vector spaces let you define vector size using an additional structure called a *norm*: for instance, you could use the Pythagorean theorem to define the norm of an element $(x, y) \in \mathbb{R}^2$ as $\sqrt{x^2 + y^2}$, and thus the distance between two elements as $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$. If this norm satisfies a few axioms, then you can use it to define the limit of a sequence. We'll discuss a few examples of spaces with norms, most importantly $\mathbb{R}^n$ and $\mathbb{C}^n$, in Chapter 9. But these structures don't exist in generic vector spaces, and there are spaces for which no norm exists: for example, vector spaces over a field with a finite number of elements, which we won't talk about much in this book but which are important in other fields of higher mathematics.

1. *An expression $c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$, where $c_1, \ldots, c_n$ are elements of the base field of V and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are elements of S. If the coefficients $c_1, \ldots, c_n$ are all 0, then this is expression is called **trivial**; if at least one coefficient is not 0, then it's **nontrivial**.*

2. *The value of this expression as an element of V.*

We won't be so pedantic as to insist on a distinction between "linear combination" and "value of a linear combination" except when context requires. We will also consider an empty sum, containing zero terms, to be a linear combination with value $\mathbf{0}$.

The trivial linear combination $0\mathbf{v}_1 + \cdots + 0\mathbf{v}_n$, of course, equals $\mathbf{0}$, the zero element of V. This is because the field element 0 times any vector is $\mathbf{0}$, and the sum of any number of terms of $\mathbf{0}$ is also $\mathbf{0}$.

Nontrivial linear combinations, however, don't necessarily have to have nonzero value. Many important properties of a set of vectors are associated with whether they can produce a nontrivial linear combination that equals $\mathbf{0}$. We'll discuss these properties for the rest of the book (including in the coming sections), as well as the question of howe we can determine, given a set of vectors, whether it can produce nontrivial linear combinations with value $\mathbf{0}$.

One last definition:

Definition. *The **span** of a set of vectors in S is the set of values of every possible linear combination in S.*

So if $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ (and usually in this book, S will be finite), the span of S is the set of all sums of the form $c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$, with the coefficients $c_1, \ldots, c_n$ drawn from the base field. By convention, the span of the empty set contains the zero vector but nothing else: $\operatorname{span}\emptyset = \{\mathbf{0}\}$.

In $\mathbb{R}^2$ and $\mathbb{R}^3$, spans have geometric interpretations. The span of a set containing just one vector (in any vector space) is the set of all scalar multiples of that vector: $\operatorname{span}\{\mathbf{v}\} = \{c\mathbf{v} : c \in \mathbb{R}\}$. In $\mathbb{R}^2$ and $\mathbb{R}^3$, therefore, the vectors in $\operatorname{span}\{\mathbf{v}\}$ are all the vectors that lie on a line through the origin of $\mathbb{R}^2$ or $\mathbb{R}^3$ that contains $\operatorname{span}\{\mathbf{v}\}$ (except, of course, if $\mathbf{v} = \mathbf{0}$, in which case $\operatorname{span}\{\mathbf{v}\} = \{\mathbf{0}\}$).

The span of a set of two vectors $\{\mathbf{u}, \mathbf{v}\}$, meanwhile, is the plane that contains the origin, $\mathbf{u}$, and $\mathbf{v}$—unless one of $\mathbf{u}$ or $\mathbf{v}$ is $\mathbf{0}$, or if $\mathbf{u}$ and $\mathbf{v}$ are scalar multiples of each other: $\mathbf{u} = k\mathbf{v}$. In this latter case, any linear combination $a\mathbf{u} + b\mathbf{v}$ can be rewritten as just $(ka + b)\mathbf{v}$ with $\mathbf{v}$ alone, so $\operatorname{span}\{\mathbf{u}, \mathbf{v}\} = \operatorname{span}\{\mathbf{v}\}$.

1.5.2 Finding sets with identical span to a given set

Given a set S, you can often find a set of simpler or easier-to-visualize vectors with the same span as S by using this result:

Lemma. *Suppose V is a vector space, S is some subset of V, and $\mathbf{v} \in V$ is some linear combination $\mathbf{v} = c_1\mathbf{s}_1 + \cdots + c_n\mathbf{s}_n$, where $\mathbf{s}_1, \ldots, \mathbf{s}_n$ are elements of S and the scalars $c_1, \ldots, c_n$ are nonzero. Then if you add $\mathbf{v}$ to S and remove any one of the vectors $\mathbf{s}_1, \ldots, \mathbf{s}_n$, you get a set with the same span as S.*

Proof. Let's prove that replacing $\mathbf{s}_1$ with $\mathbf{v}$ leaves $\operatorname{span} S$ unchanged: that is,

$$\operatorname{span}\{\mathbf{s}_1, \ldots, \mathbf{s}_n\} = \operatorname{span}\{\mathbf{v}, \mathbf{s}_2, \ldots, \mathbf{s}_n\}.$$

and, as a consequence,[5]

$$\operatorname{span}(\{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \cup S') = \operatorname{span}(\{\mathbf{v}, \mathbf{s}_2, \ldots, \mathbf{s}_n\} \cup S')$$

where S' contains the vectors of S other than $\mathbf{s}_1, \ldots, \mathbf{s}_n$. (The same argument works for any of the vectors $\mathbf{s}_i$, not just $\mathbf{s}_1$.)

We can rewrite any linear combination of $\mathbf{v}, \mathbf{s}_2, \ldots, \mathbf{s}_n$ as as a linear combination of $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ just by substituting $c_1\mathbf{s}_1 + \cdots + c_n\mathbf{s}_n$ for $\mathbf{v}$. Similarly, any linear combination involving $\mathbf{s}_1$ can be rewritten as one that involves $\mathbf{v}, \mathbf{s}_2, \mathbf{s}_3, \ldots, \mathbf{s}_n$, because if $\mathbf{v} = c_1\mathbf{s}_1 + c_2\mathbf{s}_2 + \cdots + c_n\mathbf{s}_n$, then a simple rearrangement gives

$$\mathbf{s}_1 = \frac{1}{c_1}\mathbf{v} - \frac{c_2}{c_1}\mathbf{s}_2 - \cdots - \frac{c_n}{c_1}\mathbf{s}_n.$$

As long as $c_1 \neq 0$, the expression on the right can substitute for $\mathbf{s}_1$.

□

Let's put this result into practice: starting with $S = \{(40, 100, 120), (1, 3, 4)\} \subset \mathbb{R}^3$, we'll find a simpler subset of $\mathbb{R}^3$ with the same span as S. First, we can replace $(40, 100, 120)$ with $\frac{1}{10}(40, 100, 120) = (4, 10, 12)$: a nonzero scalar multiple of a single vector $\mathbf{v}$ is certainly a linear combination that includes $\mathbf{v}$ with nonzero coefficient. So we have a new set $S' = \{(4, 10, 12), (1, 3, 4)\}$ with the same span as S.

Now note that $(4, 10, 12) - 3(1, 3, 4) = (1, 1, 0)$. This new vector $(1, 1, 0)$ can replace either vector in S' without changing its span. If we choose the first vector, then we get the set $S''\{(1, 1, 0), (1, 3, 4)\}$.

Similarly, $\frac{1}{2}(1, 3, 4) - \frac{1}{2}(1, 1, 0) = (0, 1, 2)$; this vector can replace $(1, 3, 4)$ without changing the span S'', giving $S''' = \{(1, 1, 0), (0, 1, 2)\}$. The span of S''' (which is also the span of S) is the set of all values of linear combinations $a(1, 1, 0) + b(0, 1, 2) = (a, a + b, 2b)$, where a and b are freely chosen from $\mathbb{R}$.

We can continue this process to find other sets with the same span as S. For instance, if we replace $(1, 1, 0)$ with $(1, 1, 0) - (0, 1, 2) = (1, 0, -2)$, we get $\{(1, 0, -2), (0, 1, 2)\}$, giving a general form $(a, c, -2a + 2c)$ for elements of $\operatorname{span} S$. (You can see that this form is equivalent to $(a, a + b, 2b)$ by setting $c = a + b$.)

This process is a less systematic version of an algorithm called *Gauss–Jordan elimination* developed for solving systems of linear equations. We'll present Gauss–Jordan elimination fully in Chapter 3.

Answers to key questions.

1. A linear combination of a set of vectors is an expression consisting of a sum of scalar multiples of vectors in the set.

2. A linear combination is trivial if every scalar coefficient in the linear combination is zero.

[5] In general, if A, B, C are three sets of vectors from the same space and $\operatorname{span} A = \operatorname{span} B$, then $\operatorname{span}(A \cup C) = \operatorname{span}(B \cup C)$. Any element $\mathbf{v} \in \operatorname{span}(A \cup C)$ is a linear combination of elements of $A \cup C$, which can be broken down as $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$ where $\mathbf{v}_1$ is a linear combination from A and $\mathbf{v}_2$ is a linear combination from B. But if $\operatorname{span} A = \operatorname{span} B$, then we can also write $\mathbf{v}_1$ as a linear combination from B and reuse the same linear combination for B, so their sum $\mathbf{v}_1 + \mathbf{v}_2$ is a linear combination from $B \cup C$.

3. The span of a set of vectors S is the set of all values taken by linear combinations of elements of S, with the scalar coefficients chosen freely from the base field. This always includes $\mathbf{0}$ because $\mathbf{0}$ is the value of the trivial linear combination.

4. If we alter a set S by replacing any vector $\mathbf{v}$ with a linear combination of elements of S that includes $\mathbf{v}$ with a nonzero coefficient, then the span of S doesn't change. (The linear combination doesn't necessarily have to include vectors besides $\mathbf{v}$.) So we can replace $(1, 2, 3)$ in the set $\{(1, 2, 3), (1, 1, 1)\}$ with the vector $100(1, 2, 3) = (100, 200, 300)$, giving us the set $\{(100, 200, 300), (1, 1, 1)\}$. Next we can replace $(1, 1, 1)$ with $(100, 200, 300)+5(1, 1, 1)$, giving $\{(100, 200, 300), (105, 205, 305)\}$. This set has the same span as the original set.

1.6 Subspaces

Key questions.

1. What is a subspace of a vector space? List the three axioms that a subspace must satisfy. What element must every subspace include?

2. What does it mean for a subspace to be trivial?

3. Why is the span of any set of vectors a subspace?

4. ($\star$) Is $\{(x, y) \in \mathbb{R}^2 : x = 0\}$ a subspace of $\mathbb{R}^2$? What about $\{(x, y) \in \mathbb{R}^2 : x \neq 0\}$? Why or why not?

5. Is the intersection of two subspaces always a subspace? What about the union of two subspaces?

6. What are the two definitions of the sum of two subspaces? Why are these definitions equivalent?

7. What is a spanning set of a subspace? Is every subspace a spanning set of itself?

In the last section, we introduced the concepts of *linear combinations* and *span*. As a reminder: a linear combination of a set of vectors is a sum of scalar multiples of the vectors. The span of a set S of vectors is the set of all vectors that can be written as linear combinations of elements of S—that is, the set of all vectors that you can get by using vector addition and scalar multiplication on the elements of S.

1.6.1 Spans are subspaces; definition of subspace

Now let $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ be a set of elements of a vector space V. Note two special properties of the set $\operatorname{span} S$:

1. If $\mathbf{v} = c_1\mathbf{s}_1 + c_n\mathbf{s}_n$ is in $\operatorname{span} S$, then so are all of its multiples: $k\mathbf{v} = k(c_1\mathbf{s}_1 + \cdots + c_n\mathbf{s}_n) = (kc_1)\mathbf{s}_1 + \cdots + (kc_n)\mathbf{s}_n$.

2. If $\mathbf{v}_1 = a_1\mathbf{s}_1 + \cdots + a_n\mathbf{s}_n$ and $\mathbf{v}_2 = b_1\mathbf{s}_1 + \cdots + b_n\mathbf{s}_n$ are linear combinations of elements of S, then so is their sum: $\mathbf{v}_1 + \mathbf{v}_2 = (a_1 + b_1)\mathbf{s}_1 + \cdots + (a_n + b_n)\mathbf{s}_n$.

That is, all sums and scalar multiples of elements of $\operatorname{span} S$ are also in $\operatorname{span} S$, so $\operatorname{span} S$ acts like a self-contained vector space in its own right. These properties are also true if S contains an infinite number of vectors.

There's a special term for a subset of a larger vector space V that satisfies these properties of $\operatorname{span} S$: a *subspace* of V. In general:

Definition. *A subset W of a vector space V is a **subspace** (or **vector subspace**, if the context requires us to be more specific) if it satisfies these axioms:*

1. Closure under addition*: If $\mathbf{u}$ and $\mathbf{v}$ are in W, then so is $\mathbf{u}+\mathbf{v}$.*
2. Closure under multiplication*: If $\mathbf{v}$ is in W, then so is $k\mathbf{v}$ for every scalar k.*
3. Non-emptiness*: W is not the empty set. (This means that $\mathbf{0} \in W$, because if W has at least one element $\mathbf{v}$, it must contain $0\mathbf{v} = \mathbf{0}$ by closure under multiplication.)*

We could unify axioms 1 and 2 into one sentence: *Any linear combination of elements of W is also an element of W.* This means that any subspace that includes S must contain every linear combination of S—that is, it must contain $\operatorname{span} S$. Since $\operatorname{span} S$ is also a subspace, it follows that $\operatorname{span} S$ is the smallest subspace that contains S: if W is a subspace of V that also contains S, then W must contain (or equal) $\operatorname{span} S$. Another consequence is that W is any subspace of V, then $\operatorname{span} W = W$—and, therefore, $\operatorname{span}(\operatorname{span} S) = \operatorname{span} S$ for any set S.

Every vector space V, except the one-element vector space $V = \{\mathbf{0}\}$, has at least two subspaces. The first is V itself: every vector space is a subset of itself. The second is $\{\mathbf{0}\}$. These subspaces are called "trivial subspaces." (If $V = \{\mathbf{0}\}$, then both of these trivial subspaces are the same, and V has no other subspaces.)

To solidify our understanding of subspaces, let's take a look at a few subsets of $\mathbb{R}^2$ and check whether they're subspaces.

1. $W_1 = \{(x, y) \in \mathbb{R}^2 : x + 2y = 0\}$ is a subspace. If $\mathbf{u} = (a, b)$ and $\mathbf{v} = (c, d)$ are both in W_1, then $a + 2b$ and $c + 2d$ are both zero, so $(a + b) + 2(c + d) = 0$. That is, $\mathbf{u} + \mathbf{v} = (a + c, b + d)$ is also in W_1. So W_1 satisfies the axiom of closure under addition. Similarly, $k\mathbf{u} = (ka, kb)$ is in W_1 for any scalar k, because if $a + 2b = 0$ then $ka + 2kb = 0$, so W_1 also satisfies closure under multiplication. Finally, you can check that $\mathbf{0} = (0, 0)$ is in W_1, so W_1 is nonempty.
2. $W_2 = \{(x, y) \in \mathbb{R}^2 : x + 2y = 1\}$ can't be a subspace because it doesn't include $\mathbf{0}$.
3. $W_3 = \{(x, y) \in \mathbb{R}^2 : x = 0 \text{ or } y = 0 \text{ or both}\}$ contains $\mathbf{0}$, and it satisfies closure under multiplication: any multiple of a vector of the form $(x, 0)$ or $(0, y)$ also has be of the same form. But it fails closure under addition: for instance, $(1, 0)$ and $(0, 1)$ are both in W_3, but their sum $(1, 1)$ isn't.
4. $W_4 = \{(x, y) \in \mathbb{R}^2 : x \text{ is an integer}\}$ satisfies closure under addition: if a and c are integers, then so is $a + c$; so if (a, b) and (c, d) are in W_4, then so is their sum $(a + c, b + d)$. But it fails closure under multiplication: for instance, $\mathbf{v} = (1, 0)$ is in W_4, but its scalar multiple $\frac{1}{2}\mathbf{v} = (\frac{1}{2}, 0)$ isn't.

One final bit of vocabulary.

Definition. *A set S is a **spanning set** of a subspace W of a vector space V if $\operatorname{span} S = W$. (This definition allows $W = V$.)*

1.6.2 Subspace intersections and sums

If W_1 and W_2 are subspaces of V, then their intersection $W_1 \cap W_2$ is also a subspace. Every sum or multiple of elements in W_1 is in W_1, and every sum or multiple of elements of W_2 is in W_2, so any sum or multiple of elements of *both* W_1 and W_2 must also be in both subspaces. Finally, the intersection $W_1 \cap W_2$ has to contain at least $\mathbf{0}$, so it is not empty. This line of argument generalizes to any number of subspaces, even an infinite number of subspaces. (Infinite intersections are what they sound like: an element is in an intersection of an infinite collection of sets if it's included in every single set in the collection.)

The *union* of two subspaces, on the other hand, isn't necessarily a subspace: the sum of an element of W_1 and an element from W_2 doesn't have to be W_1 or W_2. For instance, suppose $W_1 \subset \mathbb{R}^2$ contains all vectors of the form $(x, 0)$ with a zero second component, and $W_2 \subset \mathbb{R}^2$ contains all vectors of the form (y, y) with two equal components. Then $(1, 0)$ is in W_1 and $(1, 1)$ is in W_2, but their sum $(2, 1)$ isn't in W_1 or W_2. The space W_3 in the list of examples in the last section is another union of two subspaces (namely $\text{span}\{(1, 0)\}$ and $\text{span}\{(0, 1)\}$) that isn't a subspace itself.

The closest thing to unions of subspaces is *sums*. Subspace sums are defined in terms of vector sums; to be precise:

Definition. *The **sum** (or **subspace sum**, when disambiguation is necessary) $W_1 + W_2$ of two subspaces W_1, W_2 of a larger space is the set $\{\mathbf{w}_1 + \mathbf{w}_2 : \mathbf{w}_1 \in W_1, \mathbf{w}_2 \in W_2\}$ of all sums of a vector from W_1 and a vector from W_2.*

It turns out that $W_1 + W_2$ also satisfies the vector subspace axioms. To see that it satisfies closure under addition and multiplication, note that if $\mathbf{u} = \mathbf{w}_{1a} + \mathbf{w}_{2a}$ and $\mathbf{v} = \mathbf{w}_{1b} + \mathbf{w}_{2b}$ are sums of a vector from W_1 and a vector from W_2, then so is $\mathbf{u} + \mathbf{v} = (\mathbf{w}_{1a} + \mathbf{w}_{1b}) + (\mathbf{w}_{2a} + \mathbf{w}_{2b})$, as is any multiple $k\mathbf{u} = k\mathbf{w}_{1a} + k\mathbf{w}_{2a}$. (Remember that vector addition is associative and commutative, so we can reorder and reparenthesize vector sums however we want.) Subspace sums are also nonempty: since $\mathbf{0}$ is always in W_1 and W_2, so $\mathbf{0} + \mathbf{0} = \mathbf{0}$ is an element of $W_1 + W_2$.

We have another equivalent definition:

Proposition. *$W_1 + W_2$, as defined above, is the intersection of every subspace that contains both W_1 and W_2.*

Proof. Let's write $\mathcal{S}$ for the collection of subspaces that contain both W_1 and W_2, and write X for the intersection of every subspace in $\mathcal{S}$. (This means that X is a subset of every element of $\mathcal{S}$.) Now, $W_1 + W_2$ (as we've just shown) is a subspace, and it contains both W_1 and W_2, so $W_1 + W_2 \in \mathcal{S}$ and so $X \subseteq W_1 + W_2$.

Conversely, since a subspace must contain every linear combination of any of its elements, every subspace in $\mathcal{S}$ must contain any linear combination constructed from elements of $W_1 \cup W_2$. In particular, every subspace in $\mathcal{S}$ must contain the set of linear combinations made up from adding one element of W_1 to one element of W_2: that is, it must contain $W_1 + W_2$. And since $W_1 + W_2$ is a subset of everything in $\mathcal{S}$, it must also be a subset of X. So we've proved $W_1 + W_2 \subseteq X$ as well as $X \subseteq W_1 + W_2$.

□

And, finally, a third equivalent definition:

Proposition. *$W_1 + W_2$, as defined above, is the smallest subspace that contains both W_1 and W_2, in that if X is any other such subspace, then $W_1 + W_2 \subseteq X$.*

Proof. Any subspace X that contains W_1 and W_2 must be one of the elements of the subspace collection $\mathcal{S}$ that we defined in the previous proposition. But $W_1 + W_2$ is the intersection of everything in $\mathcal{S}$, and the intersection of any collection of sets must be contained in any individual set in the collection, so $W_1 + W_2 \subseteq X$.

□

If this discussion of subset sums seems too abstract, a short example may be useful. Let $W_1 \subset \mathbb{R}^4$ be the set of all vectors of the form $(x, y, 0, 0)$ and W_2 be the set of all vectors of the form $(0, y, z, 0)$. Then $W_1 + W_2$ contains all vectors of the form $(x, y, z, 0)$. Any such vector can be broken into sums such as $(x, y, 0, 0) + (0, 0, z, 0)$ or $(x, 2x + 3y, 0, 0) + (0, -x - 2y, z, 0)$, where the first term is in W_1 and the second is in W_2.

Subspace sums can also be defined for three or more subspaces. For example, $W_1 + W_2 + W_3 = \{\mathbf{w}_1 + \mathbf{w}_2 + \mathbf{w}_3 : \mathbf{w}_1 \in W_1, \mathbf{w}_2 \in W_2, \mathbf{w}_3 \in W_3\}$.

1.6.3 Subspaces, sums, and intersections of spans

Finally, a few facts about subspaces, sums, and intersections of spans. First, if $S_1 \subseteq S_2$, then $\operatorname{span} S_1 \subseteq \operatorname{span} S_2$, because every linear combination of elements of S_1 is also a combination of elements of S_2.

Second, if $\operatorname{span} S_1 = W_1$ and $\operatorname{span} S_2 = W_2$, then $\operatorname{span}(S_1 \cup S_2) = W_1 + W_2$. Any element from $W_1 + W_2$ can be broken into an element of W_1 (i.e. a linear combination from S_1) plus an element of W_2 (i.e. a linear combination from S_2, and the sum of these linear combinations is obviously an element from $S_1 \cup S_2$. Conversely, you can evaluate any linear combination from $S_1 \cup S_2$ by adding all the terms from S_1 to get an element of W_1, then adding the remaining terms in S_2 to get an element of W_2; their sum has to be in $W_1 + W_2$.

The analogous statement for intersections, though, isn't true: $\operatorname{span}(S_1 \cap S_2)$ is a subspace of, but doesn't have to equal, $W_1 \cap W_2$. For instance, define $S_1, S_2 \subset \mathbb{R}^3$ as $S_1 = \{(1, 0, 0), (0, 1, 0)\}$ and $S_2 = \{(1, 0, 0), (1, 1, 0)\}$. Then W_1, W_2, and $W_1 \cap W_2$ are all equal: they are the set of vectors of the form $(x, y, 0)$ with a third component of zero. (We could write any such vector as a linear combination $x(1, 0, 0)+y(0, 1, 0)$ of elements from S_1, or as a linear combination $(x - y)(1, 0, 0) + y(0, 1, 0)$ of elements from S_2.) But $\operatorname{span}(S_1 \cap S_2)$ is the smaller set of multiples of $(1, 0, 0)$: that is, the set of vectors of the form $(x, 0, 0)$.

Answers to key questions.

1. A subspace W of a vector space V is a subset that acts like a vector space in its own right with the same operations inherited from V. The three axioms are closure under addition (any sum of elements of W is also in W), closure under multiplication (the product of any element of W with any scalar is also in W), and non-emptiness (W has at least one element). Every subspace includes $\mathbf{0}$, because closure under addition means that if any vector $\mathbf{w}$ is in W, then so is $0\mathbf{w} = \mathbf{0}$.

2. A subspace of V is trivial if it is either V itself if $\{\mathbf{0}\}$.

3. The span of any set S of vectors is a subspace because the sum of two linear combinations $c_1\mathbf{v}_1 + \cdots c_n\mathbf{v}_n$ and $d_n\mathbf{v}_1 + \cdots + d_n\mathbf{v}_n$ of vectors in S is also a linear combination $(c_1 + d_1)\mathbf{v}_1 + \cdots + (c_n + d_n)\mathbf{v}_n$ of the same vectors in S, and so is any scalar product $k(c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n) = (kc_1)\mathbf{v}_1 + \cdots + (kc_n)\mathbf{v}_n$.

4. $\{(x, y) \in \mathbb{R}^2 : x = 0\}$, containing all elements of $\mathbb{R}^2$ with a zero in the first component, is a subspace (call it W). We can check the three axioms: it is closed under addition (if $\mathbf{w}_1 = (0, y_1)$ and $\mathbf{w}_2 = (0, y_2)$ are in W then so is $\mathbf{w}_1 + \mathbf{w}_2 = (0, y_1 + y_2)$), it's closed under multiplication (if $\mathbf{w} = (0, y) \in W$ then $k\mathbf{w} = (0, ky) \in W$), and it's non-empty ($(0, 0)$ is one element).

   On the other hand, $\{(x, y) \in \mathbb{R}^2 : x \neq 0\}$ (call this set $\mathbb{R}^2 \setminus W$) is not closed under addition or multiplication. For instance, $\mathbf{v}_1 = (1, 1)$ and $\mathbf{v}_2 = (-1, 1)$ are in $\mathbb{R}^2 \setminus W$, but $\mathbf{v}_1 + \mathbf{v}_2 = (0, 2)$ and $0\mathbf{v}_1 = (0, 0)$ are not.

5. The intersection of two subspaces is always a subspace. The union of two subspaces may not be a subspace, because it may not be closed under addition: sums of elements from different subspaces may not be in either subspace. For instance, $W_1 = \{(x, 0) : x \in \mathbb{R}\}$ and $W_2 = \{(0, y) : y \in \mathbb{R}\}$ are both subspaces of $\mathbb{R}^2$, and though $(1, 0)$ and $(0, 1)$ are both in $W_1 \cup W_2$, their sum $(1, 1)$ is not.

6. The two definitions of a sum $W_1 + W_2$ of two subspaces of V are (1) the set of all sums of an element of W_1 and an element of W_2, subspace of V that includes both W_1 and W_2 ("smallest" in the sense that a

1.7 Linear independence

Key questions.

1. What does it mean for a set to be "linearly independent"?

2. ($\star$) If S is a set of vectors, is it possible to have two vectors $\mathbf{u}, \mathbf{v} \in \operatorname{span} S$ such that only one linear combination of elements of S equals $\mathbf{u}$, but multiple linear combinations of elements of S equal $\mathbf{v}$? Give an example, or explain why not.

3. ($\star$) Prove that if S is not linearly independent, then we can remove at least one vector from S without affecting $\operatorname{span} S$.

In the last sections, we introduced the notion of a *span* of a set of vectors (that is, the set of values that you can achieve by multiplying and adding elements of S; this is always guaranteed to be a vector subspace) and a *spanning set* of a subspace (that is, any set whose span is the subspace). Besides span, the other most important property of a set of vectors is whether it is *linearly independent*. Roughly speaking, a linearly independent set S generates its span without redundancy: any element in its span is the value of one and only one linear combination of elements in S, and no element of S can be removed without shrinking its span.

1.7.1 Motivating examples

Let's start with a simple example: take two vectors $\mathbf{u} = (2, 3i)$ and $\mathbf{v} = (-4i, 6)$ in $\mathbb{C}^2$, and let $S = \{\mathbf{u}, \mathbf{v}\}$. What is $\operatorname{span} S$?

Your first answer should be that $\operatorname{span} S$ is the set of all linear combinations $a\mathbf{u} + b\mathbf{v} = (2a - 4ib, 3ia + 6b)$, where a, b are freely chosen elements of $\mathbb{C}$. But notice something peculiar about S: its elements are scalar multiples of each other—specifically, $\mathbf{v} =$

$-2i\mathbf{u}$. So any linear combination of $\mathbf{u}$ and $\mathbf{v}$ can be rewritten as a scalar multiple of $\mathbf{u}$ alone: $a\mathbf{u} + b\mathbf{v} = (a - 2ib)\mathbf{u}$, and $\operatorname{span} S$ is just the set of linear multiples of $\mathbf{u}$.

Another example: Define $S \subset \mathbb{R}^3$ as the set containing the three vectors $\mathbf{u} = (1, -2, 3)$, $\mathbf{v} = (-1, -4, 1)$, and $\mathbf{w} = (5, 8, 3)$. What is $\operatorname{span} S$? Your first answer, again, should be the set of all linear combinations $a\mathbf{u} + b\mathbf{v} + c\mathbf{w} = (a - b + 5c, -2a - 4b + 2c, 3a + 8b + 3c)$ with freely chosen coefficients $a, b, c \in \mathbb{R}$.

But we can rewrite any element of S as a linear combination of the others: for instance, $\mathbf{w} = 2\mathbf{u} - 3\mathbf{v}$. So any linear combination $a\mathbf{u} + b\mathbf{v} + c\mathbf{w}$ can be rewritten in terms of $\mathbf{u}$ and $\mathbf{v}$ alone as $(a + 2c)\mathbf{u} + (b - 3c)\mathbf{v}$. Conversely, we can write any linear combination $x\mathbf{u} + y\mathbf{v}$ as a linear combination $a\mathbf{u} + b\mathbf{v} + c\mathbf{w}$ with an appropriate choice of a, b, c (the simplest choice is just $a = x, b = y, c = 0$, but there are others). So $x\mathbf{u} + y\mathbf{w}$, where x and y are arbitrary real numbers, is also a valid general form for any element in $\operatorname{span} S$, and $\operatorname{span} S = \operatorname{span}\{\mathbf{u}, \mathbf{v}\}$. (It turns out that $\operatorname{span} S = \operatorname{span}\{\mathbf{u}, \mathbf{w}\} = \operatorname{span}\{\mathbf{v}, \mathbf{w}\}$ as well.)

1.7.2 Equivalent definitions of linear independence

We can define linear independence in a few different ways (and then prove that these ways are all equivalent). To motivate these definitions, let's look at some properties of the two sets S above that *weren't* linearly independent.

1. In each case, one of the vectors in S is a linear combination of the others, so a linear combination involving that vector can be written without it. Some proper subset of S, therefore, has the same span as S itself.

2. By rearranging the expression for one vector in terms of the others, we can get a nontrivial linear combination of the vectors in S that equals $\mathbf{0}$. For example, if $\mathbf{v} = 2\mathbf{u} + 3\mathbf{w}$, then $2\mathbf{u} - \mathbf{v} + 3\mathbf{w} = \mathbf{0}$.

3. Any vector in $\operatorname{span} S$ can be written in multiple ways just by adding a multiple of the linear combination for $\mathbf{0}$. For instance, if $2\mathbf{u} - \mathbf{v} + 3\mathbf{w} = 0$, then any other linear combination $a\mathbf{u} + b\mathbf{v} + c\mathbf{w}$ is equal to $(a + 2k)\mathbf{u} + (b - k)\mathbf{v} + (c + 3k)\mathbf{w}$ for any scalar k.

Reversing these observations gives us a definition—in fact, several definitions—for *linear independence*.

Definition. *Let S be a subset of a vector space V. Then S is **linearly independent** if it is empty, or if it is nonempty and satisfies one of these equivalent properties:*

1. Every *element in $\operatorname{span} S$ is the value of only one linear combination of the elements of S.*

2. *The only linear combination of elements of S that generates $\mathbf{0}$ is the trivial linear combination, which gives every element of S the coefficient zero.*

3. There exists *some element $\mathbf{v} \in \operatorname{span} S$ that is the value of only one linear combination of elements in S.*

4. *No element of S can be written as a linear combination of the others.*

5. *There is no strict subset S' of S (meaning that S contains at least one element not in S') such that $\operatorname{span} S' = \operatorname{span} S$. (That is, removing an element from S always shrinks the span.)*

The claim that the properties in this definition are equivalent probably isn't immediately clear, so let's prove it.

Proposition. *The five properties in the list above are equivalent for nonempty sets S.*

Proof. We'll prove the implications $1 \implies 2 \implies 3 \implies 1$, $1 \implies 4 \implies 2$, and $4 \iff 5$. I'll leave it to you to convince yourself (perhaps by drawing a diagram) that with this chain of implications proved, the truth of any one property implies the truth of the other four.

- *1 implies 2*: If every element in $\operatorname{span} S$ is equal to only one linear combination in S, then $\mathbf{0}$, which is an element of the span of every set (including the empty set), can only have one linear combination. And the trivial linear combination always equals $\mathbf{0}$, so if $\mathbf{0}$ is the value of only one linear combination, it can't be the value of any nontrivial linear combination.

- *2 implies 3*: If $\mathbf{0}$ is equal to only one linear combination in S, then there must exist some element in $\operatorname{span} S$ equal to only one linear combination in S, namely $\mathbf{0}$.

- *3 implies 1*: We'll prove the contrapositive: not-1 implies not-3. Suppose that that property 1 is false for some set S: that is, some vector $\mathbf{v} \in \operatorname{span} S$ can be written as two different linear combinations $\mathbf{v} = a_1\mathbf{s}_1 + \cdots + a_n\mathbf{s}_n = b_1\mathbf{s}_1 + \cdots + b_n\mathbf{s}_n$, where $a_i \neq b_i$ for at least one index $1 \leq i \leq n$ and $\mathbf{s}_1, \ldots, \mathbf{s}_n$ are all elements of S. Then subtracting these linear combinations gives a nontrivial linear combination $(a_1 - b_1)\mathbf{s}_1 + \cdots + (a_n - b_n)\mathbf{s}_n$ for $\mathbf{0}$. We can add this to any other linear combination $c_1\mathbf{s}_1 + \cdots + c_n\mathbf{s}_n$ to get another linear combination $(a_1 + c_1 - b_1)\mathbf{s}_1 + \cdots + (a_n + c_n - b_n)\mathbf{s}_n$. So every element of $\operatorname{span} S$ is the value of multiple linear combinations, and property 3 is false for the set S.

- *1 implies 4*. Every element $\mathbf{s} \in S$ is already a linear combination of elements of S in at least one way (namely, the linear combination that includes only the element itself, with coefficient 1). If S satisfies property 1, then $\mathbf{s}$ can't equal any other linear combinations from S, and in particular, it can't be any linear combination whose values are drawn from $S \setminus \{\mathbf{s}\}$.

- *4 implies 2*. We'll prove that not-2 implies not-4. Suppose that we have some nontrivial linear combination $c_1\mathbf{s}_1 + \cdots + c_n\mathbf{s}_n = \mathbf{0}$ of elements of S (and suppose that the coefficients $c_1, \ldots, c_n$ are all nonzero). If $n = 1$, then $\mathbf{s}_1 = 0$, and $\mathbf{0}$ can always be written as a linear combination of the other elements of S (namely, the trivial linear combination). Otherwise, we have $\mathbf{s}_1 = -\frac{c_2}{c_1}\mathbf{s}_2 - \cdots - \frac{c_n}{c_1}\mathbf{s}_n$ exhibiting $\mathbf{s}_1$ as a nontrivial linear combination of other elements of S.

- *4 implies 5.* We'll prove that not-5 implies not-4. Suppose we have a proper subset $S' \subsetneq S$ such that $\operatorname{span} S' = \operatorname{span} S$. Let $\mathbf{s}$ be some element in S but not in S'. If $\operatorname{span} S = \operatorname{span} S'$, then $\mathbf{s} \in \operatorname{span} S'$, so $\mathbf{s}$ can be written as a linear combination of elements of S'—that is, of elements of S other than itself.

- *5 implies 4.* We'll prove that not-4 implies not-5. If some element $\mathbf{s} \in S$ is a linear combination of the others, then any linear combination that includes $\mathbf{s}$ can be rewritten in terms of the other elements, so $S \setminus \{\mathbf{s}\}$ is a strict subset of S that has the same span as $\mathbf{s}$.

□

One implication of the definition that's worth pointing out explicitly: if S contains the zero vector $\mathbf{0}$, then it can't be linearly independent. In this case, $\mathbf{0}$ is the value of a nontrivial linear combination from S (namely, the one-term linear combination $k\mathbf{0}$ for any nonzero scalar k), and it's also the value of a linear combination from the elements of S not including itself (namely, the trivial linear combination). In either case, it can be eliminated without changing the span of S. (This is true even if $\mathbf{0}$ is the only element of S, because by convention, $\operatorname{span}\emptyset = \operatorname{span}\{\mathbf{0}\} = \{\mathbf{0}\}$, and a sum of zero elements counts as a trivial "linear combination" of zero vectors.)

1.7.3 Infinite linear independent sets

One final result that will be occasionally useful:

Proposition. *An infinite set of vectors is linearly independent if and only if all of its finite subsets are linearly independent.*

Proof. Let S be an infinite subset of a vector space V. If S has a finite linearly dependent subset T, then any nontrivial linear combination from T with value $\mathbf{0}$ is also a linear combination from S. The contrapositive is that if S is linearly independent, then all of its subsets have to be as well. (This argument doesn't actually depend on the infinitude of S; it's a general proof that any subset of a linearly independent set is also linearly independent.)

Conversely, suppose S is linearly dependent; that is, $c_1\mathbf{s}_1 + \cdots + c_n\mathbf{s}_n = \mathbf{0}$ for some coefficients c_i and vectors s_1 chosen out of S. Then $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ is a linearly dependent finite subset of S. The contrapositive is that if all of S's finite subsets are linearly independent, then S is as well.

□

Answers to key questions.

1. A set is linearly independent if the only linear combination from the set that equals the zero vector is the trivial combination with all coefficients equal to zero.

2. This isn't possible: if we subtract one linear combination equal to $\mathbf{v}$ from another, then we get a nontrivial linear combination equal to $\mathbf{0}$. We can add this linear combination to any linear combination for $\mathbf{u}$ to get another linear combination for $\mathbf{u}$, so $\mathbf{u}$ has multiple linear combinations.

3. If S is not linearly independent, then there's some linear combination $c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$ of vectors in S that equals $\mathbf{0}$ with at least one of the coefficients $c_1, \ldots, c_n$ not equal to zero. Suppose that $c_n \neq 0$ (we can order the vectors and coefficients if we need to). Then any linear combination that includes $\mathbf{v}_n$ can be changed to an equivalent linear combination with $\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}$ by replacing $\mathbf{v}_n$ with $-\frac{c_1}{c_n}\mathbf{v}_1 - \cdots - \frac{c_{n-1}}{c_n}\mathbf{v}_{n-1}$. This means that S and $S \setminus \{\mathbf{v}_n\}$ have the same span.

1.8 Bases

Key questions.

1. What is the relationship between the concepts *basis*, *spanning set*, and *linearly independent set*?

2. Suppose S_1 is a linearly independent set of a vector space V and S_2 is a spanning set. What is a necessary relationship between the sizes of the sets S_1 and S_2?

3. Define *dimension*. What is the dimension of the spaces $\mathbb{R}^n$ for positive integers n? Give two different bases of $\mathbb{R}^3$. If the dimension of a vector space V is n, then is every set of n vectors from V a basis?

4. ($\star$) Consider the subspace $W = \{(x-y, y-z, z-x) : x, y, z \in \mathbb{R}\}$ of $\mathbb{R}^3$. Find two elements of W that are not scalar multiples of each other, and also find an element of $\mathbb{R}^3$ that is not in W. Explain why this is enough to conclude that $\dim W = 2$.

5. What does it mean to "extend" a basis for a subspace to a basis for an entire vector space?

6. Define *codimension*. If W has dimension 3 and is a subspace of a vector space V with dimension 5, what is the codimension of W? Give an example of an infinite-dimensional vector space and a subspace with finite codimension.

1.8.1 Core definitions: basis and dimension

We'll start out with two definitions.

Definition. *A **basis** of a vector space V (which could be a subspace of a larger space) is a linearly independent spanning set of V. The **dimension** of V, denoted $\dim V$, is the size of a basis of V.*

This definition should probably get your hackles up: how can we define the dimension of a space as the size of a basis of V when we don't know if bases all have the same size—or, in fact, if bases even exist for every space? We'll get to these points soon.

First, though, a word on the practical utility of bases: if $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis of V, then every element of V can be written in the form $c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$ in exactly one way: there's a one-to-one correspondence between elements of V and lists of coefficients $c_1, \ldots, c_n$. So identifying elements of an arbitrary vector space with their coefficients in some basis can make doing calculations with them, not to mention proving many results, a lot easier.

There are two key results about bases that we need to make our definition of "dimension" above valid:

1. Every vector space has a basis; and

2. Every basis of any given vector space has the same (possibly infinite) number of vectors, called the *dimension* of the vector space.

The proof of the first result, that every vector space has a basis, requires some unintuitive technical set theory, so I'll ask you to take it on faith. The second result follows (at least when the dimension is finite) from a result called the *Steinitz exchange lemma*. The details aren't crucial, but they're a good illustration of how to apply the theory that we've developed so far. This and future optional passages will be set in smaller text.

Lemma (Steinitz exchange lemma)**.** *Suppose $I = \{\mathbf{i}_1, \ldots, \mathbf{i}_m\}$ is a linearly independent set of m vectors in some space V, and $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ is a set that spans V. Then $m \leq n$, and you can make I into a set that spans V by adding some subset of $n - m$ vectors from S.*

Proof. S spans V, so we can write $\mathbf{i}_1$ as a linear combination from S:

$$\mathbf{i}_1 = c_1\mathbf{s}_1 + \cdots + c_n\mathbf{s}_n.$$

This linear combination must be nontrivial, because I is linearly independent and so $\mathbf{i}_1 \neq \mathbf{0}$. So at least one of the scalars $c_1, \ldots, c_n$ is nonzero. Suppose that $c_1 \neq 0$ (if not, we can just renumber the elements of S, designating a different element $\mathbf{s}_1$). On page 34, we showed that a linear combination of elements of S can replace any element with a nonzero coefficient in that linear combination and leave the span of S unchanged. So S has the same span as $S_1 \coloneqq \{\mathbf{i}_1, \mathbf{s}_2, \mathbf{s}_3, \ldots, \mathbf{s}_n\}$, the set formed by replacing $\mathbf{s}_1$ with $\mathbf{i}_1$. That is, S_1 is a spanning set of V.

Now write $\mathbf{i}_2$ as a linear combination of the elements of S_1:

$$\mathbf{i}_2 = c_1\mathbf{i}_1 + c_2\mathbf{s}_2 + c_3\mathbf{s}_3 + \cdots + c_n\mathbf{i}_n$$

(the coefficients c_n in this equation are different from the ones in the equation for $\mathbf{i}_1$). I is linearly independent, so none of its elements can be written as a linear combination of the others. In particular, $\mathbf{i}_2$ can't be a scalar multiple of $\mathbf{i}_1$. So at least one of the coefficients $c_2, \ldots, c_n$ on the elements of S must be nonzero. Suppose $c_2 \neq 0$ (again, we can renumber the elements of S if we have to). Then, again by our result from page 34, we can swap $\mathbf{i}_2$ in and $\mathbf{s}_2$ out without altering the span, getting a set $S_2 \coloneqq \{\mathbf{i}_1, \mathbf{i}_2, \mathbf{s}_3, \mathbf{s}_4, \ldots, \mathbf{s}_n\}$ that also spans V.

Similarly, $\mathbf{i}_3$ can be written as a linear combination from S_2:

$$\mathbf{i}_3 = c_1\mathbf{i}_1 + c_2\mathbf{i}_2 + c_3\mathbf{s}_3 + c_4\mathbf{s}_4 \ldots + c_n\mathbf{s}_n$$

and the linear independence of I means that $\mathbf{i}_3$ can't be a linear combination of just $\mathbf{i}_1$ and $\mathbf{i}_2$, so at least one of the coefficients $c_3, \ldots, c_n$ must be nonzero. If c_3 is nonzero (renumbering S if necessary), then $S_3 \coloneqq \{\mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_3, \mathbf{s}_4, \mathbf{s}_5, \ldots, \mathbf{s}_n\}$ is a spanning set of V.

If we repeat this process, we get spanning sets S_j that contain j elements of I and $n - j$ elements of S. We'll be forced to stop in one of two ways: either we make it to set S_m, containing all elements of I and possibly some leftover elements of S (that is $m \leq n$); or we run out of elements of S first (that is, $m > n$) and stop at S_n, a proper subset of I that also spans V. But the second way is actually impossible, because if $S_n = \{\mathbf{i}_1, \ldots, \mathbf{i}_n\}$ spans V, then we can write any of the leftover elements $\mathbf{i}_{n+1}, \ldots, \mathbf{i}_m$ of I as a linear combination of $\mathbf{i}_1, \ldots, \mathbf{i}_n$, contradicting the linear independence of I. So $m \leq n$, and S_m contains I along with $n - m$ elements of S.

□

The Steinitz exchange lemma shows that the largest linearly independent subsets of a vector space are at most the size of the smallest spanning sets. So all sets that are both linearly independent and spanning—that is, bases—must be the same size.

We can go a step further. Suppose a vector space V has finite dimension n. Then:

1. Any linearly independent set I with n elements must also span V. If I didn't span V, then we could add any vector outside $\operatorname{span} I$ to I and get a linearly independent set of $n + 1$ vectors, but this is impossible.

2. Any spanning set S with n elements must also be linearly independent. If S wasn't linearly independent, then we could eliminate some element of $\operatorname{span} S$ that was included in the span of the other other elements and get a spanning set of $n - 1$ vectors. This is also impossible.

So to sum up:

1. Every vector space has a basis.
2. All bases for a vector space have the same size, called the *dimension*.
3. The largest linearly independent sets, and the smallest spanning sets, both have size equal to the dimension.
4. A set whose size equals the dimension cannot be a spanning set but not linearly independent, or vice versa: it must be either both or neither.

Finally, the trivial subspace $\{\mathbf{0}\}$ is considered to have dimension zero and the empty set $\emptyset$ as a basis.

1.8.2 Bases for $\mathbb{F}^n$; conversion between bases

Most of our examples of vector spaces have had the form $\mathbb{F}^n$—that is, fixed-length lists of elements of a field such as $\mathbb{R}$ or $\mathbb{C}$. The dimension of $\mathbb{F}^n$ is n, because one basis for it is the n-element *standard basis* containing n vectors with one component equal to 1 and the other components all 0. The standard basis vector with 1 in component number i is commonly notated $\mathbf{e}_i$. For example, $\mathbb{F}^3$ has the standard basis $\{\mathbf{e}_1 = (1,0,0), \mathbf{e}_2 = (0,1,0), \mathbf{e}_3 = (0,0,1)\}$. Any vector $(x,y,z) \in \mathbb{F}^3$ can be written as a linear combination $x\mathbf{e}_1 + y\mathbf{e}_2 + z\mathbf{e}_3$.

But $\mathbb{F}^n$ has other bases as well. For example, $\{\mathbf{e}_1 = (1,0), \mathbf{e}_2 = (0,1)\}$ is one basis for $\mathbb{R}^2$, but so is $\{\mathbf{u}_1 = (1,2), \mathbf{u}_2 = (-2,-2)\}$. (It's easy to check that $\mathbf{u}_1$ and $\mathbf{u}_2$ are not scalar multiples of each other, and any linearly independent two-element subset of $\mathbb{R}^2$ must be a basis.)

You can convert vectors between these two bases. Conversion from $\{\mathbf{u}_1, \mathbf{u}_2\}$ to the standard basis is straightforward: $a\mathbf{u}_1 + b\mathbf{u}_2 = (a-2b)\mathbf{e}_1 + (2a-2b)\mathbf{e}_2$. Converting in the other direction requires us to write $\mathbf{e}_1$ and $\mathbf{e}_2$ as linear combinations of $\mathbf{u}_1$ and $\mathbf{u}_2$. We do this by solving linear systems with one equation for each component of $\mathbf{e}_1$ or $\mathbf{e}_2$.

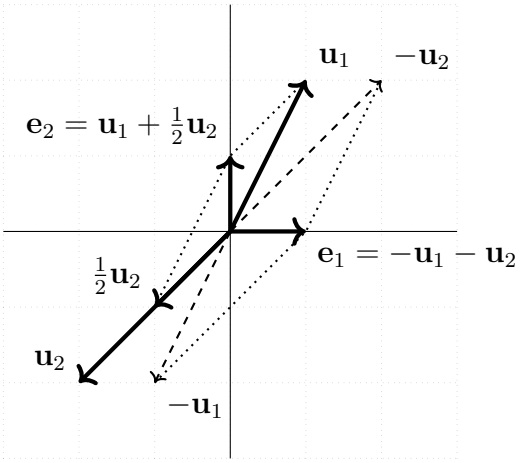
Specifically, suppose $\mathbf{e}_1 = a\mathbf{u}_1 + b\mathbf{u}_2$. The expression $a\mathbf{u}_1 + b\mathbf{u}_2$ is $a(1,2) + b(-2,-2) = (a-2b, 2a-2b)$. If this vector is $\mathbf{e}_1$, then the scalars a, b satisfy the system

$$\begin{aligned} a - 2b &= 1 \\ 2a - 2b &= 0 \end{aligned}$$

which you can solve to get $(a,b) = (-1,-1)$, so $\mathbf{e}_1 = -\mathbf{u}_1 - \mathbf{u}_2$. Similarly, if $\mathbf{e}_2 = c\mathbf{u}_1 + d\mathbf{u}_2$, then c and d satisfy

$$\begin{aligned} c - 2d &= 0 \\ 2c - 2d &= 1 \end{aligned}$$

which you can solve to get $(c,d) = (1, \frac{1}{2})$, so $\mathbf{e}_2 = \mathbf{u}_1 + \frac{1}{2}\mathbf{u}_2$. This diagram may help you visualize what's going on:



So conversion from the standard basis to the basis $\{\mathbf{u}_1, \mathbf{u}_2\}$ uses the formula $x\mathbf{e}_1 + y\mathbf{e}_2 = x(-\mathbf{u}_1 - \mathbf{u}_2) + y(\mathbf{u}_1 + \frac{1}{2}\mathbf{u}_2) = (-x + y)\mathbf{u}_1 + (-x + \frac{1}{2}y)\mathbf{u}_2$. Later on, we'll learn a standard technique for finding conversion formulas from the standard basis of $\mathbb{F}^n$ to nonstandard bases. The theory of basis changes is crucial because linear maps with complicated formulas in one basis often have much simpler forms in another basis.

One final note: the vectors $\mathbf{e}_1 = (1, 0, 0, 0, \ldots)$, $\mathbf{e}_2 = (0, 1, 0, 0, \ldots)$, $\mathbf{e}_3 = (0, 0, 1, 0, \ldots)$, and so on form a basis for $\mathbb{F}^\infty$, the set of infinite sequences within only a finite number of nonzero terms—for instance, the sequence $(-2, 1, 0, 5, 0, 0, 0, \ldots, 0, \ldots)$ is $-2\mathbf{e}_1 + \mathbf{e}_2 + 5\mathbf{e}_4$. (This space is often useful as a source of counterexamples for statements about finite-dimensional spaces that don't apply to infinite-dimensional spaces, so we'll see it a few more times.) But $\{\mathbf{e}_1, \mathbf{e}_2, \ldots\}$ is *not* a basis for $\mathbb{F}^\mathbb{N}$, the set of sequences with a possibly infinite number of nonzero terms. For instance, the constant sequence $(1, 1, 1, \ldots)$ would have to be written in terms of the standard basis vectors for $\mathbb{F}^\infty$ as $\mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3 + \mathbf{e}_4 + \cdots$, but in generic vector spaces, there is no notion of infinite sums.

1.8.3 Equivalence of finite-dimensional vector spaces and $\mathbb{F}^n$

The existence of bases lets us translate arithmetic in any finite-dimensional vector space into arithmetic in $\mathbb{F}^n$. Suppose, for example, that V is a three-dimensional space with a basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$. Any vector $\mathbf{v} \in V$ can therefore be written as $\mathbf{v} = a\mathbf{u}_1 + b\mathbf{u}_2 + c\mathbf{u}_3$. Let's represent this vector with the element (a, b, c) of $\mathbb{F}^3$. Then:

1. Addition of elements of V can be modeled by adding their representations in $\mathbb{F}^3$. That is, if $\mathbf{v} \coloneqq a\mathbf{u}_1 + b\mathbf{u}_2 + c\mathbf{u}_3$ and $\mathbf{w} \coloneqq x\mathbf{u}_1 + y\mathbf{u}_2 + z\mathbf{u}_3$ correspond to (a, b, c) and (x, y, z) respectively, then their sum $\mathbf{v} + \mathbf{w} = (a + x)\mathbf{u}_1 + (b + y)\mathbf{u}_2 + (c + z)\mathbf{u}_3$ corresponds to $(a + x, b + y, c + z)$.

2. Taking scalar multiples of any element of V corresponds to scaling multiples of its representation in $\mathbb{F}^3$. If $\mathbf{v} \coloneqq a\mathbf{u}_1 + b\mathbf{u}_2 + c\mathbf{u}_3$ corresponds to (a, b, c), then $k\mathbf{v} = ka\mathbf{u}_1 + kb\mathbf{u}_2 + kc\mathbf{u}_3$ corresponds to (ka, kb, kc).

The same idea works for dimensions besides 3: any n-dimensional vector space V over a field $\mathbb{F}$ has the same structure as $\mathbb{F}^n$. Depending on your choice of basis for V, you could make any vector in V correspond to any vector in $\mathbb{F}^n$ (except that zero vectors always have to correspond to each other). But as long as you keep your choice of basis consistent, you can model all vector space operations on V in $\mathbb{F}^n$, so any results on $\mathbb{F}^n$ will generalize to all vector spaces.

Note that for translating from V to $\mathbb{F}^n$, the order of basis vectors matters. If you choose $\{\mathbf{u}_3, \mathbf{u}_1, \mathbf{u}_2\}$ as a basis, then the element of $\mathbb{F}^3$ corresponding to $a\mathbf{u}_1 + b\mathbf{u}_2 + c\mathbf{u}_3$

is (c, a, b) rather than (a, b, c). (This means that using the curly set brackets to denote bases is a slight abuse of notation, because the elements of a set don't have an order, but it's common and shouldn't be too confusing.)

1.8.4 Bases of subspaces; codimension

Suppose V is a vector space and W is a nontrivial subspace of V. It's always possible to find a basis A of V such that some subset of W is a basis for W. In fact, even if you start with a *specific* basis B for W, you can extend this to some other basis A that includes B as a subset.

The proof of this in the general case involves technical set theory, but in the finite-dimensional case, it's straightforward. Suppose V has dimension m and W has dimension n (where, of course, $m \geq n$). The process for extending some basis B of W to a basis for A is simple:

1. Take an arbitrary vector in V that is not in $\operatorname{span} B$, and add it to B.
2. Repeat step 1 until $\operatorname{span} B = V$.

This process keeps B linearly independent at each step, and it is guaranteed to finish after adding $m - n$ vectors to B. Once B contains m vectors, its span must be all of B; otherwise, we could add some element in V but outside $\operatorname{span} B$ and get a linearly independent set of $m + 1$ elements in an m-dimensional vector space, but such a set can't exist. One corollary (which should be intuitive) is that the dimension of a subspace can't exceed the dimension of the larger space.

This idea also gives us an important new term:

Definition. *If W is a subspace of V, then the **codimension** of W is the number of vectors that we have to add to a basis of W to get a basis of V.*

When V is finite-dimensional, of course, $\operatorname{codim} W = \dim V - \dim W$. But when W and V are both infinite-dimensional, $\operatorname{codim} W$ can stil be finite.

As an example of an infinite-dimensional vector space and subspace with finite codimension, consider the space $\mathbb{F}^\infty$ of infinite sequences with a finite number of nonzero terms: an element of $\mathbb{F}$ has the form $(a_1, a_2, a_3, \ldots)$, where for each element there's some index N such that $a_n \neq 0$ for every $n \geq N$. The n$\mathbb{F}^\infty$ has a basis made up of the standard basis vectors $\mathbf{e}_1 = (1, 0, 0, 0, \ldots)$, $\mathbf{e}_2 = (0, 1, 0, 0, \ldots)$, $\mathbf{e}_3 = (0, 0, 1, 0, \ldots)$, and so forth. Now consider the subspace $W \subset \mathbb{F}^\infty$ of all such sequences whose first entry is zero: that is, sequences of the form $(0, a_2, a_3, \ldots)$, where there's some integer N such that $a_n = 0$ for all $n \geq N$. Then W has basis $\{\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \ldots\}$, and adding $\mathbf{e}_1$ makes this into a vector space basis for all of $\mathbb{F}^\infty$. So $\operatorname{codim} W = 1$.

One final remark that should go without saying: the codimension of a vector space depends on which larger vector space it's being considered a subset of. For instance, if you have three nested vector spaces $U \subset V \subset W$, then the codimension of U with respect to V is not equal to the codimension of U with respect to W. Usually, though, the larger space relative to which codimensions are defined should be clear.

(You may also be wondering whether the codimension, like the dimension, is necessarily unique. Rest assured that it is. We won't present a proof here, but later, we'll see that the codimension of a subspace W with respect to a larger space V is also the dimension of another vector space notated V/W and called the *quotient space*, which

we'll define in section 5.4. As the dimension of V/W, like the dimension of any vector space, is uniquely defined, so is the codimension of W.)

Answers to key questions.

1. A basis of a vector space is a set that is both lineraly independent and a spanning set of the entire vector vector space.

2. $|S_1| \leq |S_2|$.

3. The dimension of a space is the size of any basis for it. The dimension of $\mathbb{R}^n$ is n.

   Any set of three linearly independent elements of $\mathbb{R}^3$ must be a basis. Two possible bases are $\{(1,0,0),(0,1,0),(0,0,1)\}$ and $\{(2,0,0),(4,5,0),(0,0,-10)\}$ (there are many other possibilities).

   A set of n vectors from an n-dimensional vector space is a basis of and only if it is linearly independent, so it is not always a basis.

4. $(1,-1,0)$ (from setting $x=1, y=0, z=1$) and $(1,0,-1)$ (from setting $x=1,0, z=0$) are two elements of W that are not scalar multiples of each other (that is, they form a linearly independent set). So W must have dimension at least 2. Meanwhile, $(1,0,0)$ is not in W (to see this, note that the components of any element of W must have sum zero). So W can't be all of $\mathbb{R}^3$, which means it can't have dimension 3. And of course, a subspace of a three-dimensional vector space can't have dimension 4 or more. So W must have dimension 2.

5. The codimension of a vector subspace is the number of elements we'd need to add to a basis for the subspace to form a basis for the entire space. The codimension of a three-dimensional subspace of a five-dimensional space is $5-3=2$.

   One example of an infinite-dimensional vector space is $\mathbb{R}^{\mathbb{N}}$, the set of infinite-dimensional sequences of real numbers. One subspace of this (call it W) is the set of infinite-dimensional sequences with a zero in the first element. Any element of $\mathbb{R}^{\mathbb{N}}$ can be written uniquely as the sum of an element of W plus a multiple of the sequence $(1,0,0,0,\ldots)$, so adding $(1,0,0,0,\ldots)$ to a basis of W produces a basis of $\mathbb{R}^{\mathbb{N}}$, so W has codimension 1.

1.9 Affine spaces

Key questions.

1. If W is a subspace of V and A is a subset of W, what is the geometric intuition behind saying that A is an *affine space parallel to* or *coset of* W? Give four equivalent algebraic definitions: two of the form "for any element $\mathbf{a}_0 \in A$, ..." and two of the form "there exists an element $\mathbf{a}_0 \in A$ such that ...".

2. Explain why every subspace is a coset of itself.

3. ($\star$) Can two different cosets of the same subspace have vectors in common?

4. ($\star$) Prove or give a counterexample: if W_1 and W_2 are two subspaces of a vector space V, and W_2 contains a coset of W_1, then $W_1 \subseteq W_2$.

5. ($\star$) Show that if A is an affine subspace of V, then $\{\mathbf{a}_1 - \mathbf{a}_2 : \mathbf{a}_1, \mathbf{a}_2 \in A\}$ is a subspace of V. Is the converse also true?

6. What is the formula for the sum of cosets of two different subspaces (defined analogously to the sum of subspaces)?

7. Suppose A_1, A_2 are affine spaces of $\mathbb{R}^{10}$ with respective dimensions 6 and 7. What are the possible dimensions of $A_1 + A_2$? Is it possible for $A_1 \cap A_2$ to be the empty set? What are the possible dimensions of $A_1 \cap A_2$ if it's not the empty set?

When we defined subspaces, we noted that that every subspace always contains $\mathbf{0}$, and that in $\mathbb{R}^2$ and $\mathbb{R}^3$, we can interpret nontrivial subspaces geometrically as lines and planes through the origin. You may wonder if there's a similar concept for the lines and planes that do *not* include the origin. In fact, there is: the concept of *affine subspaces*. This may seem like an unmotivated definition, but it will come in very useful when we discuss how to solve linear systems.

Let's start with an example: the set $W \subset \mathbb{R}^2$ of pairs (x, y) such that $y = 2x$. You can check easily that W is a vector subspace of $\mathbb{R}^2$, and W is geometrically a line through the origin of a two-dimensional Cartesian plane.

Now let A be the set of pairs (x, y) such that $y = 2x + 1$—that is, those of the form $(x, 2x+1)$. You can see easily that A is not a subspace of $\mathbb{R}^2$ (for instance, you can simply note that $\mathbf{0} \notin A$, or that $(1, 3)$ is in A but not its multiple $(2, 6)$). Geometrically, A and W are parallel lines. In purely algebraic terms, A has two properties that relate it back to W, and we'll use these properties to produce a more general notion of "parallel" that applies to vector spaces without simple geometric interpretations:

1. The difference between any two elements of A is an element of W: if $\mathbf{a}_1 = (x_1, 2x_1 + 1)$ and $\mathbf{a}_2 = (x_2, 2x_2 + 1)$, then $\mathbf{a}_1 - \mathbf{a}_2 = (x_1 - x_2, 2x_1 - 2x_2)$, which is in W because its second component is double its first component.

2. Conversely, if you pick some fixed vector $\mathbf{a} \in A$, then any vector in W plus $\mathbf{a}$ is another vector in A. This is clearest if we choose $\mathbf{a} = (0, 1)$, because some arbitrary vector $(k, 2k) \in W$ plus $(0, 1)$ is $(k, 2k + 1)$, which matches the general form for elements of A. But if we choose, say, $\mathbf{a} = (-\frac{1}{2}, 0)$ instead, then $(k, 2k) + (-\frac{1}{2}, 2k) = (k - \frac{1}{2}, 2k)$, which is also of the form $(x, 2x + 1)$ (just choose $x = k - \frac{1}{2}$). Similarly, if we chose $\mathbf{a} = (1, 3)$, then $(k + 1, 2k + 3)$ also has the form $(x, 2x + 1)$ (choose $x = k + 1$).

Affine spaces are essentially sets that are "parallel to" some subspace, where we define "parallel" using the algebraic definition above.

Definition. *Let V be a vector space, W be a subspace, and A be a nonempty subset of V. A is an **affine space** parallel to W, or equivalently a **coset** of W, if any of these (logically equivalent) conditions holds.*

1. *For* all *vectors $\mathbf{a}_0 \in A$, the set $\{\mathbf{a} - \mathbf{a}_0 : \mathbf{a} \in A\}$ equals W.*

2. *There exists* at least one *$\mathbf{a}_0 \in V$ (necessarily also in A) such that $\{\mathbf{a} - \mathbf{a}_0 : \mathbf{a} \in A\} = W$. (The reason that $\mathbf{a}_0$ must be in A is that otherwise, $\{\mathbf{a} - \mathbf{a}_0 : \mathbf{a} \in A\}$ would not contain $\mathbf{0} \in W$.)*

3. *For* all *fixed vectors* $\mathbf{a}_0 \in A$*, the set* $\{\mathbf{w} + \mathbf{a}_0 : \mathbf{w} \in W\}$ *of sums of* $\mathbf{a}_0$ *and any vector in* $\mathbf{w}$ *equals* A*.*

4. *There exists* at least one *vector* $\mathbf{a}_0 \in V$ *(again, necessarily in* A*—why?) such that* $\{\mathbf{w} + \mathbf{a}_0 : \mathbf{w} \in W\} = A$*.*

Note that according to this definition, W is also an affine space parallel to (that is, it's a coset of) itself.

It's probably not immediately apparent why the four conditions in the definition are equivalent. Let's prove it.

Proposition. *The conditions in the definition of coset are logically equivalent.*

Proof. We'll prove the logical equivalences $1 \leftrightarrow 2$, $3 \leftrightarrow 4$, and $2 \leftrightarrow 4$.

Equivalence of 1 and 2. It's trivial that condition 1 implies condition 2: the inference from "for all $\mathbf{a}_0 \in A$" to "there exists $\mathbf{a}_0 \in V$" is always true if A is a nonempty subset of V. To show that 2 implies 1, suppose $\mathbf{a}_0 \in A$ satisfies $\{\mathbf{a} - \mathbf{a}_0 : \mathbf{a} \in A\} = W$, and choose an arbitrary element $\mathbf{b}_0$ of A. (This means $\mathbf{b}_0 - \mathbf{a}_0$, as the difference of two elements of A, is in W.) Write $S \coloneqq \{\mathbf{a} - \mathbf{b}_0 : \mathbf{a} \in A\}$. To show that condition 1 is true, we need to show that $S = W$. We can show this by proving inclusion two ways: that $S \subseteq W$ and $W \subseteq S$.

- *Is* S *a subset of* W*?* We know that $\mathbf{b}_0 - \mathbf{a}_0$ is in W, as is $\mathbf{a} - \mathbf{a}_0$ for any $\mathbf{a} \in A$. So their difference $(\mathbf{a} - \mathbf{a}_0) - (\mathbf{b}_0 - \mathbf{a}_0) = \mathbf{a} - \mathbf{b}_0$, which is an element of S, must also be in W, because the difference of any two elements of a vector subspace is also in the subspace.

- *Is* W *a subset of* S*?* If $\mathbf{w}$ is in W, then so is $\mathbf{w}' \coloneqq \mathbf{w} + (\mathbf{b}_0 - \mathbf{a}_0)$. Condition 2 means that we can choose some element $\mathbf{a} \in A$ such that $\mathbf{w}' = \mathbf{a} - \mathbf{a}_0$. So $\mathbf{w}$ can be written as $(\mathbf{a} - \mathbf{a}_0) - (\mathbf{b}_0 - \mathbf{a}_0) = \mathbf{a} - \mathbf{b}_0$. That is, we can write any element of W in the form of an element in S.

Equivalence of 3 and 4. Again, it's trivial that 3 implies 4: "for any" on a nonempty set implies "there exists." To prove that 4 implies 3, suppose $\{\mathbf{w} + \mathbf{a}_0 : \mathbf{w} \in W\} = A$. Choose an arbitrary vector $\mathbf{b}_0 \in A$, and choose $\mathbf{w}' \in W$ such that $\mathbf{w}' + \mathbf{a}_0 = \mathbf{b}_0$. Define $S = \{\mathbf{w} + \mathbf{b}_0 : \mathbf{w} \in W\}$. To show that 3 is true, we need to show $S = A$ for every possible choice of $\mathbf{b}_0$. Again, to show that two sets are equal, we can show that each is a subset of the other.

- *Is* S *a subset of* A*?* Evern element of S can be written as $\mathbf{w} + \mathbf{b}_0$ where $\mathbf{w}$ is an element of W. Remember that $\mathbf{w}' \in W$ satisfies $\mathbf{w}' + \mathbf{a}_0 = \mathbf{b}_0$. Then $(\mathbf{w} + \mathbf{w}') + \mathbf{a}_0 = \mathbf{w} + \mathbf{b}_0$, so $\mathbf{w} + \mathbf{b}_0$ is the sum of $\mathbf{a}_0$ and an element of W, so it has to be in A.

- *Is* A *a subset of* S*?* Any element $\mathbf{w} + \mathbf{a}_0 \in A$ can be rewritten $\mathbf{w} + \mathbf{w}' + \mathbf{b}_0$, which is the sum of $\mathbf{b}_0$ and an element $\mathbf{w} + \mathbf{w}'$ of W, so every vector in A is also in S.

2 implies 4. Let $\mathbf{a}_0$ be such that $\{\mathbf{a} - \mathbf{a}_0 : \mathbf{a} \in A\} = W$, and let $S = \{\mathbf{w} + \mathbf{a}_0 : \mathbf{w} \in W\}$. We want to prove that $A = S$. Again, we prove:

- *Is* S *a subset of* A*?* Let $\mathbf{w} + \mathbf{a}_0$ be an arbitrary element of S. $\mathbf{w}$ must equal $\mathbf{a} - \mathbf{a}_0$ for some $\mathbf{a} \in A$, so $\mathbf{w} + \mathbf{a}_0 = \mathbf{a} \in A$. So $S \subseteq A$.

- *Is* A *a subset of* S*?* If $\mathbf{a} \in A$, then $\mathbf{a} - \mathbf{a}_0 \in W$, and thus $(\mathbf{a} - \mathbf{a}_0) + \mathbf{a}_0 = \mathbf{a} \in S$.

4 implies 2. Let $\mathbf{a}_0$ be such that $\{\mathbf{w} + \mathbf{a}_0 : \mathbf{w} \in W\} = A$, and let $S = \{\mathbf{a} - \mathbf{a}_0 : a \in A\}$. We want to prove that $S = W$.

- *Is* S *a subset of* W*?* If $\mathbf{s} \coloneqq \mathbf{a} - \mathbf{a}_0$ is any element of S, then $\mathbf{s} + \mathbf{a}_0 = \mathbf{a}$ is an element of A. Every element $\mathbf{a}$ of A can be written $\mathbf{a} = \mathbf{w} + \mathbf{a}_0$ for some $\mathbf{w} \in W$. So $\mathbf{w} + \mathbf{a}_0 = \mathbf{s} + \mathbf{a}_0$, so $\mathbf{w} = \mathbf{s}$ and $\mathbf{s} \in W$.

- *Is* W *a subset of* S*?* If $\mathbf{w} \in W$, then $\mathbf{a} \coloneqq \mathbf{w} + \mathbf{a}_0 \in A$, so $\mathbf{a} - \mathbf{a}_0 = \mathbf{w} \in S$.

□

Note: If A is an affine subspace, then the set $\{\mathbf{a}_1 - \mathbf{a}_2 : \mathbf{a}_1, \mathbf{a}_2 \in A\}$ of differences between elements of A is a vector subspace. But $\{\mathbf{a}_1 - \mathbf{a}_2 : \mathbf{a}_1, \mathbf{a}_2 \in A\}$ can be a vector subspace even if A is just a subset of an affine space but not an affine space itself. Consider, for instance, $A = \{(x, 1) : x \in \mathbb{R}, x \geq 0\} \subset \mathbb{R}^2$, containing all points on the line $y = 1$ with a non-negative x-coordinate. Then $\{\mathbf{a}_1 - \mathbf{a}_2 : \mathbf{a}_1, \mathbf{a}_2 \in A\} = \{(x, 0) : x \in \mathbb{R}\} = \mathrm{span}\{\mathbf{e}_1\}$, which is a subspace of $\mathbb{R}^2$, but A is not a coset of $\mathrm{span}\{\mathbf{e}_1\}$.

If W is a subspace and A is a parallel affine space $\{\mathbf{w} + \mathbf{a}_0 : \mathbf{w} \in W\}$, then we can denote A as $\mathbf{a}_0 + W$. Note again, finally, that the choice of base point $\mathbf{a}_0$ is arbitrary: any point in A gives the same set. In particular, if $\mathbf{a}_0 - \mathbf{b}_0 \in W$, then $\mathbf{a}_0 + W = \mathbf{b}_0 + W$. This point is key to making algebraic manipulations with affine subspaces.

One final proposition:

Proposition. *Suppose A is an affine subspace of a vector space V, and W is the vector subspace parallel to A. Then any vector subspace of V that contains A also contains W.*

Proof. Every element of W can be written as the difference of two elements in A. The difference of two elements of a vector subspace is also in that subspace, so any subspace that contains A must also contain the set of differences of elements of A; that is, it must contain W.

□

Chapter 2

Linear maps

Overview. A linear map is a function between two vector spaces that preserves the vector space operations: adding the inputs to a linear map corresponds to adding the outputs, and likewise for multiplying an input by a scalar. The study of linear maps is important because in many other fields of mathematics (not to mention physics and engineering), many commonly used functions are linear, or at least can be approximated by linear maps. (For instance, in multivariable calculus, if f is a real-valued function of n real variables, then the derivative of f at a point $(x_1, \ldots, x_n)$ is the linear map $T : \mathbb{R}^n \to \mathbb{R}$ such that $T(h_1, \ldots, h_n)$ is the closest possible approximation to $f(x_1 + h_1, \ldots, x_n + h_n) - f(x_1, \ldots, x_n)$ when $h_1, \ldots, h_n$ are small).

This chapter presents the most basic concepts of linear maps that are required to understand core matrix algorithms. Section 2.1 presents the core axioms of linear maps and provides some examples. In introductory linear algebra, the most important class of linear maps are those from one finite-dimensional vector space to another. In Section 2.2, we present a general form for maps between the prototypical finite-dimensional vector spaces $\mathbb{F}^m$ and $\mathbb{F}^n$: each component of an output vector is a weighted sum of the components of the input vector, multiplied by fixed coefficients.

The set of maps between two vector spaces can be made into a vector space itself, with the sum of two different maps T_1, T_2 being defined as the map that takes an input vector $\mathbf{v}$ to the output $T_1\mathbf{v} + T_2\mathbf{v}$ (and scalar multiplication is similar). Section 2.3 presents this idea and also discusses the dimension of the resulting vector space.

Every linear map defines two important vector subspaces: its *image* or set of values, which is a subspace of its codomain; and its *kernel* or the set of input vectors that get mapped to zero, wihch is a subspace of its domain. Section 2.4 defines these spaces, proves that they are in fact vector subspaces, and points out a few consequences. Most importantly: a linear map is injective if and only if its kernel contains only the zero vector, and the image of any injective map has the same dimension as its domain.

Section 2.5 presents a further generalization that will be a crucial result for working with finite-dimensional vector spaces, including matrix algorithms: the *rank–nullity theorem*. This states, to put it intuitively, that any linear map collapses the dimensions contained in its kernel, and maps any remaining dimensions of the domain to independent dimensions in the image—that is, the codimension of any map's kernel equals the domain of its image. When the domain is finite-dimensional, this means that the sum of the kernel and image dimensions equals the domain dimension. It's hard to appreciate how important this result is, but it will crop up frequently in future discussions.

Section 2.6 presents various results building on the rank–nullity theorem, relating

the dimensions of the images and preimages of all affine subspaces of V and W under a linear map $T : V \to W$ to the dimension of T's kernel (that is, the preimage of $\{\mathbf{0}_W\}$) and its image (that is, the image of all of V). This section is something of a grab-bag of results, but two findings are especially important: first, the preimage of any single-vector subset $\{\mathbf{w}\}$ of W has a preimage that is either empty or an affine space parallel to T's kernel (this observation is crucial for understanding how matrices can be used to solve linear systems); and second, the dimension of the kernel of a composition of several maps is at most the sum of the kernel dimensions of the original maps.

Finally, Section 2.7 is a brief mention of another important fact: the inverse function to any linear map (if the inverse is defined) is also a linear map. This is preparation for the algorithms presented in the next chapter, which will give us practical ways of computing map inverses.

2.1 Basic definitions and examples

Key questions.

1. What two properties must a linear map satisfy?

2. Suppose $T : \mathbb{R}^2 \to \mathbb{R}^2$ is a linear map such that $T(1, 1) = (2, -2)$ and $T(0, 1) = (3, 4)$. What is $T(3, 2)$?

3. Is the composition of two linear maps always linear?

4. Give an example of a nonlinear map $T : \mathbb{R}^2 \to \mathbb{R}^2$ such that $T(0, 0) = (0, 0)$.

Vector spaces by themselves aren't that interesting: the interesting things to study are linear maps from one vector space to another. If V and W are vector spaces over the same field $\mathbb{F}$ (we can't define linear maps between vector spaces over different fields!), then a function $f : V \to W$ is *linear* if it satisfies these criteria:

1. *Respect for addition*: For any two elements $\mathbf{v}_1, \mathbf{v}_2 \in V$, $f(\mathbf{v}_1 + \mathbf{v}_2) = f(\mathbf{v}_1) + f(\mathbf{v}_2)$. Note that the addition on the left-hand side of this equation happens in V, and the addition on the right-hand side happens in W.

   This axiom implies its own generalization to sums of three or more terms: for instance $f(\mathbf{v}_1+\mathbf{v}_2+\mathbf{v}_3) = f((\mathbf{v}_1+\mathbf{v}_2)+\mathbf{v}_3) = f(\mathbf{v}_1+\mathbf{v}_2)+f(\mathbf{v}_3) = f(\mathbf{v}_1)+f(\mathbf{v}_2)+f(\mathbf{v}_3)$.

2. *Respect for scalar multiplication*: For any element $\mathbf{v} \in V$ and any scalar $k \in \mathbb{F}$, $f(k\mathbf{v}) = kf(\mathbf{v})$. (Note that the left-hand side of this equation is scalar multiplication in V and the right-hand side is scalar multiplication in W. This is why we can't define linear maps between vector spaces over two different fields.)

   One consequence is that $f(\mathbf{0}_V) = \mathbf{0}_W$, which follows from setting $k = 0$ and $\mathbf{v}$ to an arbitrary vector in V; this fact can be useful for quickly showing that a map is not linear. (We'll use subscripts on the symbol $\mathbf{0}$ to clarify which subspace's zero vector we're referring to.)

We can unify these two axioms into one formula: if f is linear, then $f(c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n) = c_1 f(\mathbf{v}_1) + \cdots + c_n f(\mathbf{v}_n)$ for any set of scalars $c_1, \ldots, c_n \in \mathbb{F}$ and vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$. That is, if you know the values of f on any set of vectors $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset V$, then you know its values on the entire span of S.

As always, a couple of examples may make things clearer.

1. The map $T : \mathbb{C}^2 \to \mathbb{C}$ given by $T(z, w) = z+(3+i)w$ is linear. If $k \in \mathbb{C}$ is any scalar, for instance, then $T(kz, kw) = kz + (3 + i)kw = k(z + (3 + i)w) = kT(z, w)$, so T respects multiplication. Similarly, $T(z_1+z_2, w_1+w_2) = z_1+(3+i)w_1+z_2+(3+i)w_2 = T(z_1, w_1) + T(z_2, w_2)$, so T respects scalar addition.

2. The map $T : \mathbb{R}^3 \to \mathbb{R}^4$ given by $T(x, y, z) = (y - 2z, x + 3z, y, -y)$ is also linear. You might want to try to show this yourself.

3. The map $T : \mathbb{R}^2 \to \mathbb{R}^2$ given by $T(x, y) = (x^2, 2y^2)$, however, is *not* linear. You can show this by noting that $T(1, 1) = (1, 2)$ but $T(2, 2) = (4, 8)$. Any linear map T, however, has to satisfy $T(2, 2) = 2T(1, 1)$ because it has to preserve scalar multiplication.

4. The map $T : \mathbb{R}^2 \to \mathbb{R}^2$ given by $T(x, y) = (x + 1, y + 1)$ is also not linear. You can show this by simply noting that $T(0, 0) \neq (0, 0)$, but every linear map has to take the zero vector in its domain to the zero vector in its codomain.

We usually denote linear maps with capital letters such as T and drop the parentheses that indicate function application (unless we need them to indicate correct order of operations, such as distinguishing $T(\mathbf{u} + \mathbf{v})$ from $T(\mathbf{u}) + \mathbf{v}$): ordinarily, we'll write the value of a map T on a vector $\mathbf{v}$ as just $T\mathbf{v}$. We also won't write function composition symbols: the map $T_1 : U \to V$ composed with $T_2 : V \to W$ is just written T_2T_1, not $T_2 \circ T_1$. (As always with function composition, the first function to be applied goes on the right.)

You may think that this notation makes application and composition of linear maps look like a sort of multiplication. This is intentional. It turns out that over finite-dimensional spaces at least, linear maps can be represented by *matrices* containing the coefficients in their formula, and vectors can be represented by *column vectors* (that is, matrices with only one column). Applying linear maps to vectors, and composing linear maps, turns out to be equivalent to multiplying their matrix representations using standard matrix multiplication.

One final result, not hard to prove but stil worth stating explicitly:

Proposition. *The composition of two linear maps $T_1 : U \to V$ and $T_2 : V \to W$ is itself a linear map from U to V.*

Proof. To show that T_2T_1 respects addition, note that for any vectors $\mathbf{u}_1, \mathbf{u}_2$ we have $T_2T_1(\mathbf{u}_1 + \mathbf{u}_2) = T_2(T_1\mathbf{u}_1 + T_1\mathbf{u}_2)$ (because T_1 by itself respects addition), which equals $T_2T_1\mathbf{u}_1 + T_2T_1\mathbf{u}_2$ (because T_2 by itself respects addition). The argument that T_2T_1 respects multiplication is similar.

□

Answers to key questions.

1. A linear map must satisfy *respect for addition* (its value on a sum of two inputs is the sum of its values on the individual inputs) and *respect for multiplication* (multiplying the input multiplies the output by the same factor).
2. As $(3,2) = 3(1,1) - (9,1)$, so $T(3,2) = 3(2,-2) - (3,4) = (3,-10)$.
3. Yes, the composition of linear maps is always linear.
4. There are many possibilities, for instance $T(x,y) = (x^2+y^2, \sqrt[3]{x})$.

2.2 General form of linear maps from $\mathbb{F}^m$ to $\mathbb{F}^n$

Key questions.

1. If $T : \mathbb{R}^3 \to \mathbb{R}^2$ is a linear map such that $T(1,0,0) = (a,b)$, $T(1,1,0) = (c,d)$, and $T(0,1,2) = (e,f)$, then what is $T(x,y,z)$ in terms of a,b,c,d,e,f,x,y,z?
2. In general, a linear map from an m-dimensional vector space to an n-dimensional space is determined by how many coefficients?

We mentioned in the last section that if you know the values of a linear map on any set of vectors S, then you know its values on all of $\operatorname{span} S$. So if you know a linear map's values on a basis of its domain, then you know its value on its entire domain. We can use this to work out a general form for linear maps between any two finite-dimensional vector spaces.

As an example, let's work out a general form for linear maps from $\mathbb{R}^2$ to $\mathbb{R}^3$. Remember the notation $\mathbf{e}_1 = (1,0)$, $\mathbf{e}_2 = (0,1)$.

Suppose $T : \mathbb{R}^2 \to \mathbb{R}^3$ is some linear map, and we know the values that T takes on the standard basis of $\mathbb{R}^2$: namely, $T\mathbf{e}_1 = (a,b,c)$ and $T\mathbf{e}_2 = (d,e,f)$ for some known constants $a,b,c,d,e,f \in \mathbb{R}$. Then for any other vector $(x,y) \in \mathbb{R}^2$, we know that $(x,y) = x\mathbf{e}_1 + y\mathbf{e}_2$, so $T(x,y) = T(x\mathbf{e}_1 + y\mathbf{e}_2) = x(T\mathbf{e}_1) + y(T\mathbf{e}_2) = (ax+dy, bx+ey, cx+fy)$. You can easily show that any choice of the constants a,b,c,d,e,f makes the map $T(x,y) = (ax+dy, bx+ey, cx+fy)$ a valid linear map, so the set of linear maps is represented completely by six coefficients.[1]

We can use this fact to find a general formula for linear maps $T : \mathbb{R}^2 \to \mathbb{R}^3$ if we're given their values on a basis of $\mathbb{R}^2$ that's not the standard basis $\{\mathbf{e}_1, \mathbf{e}_2\}$. Suppose, for instance, we knew that $T : \mathbb{R}^2 \to \mathbb{R}^3$ was a linear map such that $T(1,1) = (0,2,-1)$ and $T(-2,-4) = (2,0,2)$, and want to find a formula for T. We can do this by using linearity properties to find the values of $T\mathbf{e}_1$ and $T\mathbf{e}_2$. In this case, $4(1,1) + (-2,-4) = 2\mathbf{e}_1$, so $4T(1,1) + T(-2,-4) = 2T(\mathbf{e}_1) = (2,8,-2)$, so $T\mathbf{e}_1 = (1,4,-1)$. Similarly, you can use $2(1,1) + (-2,-4) = -2\mathbf{e}_2$ to find that $-2T\mathbf{e}_2 = (2,4,0)$, so $T\mathbf{e}_2 = (-1,-2,0)$.

[1] As a preview of matrix representation of linear maps: if we represent element (x,y) and (x,y,z) of $\mathbb{R}^2$ and $\mathbb{R}^3$ as one-column matrices $\begin{bmatrix} x \\ y \end{bmatrix}$ and $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$, then we can represent the matrix T as $\begin{bmatrix} a & d \\ b & e \\ c & f \end{bmatrix}$. Then $T(x,y)$ is the matrix product $\begin{bmatrix} ax+dy \\ bx+ey \\ cx+fy \end{bmatrix} = \begin{bmatrix} a & d \\ b & e \\ c & f \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$, with standard matrix multiplication.

The only linear map $T : \mathbb{R}^2 \to \mathbb{R}^3$ with these values on the standard basis vectors is $T(x, y) = (x - y, 4x - 2y, -x)$.

(Later, when we discuss change-of-basis matrices, we'll discuss algorithms for finding combinations of specified vectors that add up to the standard basis vectors.)

This reasoning generalizes to all maps $T : \mathbb{F}^m \to \mathbb{F}^n$ over spaces with generalized dimension and base field. Each of the n components of the value of $T\mathbf{v}$ is a linear combination, with fixed coefficients chosen freely from $\mathbb{F}$, of the m components of the input vector $\mathbf{v}$. Thus, T is determined by a choice of mn coefficients.

Answers to key questions.

1. We can work out that $T(0, 1, 0) = T(1, 1, 0) - T(1, 0, 0) = (c - a, d - b)$ and similarly $T(0, 0, 1) = \frac{1}{2}(T(0, 1, 2) - T(0, 1, 0)) = \frac{1}{2}(a - c + e, b - d + f)$. So
$$T(x, y, z) = \left(ax + (c - a)y + \frac{1}{2}(a - c + e)z, bx + (d - b)y + \frac{1}{2}(b - d + f)z \right).$$
2. mn coefficients.

2.3 The set of linear maps is a vector space

Key questions.

1. How can you define the sum of two linear maps? What about the product of a linear map by a scalar?
2. What does the notation $\mathrm{Hom}(V, W)$ mean? If V has dimension 5 and W has dimension 6, what is $\dim \mathrm{Hom}(V, W)$?

If V, W are two vector spaces over the same field $\mathbb{F}$, then we can define addition and scalar multiplication on the set of linear maps from V to W, making the set of linear maps itself a vector space. $T_1 + T_2$ is the map that sends every $\mathbf{v} \in V$ to to $T_1\mathbf{v} + T_2\mathbf{v}$, and cT is the map that sends $\mathbf{v}$ to $c(T\mathbf{v})$. This space is often denoted $\mathrm{Hom}(V, W)$ (short for "homomorphism," a term with a more general definition in abstract algebra).

We still have to check that these definitions satisfy the axioms for vector space operations. In particular, addition on $\mathrm{Hom}(V, W)$ must satisfy the abelian group axioms of associativity, commutativity, identity, and inverses; scalar multiplication must distribute with both vector space and field addition as well as satisfy the pseudo-associative law $a(bT) = (ab)T$; and scalar multiplication by 1 leaves a map unchanged. (Review section 1.4 if you need a reminder.) Most of these axioms, though, follow quickly from the fact that the same axioms are also true for W.

1. Associativity of addition holds because $((T_1 + T_2) + T_3)\mathbf{v} = (T_1 + T_2)\mathbf{v} + T_3\mathbf{v}$ by definition of the sum of the maps $(T_1 + T_2)$ and T_3, and you can further expand this into $(T_1\mathbf{v} + T_2\mathbf{v}) + T_3\mathbf{v}$ by definition of the sum of the maps T_1 and T_2. You can similarly expand $(T_1 + (T_2 + T_3))\mathbf{v}$ out into $T_1\mathbf{v} + (T_2\mathbf{v} + T_3\mathbf{v})$, which equals $(T_1\mathbf{v} + T_2\mathbf{v}) + T_3\mathbf{v}$ because W is a vector space and sums of is elements (such as $T_1\mathbf{v}$, $T_2\mathbf{v}$, and $T_3\mathbf{v}$) must be associative.
2. Showing commutativity of addition is similar.

3. The map that takes every vector in V to $\mathbf{0}_W$ is a zero element of $\mathrm{Hom}(V, W)$, and the sum of T and the negative map $\mathbf{v} \mapsto -T\mathbf{v}$ is the zero element, so every map has an additive inverse.

4. To show that scalar multiplication distributes over map addition, note that we can expand $k(T_1 + T_2)\mathbf{v}$ into $k(T_1\mathbf{v} + T_2\mathbf{v})$ by the definition of the sum of linear maps, and then into $k(T_1\mathbf{v}) + k(T_2\mathbf{v})$ by the distributivity of addition. Meanwhile, $(kT_1)\mathbf{v} = k(T_1\mathbf{v})$ and $(kT_2)\mathbf{v} = k(T_2\mathbf{v})$ by the definition of scalar multiples of maps, so $(kT_1 + kT_2)\mathbf{v} = k(T_1\mathbf{v}) + k(T_2\mathbf{v})$. So $k(T_1 + T_2) = kT_1 + kT_2$.

5. Showing that scalar addition distributes over scalar multiplication, i.e. $(k_1 + k_2)T = k_1T + k_2T$, is similar.

6. The pseudo-associativity axiom $(ab)T = a(bT)$ holds because by our definitions, $(ab)T$ is the map from $\mathbf{v}(ab)(T\mathbf{v})$ and $a(bT)$ is a times the map $\mathbf{v} \mapsto b(T\mathbf{v})$ (i.e. $\mathbf{v} \mapsto a(b(T\mathbf{v}))$, and the equality of $(ab)(T\mathbf{v})$ and $a(b(T\mathbf{v}))$ is just the pseudo-associativity axiom on W.

7. The product of any map with the scalar 1 is the map itself, because $(1T)\mathbf{v} = 1(T\mathbf{v})$ by definition of map multiplication, and 1 times any element of W (such as $T_\mathbf{v}$) is the same element of W by the vector space axioms on W.

If $\dim V = m$ and $\dim W = n$, then the dimension of $\mathrm{Hom}(V, W)$ is mn. Here's one possible choice of basis: choose bases $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ of V and $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ of W. Then for each pair of integers $1 \leq i \leq m$ and $1 \leq j \leq n$, put the linear map that sends $\mathbf{v}_i$ to $\mathbf{w}_j$, and sends every other vector $\mathbf{v}_k$ for $k \neq i$ to $\mathbf{0}$ into this basis.

In the case of $\mathrm{Hom}(\mathbb{R}^2, \mathbb{R}^3)$, for example, the six basis maps constructed in this fashion, using the standard bases for $\mathbb{R}^2$ and $\mathbb{R}^3$, are $(x, y) \mapsto (x, 0, 0)$, $(x, y) \mapsto (0, x, 0)$, $(x, y) \mapsto (0, 0, x)$, $(x, y) \mapsto (y, 0, 0)$, $(x, y) \mapsto (0, y, 0)$, and $(x, y) \mapsto (0, 0, y)$. Each of these maps corresponds to setting one of the coefficients a, b, c, d, e, f in the generic formula $T(x, y) = (ax+dy, bx+ey, cx+fy)$ to 1, and the others to 0. It's easy to prove that adding two maps, or multiplying one map by a scalar, is equivalent to adding (or scaling) their defining coefficients.

Answers to key questions.

1. The sum of two maps $T_1 + T_2$ is the map that takes an input $\mathbf{v}$ to $T_1\mathbf{v} + T_2\mathbf{v}$. The multiple kT is the map that takes $\mathbf{v}$ to $k(T\mathbf{v})$.

2. $\mathrm{Hom}(V, W)$ is the vector space of linear maps between two vector spaces V and W. If $\dim V = 5$ and $\dim W = 6$, then $\dim \mathrm{Hom}(V, W) = 5 \times 6 = 30$.

2.4 Kernel and image

Key questions.

1. What is the *kernel* of a linear map $T : V \to W$? What is its *image*? Which one of the kernel and image can we denote $T(V)$, and which can we denote $T^{-1}(\{\mathbf{0}_W\})$?

2. If B is a basis of V and $T \in \mathrm{Hom}(V, W)$, then what set is always a spanning set for $\mathrm{im}\, T$?

3. Define *bijective*, *injective*, and *surjective*.

4. If $T : \mathbb{R}^3 \to \mathbb{R}^4$ is an injective map, what are the possible dimensions of $\ker T$?

5. Suppose $T \in \mathrm{Hom}(\mathbb{R}^3, \mathbb{R}^5)$. Could T be injective? Could it be surjective? Could it be bijective? Give an example for each possible case. Repeat for $T \in \mathrm{Hom}(\mathbb{R}^5, \mathbb{R}^3)$ and $T \in \mathrm{Hom}(\mathbb{R}^3, \mathbb{R}^3)$.

2.4.1 Definitions

Any map $T : V \to W$ defines two vector subspaces, one of V and one of W, whose relationship is a foundational result of linear algebra.

Definition. *Let $T : V \to W$ be a linear map between two vector spaces. The **image** of T, denoted $\mathrm{im}\, T \subseteq W$, is the set of values of T. That is, $\mathbf{w} \in W$ is in $\mathrm{im}\, T$ if and only if there's some $\mathbf{v} \in V$ such that $T\mathbf{v} = \mathbf{w}$. (This is the same definition of* image *as the set-theoretic definition for arbitrary functions between two sets.)*

*The **kernel** of T, denoted $\ker T \subseteq V$, is the set of elements $\mathbf{v} \in V$ such that $T\mathbf{v} = \mathbf{0}_W$ (or, in our more concise set-theoretic notation, the preimage $T^{-1}(\{\mathbf{0}_W\})$).*

It's easy to prove that these spaces satisfy the three subspace axioms of closure under addition, closure under multiplication, and non-emptiness.

Proposition. *The kernel and image of any linear map $T : V \to W$ are subspaces of (respectively) V and W.*

Proof. We'll prove that each set in turn satisfies the three subspace axioms:

- *Image*: Suppose $\mathbf{w} \in \mathrm{im}\, T$, and choose $\mathbf{v} \in V$ such that $T\mathbf{v} = \mathbf{w}$. Then $T(k\mathbf{v}) = k\mathbf{w}$ for any scalar k, so $k\mathbf{w} \in \mathrm{im}\, T$. So $\mathrm{im}\, T$ is closed under multiplication. Similarly, if $\mathbf{w}_1 = T\mathbf{v}_1 \in \mathrm{im}\, T$ and $\mathbf{w}_2 = T\mathbf{v}_2 \in \mathrm{im}\, T$, then $\mathbf{w}_1 + \mathbf{w}_2 = T(\mathbf{v}_1 + \mathbf{v}_2) \in \mathrm{im}\, T$, so $\mathrm{im}\, T$ is closed under addition. Finally, $\mathrm{im}\, T$ cannot be empty, because every function whose domain contains at least one element (and vector spaces must contain $\mathbf{0}$ at least) has to have at least one value.

- *Kernel*: If $T\mathbf{v} = \mathbf{0}$, then $T(k\mathbf{v}) = k(T\mathbf{v}) = k\mathbf{0} = \mathbf{0}$, so $\ker T$ is closed under multiplication. Similarly, if $T\mathbf{v}_1 = T\mathbf{v}_2 = \mathbf{0}$, then $T(\mathbf{v}_1 + \mathbf{v}_2) = T\mathbf{v}_1 + T\mathbf{v}_2 = \mathbf{0} + \mathbf{0} = \mathbf{0}$, so $\ker T$ is closed under addition. Finally, $\ker T$ has to include at least $\mathbf{0}_V$, so it can't be empty.

□

These results still apply for maps whose domains are subspaces of some larger vector space. That is, if U is a subspace of V and $T : V \to W$ is a map, then the image and kernel of the restricted map $T|_U$ are also subspaces: $\mathrm{im}\, T|_U = \{T\mathbf{u} : \mathbf{u} \in U\}$ is a subspace of W, and $\ker T|_U = \{\mathbf{u} \in U : T\mathbf{v} = 0\}$ is a subspace of U (and, therefore, of V). In particular, $\ker T|_U = U \cap \ker T$. The point about images is enough to denote a proposition:

Proposition. *Linear maps send subspaces to subspaces. That is, the image of any subspace of a linear map's range is a subspace of its domain.*

Proof. Just given.

□

The values of $T : V \to W$ on a basis of V provide a spanning set (not necessarily a basis!) for W. If $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis of V, and we define $\mathbf{w}_i \coloneqq T\mathbf{v}_i$ for $1 \leq i \leq n$, then any element of $\operatorname{im} T$ can be written as $T(c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n) = c_1\mathbf{w}_1 + \cdots + c_n\mathbf{w}_n$, which is in $\operatorname{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$. This, incidentally, means that linear maps *cannot increase the dimension of their inputs*—at most, they can keep the dimension the same. A further corollary is that *no surjective linear map can exist from a lower- to a higher-dimensional space*.

2.4.2 Injectivity, surjectivity, isomorphism

A reminder of vocabulary from generic set theory: a function $f : X \to Y$ from a set X to a set Y is:

- *injective* if there are no two elements of the domain x_1, x_2 for which $x_1 \neq x_2$ but $f(x_1) = f(x_2)$.
- *surjective* if its image is all of Y: for every $y \in Y$ there's some $x \in X$ such that $f(x) = y$.
- *bijective* if it's both injective and surjective.

The axioms for a linear map are restrictive enough that knowing the kernel of a linear map is enough to tell you whether it's injective: if a linear map T maps any two elements to the same element, then it also must map multiple elements to $\mathbf{0}$. This is a crucial result whose importance is out of proportion to how easy it is to prove:

Proposition. *A linear map $T : V \to W$ is injective if and only if $\ker T = \{\mathbf{0}_V\}$.*

Proof. If $\ker T$ contains some element $v \neq 0$, then $T\mathbf{v} = T\mathbf{0}_V = \mathbf{0}_W$, so T isn't injective. Conversely, suppose T isn't injective: that is, there are two different elements $\mathbf{v}_1, \mathbf{v}_2 \in V$ such that $T\mathbf{v}_1 = T\mathbf{v}_2$. Then $T(\mathbf{v}_1 - \mathbf{v}_2) = T\mathbf{v}_1 - T\mathbf{v}_2 = \mathbf{0}_W$, so $\mathbf{v}_1 - \mathbf{v}_2$ is a nonzero element of $\ker T$.

□

If a linear map $T : V \to W$ is injective and also surjective, then it establishes what we'll call an *isomorphism*[2] between V and W. That is, it shows that V and W have identical structure: T pairs every element of V with another element with W such that the result of any arithmetic operation on V matches the result of the same operation on the paired elements of W. One consequence: the existence of a surjective map from V to W with kernel $\{\mathbf{0}_V\}$ establishes that V and W have the same dimension. This is an important enough result that we'll prove it explicitly, with two pairs of proposition and corollary:

Proposition. *If $T : V \to W$ is bijective, then V and W have the same dimension.*

[2] *Isomorphism* is one of the words that keeps coming up in higher mathematics in many different contexts: there are many other kinds of structures besides vector spaces that also have some notion of isomorphism.

Proof. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a basis for V. Then $\{T\mathbf{v}_1, \ldots, T\mathbf{v}_n\}$ is a spanning set of $\operatorname{im} T$; and T is surjective, so $\operatorname{im} T = W$. Furthermore, this set must be linearly independent. To see why, suppose to the contrary that there was some nontrivial linear combination $c_1 T\mathbf{v}_1 + \cdots + c_n T\mathbf{v}_n = \mathbf{0}_W$. Then the corresponding linear combination $c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$ must be in $\ker T$, but couldn't have a value of zero (because $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is linearly independent). But T is bijective (and thus injective), so $\ker T$ can't have nonzero elements, a contradiction.

□

Corollary. *There are no injective linear maps from a higher-dimensional space to a lower-dimensional space.*

Proof. If $T : V \to W$ is injective, then it's bijective from V to $\operatorname{im} T$, so $\dim V = \dim \operatorname{im} T$. Furthermore, $\operatorname{im} T$ is contained in W, so $\dim \operatorname{im} T \leq \dim W$.

□

Proposition. *Let $T : V \to W$ be a linear map, and let S be a linearly independent subset of $\operatorname{im} T$. Construct a subset R of V that contains, for each vector $\mathbf{s}_i \in S$, one arbitrary vector $\mathbf{r}_i \in V$ such that $T\mathbf{r}_i = \mathbf{s}_i$.*

Then R is also linearly independent, and the restricted map[3] *$T|_{\operatorname{span} R} : \operatorname{span} R \to \operatorname{span} S$ is a bijection.*

Proof. We need to prove four statements:

1. *R is linearly independent*: if it weren't linearly independent, then there would be some nontrivial linear combination $\mathbf{0}_V = c_1\mathbf{r}_1 + \cdots + c_n\mathbf{r}_n$ whose image under T would be $\mathbf{0}_W = c_1\mathbf{s}_1 + \cdots + c_n\mathbf{s}_n$, because $T\mathbf{0}_V = \mathbf{0}_W$ for any linear map $T : V \to W$. But this contradicts the linear independence of S.

2. *$T|_{\operatorname{span} R}$ is surjective onto $\operatorname{span} S$*: any element $c_1\mathbf{s}_1 + \cdots + c_n\mathbf{s}_n$ of $\operatorname{span} S$ equals $T(c_1\mathbf{r}_1 + \cdots + c_n\mathbf{r}_n)$ and so is in the image of $T|_{\operatorname{span} R}$.

3. *$T|_{\operatorname{span} R}$ is injective*: if $\ker T|_{\operatorname{span} R}$ contained a nonzero element $c_1\mathbf{r}_1 + \cdots + c_n\mathbf{r}_n$, then its image under T would be a nontrivial nontrivial linear combination $\mathbf{0}_W = c_1\mathbf{s}_1 + \cdots + c_n\mathbf{s}_n$ that equaled $\mathbf{0}_W$, a contradiction of the linear independence of S.

4. *$T|_{\operatorname{span} R}$ doesn't have any values outside $\operatorname{span} S$*: if $c_1\mathbf{r}_1 + \cdots + c_n\mathbf{r}_n$ is an arbitrary element of $\operatorname{span} R$, then its image $T(c_1\mathbf{r}_1 + \cdots + c_n\mathbf{r}_n) = c_1\mathbf{s}_1 + \cdots + c_n\mathbf{s}_n$ must be in $\operatorname{span} S$.

□

Corollary. *There are no surjective linear maps from a lower-dimensional space to a higher-dimensional space.*

Proof. If $T : V \to W$ is surjective and S is a basis of W, then the above lemma guarantees the existence of a linearly independent set $R \subset V$ the same size as S. The size of a linearly independent set can be at most the dimension of its enclosing space, so $\dim V \geq |R| = |S| = \dim W$.

□

[3] Reminder: if $f : X \to Y$ is any function from a set X to a set Y, and Z is a subset of X, then the *restricted* function $f|_Z : Z \to Y$ satisfies $f|_Z(x) = f(x)$ whenever $x \in Z$ and is undefined for inputs outside of Z.

Answers to key questions.

1. The *kernel* of T is the subset of V containing all elements that T sends to $\mathbf{0}_W$, or $T^{-1}(\{\mathbf{0}_W\})$. The *image* is the set of all elements of W to which T sends some element of V, or $T(V)$.

2. $\operatorname{im} T$ is always spanned by $T(B) = \{T\mathbf{b} : \mathbf{b} \in B\}$, the set containing the image of every element of B.

3. *Injective* = no two elements of the domain map to the same element. *Surjective* = every element in the codomain is the image of something in the domain. *Bijective* = both injective and surjective.

4. $\dim \ker T$ can only equal zero (as is the case for injective maps on spaces of any, including infinite, dimension).

5. A linear map $T \in \operatorname{Hom}(\mathbb{R}^3, \mathbb{R}^5)$ could be injective and not surjective (for example, $T(a, b, c) = (a, b, c, 0, 0)$), but it can't be surjective (or, therefore, bijective).

   A linear map $T \in \operatorname{Hom}(\mathbb{R}^5, \mathbb{R}^3)$ could be surjective and not injective (for example, $T(a, b, c, d, e) = (a, b, c)$, but it can't be injective (or, therefore, bijective).

   A linear map $T \operatorname{im} \operatorname{Hom}(\mathbb{R}^3, \mathbb{R}^3)$ can be bijective (and thus both injective and surjective); for example, $T(a, b, c) = (a, b, c)$.

2.5 Rank–nullity theorem

Key questions.

1. State the *rank–nullity theorem*. What alternate statement of the rank–nullity theorem is valid in finite-dimensional vector spaces?

2. ($\star\star$) Let $\mathbb{R}^\infty$ be the set of infinite sequences of real numbers with only a finite number of nonzero entries, and let $T : \mathbb{R}^\infty \to \mathbb{R}^3$ be the map $T(a_1, a_2, a_3, a_4, \ldots) = (a_1, a_1 + a_2, 0)$. What are $\operatorname{im} T$ and $\ker T$? Show that $\dim \operatorname{im} T = \operatorname{codim} \ker T$ by finding explicit bases for $\operatorname{im} T$ and $\ker T$, and a set of elements of $\mathbb{R}^\infty$ that you can use to extend the basis of $\ker T$ to a basis of $\mathbb{R}^\infty$.

The result that we established in the last chapter, that any bijective map has the same dimensions of domain and image, is a special case of a vital theorem called the *rank–nullity theorem*, which relates the dimension of a map's image (sometimes called its *rank*) and the dimension of its kernel (sometimes called its *nullity*) to the dimension of its domain.

This theorem is important enough that at least one other textbook calls it the "Fundamental Theorem of Linear Algebra." Learn it well!

Theorem (Rank–nullity). *If $T : V \to W$ is a linear map, then $\dim \operatorname{im} T = \operatorname{codim} \ker T$. (If $\dim V$ is finite and so $\operatorname{codim} \ker T = \dim V - \dim \ker T$, then this means $\dim \operatorname{im} T + \dim \ker T = \dim V$.)*

Proof. Write $m \coloneqq \operatorname{codim} \ker T$. (Remember: the codimension of a subspace is the number of vectors you have to add to a basis for the subspace to get a basis for the larger space.) Let B be a basis of $\ker T$, and let $\mathbf{v}_1, \ldots, \mathbf{v}_m$ be vectors such that $B \cup \{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is a basis of V. Our notation assumes that $\operatorname{codim} \ker T$ is finite, but similar reasoning works when $\operatorname{codim} \ker T = \infty$.

We claim that $S \coloneqq \{T\mathbf{v}_1, \ldots, T\mathbf{v}_m\}$ forms a basis for $\operatorname{im} T$, so $\dim \operatorname{im} T = m$. Proving this requires proving two claims:

1. *S is a spanning set of* $\operatorname{im} T$: We can write any element of V as $\mathbf{u} + c_1\mathbf{v}_1 + \cdots + c_m\mathbf{v}_m$ where $\mathbf{u} \in \ker T$, so we can write every element of $\operatorname{im} T$ as $T(\mathbf{u} + c_1\mathbf{v}_1 + \cdots + c_m\mathbf{v}_m) = c_1T\mathbf{v}_1 + \cdots + c_1T\mathbf{v}_m$. So $\operatorname{im} T$ must be spanned by S.

2. *S is linearly independent*: Suppose to the contrary that there's a nontrivial linear combination $c_1T\mathbf{v}_1 + \cdots + c_mT\mathbf{v}_m = \mathbf{0}_W$. Then $\mathbf{u} \coloneqq c_1\mathbf{v}_1 + \cdots + c_m\mathbf{v}_m$ is an element of $\ker T$. But we can also write $\mathbf{u}$, just like any element of $\ker T$, as a linear combination of elements of B. But this implies that $B \cup S$ is not linearly independent, a contradiction, since we defined S to make $B \cup S$ a basis of V.

□

Remark. One mnemonic for remembering the formula $\dim \operatorname{im} T = \operatorname{codim} \ker T$ is to note that *codimension* and *kernel* alliterate (both begin with a K sound), and *dimension* and *image* have rhyming first syllables, so the pairs that sound similar go on the same side of the equation.

An intuitive way of thinking about the rank–nullity theorem might be this: every linear map T collapses the dimensions of $\ker T$ to a single point and "removes" $\dim \ker T$ dimensions from the output.

A few corollaries:

1. As we already remarked, a linear map $T : V \to W$ is injective if and only if $\ker T = \{\mathbf{0}_V\}$; that is, if $\dim \ker T = 0$. The rank–nullity theorem thus implies that $T : V \to W$ is injective if and only if $\dim \operatorname{im} T = \dim V$.

2. A linear map between two spaces of the same finite dimension n can be bijective, or it can be neither injective nor surjective. There are no other options.

Answers to key questions.

1. The rank–nullity theorem is that $\dim \operatorname{im} T = \operatorname{codim} \ker T$ for any linear map $T : V \to W$. In the case when V is finite, we can write $\dim \operatorname{im} T + \dim \ker T = \dim V$.

2. $\operatorname{im} T = \{(x, y, 0) : z \in \mathbb{R}^3\}$, which has basis $\{\mathbf{e}_1, \mathbf{e}_2\} \subset \mathbb{R}^3$, so $\dim \operatorname{im} T = 2$. And $\ker T$ is the set of sequences $(a_1, a_2, a_3, a_4, \ldots)$ such that $a_1 = a_2 = 0$. This has basis $\{\mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5, \ldots\}$ where $\mathbf{e}_n$ is the sequence with 1 in the nth position and 0 elsewhere. We can extend this to a basis of $\mathbb{R}^\infty$ by adding $\mathbf{e}_1$ and $\mathbf{e}_2$, so $\operatorname{codim} \ker T = 2$.

2.6 Subspace and affine space images and preimages

Key questions.

1. ($\star\star$) Suppose $T : \mathbb{C}^{10} \to \mathbb{C}^7$ is a surjective linear map, and $A \subset \mathbb{C}^{10}$ is an affine subspace of dimension 3. What are the possible dimensions of $T(A)$? What about if A has dimension 9? How do the answers change if we allow T to be non-surjective?

2. ($\star\star$) Suppose $T : \mathbb{R}^4 \to \mathbb{R}^8$ is an injective linear map and $A \subset \mathbb{R}^8$ is an affine subspace of dimension 2. What are the possible dimensions of $T^{-1}(A)$, assuming it's non-empty? What about if A has dimension 6? What if we allow T to be non-injective?

Reminder: $f : X \to Y$ is any function and S is a subset of Y, the *preimage* $f^{-1}(S)$ of S is defined as $\{x \in X : f(x) \in S\}$. If we have two maps $f : X \to Y$ and $g : Y \to Z$, then $(g \circ f)^{-1}(S) = f^{-1}(g^{-1}(S)$. Remember that the kernel of a linear map $T : V \to W$ is a preimage of a one-point set: namely, $\ker T = T^{-1}(\{\mathbf{0}_W\})$.

Similarly, the *image* $f(S)$ of any subset S of X is the set $\{f(x) : x \in S\}$ of images of all elements of S—equivalently, it's the image of the restricted function $f|_S$. And likewise, $(g \circ f)(S) = g(f(S))$.

2.6.1 Images of affine spaces

We've mentioned the geometric intuition that a linear map $T : V \to W$ collapses the dimensions of $\ker T$ but preserves the other dimensions. This intuition also applies to the following result, which you can interpret as saying that T collapses dimensions of an affine space that are parallel to $\ker T$ and preserves other dimensions.

Theorem. *Let U be a subspace of a vector space V, and let A be a coset of U (remember that U is a coset of itself). Let $T : V \to W$ be a linear map. Then $T(A)$ is an affine subspace of W with dimension $\dim U - \dim(U \cap \ker T)$.*

Proof. First, let's consider the case $A = U$. Since $T(U)$ is the image of the restricted linear map $T|_U : U \to V$, it must be a vector subspace (and therefore an affine subspace) of W. The formula $\dim T(U) = \dim U - \dim(U \cap \ker T)$ by applying the rank–nullity theorem to the restricted map $T|_U : U \to W$, because $\ker T|_U = U \cap \ker T$.

Now let's look at the general case where A is a coset of U, and write $X = T(U)$. We just proved that X has dimension $\dim U - \dim(U \cap \ker T)$, so we only need to show that $T(A)$ is a coset of X.

If $\mathbf{v}_1, \mathbf{v}_2$ are any two elements of A, then $\mathbf{v}_1 - \mathbf{v}_2 \in U$ and so $T\mathbf{v}_1 - T\mathbf{v}_2 = T(\mathbf{v}_1 - \mathbf{v}_2) \in X$. That is, the the difference of any two elements in $T(A)$ is in X, so $T(A)$ must be contained in a coset of X. Write this coset as $T\mathbf{a}_0 + X$, where $\mathbf{a}_0$ is some arbitrary element of A. We'll prove that $T(A)$ is actually the entire coset $T\mathbf{a}_0 + X$.

To prove this, we need to show that $T\mathbf{a}_0 + X \subseteq T(A)$: that is, for every element $\mathbf{w} \in T\mathbf{a}_0 + X$, there's some $\mathbf{a} \in A$ such that $T\mathbf{a} = \mathbf{w}$. First, define $\mathbf{x} = \mathbf{a}_0 - \mathbf{w}$. Since $\mathbf{x}$ is the difference of two elements of a coset of X, it must be in X. Since $T(U) = X$ by definition, there's some element $\mathbf{u} \in U$ such that $T\mathbf{u} = \mathbf{x}$. Then $\mathbf{a}_0 + \mathbf{u} \in T(A)$ and $T(\mathbf{a}_0 + \mathbf{u}) = T\mathbf{a}_0 + \mathbf{x} = \mathbf{w}$, so $\mathbf{w}$ is the image of some element of A. This completes the proof.

□

2.6.2 Preimages of points and affine spaces

The following two results on dimensions of preimages of affine spaces are mainly useful to prove a result on the dimension of the kernel of compositions of linear maps. The first presents a result for single-point sets (that is, affine spaces of dimension zero); this is the more important result and will prove to be a key part of our theory of linear systems. The second result generalizes it.

Proposition. *If $T : V \to W$ is a linear map and $\mathbf{w} \in \operatorname{im} T$, then $T^{-1}(\{\mathbf{w}\})$ is a coset of $\ker T$. Furthermore, each different choice of $\mathbf{w}$ produes a different coset $T^{-1}(\{\mathbf{w}\})$.*

Proof. Take any vector $\mathbf{w} \in \operatorname{im} T$, and choose some $\mathbf{v} \in V$ such that $T\mathbf{v} = \mathbf{w}$. Then for any other vector $\mathbf{u} \in V$, $T(\mathbf{v} + \mathbf{u}) = \mathbf{w}$ if and only if $T\mathbf{u} = \mathbf{0}_W$: that is, if $\mathbf{u} \in \ker T$. That is, $T^{-1}(\mathbf{w})$ precisely equals $\mathbf{v} + \ker T$. This correspondence between elements $\mathbf{w} \in \operatorname{im} T$ and cosets $\mathbf{v} + \ker T$ is bijective because if two elements $\mathbf{v}_1, \mathbf{v}_2 \in V$ are in different cosets of $\ker T$, then their difference $\mathbf{v}_1 - \mathbf{v}_2$ is not in $\ker T$ and so $T\mathbf{v}_1 - T\mathbf{v}_2 = T(\mathbf{v}_1 - \mathbf{v}_2) \neq 0_W$; that is, $T\mathbf{v}_1 \neq T\mathbf{v}_2$.

□

A generalization (remember that single-vector sets are cosets of $\{\mathbf{0}\}$):

Theorem. *Let $T : V \to W$ be a linear map, X be a subspace of W, and A be a coset of X (possibly $A = X$) contained in $\operatorname{im} T$. Then $T^{-1}(X)$ is a subspace of V with dimension $\dim X + \dim \ker T$, and $T^{-1}(A)$ is a coset of $T^{-1}(X)$.*

Proof. The theorem conclusion comprises three statements:

1. *$T^{-1}(X)$ is a vector subspace of V.* We need to check that $T^{-1}(X)$ satisfies the three subspace axioms.

   (a) *Closure under addition*: Let $\mathbf{v}_1, \mathbf{v}_2$ be arbitrary elements of $T^{-1}(X)$, and define $\mathbf{x}_1 \coloneqq T\mathbf{v}_1, \mathbf{x}_2 \coloneqq \mathbf{v}_2$. Then $T(\mathbf{v}_1 + \mathbf{v}_2) = T\mathbf{v}_1 + T\mathbf{v}_2 = \mathbf{x}_1 + \mathbf{x}_2$ is the sum of two elements of X, so it is also in X because X, as a vector subspace, is also closed under addition. So $\mathbf{v}_1 + \mathbf{v}_2 \in T^{-1}(X)$.

   (b) *Closure under multiplication*: Let $\mathbf{v}$ be an arbitrary element of $T^{-1}(X)$, define $\mathbf{x} \coloneqq T\mathbf{v}$, and let k be an arbitrary scalar. Then $T(k\mathbf{v}) = k\mathbf{x} \in X$ because X is closed under multiplication, so $k\mathbf{v} \in T^{-1}(X)$).

   (c) *Non-emptiness*: Any subspace $X \subseteq W$ includes $\mathbf{0}_W$, so $T^{-1}(X)$ includes $\mathbf{0}_V$.

2. *$T^{-1}(X)$ has dimension $\dim X + \dim \ker T$.* Let B_X be a basis of X. Define a set B_U that contains one vector $\mathbf{u} \in V$ such that $T\mathbf{u} = \mathbf{x}$ for each vector in $\mathbf{x}$. Define $U = \operatorname{span} B_U$.

   By page 61, B_U is linearly independent, and the restricted map $T|_U$ is a bijection from U to X. In particular, $\ker T|_U$ (which equals $U \cap \ker T$) is just $\{\mathbf{0}_V\}$ and the subspace sum $U + \ker T$ is direct.

   We claim that $U \oplus \ker T = T^{-1}(X)$. We'll prove this set equality by proving inclusion in each direction:

   (a) $U \oplus \ker T \subseteq T^{-1}(X)$: Let $\mathbf{u} \oplus \mathbf{k}$ be an arbitrary element of $U \oplus \ker T$, with $\mathbf{u} \in U$ and $\mathbf{k} \in \ker T$. Then $T(\mathbf{u}+\mathbf{k}) = T\mathbf{u}+T\mathbf{k} = T\mathbf{u}+\mathbf{0}_W = T\mathbf{u}$. This vector must be in X, because $T|_U$ is a bijection from U to X. So $\mathbf{u} + \mathbf{v} \in T^{-1}(X)$.

(b) $T^{-1}(X) \subseteq U \oplus \ker T$: Let $\mathbf{v}$ be an arbitrary element of $\mathbf{v} \in T^{-1}(X)$. Let $\mathbf{u}$ be the (necessarily unique) element of U such that $T\mathbf{u} = T\mathbf{v}$ (as $T|_U$ gives a bijection between U and X, so $\mathbf{u}$ must exist and be unique). Then $T\mathbf{v} - T\mathbf{u} = T(\mathbf{v} - \mathbf{u}) = \mathbf{0}_W$, so $\mathbf{v}$ is the sum of an element $\mathbf{u}$ of U and an element $\mathbf{v} - \mathbf{u}$ of $\ker T$.

3. *$T^{-1}(A)$ is a coset of $T^{-1}(X) = U \oplus \ker T$, with U defined as in statement 2.* Choose some arbitrary base point $\mathbf{a}_0 \in A$, and choose $\mathbf{v}_0 \in V$ such that $T\mathbf{v}_0 = \mathbf{a}_0$ ($\mathbf{v}_0$ must exist because $A \subseteq \operatorname{im} T$). We claim that $T^{-1}(A) = \mathbf{v}_0 + (U \oplus \ker T)$. Again, proving set equality requires proving inclusion two ways.

   (a) $\mathbf{v}_0 + (U \oplus \ker T) \subseteq T^{-1}(A)$: We can write an arbitrary element $\mathbf{v}$ of $\mathbf{v}_0 + (U \oplus \ker T)$ as $\mathbf{v} = \mathbf{v}_0 + \mathbf{u} + \mathbf{k}$, where $\mathbf{u} \in U$ and $\mathbf{k} \in \ker T$. So $T\mathbf{v} = T\mathbf{v}_0 + T\mathbf{u} + T\mathbf{k} = \mathbf{a}_0 + T\mathbf{u}$. We know that $T\mathbf{u} \in X$, so $T\mathbf{v} \in \mathbf{a}_0 + X = A$, so $\mathbf{v} \in T^{-1}(A)$.

   (b) $T^{-1}(A) \subseteq \mathbf{v}_0 + (U \oplus \ker T)$: To show this, it's enough to show that $\mathbf{v}_0 \in T^{-1}(A)$ (which is true because $T\mathbf{v}_0 = \mathbf{a}_0 \in T$ by definition), and that the difference between any two elements of $T^{-1}(A)$ is in $U \oplus \ker T$ (that is, $T^{-1}(A)$ does not contain elements from two different cosets of $U \oplus \ker T$). If we take arbitrary elements $\mathbf{v}_1, \mathbf{v}_2 \in T^{-1}(A)$, then $T(\mathbf{v}_2 - \mathbf{v}_1) = T\mathbf{v}_2 - T\mathbf{v}_1$ is the difference of two elements in A; that is, it is in X. Therefore, $\mathbf{v}_2 - \mathbf{v}_1 \in T^{-1}(X)$.

Note that we've never assumed that any of the spaces or subspaces in the theorem statement has finite dimension.

□

There are two corollary results that will be useful later.

Corollary. *If $T : V \to W$ is linear and A is an affine subspace of W that is not necessarily contained in $\operatorname{im} T$, then $T^{-1}(A)$ is either empty or an affine subspace of V with dimension* at most $\dim A + \dim \ker T$.

Proof. Elements of A outside $\operatorname{im} T$ don't contribute to the preimage $T^{-1}(A)$, so $T^{-1}(A) = T^{-1}(A \cap \operatorname{im} T)$. So if $T^{-1}(A)$ is not empty, then $\dim T^{-1}(A) = \dim \ker T + \dim(A \cap \operatorname{im} T)$, and $\dim(A \cap \operatorname{im} T) \leq \dim A$.

□

Corollary. *Suppose $T_1 : U \to V$ and $T_2 : V \to W$ are two linear maps. Then* $\dim \ker T_1 \leq \dim \ker(T_2 T_1) \leq \dim \ker T_1 + \dim \ker T_2$.

Proof. The lower bound $\dim \ker(T_2 T_1) \geq \dim \ker T_1$ follows from noting that if $\mathbf{u} \in \ker T_1$, then $T_2 T_1 \mathbf{u} = T_2 \mathbf{0}_V = \mathbf{0}_W$, so $\ker T_1 \subseteq \ker(T_2 T_1)$.

For the upper bound $\dim \ker(T_2 T_1) \leq \dim \ker T_1 + \dim \ker T_2$, note that $\ker(T_2 T_1) = T_1^{-1}(\ker T_2)$: that is, if T_1 takes some element of U into $\ker T_2$, then $T_2 T_1$ takes that element to $\mathbf{0}$. Thus, by the previous corollary, $\dim \ker(T_2 T_1) = \dim \ker T_1 + \dim(\ker T_2 \cap \operatorname{im} T_1) \leq \dim \ker T_1 + \dim \ker T_2$.

□

Remark. This result generalizes to compositions of three or more maps: for instance, $\dim \ker(T_3 T_2 T_1) \leq \dim \ker T_3 + \dim \ker(T_2 T_1) \leq \dim \ker T_1 + \dim \ker T_2 + \dim \ker T_3$.

Answers to key questions.

1. The map T reduces the dimension of an affine space A by $\dim(W \cap \ker T)$ where W is the vector subspace of which A is a coset. If T is surjective, then $\ker T = 3$ by rank–nullity and the intersection of two dimension-3 spaces in a space of dimension 6 can be anywhere from 0 to 3. So $0 \leq \dim T(A) \leq 3$.

   If A and thus W have dimension 9, then $W \cap \ker T$ has dimension either 2 or 3, because $\operatorname{codim}(W \cap \ker T) \leq \operatorname{codim} A + \operatorname{codim} \ker T = 1 + 7 = 8$. So $T(A)$ has dimension either 6 or 7.

   If we allow T to be non-surjective, then $\ker T$ could have any dimension ≥ 3 (and thus any codimension ≤ 7). So if $\dim A = \dim W = 3$, then $\dim(W \cap \ker T)$ also has dimension between 0 and 3, and the answer doesn't change. If $\dim A = 9$, then $W \cap \ker T$ could have any dimension between 2 and 9, so $T(A)$ could have any dimension between 0 and 7.

2. If T is injective, then $\dim \operatorname{im} T = 4$ and $\dim \ker T = 0$, and $\dim T^{-1}(A) = \dim(A \cap \operatorname{im} T)$. If $\dim A = 2$, then $A \cap \operatorname{im} T$ could have any dimension between 0 and 2, or be empty, and this is also the set of possible dimensions of $T^{-1}(A)$.

   If $\dim A = 6$, then $A \cap \operatorname{im} T$ could have any dimension between 2 and 4, or be empty. This is also the set of possible dimensions of $T^{-1}(A)$.

   If we allow T to be non-injective, then $\dim \ker T$ and $\dim \operatorname{im} T$ could be anything, provided they add to 4. If $\dim A = 2$, then the maximum possible dimension of $T^{-1}(A)$ is 4, given when $\operatorname{im} T$ is completely contained in A (which requires $\dim \operatorname{im} T < 2$.

   If $\dim A = 6$ and $\dim \operatorname{im} T = n$, $\dim \ker A = 4 - n$ for $0 \leq n \leq 4$, then $\dim(A \cap \operatorname{im} T)$ (provided it's not empty) could be anything from 0 to n for $n = 0, 1, 2$, and anything from $n - 2$ to n for $n = 3, 4$. The corresponding dimensions of $T^{-1}(A)$ are $4 - n$ to 4 for $n = 0, 1, 2$ and 2 to 4 for $n = 3, 4$. Combining these results for all N means that $T^{-1}(A)$ could have any dimension between 2 and 4, the same as if T is required to be injective.

2.7 Map inverses

If $T : V \to W$ is a bijective linear map, then we can define the *inverse* map to T, notated $T^{-1} : W \to V$, which takes every $\mathbf{w} \in W$ to the vector $\mathbf{v}$ such that $T\mathbf{v} = \mathbf{w}$. This is just the ordinary definition of an inverse function from set theory.

It is easy to prove that if T is linear, then T^{-1} is linear as well:

1. If $T\mathbf{v}_1 = \mathbf{w}_1$ and $T\mathbf{v}_2 = \mathbf{w}_2$, then $T(\mathbf{v}_1 + \mathbf{v}_2) = \mathbf{w}_1 + \mathbf{w}_2$. Thus, $T^{-1}(\mathbf{w}_1 + \mathbf{w}_2) = \mathbf{v}_1 + \mathbf{v}_2 = T^{-1}\mathbf{w}_1 + T^{-1}\mathbf{w}_2$, so T preserves addition.

2. If $T\mathbf{v} = \mathbf{w}$, then $T(k\mathbf{v}) = k\mathbf{w}$. Thus, $T^{-1}(k\mathbf{w}) = k\mathbf{v} = kT^{-1}\mathbf{w}$.

There are algorithms for computing the formula for a linear map's inverse given the formula of the forward map; in the following chapter, you'll see them.

Chapter 3

Matrices

Overview. Matrices are a way of representing maps from one finite-dimensional vector space to another. The input vectors must be represented in a particular form called *column vectors*, or single-column matrices (the representation works by choosing a basis for the inputs, then writing the coefficients for an input vector with respect to that basis in a column), and multiplying a matrix representing a map by the column vector. Sections 3.1 and 3.2 spell out these basic notions and present the formula and some key algebraic properties of matrix multiplication. In particular, matrix multiplication is associative but *not* commutative: AB does not generally equal BA.

Section 3.3 presents more details on matrix multiplication. In particular, if M is a matrix, then the map that sends a column vector $\mathbf{c}$ to the product $M\mathbf{c}$ is itself a linear map from one vector space of column vectors to another. Like any map, this map has a kernel (called the *null space* of the matrix M) as well as an image, which turns out to be the space spanned by the matrix's columns when taken as individual column vectors, and is called the *column space*. The rank–nullity theorem applies to these maps just as it does to any map. The space spanned by a matrix's rows is called the *row space*; it's not as immediately important, but a few of its properties will later prove useful. Finally, we present the concept of *inverse matrices* as well as a way to intrepret the matrix product AB as a recombination of the rows of B according to a formula specified by the entries of A. This perspective is crucial to understanding Gauss–Jordan elimination, a method for using matrices to solve linear systems of equations and the main topic of the next chapter.

3.1 Definitions

Key questions.

1. Define the following terms: *matrix*, *column vector*, *row vector*, *row space*, *column space*.

2. What is the notation for the set of matrices with 3 rows, 4 columns, and rational entries? (Remember: the field of rational numbers is denoted $\mathbb{Q}$.) What operations make this into a vector space over $\mathbb{Q}$?

In the last chapter, we considered linear maps in a relatively abstract setting, presenting findings that could apply to maps over any vector spaces. Most books, however, start by presenting linear maps over finite-dimensional vector spaces in a very

concrete representation: a rectangular block of numbers called a matrix. The advantage of putting some more abstract considerations first is that it's easier to see through cumbersome matrix notation and understand what an algorithm is really doing when we have the more concise notation of abstract linear maps, as well as a firmly established sense that matrices are just shorthand for more abstract operations, to guide us.

Some preliminary notation. A *matrix* is a rectangular block of numbers. Matrix dimensions are listed as rows first, then columns second: for instance, a "two-by-three" or 2×3 matrix has two rows and three columns. In a bit, we'll see why this convention is arguably backwards, but we're all stuck with it now.

We'll write $\mathrm{Mat}_{r\times c}(\mathbb{F})$ for the set of matrices with r rows and c columns with entries in the field $\mathbb{F}$. For instance, one element of $\mathrm{Mat}_{2\times 3}(\mathbb{C})$ is

$$\begin{bmatrix} 4+2i & \sqrt{17}+\pi^2 i & -5 \\ 2i & 3-0.17i & 0 \end{bmatrix}.$$

Matrix rows are numbered starting with 1 at the top, and columns are numbered starting from 1 at the left, so (for example) $2i$ in the above matrix is in row 2 and column 1. The elements with equal row and column numbers—that is, the elements in a diagonal that descends from top left toward lower right—are the *diagonal*. (So in the above example, the diagonal entries are $4+2i$ and $3-0.17i$.) A matrix that has all zero entries off the diagonal is called, naturally enough, a *diagonal matrix*.

You can make $\mathrm{Mat}_{r\times c}(\mathbb{F})$ into a vector space in the natural way: matrices with the same size can be added by adding corresponding elements, and scalar multiplication works by multiplying every individual matrix element by the scalar, as in the following formula for 2×2 matrices:

$$k_1 \begin{bmatrix} a_1 & b1 \\ c1 & d1 \end{bmatrix} + k_2 \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix} = \begin{bmatrix} k_1 a_1 + k_2 a_2 & k_1 b_1 + k_2 b_2 \\ k_1 c_1 + k_2 c_2 & k_1 d_1 + k_2 d_2 \end{bmatrix}$$

There's a natural choice of rc basis elements for $\mathrm{Mat}_{r\times c}(\mathbb{F})$: choose one of the rc entries of the matrix to be 1 and set all the others to 0.

Matrices with the same number of rows and columns are called, naturally enough, *square*.

One important subclass of matrices is $\mathrm{Mat}_{n\times 1}(\mathbb{F})$, matrices with n rows but only one column. These are called *column vectors*. In this book, we'll denote these as $\mathrm{Col}_n(\mathbb{F})$ (this isn't standard notation). One element of $\mathrm{Col}_3(\mathbb{R})$, for instance, is $\begin{bmatrix} 1+\sqrt[5]{11} \\ \pi^{1000}\arctan(\sqrt{10}-1) \\ -0.000001 \end{bmatrix}$.

$\mathrm{Col}_n(\mathbb{F})$ is, of course, an n-dimensional vector space over $\mathbb{F}$, and some books actually define $\mathbb{F}^n$ to be what we're calling $\mathrm{Col}_n(\mathbb{F})$. It will make things clearer, though, if you keep these two concepts separate and think of elements of $\mathbb{F}^n$ just as an ordered list of n numbers without any matrix structure. Usually $\mathbb{F}^n$ is the vector space that we want to work over, and $\mathrm{Col}_n(\mathbb{F})$ is used for a matrix representation of elements in $\mathbb{F}^n$. Keeping spaces and their representations conceptually separate will help you avoid a lot of confusions, especially since a lot of matrix theory involves translating between different representations of the same space. (Soon, we'll explain precisely what we mean by "representation.")

Symmetrically, *row vectors* are the set of matrices with 1 row and an arbitrary number of columns. We'll denote the set of row vectors with n entries in a field $\mathbb{F}$ as $\mathrm{Mat}_{1\times n}(\mathbb{F})$ or as $\mathrm{Row}_n(\mathbb{F})$.

One final bit of shorthand: we'll say the element in "position (r, c)" of a matrix to mean the element in row r and column c.

3.2 Matrix multiplication

Key questions.

1. How is matrix multiplication defined? Compute the product $\begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 5 & 4 & 3 \\ 2 & 1 & 0 \end{bmatrix}$.

2. Suppose A is a 3×4 matrix. What matrix dimensions can B have if the matrix product AB is defined? What dimensions can it have if the products AB and BA are both defined?

3. Consider the set of $n \times n$ matrices for some integer $n \geq 2$, other than the zero matrix, together with the operation of multiplication (and ignoring all other matrix operations). Which abelian group axioms does this structure satisfy? Which does it not satisfy?

4. ($\star$) Does $\begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix}$ have an inverse? How about $\begin{bmatrix} 0 & 1 \\ 0 & 3 \end{bmatrix}$? How about $\begin{bmatrix} 0 & 1 & 2 \\ 10 & 11 & 12 \\ 20 & 21 & 22 \end{bmatrix}$?

5. What are the standard basis column vectors? If $\mathbf{e}_i$ is the ith standard basis column vector of $\mathrm{Col}_c(\mathbb{F})$ and M is an $r \times c$ matrix, then what is $M\mathbf{e}_i$?

6. What does it mean that in the matrix product AB, A "acts on B by rows" and B "acts on A by columns"? If $A = \begin{bmatrix} 1 & 3 & -1 \\ -2 & 0 & 4 \end{bmatrix}$, then how are the rows of the matrix product AB related to the rows of B (assuming B has compatible dimensions)? How are the columns of BA related to the columns of B?

3.2.1 Definitions

There's one final operation possible between pairs of matrices: *matrix multiplication*—matrices can be multiplied not just by scalars but also, if their dimensions are right, with each other. In high school mathematics, you may have learned the algorithm for matrix multiplication (though probably not what it's actually useful for). But as a review (or if you haven't learned it) the matrix product AB is only defined between matrices with compatible dimensions. A must have dimensions $m \times n$ and B has dimensions $n \times p$: that is, A has exactly as many columns as B has rows. The entry in position (r, c) of the matrix product AB is the dot product of row r of A and column c of B—that is, calculate the products of entries in corresponding positions in row r of A (starting from the left) and column c of B (starting from the top), then add all these products.

For instance, the product of a 3×4 matrix and a 4×2 matrix is a 3×2 matrix. The general formula for multipling matrices with these dimensions is

$$\begin{bmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & \ell \end{bmatrix} \begin{bmatrix} s & t \\ u & v \\ w & x \\ y & z \end{bmatrix} = \begin{bmatrix} as + bu + cw + dy & at + bv + cx + dz \\ es + fu + gw + hy & et + fv + gx + hz \\ is + ju + kw + \ell y & it + jv + kx + \ell z \end{bmatrix}$$

3.2.2 The identity matrix

The $n \times n$ *identity matrix* (which some books denote I_n) is a matrix with entries of 1 on the diagonal and 0 everywhere else. The 3×3 identity matrix, for instance, is

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

It's called the identity matrix because multiplication of any matrix on the left or right leaves the matrix unchanged: if A is an $m \times n$ matrix, then $I_m A = A I_n = A$. (You may want to convince yourself of this more fully.)

3.2.3 Associativity

It's straightforward (though a bit tedious) to show that matrix multiplication also distributes over matrix addition and is compatible with scalar multiplication: that is, $A(B + C) = AB + AC$ and $(kA)B = A(kB) = k(AB)$ for all scalars k and matrices A, B, C with compatible dimensions. The only remaining property that's mildly tricky to prove is associativity: $A(BC) = (AB)C$ (as with any associative operation, proving this axiom proves further that any series of four or more terms can be parenthesized in any order). The proof is not particularly interesting:

Proposition. *Matrix multiplication is associative.*

Proof. Let A, B, C be matrices with respective dimensions $m \times n$, $n \times p$, and $p \times q$ and entries in the same field. Denote the entry in row i and column j of A as a_{ij} (and similarly for B and C). Let $\alpha_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$ and $\beta_{ij} = \sum_{k=1}^{p} b_{ik} c_{kj}$ designate the entry in row i and column j of AB (which has size $m \times p$) and BC (which has size $n \times q$).

Pick some k, ℓ, where $1 \leq k \leq m$ and $1 \leq \ell \leq q$. Then the entry in row k and column ℓ of $(AB)C$ is

$$\begin{aligned} \sum_{i=1}^{p} \alpha_{ki} c_{i\ell} &= \sum_{i=1}^{p} \left[\left(\sum_{j=1}^{n} a_{kj} b_{ji} \right) c_{i\ell} \right] \\ &= \sum_{i=1}^{p} \sum_{j=1}^{n} a_{kj} b_{ji} c_{i\ell} \end{aligned}$$

while the entry in row k and column ℓ of $A(BC)$ is

$$\begin{aligned} \sum_{j=1}^{n} a_{kj} \beta_{j\ell} &= \sum_{j=1}^{n} \left[a_{kj} \left(\sum_{i=1}^{p} b_{ji} c_{i\ell} \right) \right] \\ &= \sum_{j=1}^{n} \sum_{i=1}^{p} a_{kj} b_{ji} c_{i\ell} \end{aligned}$$

and these sums are evidently equal (it doesn't matter if the sum over i or the sum over j comes first, because we can swap finite sums without a problem).

□

3.2.4 Noncommutativity

Importantly, matrix multiplication is *not* commutative: AB is not necessarily equal to BA. For one, A and B may have dimensions such that BA is not even defined, even if AB is defined. Even if both products are defined—that is, A has dimension $m \times n$ and B has dimension $n \times m$—then AB is an $m \times m$ matrix and BA is an $n \times n$ matrix, and these can't possibly be equal unless $m = n$.

And even if A and B are square matrices of the same dimension, AB is not guaranteed to equal BA—in fact, $AB = BA$ only if A and B satisfy some specific conditions that we'll discuss later. Here's a simple example of noncommutative square matrices:

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 2 & 0 \end{bmatrix}$$
$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -2 \\ 1 & 0 \end{bmatrix}$$

3.3 More on matrix multiplication

3.3.1 Matrices as maps on column vectors

Every $r \times c$ matrix M defines a map from $\mathrm{Col}_c(\mathbb{F})$ to $\mathrm{Col}_r(\mathbb{F})$ that sends the c-entry column vector $\mathbf{c}$ to the matrix product (and r-entry column vector) $M\mathbf{c}$. Let's temporarily denote this map $T_M : \mathrm{Col}_c(\mathbb{F}) \to \mathrm{Col}_r(\mathbb{F})$. (It may seem pedantic to insist that matrices and the multiplication maps that they produce are different objects, but this will help with conceptual clarity.)

This map T_M must be linear, since matrix-by-column-vector multiplication (which is just a subclass of matrix-by-matrix multiplication) distributes over matrix addition and is compatible with scalar-by-matrix multiplication: that is, $M(\mathbf{c}_1+\mathbf{c}_2) = M\mathbf{c}_1+M\mathbf{c}_2$ and $M(k\mathbf{c}) = kM\mathbf{c}$. And conversely, from dimensional considerations, every linear map $\mathrm{Col}_c(\mathbb{F}) \to \mathrm{Col}_r(\mathbb{F})$ must be multiplication by some element of $\mathrm{Mat}_{r\times c}(\mathbb{F})$. (Remember our remarks in Section 2.3 that the space of linear maps from a c-dimensional space such as $\mathrm{Col}_c(\mathbb{F})$ to an r-dimensional space such as $\mathrm{Col}_r(\mathbb{F})$ has dimension cr, the same as the dimension of $\mathrm{Mat}_{r\times c}(\mathbb{F})$ as a vector space.) This is worth making a conspicuous proposition:

Proposition. *Every linear map from* $\mathrm{Col}_c(\mathbb{F})$ *to* $\mathrm{Col}_r(\mathbb{F})$ *is the multiplication map* $\mathbf{c} \mapsto M\mathbf{c}$*, where* M *is some matrix in* $\mathrm{Mat}_{r\times c}(\mathbb{F})$*.*

Proof. Just given.

□

In fact, given any map $T : \mathrm{Col}_c(\mathbb{F}) \to \mathrm{Col}_r(\mathbb{F})$, you can find the matrix $M \in M_{r\times c}(\mathbb{F})$ such that $T = T_M$ if you know the value of T on the standard basis vectors

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

with a single entry of 1 and other entries of zero. For any $r \times c$ matrix M, the matrix product $M\mathbf{e}_i$ (as you can check for yourself) is the ith column of M, so the matrix M such that $T_M = T$ must have $T\mathbf{e}_1, \ldots, T\mathbf{e}_c$ as column vectors. This fact is also worth proposition status:

Proposition. *Column i of a matrix $M \in \mathrm{Mat}_{r\times c}(\mathbb{F})$ is the image of $\mathbf{e}_i \in \mathrm{Col}_c(\mathbb{F})$ under the multiplication map $\mathrm{Col}_c(\mathbb{F}) \to \mathrm{Col}_r(\mathbb{F})$ that M produces.*

Proof. Just given.

□

Since $\{\mathbf{e}_1, \ldots, \mathbf{e}_c\}$ is a spanning set of the domain of any map $T : \mathrm{Col}_c(\mathbb{F}) \to \mathrm{Col}_r(\mathbb{F})$, so $\{T\mathbf{e}_1, \ldots, T\mathbf{e}_c\}$ must be a spanning set of the image of T. This gives us another important result.

Corollary. *The image of the multiplication map created by a matrix M is the span of M's columns.*

3.3.2 Matrix multiplication is map composition

The associativity of matrix multiplication $(AB)C = A(BC)$ also applies when C is a column vector $\mathbf{c}$, in which case $(AB)\mathbf{c} = A(B\mathbf{c})$. This fact implies an extremely important result: *multiplying matrices is equivalent to composing their multiplication maps on column vectors*: that is, $T_{AB} = T_A \circ T_B$. To see this, note that $(AB)\mathbf{c}$ means applying T_{AB} to $\mathbf{c}$, while $A(B\mathbf{c})$ means first applying T_B to $\mathbf{c}$ and then applying T_A to the resulting column vector.

The correspondence between matrix multiplication and composition of maps extends, in fact, from maps on column vectors to maps on any more abstract space that we're using column vectors to represent. We'll talk more about this later.

3.3.3 Null, row, and column spaces

Finally, now that we've seen examples of multiplication of matrices by columns, we can define a few special spaces associated with a matrix:

1. The *nullspace* or *null space* of an $r \times c$ matrix A with entries in a field $\mathbb{F}$, which we'll denote $\mathrm{nullsp}\, A$, is the subset of $\mathrm{Col}_c(\mathbb{F})$ containing all the column vectors $\mathbf{c}$ such that $A\mathbf{c} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$. This definition should remind you strongly of the definition for the kernel of a linear map, and indeed, some books use the same terminology and notation (either "kernel" or "nullspace") for both matrices and linear maps. But in the interest of keeping a clean conceptual line between linear maps and their matrix representations, we'll use "kernel" for linear maps and "nullspace" for matrices.

2. The *columnspace* or *column space* of A, which we'll denote $\mathrm{colsp}\, A$, is the subspace of $\mathrm{Col}_r(\mathbb{F})$ spanned by the columns of A. The maximum possible dimension of this space is $\min(r, c)$, because it's the span of c elements of the r-dimensional

space $\mathrm{Col}_r(\mathbb{F})$. For instance, the column space of the real matrix $\begin{bmatrix} 2 & 1 & 5 \\ 0 & 4 & -2 \end{bmatrix}$ is $\mathrm{span}\left\{\begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \begin{bmatrix} 5 \\ -2 \end{bmatrix}\right\}$, which it's not hard to prove is all of $\mathrm{Col}_2(\mathbb{R})$ (because any subspace of $\mathrm{Col}_2(\mathbb{R})$ can have dimension at most 2, and the column space of this particular matrix can't have dimension 1 because its columns are not all scalar multiples of each other).

As we discussed at the end of section 3.3.1, the column space of A is also the image of the multiplication map that it creates from $\mathrm{Col}_c(\mathbb{F})$ to $\mathrm{Col}_r(\mathbb{F})$: there is a solution $\mathbf{x}$ to the matrix equation $A\mathbf{x} = \mathbf{b}$ if and only if $\mathbf{b} \in \mathrm{colsp}\, A$.

3. The *rowspace* or *row space* of A, which we'll denote $\mathrm{rowsp}\, A$, is the subspace of $\mathrm{Row}_c(\mathbb{F})$ spanned by the rows of A. The rowspace isn't generally as useful as the other two spaces, and it also doesn't have an easy-to-see equivalent in the language of abstract linear maps. But we will occasionally need to discuss it.

4. The *rank* of a matrix is the dimension of its column space. If the rank of an $r \times c$ matrix is $\min(r, c)$ (that is, the largest rank that any $r \times c$ matrix could possibly have), then we say that it has *full rank*.

   It turns out the rank of a matrix is also the dimension of its row space: that is, every matrix's row and column spaces have the same dimension. This is not at all an obvious claim, and we'll have to prove it later in the chapter.

Since multiplication by an $r \times c$ matrix gives a map from $\mathrm{Col}_c(\mathbb{F})$ to $\mathrm{Col}_r(\mathbb{F})$ whose kernel is the matrix's nullspace and whose image is the matrix's column space, we have:

Proposition (Rank–nullity theorem for matrices)**.** *If M is an $r \times c$ matrix, then* $\dim \mathrm{nullsp}\, M + \dim \mathrm{colsp}\, M = c$.

Proof. Immediate corollary of the rank–nullity theorem for maps.

□

3.3.4 Inverse matrices

The multiplication map created by a square $n \times n$ matrix M is an operator on $\mathrm{Col}_n(\mathbb{F})$. The image of this operator is $\mathrm{colsp}\, M$, so the operator is bijective if and only if the columns of M are linearly independent (that is, if $\dim \mathrm{colsp}\, M = n$).

If M has linearly independent columns, then the multiplication operator must have an inverse operator which is also necessarily linear (as the inverse of any linear map is linear: section 2.7) and thus must also be multiplication by some other matrix that we'll denote M^{-1}. The product of M and M^{-1}, in either order, must be the matrix whose multiplication operator on $\mathrm{Col}_n(\mathbb{F})$ is the identity operator (i.e. the composition of any multiplication map and its inverse). This matrix is the identity matrix

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

These observations don't yet give us a practical way to compute M^{-1} given M, but we'll soon see multiple methods for computing matrix inverses.

3.3.5 Matrix multiplication as modification of rows and columns

It's often useful to consider matrix multiplication as a way in which each matrix operates on entire rows or columns of the other, specifying the coefficients of a linear combination that produces each row or column of the result matrix. To be more precise, consider the matrix product AB. Then:

1. Each row of AB is a linear combination of the rows of B, with coefficients specified by the corresponding row of A.

2. Each column of AB is a linear combination of the columns of A, with coefficients specified by the corresponding column of B.

Mathematicians sometimes summarize this by saying that in the matrix product AB, A "acts on B by rows" while B "acts on A by columns." As an example, consider the matrix product

$$AB = \begin{bmatrix} 1 & 3 \\ -4 & 2 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 30 & 50 \end{bmatrix} = \begin{bmatrix} 92 & 153 \\ 52 & 88 \end{bmatrix}.$$

We can look at this product in two ways:

1. The top row of AB, namely $(92, 153)$, equals the top row of B plus three times the bottom row. The bottom row of the product, $(52, 88)$, equals -4 times the top row of B plus 2 times the bottom row. These coefficients are given in the corresponding rows of A.

2. The left column of AB, namely $(92, 52)$, equals 2 times the left column of A plus 30 times the right column. The right column of AB, namely $(153, 88)$, equals 3 times the left column of A plus 50 times the right column. These coefficients are given in the coresponding columns of B.

This perspective is especially useful for *permutation matrices*, which contain all zeros except for an entry of 1 in each row and column—for example,

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Multiplying any matrix A on the left by a permutation matrix P rearranges the columns: if P has an entry of 1 in row i and column j, then row j of A becomes row i of PA. Multiplying on the right, though, rearranges the columns: column i of A becomes column

j of AP. For example, with the permutation matrix above:

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a & b & c & d & e \\ f & g & h & i & j \\ k & \ell & m & n & o \\ p & q & r & s & t \\ u & v & w & x & y \end{bmatrix} = \begin{bmatrix} p & q & r & s & t \\ u & v & w & x & y \\ a & b & c & d & e \\ k & \ell & m & n & o \\ f & g & h & i & j \end{bmatrix}$$

$$\begin{bmatrix} a & b & c & d & e \\ f & g & h & i & j \\ k & \ell & m & n & o \\ p & q & r & s & t \\ u & v & w & x & y \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} c & e & d & a & b \\ h & j & i & f & g \\ m & o & n & k & \ell \\ r & t & s & p & q \\ w & y & x & u & v \end{bmatrix}$$

Left-multiplication by this example permutation matrix permutes the rows of the other matrix in the pattern $1 \to 3 \to 4 \to 1$ and $2 \leftrightarrow 5$, whereas right-multiplication permutes the columns in the inverse pattern: $1 \to 4 \to 3 \to 1$ and $2 \leftrightarrow 5$.

One related crucial point: in a matrix product AB, we can imagine that A works on each column of B as a separate column vector. That is, suppose that

$$B = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_p \\ | & | & \cdots & | \end{bmatrix}$$

where the symbols $\mathbf{b}_j$ represent columns of B interpreted as individual column vectors, then

$$AB = \begin{bmatrix} | & | & \cdots & | \\ A\mathbf{b}_1 & A\mathbf{b}_2 & \cdots & A\mathbf{b}_p \\ | & | & \cdots & | \end{bmatrix}$$

where $A\mathbf{b}_i$ equals the matrix A times the column vector $\mathbf{b}_i$.

3.4 Matrices as vector and map representations

Key questions.

1. Matrix multiplication is equivalent to what operation on the maps that the matrices represent?

2. Let $\mathbf{v}_1 = (1, 0, 0)$, $\mathbf{v}_2 = (1, 1, 0)$, and $\mathbf{v}_3 = (0, 0, 2) \in \mathbb{R}^3$. What is the column vector representation of $(6, 4, 10)$ with respect to the basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$? What about with respect to the basis $\{\mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_1\}$?

3.4.1 Representation of vectors

We've seen that every $r \times c$ matrix M creates a map (via matrix multiplication) from $\mathrm{Col}_c(\mathbb{F})$ to $\mathrm{Col}_r(\mathbb{F})$. We can generalize this even further: matrices and column vectors can model operations on *any* finite-dimensional vector space, as long as we choose a basis for translating elements of the abstract vector space into column vectors. This is what makes matrices useful.

What we mean by "model operations" will need some explanation. Suppose we have two vector spaces V and W over the same field $\mathbb{F}$. Suppose V has dimension c and W has dimension r, and choose a basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_c\}$ of V and $\{\mathbf{w}_1, \ldots, \mathbf{w}_r\}$ of W.

With respect to these bases, any element $\mathbf{v} \in V$ can be represented as a c-entry column vector in $\mathrm{Col}_c(\mathbb{F})$ as follows: first write $\mathbf{v} = k_1\mathbf{v}_1 + \cdots + k_c\mathbf{v}_c$ for some unique choice of coefficients $k_1, \ldots, k_c$. Then the representation of $\mathbf{v}$, *relative to the basis* $\{\mathbf{v}_1, \ldots, \mathbf{v}_c\}$, is $\begin{bmatrix} k_1 \\ \vdots \\ k_c \end{bmatrix}$, the column vector containing the coefficients of $\mathbf{v}$.

It can't be emphasized enough that *a representation for a vector depends on the choice of basis*, including the order of basis elements. To illustrate this, let's represent the element $\mathbf{v}(3, 0, -5) \in \mathbb{R}^3$ as an element of $\mathrm{Col}_3(\mathbb{R})$. One natural choice of basis is, of course, the standard basis $\{\mathbf{e}_1 = (1, 0, 0), \mathbf{e}_2 = (0, 1, 0), \mathbf{e}_3 = (0, 0, 1)\}$. Since $\mathbf{v} = 3\mathbf{e}_1 - 5\mathbf{e}_3$, the column vector representation of $\mathbf{v}$ relative to the standard basis is $\begin{bmatrix} 3 \\ 0 \\ -5 \end{bmatrix}$. But we could reorder the standard basis as, say, $\{\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_1\}$, and then the representation of $\mathbf{v}$ would be $\begin{bmatrix} 0 \\ -5 \\ 3 \end{bmatrix}$.

What if we chose an entirely different basis for $\mathbb{R}^3$? Let's take, for instance, the three vectors $\mathbf{v}_1 = (1, 3, -2)$, $\mathbf{v}_2 = (2, -1, 3)$, and $\mathbf{v}_3 = (1, 5, 4)$. It's not obvious that these are linearly independent (though if you try to plot them in 3D space, you may be able to see that their three endpoints and $(0, 0, 0)$ don't lie in the same plane), so you can take it on faith that they are.[1] How do we represent vectors with respect to this basis? If we have some vector $(a_1, a_2, a_3) \in \mathbb{R}^3$, then the coefficients x, y, z such that $x\mathbf{v}_1 + y\mathbf{v}_2 + z\mathbf{v}_3 = (a_1, a_2, a_3)$ are the solution to the system

$$\begin{aligned} x + 2y + z &= a_1 \\ 2x - y + 3z &= a_2 \\ -2x + 3y + 4z &= a_3 \end{aligned}$$

and the solution to this system for $(a_1, a_2, a_3) = (3, 0, -5)$ is $(x, y, z) = (2, 1, -1)$; that is, $\mathbf{v} = 2\mathbf{v}_1 + \mathbf{v}_2 - \mathbf{v}_3$. So $\mathbf{v}$, relative to the basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$, has the column vector representation $\begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}$.

3.4.2 Representation of maps

We can also, of course, represent any element of our other vector space W as a column vector with r entries relative to a chosen basis of W. This leads us to a natural question: how can we represent linear maps $T : V \to W$ if the representation needs to take two bases into account?

[1] Soon enough, we'll develop and prove the correctness of an algorithm for checking if a set of vectors is linearly independent. Or, if you want to right now, you can use the fact that replacing a vector with a linear combination that includes it leaves the span of a set of vectors unchanged to prove that $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ has the same span as $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, though it may take some trial and error to find a sequence of vector replacements that works.

The answer is to use matrices. Any such map has a matrix representation M, relative to any basis for V and any basis for W, such that if $\mathbf{c} \in \mathrm{Col}_c(\mathbb{F})$ represents some vector $\mathbf{v}$ relative to the chosen basis of V, then the matrix product $M\mathbf{c}$ represents $T\mathbf{v}$ relative to the chosen basis of W. (Again, it should go without saying, the matrix M generally depends on the choice of bases.)

To find the formula for M, first we'll define a total of rc coefficients referred to by two indices a_{ij} with $1 \leq i \leq r$ and $1 \leq j \leq c$. Suppose we have our chosen bases $\{\mathbf{v}_1, \ldots, \mathbf{v}_c\}$ for V and $\{\mathbf{w}_1, \ldots, \mathbf{w}_r\}$ for W. Then define the quantities $a_{11}, a_{21}, \ldots, a_{r1}$ to be the coefficients of the image of $\mathbf{v}_1$ in the basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_r\}$; that is, such that

$$T\mathbf{v}_1 = a_{11}\mathbf{w}_1 + a_{21}\mathbf{w}_2 + \cdots + a_{r1}\mathbf{w}_r.$$

Likewise define $a_{1c}, \ldots a_{rc}$ for every value of c to be the coefficients of $T\mathbf{v}_c$.

Then for an arbitrary element $\mathbf{v} = k_1\mathbf{v}_1 + \cdots + k_c\mathbf{v}_c \in V$, we have the formula

$$\begin{aligned}
T\mathbf{v} =& k_1 T\mathbf{v}_1 + k_2 T\mathbf{v}_2 \cdots + k_c T\mathbf{v}_c \\
=& k_1(a_{11}\mathbf{w}_1 + a_{21}\mathbf{w}_2 \cdots + a_{r1}\mathbf{w}_r) + k_2(a_{12}\mathbf{w}_1 + a_{22}\mathbf{w}_2 \cdots + a_{r2}\mathbf{w}_r) \\
&+ \cdots + k_c(a_{1c}\mathbf{w}_1 + a_{2c}\mathbf{w}_2 + \cdots + a_{rc}\mathbf{w}_r) \\
=& (k_1a_{11} + k_2a_{12} + \cdots + k_ca_{1c})\mathbf{w}_1 + (k_1a_{21} + k_2a_{22} + \cdots + k_ca_{2c})\mathbf{w}_2 \\
&+ \cdots + (k_1a_{r1} + k_2a_{r2} + \cdots + k_ca_{rc})\mathbf{w}_r
\end{aligned}$$

This vector equation is represented by the matrix equation

$$\begin{bmatrix} k_1a_{11}k_2a_{12} + \cdots + k_ca_{1c} \\ k_1a_{21} + k_2a_{22} + \cdots + k_ca_{2c} \\ \vdots \\ k_1a_{r1} + k_2a_{r2} + \cdots + k_ca_{2c} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1c} \\ a_{21} & a_{22} & \cdots & a_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rc} \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_c \end{bmatrix}$$

where the column vector on the left represents $T\mathbf{v}$, the column vector on the right represents $\mathbf{v}$, and the matrix (which we'll call M) represents T. Note that in M:

1. *Row* number i specifies a linear combination whose value is the *coefficient of the ith codomain basis vector* $\mathbf{w}_i$.

2. *Column* number j represents the *image of the jth domain basis vector* $\mathbf{v}_j$.[2]

So any map from a c-dimensional space to an r-dimensional space can be represented by a matrix with dimension $r \times c$. Be careful: when we specify the dimensions of a matrix, we give the number of rows *before* the number of columns, which is backwards from the order that we specify the domain and codomain of a map (i.e. with the domain first and the codomain second). If we had used the convention of representing

[2]Convince yourself that multiplying any $r \times c$ matrix with the column vector $\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathrm{Col}_c(\mathbb{F})$, with a single entry of 1 at the top, gives the unmodified first column of the matrix, and more generally that multiplying by a column vector with only one entry of 1 and all other entries zero gives the corresponding matrix column.

linear maps by multiplying matrices with row vectors on the left rather than column vectors on the right, we wouldn't have this annoyance, but we're stuck with it now.[3]

A few more important observations:

1. Every column in a matrix represents the image of one basis vector of the domain of the underlying linear map, so the image of any element in the domain is represented by the corresponding linear combination of the matrix columns. So *the column space of a matrix represents the image of the underlying linear transformation*, just like how the column space of a matrix actually *is* the image of the multiplication map $\mathrm{Col}_r(\mathbb{F}) \to \mathrm{Col}_c(\mathbb{F})$.

2. Similarly, *the nullspace of a matrix represents the kernel of the underlying linear transformation*. Remember that the nullspace of an $r \times c$ matrix A is the set of column vectors $\mathbf{v} \in \mathrm{Col}_c(\mathbb{F})$ such that $A\mathbf{v} = \mathbf{0} \in \mathrm{Col}_r(\mathbb{F})$, and the zero vector in any r-dimensional vector space can only ever be represented, in any basis, by the zero column vector.

3. If a linear operator $T : V \to V$ is bijective, then there is an inverse operator T^{-1} such that $T^{-1} \circ T$ and $T \circ T^{-1}$ are both the identity map. If T and T^{-1} have matrix representations M and M' relative to some choice of basis for V (with the same basis serving both for domain and for codomain), then since matrix multiplication is equivalent to map composition, the matrix products MM' and $M'M$ must be the identity matrix

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

   We'll call M' the *inverse* matrix of M, and write it as M^{-1}, not actually M'.

   A square matrix with an inverse is called *invertible* or *nonsingular*; a square matrix without an inverse is *singular*. The following conditions on an $n \times n$ matrix M are equivalent by the matrix rank–nullity theorem: (1) M has rank n; (2) the nullspace of M contains only $\mathbf{0} \in \mathrm{Col}_n(\mathbb{F})$; (3) the multiplication operator $\mathbf{v} \mapsto M\mathbf{v}$ on $\mathrm{Col}_n(\mathbb{F})$ is injective; (4) $\mathbf{v} \mapsto M\mathbf{v}$ is surjective; (5) any linear map $T : V \to W$ (where $\dim V = \dim W = n$) representable by M is injective; (6) any such linear map is surjective; (7) M has a matrix inverse.

For an actual example of representing linear transformations by matrices, let $\mathcal{P}_n(\mathbb{R})$ be the $(n+1)$-dimensional vector space over $\mathbb{R}$ of polynomials with real coefficients and degree at most n. Let $T : \mathcal{P}_3(\mathbb{R}) \to \mathcal{P}_2(\mathbb{R})$ be the map $Tp(x) = p'(x-1) - 2xp''(x)$, where p' and p'' mean first and second derivatives. (You may want to verify for yourself that this is a linear map.)

Choose bases $\{x^3, x^2, x, 1\}$ for $\mathcal{P}_3(\mathbb{R})$ and $\{x^2, x, 1\}$ for $\mathcal{P}_2(\mathbb{R})$. Then we can compute the value that T takes on the elements of our chosen basis for $\mathcal{P}_3(\mathbb{R})$:

[3] This is not the only area in which the convention of writing functions to the left of the values of their domain makes things difficult—remember also that in the composition $(A \circ B)(x)$, it's function B that gets applied first.

$$\begin{aligned}
Tx^3 &= \frac{d}{dx}(x-1)^3 - 2x\frac{d^2}{dx^2}x^3 \\
&= 3(x-1)^2 - 12x^2 \\
&= -9x^2 - 6x + 3 \\
Tx^2 &= \frac{d}{dx}(x-1)^2 - 2x\frac{d^2}{dx^2}x^2 \\
&= 2(x-1) - 4x \\
&= -2x - 2 \\
Tx &= \frac{d}{dx}(x-1) - 2x\frac{d^2}{dx^2}x \\
&= 1 \\
T1 &= \frac{d}{dx}1 - 2x\frac{d^2}{dx^2}1 \\
&= 0
\end{aligned}$$

which gives us the matrix representation

$$\begin{bmatrix} -9 & 0 & 0 & 0 \\ -6 & -2 & 0 & 0 \\ 3 & -2 & 1 & 0 \end{bmatrix}.$$

We could use this matrix to compute the value that T takes on various elements of $P_3(X)$. For instance, if $p(x) = 2x^3 - 3x^2 + 4$, then Tp is represented by the matrix computation

$$\begin{bmatrix} -9 & 0 & 0 & 0 \\ -6 & -2 & 0 & 0 \\ 3 & -2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ -3 \\ 0 \\ 4 \end{bmatrix} = \begin{bmatrix} -18 \\ -6 \\ 0 \end{bmatrix}$$

which represents $-18x^2 - 6x$.

Chapter 4

Linear systems

Overview. Linear algebra began as a set of algorithms for solving systems of linear equations—that is, systems in which several linear combinations of unknown variables with known coefficients must equal some known target values. Section 4.1 describes the simple method for recasting linear systems as matrix equations, which turns solving a linear system into finding preimages of a matrix multiplication map from column vectors to column vectors.

The solution method for linear systems rests on *Gauss–Jordan elimination*, an algorithm for transforming a matrix into another matrix that shares many key properties with the original, but is in a simpler form called *reduced row-echelon form* or RREF. The steps of the algorithm are a set of *elementary row operations*, which modify one or two rows at a time based on simple formulas. Row operations preserve several key properties of a matrix, in particular, the null space, column space, and dimension of the row space; as a result, every matrix shares these properties with its RREF, and RREF matrices are much easier to reason about than general matrices. Section 4.2 presents the three types of elementary row operations and proves key properties about them, section 4.3 defines reduced row-echelon form, and section 4.4 outlines the process of Gauss–Jordan elimination.

Section 4.5 completes the the discussion of Gauss–Jordan elimination by proving two key propositions about general matrices: first, every matrix has row and column spaces of equal dimension (something that is manifestly true for matrices in reduced row–echelon form, and that can facilitate computing the dimension of a matrix's column space); and second, although the process of Gauss–Jordan elimination may require making arbitrary choices at some junctures, the end result is always the same: every matrix has exactly one reduced row-echelon form. Finally, section 4.6 presents and proves a method for computing nullspaces of matrices in RREF; combining this method with Gauss–Jordan elimination gives a way to compute the nullspace of any matrix.

The next few sections discuss how to use Gauss–Jordan elimination on a matrix representation of a linear system to find the solution to the system. The simplest case, for linear systems with exactly as many equations as variables and a unique solution, is presented in Section 4.7; in these cases, the solution can simply be read off of one column of the RREF of the system's matrix representation. More complex cases in which systems may have multiple solutions, or no solution at all, are discussed in Section 4.8 and 4.9. The discussion of Gauss–Jordan elimination closes with a presentation of its use for computing matrix inverses in Section 4.10.

The concept of *triangular matrices*—matrices that have all of their nonzero entries on one side of the diagonal from top right to bottom left—is introduced in Section 4.11; systems that can be represented with triangular matrices have an especially efficient method of solution. This method can be further generalized to an alternate method of solving general linear systems that uses *LU decomposition*, a factoring of a matrix into a product of two triangular matrices. This method, and its potential computational advantages over Gauss–Jordan elimination, is discussed in the chapter's final section, 4.12.

4.1 Introduction

The observation that lets us apply matrix theory to linear systems is that we can rewrite a linear system as a matrix equation, with a column vector of unknown variables, another column vector of equation values, and a matrix of coefficients. For example, the system

$$\begin{aligned} x + 2y + 3z &= 0 \\ 3x - y - 4z &= 6 \\ -x + y + z &= -1 \end{aligned}$$

is equivalent to the matrix form

$$\begin{bmatrix} 1 & 2 & 3 \\ 3 & -1 & -4 \\ -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 6 \\ -1 \end{bmatrix}.$$

Any system of e linear equations in v variables can be written as a matrix equation $A\mathbf{x} = \mathbf{b}$, where $\mathbf{x}$ is an unknown column vector of v variables, A is an $e \times v$ matrix that takes column vectors of v variables to column vectors of e equation values, and $\mathbf{b}$ is a column vector of the e specified values of the equation. A system with a solution is *consistent*; otherwise, it's *inconsistent*.

Our method of solving linear systems rests on two observations:

1. The solutions to $A\mathbf{x} = \mathbf{b}$ are the *preimage* of the column vector $\mathbf{b} \in \mathrm{Col}_e(\mathbb{F})$ under the map $\mathrm{Col}_v(\mathbb{F}) \to \mathrm{Col}_e(\mathbb{F})$ induced by multiplication by A, and this map has kernel $\mathrm{nullsp}\, A$. So the set of solutions is either empty or a coset of $\mathrm{nullsp}\, A$. (Remember from section 2.6.2 that for a linear map $T : V \to W$, the preimage $T^{-1}(\{\mathbf{w}\})$ of any single-point subset of W is either a coset of $\ker T$ or empty.)

2. Suppose R is some $v \times v$ matrix that gives a bijective multiplication map from $\mathrm{Col}_v(\mathbb{F})$ to itself: that is, for two vectors $\mathbf{v}_1, \mathbf{v}_1 \in \mathrm{Col}_v(\mathbb{F})$ we have $R\mathbf{v}_1 = R\mathbf{v}_2$ if and only if $\mathbf{v}_1 = \mathbf{v}_2$. Then $A\mathbf{x} = \mathbf{b}$ if and only if $RA\mathbf{x} = R\mathbf{b}$; that is, the two systems $A\mathbf{x} = \mathbf{b}$ and $RA\mathbf{x} = R\mathbf{b}$ have the same solution sets. And a good choice of R could give a matrix RA with a simpler structure than A, whose kernel and preimages are easier to compute.

   Ideally, we can choose R to make RA equal the identity matrix I, so $R = A^{-1}$ and $RA\mathbf{x} = R\mathbf{b}$ becomes $\mathbf{x} = R\mathbf{b}$. This won't always be the case (and, if A isn't square, it will never be the case), but we can at least choose R such that RA is a matrix in a standard form called *reduced row-echelon form*, and systems with matrices in this form are especially easy to solve. In the next few sections, we'll see how to find the best choice of R.

4.2 Elementary row operations

4.2.1 Defined

We'll build R as a product (that is, map composition) of three basic types of matrices. These basic matrices are sometimes called *elementary row operation matrices*, because multiplying a elementary row operation matrix R_e on the left with another matrix A on the right produces a product R_eA in which all but one or two rows are the same as in A, and the changed rows are changed in one of three simple ways. Every elementary row operation can be reversed with another elementary row operation, so the matrices that represent them must be invertible, and any product of the matrix representations of elementary row operations is also invertible (and thus creates a bijective multiplication map on column vectors).

Specifically, there are three types of elementary row operation, which we'll demonstrate on the 3×2 example matrix $M = \begin{bmatrix} 1 & 2 \\ 10 & 30 \\ e & \pi \end{bmatrix}$:

1. *Multiply a row by a nonzero scalar*. Let's denote multiplying row i by λ as $(\mathbf{r}_i \mapsto \lambda\mathbf{r}_i)$. The matrix that represents this is the identity matrix with the 1 in row i and column j replaced by λ. For example,

$$(\mathbf{r}_1 \mapsto \lambda\mathbf{r}_1)M = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 10 & 30 \\ e & \pi \end{bmatrix} = \begin{bmatrix} 2 & 6 \\ 10 & 30 \\ e & \pi \end{bmatrix}$$

   The inverse operation to $\mathbf{r}_i \mapsto \lambda\mathbf{r}_i$ is $\mathbf{r}_i \mapsto \lambda^{-1}\mathbf{r}_i$. Applying these operations in sequence (in either order) leaves the original matrix unchanged, and multiplying the matrix representations of these operations gives the identity matrix.

2. *Swap any two rows*. Denote swapping rows i and j by $(\mathbf{r}_i \leftrightarrow \mathbf{r}_j)$. In matrix form, this is the identity matrix except that the entries at positions (i, j) (that is, row i and column j) and (j, i) are 1, and the diagonal entries at positions (i, i) and (j, j) are zero. For example,

$$(\mathbf{r}_1 \leftrightarrow \mathbf{r}_3)M = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 10 & 30 \\ e & \pi \end{bmatrix} = \begin{bmatrix} e & \pi \\ 10 & 30 \\ 1 & 2 \end{bmatrix}$$

   (You can get any reordering of the rows that you want by putting together enough individual row swaps.)

   The operation $\mathbf{r}_i \leftrightarrow \mathbf{r}_j$ is its own inverse.

3. *Add a multiple of any row i to any other row $j \neq i$*, producing a new row j but leaving row i unchanged. Let's denote adding λ times row i to row j by by $\mathbf{r}_i \mapsto \mathbf{r}_i + \lambda\mathbf{r}_j)$. The matrix that represents this transformation is the identity matrix with an additional entry of λ, insead of 0, in row i and column j. For example,

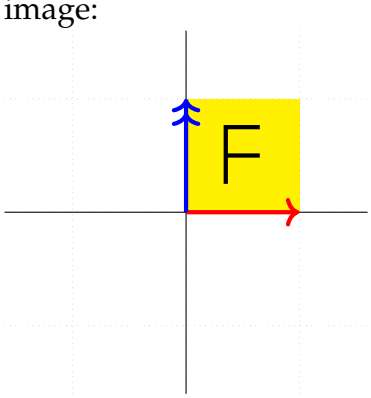
$$(\mathbf{r}_1 \mapsto \mathbf{r}_1 - 2\mathbf{r}_2)M = \begin{bmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 10 & 30 \\ e & \pi \end{bmatrix} = \begin{bmatrix} -19 & -58 \\ 10 & 30 \\ e & \pi \end{bmatrix}$$

The inverse operation to $\mathbf{r}_i \mapsto \mathbf{r}_i + \lambda\mathbf{r}_j$ is $\mathbf{r}_i \mapsto \mathbf{r}_i - \lambda\mathbf{r}_j$.

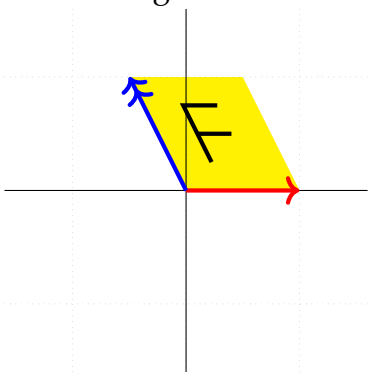
4.2.2 Names

We'll give these operations the respective names *scale*, *swap*, and *shear*.[1] The first two names should have clear motivations; the second may be harder to understand. The reason for the name is the geometric result when the matrix that represents the shear operation is interpreted as a linear operator on $\mathbb{R}^n$ relative to the standard basis: the operation $\mathbf{r}_i \mapsto \mathbf{r}_i + \lambda\mathbf{r}_j$ applied to the identity matrix produces a "shear" matrix that takes the standard basis vector $\mathbf{e}_j$ to a slanted image $\lambda\mathbf{e}_i + \mathbf{e}_j$ while all other vectors remain orthogonal to each other. This operation maps the unit square in $\mathbb{R}^2$ to a parallellogram with one side along an axis, and maps the unit (hyper)cube in $\mathbb{R}^n$ for $n \geq 3$ to a (hyper)prism with a parallelogram base.

For instance, the row shear operation $\mathbf{r}_1 \mapsto \mathbf{r}_1 - \frac{1}{2}\mathbf{r}_2$ on a $2 \times c$ matrix has the matrix representation $R = \begin{bmatrix} 1 & -\frac{1}{2} \\ 0 & 1 \end{bmatrix}$. Relative to the standard basis on $\mathbb{R}^2$, this matrix R represents a map of the form $R(x, y) = (x - \frac{1}{2}y, y)$ that takes $\mathbf{e}_1$ to itself and $\mathbf{e}_2$ to $-\frac{1}{2}\mathbf{e}_1 + \mathbf{e}_2$. Graphically, it takes the coordinate plane that we can represent by this image:



to this image:



(The letter F in the diagrams has no special meaning; it's just there as an aid to visualization.)

Since matrix multiplication is function composition, the matrix equivalent of a sequence of elementary row operations is the product of the matrices for the individual

[1] Strictly speaking, swap operations are unnecessary: the swap $\mathbf{r}_i \leftrightarrow \mathbf{r}_j$ is equivalent to the sequence $\mathbf{r}_j \mapsto \mathbf{r}_j + \mathbf{r}_i$ (shear), $\mathbf{r}_i \mapsto \mathbf{r}_i - \mathbf{r}_j$ (shear), $\mathbf{r}_i \mapsto -\mathbf{r}_i = (-1) \times \mathbf{r}_i$ (scale), $\mathbf{r}_j \mapsto \mathbf{r}_j - \mathbf{r}_i$ (shear), so we can transform any sequence of elementary operations into an equivalent sequence of scales and shears alone. But it's easier to formulate algorithms if we can use all three operations as basic steps.

operations—arranged, of course, from the first operation on the right to the last operation on the left. For instance, the swap $\mathbf{r}_1 \leftrightarrow \mathbf{r}_2$ followed by the shear $\mathbf{r}_3 \mapsto \mathbf{r}_3 + 4\mathbf{r}_2$ has matrix representation

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 4 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 4 & 0 & 1 \end{bmatrix}.$$

4.2.3 Properties preserved by row operations

Three facts about row operations will be crucial for relating properties of matrices produced by row operations to properties of the original matrix:

1. Row operations preserve the row space of a matrix. The spanning set for the row space of a matrix is its rows, and scaling an element of a set of vectors, changing the order of elements in a set, and replacing an element by a linear combination of it with other elements (recall section 1.5.2) do not change the span.

2. They preserve the nullspace of a matrix: since multiplication by an elementary row operation matrix R is bijective, $RA\mathbf{x} = \mathbf{0}$ if and only if $A\mathbf{x} = \mathbf{0}$.

3. They preserve the *dimension* of the column space, though not necessarily the column space itself. Remember that if A represents a linear transformation $T : V \to W$ relative to some bases of V and W, then the nullspace and column space of A represent the kernel and image of T. As $\dim \operatorname{im} T + \dim \ker T = \dim V$ for any linear transformation $T : V \to W$ by the rank–nullity theorem, so any operation on A that keeps the nullspace (or even the dimension of the nullspace) the same has to keep the dimension of the column space the same as well.

As any single row operation preserves these quantities for an arbitrary input matrix, it follows that any *sequence* of row operations must also preserve them.

4.3 Reduced row-echelon form

By applying row operations to a matrix, we can turn it into a matrix with a form called *reduced row-echelon form* (RREF). A matrix's RREF is produced from the original matrix via elementary row operations, so any matrix properties that are preserved by elementary row operations must be equal in the original matrix and its RREF. Three invariant properties in particular will be important to us: the row space, the nullspace, and the dimension of the column space (but not the column space itself).

To be more precise, for every $r \times c$ matrix M, there is a unique RREF matrix E and a product R of elementary row operation matrices[2] such that $E = RM$. It turns out that if M is a square bijective matrix, then E will be the identity matrix and $R = M^{-1}$, so an algorithm to find a matrix's RREF can also find its inverse matrix.

At this point, you may be wondering what reduced row-echelon form actually is. A matrix is in RREF if it satisfies the following criteria:

[2]One important subtlety: E as a matrix is unique, but it could be derived from many different sequences of operations with potentially different matrix representations R—to take a trivial example, if E has two rows of zeros, then we could modify R by adding in a swap of those two rows.

1. If a row is not entirely zeros, then its first (i.e. leftmost) nonzero entry is 1. This entry is called the *pivot*.

2. Each pivot is the only nonzero entry in its column.

3. Each pivot is further right than all pivots in rows above it.

4. Rows of all zeros, if any, are at the bottom of the matrix: there's no row of all zeros above a row with at least one nonzero.

One example of a matrix in reduced row-echelon form, with pivot elements in boxes, is

$$\begin{bmatrix} \boxed{1} & 0 & -3 & 0 & 0 & 5 \\ 0 & \boxed{1} & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \boxed{1} & 0 & -7 \\ 0 & 0 & 0 & 0 & \boxed{1} & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

4.4 Gauss–Jordan elimination

4.4.1 Definitions

The procedure to generate a matrix in reduced row-echelon form is called *Gauss–Jordan elimination* (henceforth, GJE). It's straightforward and completely mechanical.

1. Choose the leftmost column that isn't all zeros, if such a column exists. If the first row has a zero in this column, swap it with any row that doesn't.

2. Divide the first row by its first nonzero entry. This entry is now 1 and will be our first pivot.

3. Add multiples of the first row to every other row so that every non-pivot entry in the first pivot's column is now zero.

4. Repeat for the second row. Find the leftmost column to the right of the first pivot column that has a nonzero entry *somewhere other than the first row*. If the second row has a zero entry in this column, swap it with one that doesn't.

5. Divide the new second row by its first nonzero entry so it also has a pivot 1.

6. Add multiples of the new second row to all other rows, including the first row if necessary, to cancel all elements in the same column as the second pivot.

7. Find the next column to the right that has a nonzero entry in the third row or below. Repeat working down and to the right until there are no columns or nonzero rows left.

4.4.2 Example

Consider the matrix

$$\begin{bmatrix} 2 & 4 & 0 & -1 & 8 & 2 \\ 3 & 6 & 0 & -2 & 5 & 1 \\ 1 & 0 & 4 & 2 & -3 & 0 \\ 0 & -3 & 6 & 1 & -26 & -1 \end{bmatrix}$$

Start with $\mathbf{r}_1 \mapsto \frac{1}{2}\mathbf{r}_1$ to get

$$\begin{bmatrix} 1 & 2 & 0 & -\frac{1}{2} & 4 & 1 \\ 3 & 6 & 0 & -2 & 5 & 1 \\ 1 & 0 & 4 & 2 & -3 & 0 \\ 0 & -3 & 6 & 1 & -26 & -1 \end{bmatrix}$$

Now conduct $\mathbf{r}_2 \mapsto \mathbf{r}_2 - 3\mathbf{r}_1$ and $\mathbf{r}_3 \mapsto \mathbf{r}_3 - \mathbf{r}_1$, setting the non-pivot entries in the first column to zero.

$$\begin{bmatrix} 1 & 2 & 0 & -\frac{1}{2} & 4 & 1 \\ 0 & 0 & 0 & -\frac{1}{2} & -7 & -2 \\ 0 & -2 & 4 & \frac{5}{2} & -7 & -1 \\ 0 & -3 & 6 & 1 & -26 & -1 \end{bmatrix}$$

The second column has nonzero entries below the first row, so it should be the second pivot column. But the second row has an entry of zero in the second column, so we need to swap it with the third or fourth row. Let's choose the third row. (You can check for yourself that choosing the fourth row gives the same final result, and we'll give a general proof soon that it doesn't matter which row you choose.) Swap rows $\mathbf{r}_2 \leftrightarrow \mathbf{r}_3$, then divide the new second row by its leading term -2 (that is, $\mathbf{r}_2 \mapsto -\frac{1}{2}\mathbf{r}_2$) to get

$$\begin{bmatrix} 1 & 2 & 0 & -\frac{1}{2} & 4 & 1 \\ 0 & 1 & -2 & -\frac{5}{4} & \frac{7}{2} & \frac{1}{2} \\ 0 & 0 & 0 & -\frac{1}{2} & -7 & -2 \\ 0 & -3 & 6 & 1 & -26 & -1 \end{bmatrix}$$

and clear all non-pivot entries in the second column with $\mathbf{r}_1 \mapsto \mathbf{r}_1 - 2\mathbf{r}_2$ and $\mathbf{r}_4 \mapsto \mathbf{r}_4 + 3\mathbf{r}_2$:

$$\begin{bmatrix} 1 & 0 & 4 & 2 & -3 & 0 \\ 0 & 1 & -2 & -\frac{5}{4} & \frac{7}{2} & \frac{1}{2} \\ 0 & 0 & 0 & -\frac{1}{2} & -7 & -2 \\ 0 & 0 & 0 & -\frac{11}{4} & -\frac{31}{2} & \frac{1}{2} \end{bmatrix}$$

The third column has no nonzero entries below row 2, so we'll have to leave it without a pivot and make the fourth column into the third pivot. To get a leading 1 in the first row, carry out $\mathbf{r}_3 \mapsto -2\mathbf{r}_3$:

$$\begin{bmatrix} 1 & 0 & 4 & 2 & -3 & 0 \\ 0 & 1 & -2 & -\frac{5}{4} & \frac{7}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 14 & 4 \\ 0 & 0 & 0 & -\frac{11}{4} & -\frac{31}{2} & \frac{1}{2} \end{bmatrix}$$

To clear the non-pivot entries in the fourth column, we can do $\mathbf{r}_1 \mapsto \mathbf{r}_1 - 2\mathbf{r}_3$, $\mathbf{r}_2 \mapsto \mathbf{r}_2 + \frac{5}{4}\mathbf{r}_3$, and $\mathbf{r}_4 \mapsto \mathbf{r}_4 + \frac{11}{4}\mathbf{r}_3$, yielding:

$$\begin{bmatrix} 1 & 0 & 4 & 0 & -31 & -8 \\ 0 & 1 & -2 & 0 & 21 & \frac{11}{2} \\ 0 & 0 & 0 & 1 & 14 & 4 \\ 0 & 0 & 0 & 0 & 23 & \frac{23}{2} \end{bmatrix}$$

Finally, since the fifth column has a nonzero entry in the fourth row, we can make it into a fourth pivot column. The procedure here is simple: divide the fourth row by 23 to make it a pivot entry:

$$\begin{bmatrix} 1 & 0 & 4 & 0 & -31 & -8 \\ 0 & 1 & -2 & 0 & 21 & \frac{11}{2} \\ 0 & 0 & 0 & 1 & 14 & 4 \\ 0 & 0 & 0 & 0 & 1 & \frac{1}{2} \end{bmatrix}$$

and then eliminate the other entries in the fifth column with $\mathbf{r}_1 \mapsto \mathbf{r}_1 + 31\mathbf{r}_4$, $\mathbf{r}_2 \mapsto \mathbf{r}_2 - 21\mathbf{r}_4$, and $\mathbf{r}_3 \mapsto \mathbf{r}_3 - 14\mathbf{r}_4$. This gives the RREF:

$$\begin{bmatrix} 1 & 0 & 4 & 0 & 0 & \frac{15}{2} \\ 0 & 1 & -2 & 0 & 0 & -5 \\ 0 & 0 & 0 & 1 & 0 & -3 \\ 0 & 0 & 0 & 0 & 1 & \frac{1}{2} \end{bmatrix}$$

4.5 More on Gauss–Jordan elimination

4.5.1 Equality of row and column space dimensions

First, though, we can conclude a lot about properties of general matrices just from the fact that elementary row operations can turn any matrix into RREF, even without proving RREF uniqueness. Some notation: write $\mathbf{e}_1, \ldots, \mathbf{e}_c$ for the standard basis vectors of $\mathrm{Col}_c(\mathbb{F})$, where the dimension c will usually be clear from context. (For instance, if $c = 2$, then $\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.) Write $\mathbf{e}_i^T$ for the ith standard basis vector of $\mathrm{Row}_r(\mathbb{F})$. (The T stands for *transpose*; we'll discuss the more general concept of matrix transposes later.)

Two observations that are almost trivial in themselves give us an important theorem on matrix subspaces:

1. *The dimension of the column space of any matrix in RREF equals the number of pivots (equivalently, the number of nonzero rows).* If a matrix E in RREF has p pivots, then $\mathbf{e}_1, \ldots, \mathbf{e}_p$ all as columns of E, so $\mathrm{colsp}\, E$ includes at least $\mathrm{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_p\}$. But no column of E has a nonzero entry below row p, so $\mathrm{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_p\}$ is the entirety of $\mathrm{colsp}\, E$.

2. *The nonzero rows of any matrix in RREF are linearly independent.* If $\mathbf{r}_1, \ldots, \mathbf{r}_p$ are the nonzero rows of E, then any linear combination $k_1\mathbf{r}_1 + \cdots + k_p\mathbf{r}_p$ has to have an entry k_i in the column with row i's pivot, because $\mathbf{r}_i$ has 1 in that column and every other row has zero. So $k_1\mathbf{r}_1 + \cdots + k_p\mathbf{r}_p = \begin{bmatrix} 0 & 0 & \cdots & 0 \end{bmatrix}$ if and only if $k_1 = \cdots = k_p = 0$.

A corollary that's important enough to label a theorem:

Theorem. *Every matrix has a row and column space of the same dimension.*

Proof. Every matrix in RREF has row and column spaces of the same dimension (namely, the number of pivots), and every matrix can be put into RREF with row operations that preserve the row space and the dimension of the column space.

□

4.5.2 RREF existence and uniqueness

GJE gives one possible RREF of a matrix, but you may be wondering if there's more than one possibility, because our choice of steps wasn't uniquely determined. In our example, after all, we chose to swap row 2 with row 3 instead of row 4, and it's not obvious that this choice didn't matter for the end result.

But it turns out that GJE may have multiple possible sequences of steps for a given input matrix, but it only has one possible end result: every matrix has one and only one RREF. Let's prove this. The proof is a bit complicated and not that important to memorize, but you may find it interesting.

Proposition. *Every matrix can be transformed into exactly one matrix in RREF via elementary row operations.*

Proof. GJE gives us one RREF matrix; we want to prove that there can't be two.

If an $r \times c$ matrix M can be transformed into two other matrices E and E' via elementary row operations, then reversing the steps from M to E and then following the steps from M to E' gives a sequence of elementary row operations that transforms E into E'. Let R be the $r \times r$ matrix representation of this sequence, so $RE = E'$ and $R^{-1}E' = E$ (remember that R must be invertible).

We claim that if R and E are any $r \times r$ and $r \times c$ matrices such that R is invertible and both E and $E' \coloneqq RE$ are in RREF, then $E = E'$. Write $\mathbf{c}_i$ and $\mathbf{c}'_i$ for the ith columns of E and E', and remember that $\mathbf{c}'_i = R\mathbf{c}_i$ (see page 77).

Since R has trivial nullspace, multiplication by R is a bijection from $\mathrm{Col}_r(\mathbb{F})$ to itself, so the the image of any subspace of $\mathrm{Col}_r(\mathbb{F})$ under multiplication by R must be another subspace of $\mathrm{Col}_r(\mathbb{F})$ with the same dimension. As multiplication by R sends each $\mathbf{c}_i$ to the corresponding $\mathbf{c}'_i$, it also sends the subspace $\mathrm{span}\{\mathbf{c}_1, \ldots, \mathbf{c}_i\}$ to $\mathrm{span}\{\mathbf{c}'_1, \ldots, \mathbf{c}'_i\}$, so these two subspaces must have equal dimension for every integer $1 \leq i \leq c$.

In a matrix in RREF, the first entry in each row must be a pivot, so a non-pivot column can only have a nonzero entry in row j if $\mathbf{e}_j$ was a pivot column somewhere to its left. This means that every non-pivot column can be written as a linear combination of the pivot columns to its left, so $\dim \mathrm{span}\{\mathbf{c}_1, \ldots, \mathbf{c}_i\}$ is the number of pivot columns in E in positions 1 through i, and likewise $\dim \mathrm{span}\{\mathbf{c}'_1, \ldots, \mathbf{c}'_i\}$ counts the pivot columns in positions 1 through i of E'. These quantities are equal for every integer i, so E and E' must have pivot columns in the same positions.

The pivot columns in corresponding positions of E and E', furthermore, must be equal: in any $r \times c$ matrix in RREF, the first pivot column is always the first standard basis vector $\mathbf{e}_1 \in \mathrm{Col}_r(\mathbb{F})$, the second pivot column is always $\mathbf{e}_2$, and so forth. If E has rank k, then the column basis vectors $\mathbf{e}_1, \ldots, \mathbf{e}_k$ all occur at corresponding positions in E and E', so $R\mathbf{e}_i = \mathbf{e}_i$ for all integers $1 \leq i \leq k$ (because multiplication by R maps columns in E to corresponding columns in E'). This means that $R\mathbf{v} = \mathbf{v}$ for any column vector $\mathbf{v} \in \mathrm{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$. And every column of E must be in $\mathrm{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$, because rows 1 to k are pivot rows and and the other rows are all zeros. So the non-pivot columns of E and E' must also be identical, and $E = E'$.

□

So RREF is unique, and any choice of row exchanges, as long as it avoids division by zero, gives the same result. Computer implementations of GJE use row exchanges

to avoid not only zero pivots but also very small pivots, because division by small numbers can introduce large rounding errors.

Two final notes:

1. Since row operations preserve row space, one way to find a basis for a subspace $W \subset \mathbb{F}^n$, given a spanning set $\{\mathbf{w}_1, \ldots, \mathbf{w}_r\}$, is to write $\mathbf{w}_1, \ldots, \mathbf{w}_r$ as rows of an $r \times n$ matrix, and then carry out row reduction. The nonzero rows of the resulting matrix provide a basis for W.

2. The only $n \times n$ RREF matrix with rank n is the identity matrix. So if A is a square invertible matrix, then its inverse A^{-1} is the matrix representation of the sequence of elementary row operations in GJE that reduces A to I. This also means that since every invertible matrix M is itself the inverse of another matrix (namely M^{-1}), every invertible matrix M can be decomposed into a product of elementary row operation matrices.

4.6 Nullspaces of RREF matrices

Every matrix has the same nullspace as its RREF, so if we can find the nullspace of a matrix in RREF, then GJE lets us find the nullspace of any matrix. It turns out that finding the nullspace of a matrix in RREF is simple with the following procedure:

1. Add rows of zeros between the pivot rows and delete rows of zeros from the bottom of the matrix so that all pivots on the diagonal from top left to bottom right and all non-pivot diagonal entries are zero. (Adding rows of zero to a matrix M means adding entries of zero to every matrix product $M\mathbf{c}$ of M with a column vector $\mathbf{c}$, so it doesn't change $\operatorname{nullsp} M$.)

2. Change all entries of zero on the diagonal to -1.

The non-pivot columns in the result are a basis for the nullspace of the original matrix (that is, the matrix before we changed the diagonal entries).

Let's illustate this with our previous example:

$$\operatorname{rref}\begin{bmatrix} 2 & 4 & 0 & -1 & 8 & 2 \\ 3 & 6 & 0 & -2 & 5 & 1 \\ 1 & 0 & 4 & 2 & -3 & 0 \\ 0 & -3 & 6 & 1 & -26 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 4 & 0 & 0 & \frac{15}{2} \\ 0 & 1 & -2 & 0 & 0 & -5 \\ 0 & 0 & 0 & 1 & 0 & -3 \\ 0 & 0 & 0 & 0 & 1 & \frac{1}{2} \end{bmatrix}.$$

After inserting zero rows to produce a square matrix with pivots on the diagonal, and changing any non-pivot entries on the diagonal to -1, this matrix becomes

$$\begin{bmatrix} 1 & 0 & 4 & 0 & 0 & \frac{15}{2} \\ 0 & 1 & -2 & 0 & 0 & -5 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -3 \\ 0 & 0 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}$$

so a basis for the nullspace is $\left\{ \begin{bmatrix} 4 \\ -2 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{15}{2} \\ -5 \\ 0 \\ -3 \\ \frac{1}{2} \\ -1 \end{bmatrix} \right\}$. Here's a proof of correctness:

Proposition. *This construction gives a basis for the nullspace of a matrix in RREF.*

Proof. Let M be the square matrix created from the original RREF matrix by adding or deleting rows of zero (which, as we just mentioned, does not change its nullspace), but without changing entries of zero on the diagonal to -1. Let n be the number of rows or columns of M. Let m_{ij} be the element in row i and column j of M, let $\mathbf{r}_i \in \mathrm{Row}_n(\mathbb{F})$ be the ith row of M, and let $\mathbf{c}_j \in \mathrm{Col}_n(\mathbb{F})$ be the jth column of M. Let B be the set of putative nullspace basis vectors that we defined above (i.e. the set of non-pivot columns of M with diagonal entries changed to -1), and write $|B|$ for the size of B. Call an integer $k \in \{1, \ldots, n\}$ a "pivot index" if $\mathbf{c}_k$ is a pivot column of M (equivalently, if $\mathbf{r}_k$ is an original row of the matrix rather than an added row of zeros), and a "non-pivot index" otherwise. Every element of B has the form $\mathbf{c}_k - \mathbf{e}_k$, where $\mathbf{e}_k$ is the kth standard basis vector of $\mathrm{Col}_n(\mathbb{F})$ and k is a non-pivot index. Write $\mathbf{b}_k \coloneqq \mathbf{c}_k - \mathbf{e}_k$.

We'll prove three claims that together imply that B is a basis for $\mathrm{nullsp}\, M$:

1. $\dim \mathrm{nullsp}\, M = |B|$.
2. B is linearly independent.
3. Every element of B is in $\mathrm{nullsp}\, M$.

(Why are these claims sufficient? Claims 2 and 3 mean that B spans a $|B|$-dimensional space contained in $\mathrm{nullsp}\, M$, so if $\mathrm{nullsp}\, M$ has dimension $|B|$ by claim 1, it must actually equal $\mathrm{span}\, B$.)

Claim 1 follows from the rank–nullity theorem for matrices (page 75) and the linear independence of an RREF matrix's pivot columns: the $n - |B|$ pivot columns of M are a basis for the image of the map $\mathrm{Col}_n(\mathbb{F}) \to \mathrm{Col}_n(\mathbb{F})$ produced by multiplication by M, so $\dim \mathrm{nullsp}\, M = n - \dim \mathrm{colsp}\, M = n - (n - |B|) = |B|$.

To prove claim 2, note that for every non-pivot index k, $\mathbf{b}_k$ is the only element of B with a nonzero entry (namely -1) in row k, because row k of M is one of the added rows of (originally) zeros. So any linear combination from B must have a nonzero kth entry whenever $\mathbf{b}_k$ has a nonzero coefficient, because no other element of B could cancel it. So the only linear combination from B with all entries zero is the trivial linear combination.

To prove claim 3, let k be a non-pivot index. Then $M\mathbf{b}_k = M(\mathbf{c}_k - \mathbf{e}_k)$ is a column vector whose rth entry is

$$\mathbf{r}_r \mathbf{c}_k - \mathbf{r}_r \mathbf{e}_k = \left(\sum_{i=1}^{n} m_{ri} m_{ik} \right) - m_{rk}.$$

We want to prove that this expression is zero for every row index r. (Remember that the product of a $1 \times n$ row vector with an $n \times 1$ column vector is a 1×1 matrix, which we can consider equivalent to a scalar.) If r is a non-pivot index, then $\mathbf{r}_r$ is all zeros and $\mathbf{r}_r(\mathbf{c}_k - \mathbf{e}_k) = 0$, so we just have to consider the case when r is a pivot index.

If i is a pivot index other than r, then $m_{ri} = 0$, because the only nonzero entry in a pivot column is the pivot itself. And if i is a non-pivot index, then $m_{ik} = 0$, because non-pivot rows of M are all zero. So the only nonzero term in the sum $\sum_{i=1}^{n} m_{ri} m_{ik}$ is $m_{rr} m_{rk}$ for $i = r$. And since r is a pivot index, so $m_{rr} = 1$. So $\mathbf{r}_r(\mathbf{c}_k - \mathbf{e}_k) = m_{rk} - m_{rk} = 0$ when r is a pivot index as well.

□

4.7 Solving systems with Gauss–Jordan elimination

We've mentioned that if we have a linear system $A\mathbf{x} = \mathbf{b}$ with unknown variables $\mathbf{x}$, and R is any bijective square matrix with the same number of rows as A, then $A\mathbf{x} = \mathbf{b}$ if

and only if $RA\mathbf{x} = R\mathbf{b}$. In particular, R could represent the sequence of row operations that transforms A to its RREF. We can compute $R\mathbf{b}$ without explicitly computing R if, whenever we apply a row operation to A, we also apply it to $\mathbf{b}$. We can do this with standard GJE on a special *augmented matrix*, which has the matrix of coefficients A on the left and an extra column with the equation values $\mathbf{b}$ on the right.

First, let's make the connection between GJE and linear systems more explicit. Every row operation in GJE corresponds to a step in solving a corresponding linear system. Consider, for example, the system

$$\begin{aligned} y - 4z &= 2 \\ 2x - 4y + 2z &= 8 \\ 2x + 3y - z &= -3. \end{aligned}$$

You might solve this with the following steps:

1. Swap the first and second equations:

$$\begin{aligned} 2x - 4y + 2z &= 8 \\ y - 4z &= 2 \\ 2x + 3y - z &= -3 \end{aligned}$$

2. Divide the first equation by 2:

$$\begin{aligned} x - 2y + z &= 4 \\ y - 4z &= 2 \\ 2x + 3y - z &= -3 \end{aligned}$$

3. Subtract twice the first equation from the third, to eliminate x:

$$\begin{aligned} x - 2y + z &= 4 \\ y - 4z &= 2 \\ 7y - 3z &= -11 \end{aligned}$$

4. Add twice the second equation to the first, to eliminate y:

$$\begin{aligned} x - 7z &= 10 \\ y - 4z &= 2 \\ 7y - 3z &= -11 \end{aligned}$$

5. Subtract seven times the second equation from the third, to eliminate y:

$$\begin{aligned} x - 7z &= 8 \\ y - 4z &= 2 \\ 25z &= -25 \end{aligned}$$

6. Divide the last equation by 25:

$$\begin{aligned} x - 7z &= 8 \\ y - 4z &= 2 \\ z &= -1 \end{aligned}$$

7. Add seven times the third equation to the first:

$$\begin{aligned} x &= 1 \\ y - 4z &= 2 \\ z &= -1 \end{aligned}$$

8. Add four times the third equation to the second.

$$\begin{aligned} x &= 1 \\ y &= -2 \\ z &= -1 \end{aligned}$$

Each of these steps is equivalent to a row operation on an augmented matrix of the system coefficients and an extra column of equation values. We write augmented matrices with a line separating the original matrix from the extra column, like this:

$$\left[\begin{array}{ccc|c} 0 & 1 & -4 & 2 \\ 2 & -4 & 2 & 8 \\ 2 & 3 & -1 & -3 \end{array}\right]$$

Conducting the equivalent row operations on the augmented matrix looks like this:

$$\left[\begin{array}{ccc|c} 0 & 1 & -4 & 2 \\ 2 & -4 & 2 & 8 \\ 2 & 3 & -1 & -3 \end{array}\right] \xrightarrow{\mathbf{r}_1 \leftrightarrow \mathbf{r}_2} \left[\begin{array}{ccc|c} 2 & -4 & 2 & 8 \\ 0 & 1 & -4 & 2 \\ 2 & 3 & -1 & -3 \end{array}\right] \xrightarrow{\mathbf{r}_1 \mapsto \mathbf{r}_1/2} \left[\begin{array}{ccc|c} 1 & -2 & 1 & 4 \\ 0 & 1 & -4 & 2 \\ 2 & 3 & -1 & -3 \end{array}\right]$$

$$\xrightarrow{\mathbf{r}_3 \mapsto \mathbf{r}_3 - 2\mathbf{r}_1} \left[\begin{array}{ccc|c} 1 & -2 & 1 & 4 \\ 0 & 1 & -4 & 2 \\ 0 & 7 & -3 & -11 \end{array}\right] \xrightarrow{\mathbf{r}_1 \mapsto \mathbf{r}_1 + 2\mathbf{r}_2} \left[\begin{array}{ccc|c} 1 & 0 & -7 & 8 \\ 0 & 1 & -4 & 2 \\ 0 & 7 & -3 & -11 \end{array}\right] \xrightarrow{\mathbf{r}_3 \mapsto \mathbf{r}_3 - 7\mathbf{r}_2} \left[\begin{array}{ccc|c} 1 & 0 & -7 & 8 \\ 0 & 1 & -4 & 2 \\ 0 & 0 & 25 & -25 \end{array}\right]$$

$$\xrightarrow{\mathbf{r}_3 \mapsto \mathbf{r}_3/25} \left[\begin{array}{ccc|c} 1 & 0 & -7 & 8 \\ 0 & 1 & -4 & 2 \\ 0 & 0 & 1 & -1 \end{array}\right] \xrightarrow{\mathbf{r}_1 \mapsto \mathbf{r}_1 + 7\mathbf{r}_3} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & -4 & 2 \\ 0 & 0 & 1 & -1 \end{array}\right] \xrightarrow{\mathbf{r}_2 \mapsto \mathbf{r}_2 + 4\mathbf{r}_3} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -2 \\ 0 & 0 & 1 & -1 \end{array}\right].$$

The augmented matrix that we started with contained A and $\mathbf{b}$; the matrix that we ended with had RA and $R\mathbf{b}$. We've shown that $\begin{bmatrix} 0 & 1 & -4 \\ 2 & -4 & 2 \\ 2 & 3 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ 8 \\ -3 \end{bmatrix}$ if and only if $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \\ -1 \end{bmatrix}$.

This example system is, in many ways, the simplest possible case: the matrix of coefficients is square, with the same number of variables as coefficients. After Gauss–Jordan elimination on the augmented matrix, the coefficient matrix became just the identity, and the solutions to the system could be read off from the column of equation values. Systems of equations, however, aren't always this nice. We'll see examples thereof later.

4.8 Underdetermined systems

For a system with n equations in n variables, GJE yields the identity matrix on the left and a column vector of solutions on the right, unless the left-hand side of one equation is a redundant linear combination of the left-hand sides of the others and the matrix of coefficients doesn't have full rank (a complication that we'll address in the next section). But for systems with more variables than equations (called *underdetermined* systems), the setup is a bit more complicated. Remember that the solution set to $A\mathbf{x} = \mathbf{b}$ is either empty or a coset of $\text{nullsp}\, A$. If A is a $e \times v$ matrix giving a linear map from a v-dimensional space of variable values to a e-dimensional set of equation values, and $v \geq e$, then multiplication by A cannot be injective from dimensional considerations alone, and A's nullspace has to contain more than just the zero vector.

Let's take this system of two equations in four variables as an example:

$$\begin{aligned} 4w + 2x - 3y - 4z &= 20 \\ 2w + x - 3y + z &= -5 \end{aligned}$$

Gauss-Jordan elimination on the augmented matrix runs like this:

$$\begin{bmatrix} 4 & 2 & -3 & -4 & 20 \\ 2 & 1 & -3 & 1 & -5 \end{bmatrix} \xrightarrow{\mathbf{r}_1 \mapsto \mathbf{r}_1/4} \begin{bmatrix} 1 & \frac{1}{2} & -\frac{3}{4} & -1 & 5 \\ 2 & 1 & -3 & 1 & -5 \end{bmatrix}$$

$$\xrightarrow{\mathbf{r}_2 \mapsto \mathbf{r}_2 - 2\mathbf{r}_1} \begin{bmatrix} 1 & \frac{1}{2} & -\frac{3}{4} & -1 & 5 \\ 0 & 0 & -\frac{3}{2} & 3 & -15 \end{bmatrix} \xrightarrow{\mathbf{r}_2 \mapsto -2/3\mathbf{r}_2} \begin{bmatrix} 1 & \frac{1}{2} & -\frac{3}{4} & -1 & 5 \\ 0 & 0 & 1 & -2 & 10 \end{bmatrix}$$

$$\xrightarrow{\mathbf{r}_1 \mapsto \mathbf{r}_1 + \frac{3}{4}\mathbf{r}_2} \begin{bmatrix} 1 & \frac{1}{2} & 0 & -\frac{5}{2} & \frac{25}{2} \\ 0 & 0 & 1 & -2 & 10 \end{bmatrix}$$

So the original system, in matrix form

$$\begin{bmatrix} 4 & 2 & -3 & -4 \\ 2 & 1 & -3 & 1 \end{bmatrix} \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 20 \\ -5 \end{bmatrix}$$

is equivalent to the reduced system

$$\begin{aligned} w + \frac{1}{2}x - \frac{5}{2}z &= \frac{25}{2} \\ y - 2z &= 10 \end{aligned}$$

or in matrix form $RA\mathbf{x} = R\mathbf{b}$, as

$$\begin{bmatrix} 1 & \frac{1}{2} & 0 & -\frac{5}{2} \\ 0 & 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \frac{25}{2} \\ 10 \end{bmatrix}$$

The solution to this system is a coset of $\text{nullsp}\, RA$ (which, remember, also equals $\text{nullsp}\, A$), so we need to find two things: some arbitrary solution $\mathbf{x}$ to $RA\mathbf{x} = R\mathbf{b}$ to be a base point for the coset, and a basis of $\text{nullsp}\, RA$. To find a base point for the coset, we can note that the variables w and y—sometimes called "pivot variables," as their corresponding columns in RA are pivot columns—occur only in one equation each, with coefficient 1, and only in combination with non-pivot variables (sometimes called "free variables"). This is true of all linear systems with RREF matrices: each equation will have a different single pivot variable, with coefficient 1, and then zero or more free variables. One solution to this system comes from setting each pivot variable to the value of the only equation that it appears in, and setting the free variables to zero. In this example, we have $(w, x, y, z) = (25/2, 0, 10, 0)$. If we take an augmented matrix for the system and insert rows of zeros to align the pivots on the diagonal, then the right-hand column of the result gives this solution as a column vector.

$$\left[\begin{array}{cccc|c} 1 & \frac{1}{2} & 0 & -\frac{5}{2} & \frac{25}{2} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & 10 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

In section 4.6, we showed how to find $\text{nullsp}\, A$: take this padded matrix (without the equation values column) and change the non-pivot diagonal entries on the left-hand side to -1, to get

$$\begin{bmatrix} 1 & \frac{1}{2} & 0 & -\frac{5}{2} \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

then take the non-pivot columns $\left\{ \begin{bmatrix} \frac{1}{2} \\ -1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -\frac{5}{2} \\ 0 \\ -2 \\ -1 \end{bmatrix} \right\}$ as a basis for $\text{nullsp}\, A$.

By combining our chosen base point and nullspace basis, we get a general formula for the solution set:

$$\begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \frac{25}{2} \\ 0 \\ 10 \\ 0 \end{bmatrix} + c_1 \begin{bmatrix} \frac{1}{2} \\ -1 \\ 0 \\ 0 \end{bmatrix} + c_2 \begin{bmatrix} -\frac{5}{2} \\ 0 \\ -2 \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{25}{2} + \frac{1}{2}c_1 - \frac{5}{2}c_2 \\ -c_1 \\ 10 - 2c_2 \\ -c_2 \end{bmatrix}$$

This is not the only way to express the solution set. First, we don't have to use $(\frac{25}{2}, 0, 10, 0)$ as the base point. If we set $c_1 = 0$ and $c_2 = 5$, for example, we get an integer solution $(w, x, y, z) = (0, 0, 0, -5)$ that we could use as the base point instead. Second, we could choose a different basis for $\text{nullsp}\, A$—for instance, we could scale the basis vectors to have integer entries, say $\left\{ \begin{bmatrix} 1 \\ -2 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 5 \\ 0 \\ 4 \\ 2 \end{bmatrix} \right\}$. These two changes together give an alternate (but equivalent!) general form for the solution set

$$\begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} c_1 + 5c_2 \\ -2c_1 \\ 4c_2 \\ -5 + 2c_2 \end{bmatrix}$$

where, again, c_1 and c_2 can be any real coefficients.

In summary, to solve an underspecified system $A\mathbf{x} = \mathbf{b}$:

1. Reduce the augmented matrix $\left[A \mid \mathbf{b} \right]$ to RREF.
2. Add and remove rows of zero from $\text{rref} \left[A \mid \mathbf{b} \right]$ so that its left part is square and has all pivots on the diagonal.
3. Take the extra column of the resulting matrix as a representative solution $\mathbf{x}$.
4. Change all zero entries on the diagonal of the resulting matrix to -1 and take the columns that include these entries as a basis of $\text{nullsp}\, A$.

The solution to the system is $\mathbf{x} + \text{nullsp}\, A$.

4.9 Singular and overdetermined systems

Sometimes the left-hand side of one equation in a system is just a linear combination of the left-hand sides of some other equations. We call systems like this *singular*. One large class of linear systems that must be singular are *overdetermined* systems: those with more equations than variables.

Reducing the matrix of coefficients for a singular system to RREF has to leave a row of all zeros, which corresponds to eliminating all the variables in one of the equations in the system and leaving a left-hand side of zero. The original system has a solution if and only if the *values* of these equations with left-hand side zeros in the reduced system are also zero: if we can manipulate a system of equations with steps that preserve the set of solutions to the system and get an equation like $0x + 0y = 1$, then the system could not have had any solutions to begin with.

Consider, for instance, the system of three equations in two variables

$$\begin{aligned} x + y &= 3 \\ 2x - 3y &= -4 \\ 4x - y &= k \end{aligned}$$

where k is some constant. The LHS of the third equation is the LHS of the second plus twice the LHS of the first. If $x + y = 3$ and $2x - 3y = -4$, furthermore, then

$4x - y = 2(x + y) + (2x - 3y) = 2 \times 3 - 4 = 2$. So this system is solvable, with solution $(x, y) = (1, 2)$, only if $k = 2$. Otherwise, the third equation cannot hold simultaneously with the first and second, and the system is inconsistent. (If you plot the lines $x+y = 3$, $2x - 3y = -4$, and $4x - y = k$ on the same graph, then you will see that they have a common point of intersection only if $k = 2$..)

Let's translate these observations on this example system into matrix language. First, look at the consistent system with $k = 2$. GJE of the augmented matrix is

$$\left[\begin{array}{cc|c} 1 & 1 & 3 \\ 2 & -3 & -4 \\ 4 & -1 & 2 \end{array}\right] \xrightarrow{\mathbf{r}_2 \mapsto \mathbf{r}_2 - 2\mathbf{r}_1} \left[\begin{array}{cc|c} 1 & 1 & 3 \\ 0 & -5 & -10 \\ 4 & -1 & 2 \end{array}\right] \xrightarrow{\mathbf{r}_3 \mapsto \mathbf{r}_3 - 4\mathbf{r}_1} \left[\begin{array}{cc|c} 1 & 1 & 3 \\ 0 & -5 & -10 \\ 0 & -5 & -10 \end{array}\right]$$
$$\xrightarrow{\mathbf{r}_2 \mapsto -\mathbf{r}_2/5} \left[\begin{array}{cc|c} 1 & 1 & 3 \\ 0 & 1 & 2 \\ 0 & -5 & -10 \end{array}\right] \xrightarrow{\mathbf{r}_1 \mapsto \mathbf{r}_1 - \mathbf{r}_2} \left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 0 & -5 & -10 \end{array}\right] \xrightarrow{\mathbf{r}_3 \mapsto \mathbf{r}_3 + 5\mathbf{r}_2} \left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{array}\right]$$

which is equivalent to the system

$$\begin{aligned} x &= 1 \\ y &= 2 \\ 0 &= 0 \end{aligned}$$

The last line $0 = 0$ is, of course, always true, and the augmented matrix has rank 2.

Now if we took an inconsistent system, say

$$\begin{aligned} x + y &= 3 \\ 2x - 3y &= -4 \\ 4x - y &= 4 \end{aligned}$$

then GJE of the augmented matrix looks similar at first, but then diverges and results in a matrix with rank 3:

$$\left[\begin{array}{cc|c} 1 & 1 & 3 \\ 2 & -3 & -4 \\ 4 & -1 & 4 \end{array}\right] \xrightarrow{\mathbf{r}_2 \mapsto \mathbf{r}_2 - 2\mathbf{r}_1} \left[\begin{array}{cc|c} 1 & 1 & 3 \\ 0 & -5 & -10 \\ 4 & -1 & 4 \end{array}\right] \xrightarrow{\mathbf{r}_3 \mapsto \mathbf{r}_3 - 4\mathbf{r}_1} \left[\begin{array}{cc|c} 1 & 1 & 3 \\ 0 & -5 & -10 \\ 0 & -5 & -8 \end{array}\right]$$
$$\xrightarrow{\mathbf{r}_2 \mapsto -\mathbf{r}_2/5} \left[\begin{array}{cc|c} 1 & 1 & 3 \\ 0 & 1 & 2 \\ 0 & -5 & -8 \end{array}\right] \xrightarrow{\mathbf{r}_1 \mapsto \mathbf{r}_1 - \mathbf{r}_2} \left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 0 & -5 & -8 \end{array}\right] \xrightarrow{\mathbf{r}_3 \mapsto \mathbf{r}_3 + 5\mathbf{r}_2} \left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 2 \end{array}\right]$$
$$\xrightarrow{\mathbf{r}_3 \mapsto \mathbf{r}_3/2} \left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{array}\right] \xrightarrow{\mathbf{r}_1 \mapsto \mathbf{r}_1 - \mathbf{r}_3} \left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{array}\right] \xrightarrow{\mathbf{r}_2 \mapsto \mathbf{r}_2 - 2\mathbf{r}_1} \left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right]$$

The extra column in the augmented matrix is a a pivot column, and we have reduced the inconsistent system to the system

$$\begin{aligned} x &= 0 \\ y &= 0 \\ 0 &= 1. \end{aligned}$$

This system has no solution: no assignment of variables can make 0 equal 1. In general, if one equation in a system is incompatible with those that came before it, then Gauss–Jordan elimination turns the augmented matrix row for that equation into a pivot row with a pivot in the rightmost column, corresponding to the equation $0 = 1$. This is the essence of a result called the *Rouché–Capelli theorem*, which states that a system has a solution if and only if its matrix of variable coefficients has the same rank as the augmented matrix.

Another proof of the Rouché–Capelli theorem: $A\mathbf{x} = \mathbf{b}$ has a solution if and only if $\mathbf{b}$ is in the column space (that is, image of the multiplication map) of A—that is, if and only if appending $\mathbf{b}$ to A as a new column leaves the column space and its dimension (that is, the rank of A) unchanged.

4.10 Matrix inversion by Gauss–Jordan elimination

The only $n \times n$ matrix in RREF with rank n is the identity matrix I, so if A is any $n \times n$ matrix that also has full rank (and, therefore, has an inverse matrix), then GJE reduces A to I. If R is the matrix encoding the row operations that reduce A to RREF, then $RA = I$. That is, R is the inverse matrix of A.

As $R = RI$, we can find R by performing row operations on A and, in parallel, performing the same row operations on I. To do this, set up an augmented matrix with A on the left side and n extra rows constituting the identity matrix I on the right. After row reduction is complete, the left side contains I and the right side contains $R = A^{-1}$.

Let's illustrate with the example $A = \begin{bmatrix} 7 & 2 & 1 \\ 0 & 3 & -1 \\ -3 & 4 & -2 \end{bmatrix}$. Setting up the augmented matrix and carrying out row reduction gives

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

$$\xrightarrow{\mathbf{r}_1 \mapsto \mathbf{r}_1/7} \left[\begin{array}{ccc|ccc} 1 & \frac{2}{7} & \frac{1}{7} & \frac{1}{7} & 0 & 0 \\ 0 & 3 & -1 & 0 & 1 & 0 \\ -3 & 4 & -2 & 0 & 0 & 1 \end{array}\right]$$

$$\xrightarrow{\mathbf{r}_3 \mapsto \mathbf{r}_3 + 3\mathbf{r}_1} \left[\begin{array}{ccc|ccc} 1 & \frac{2}{7} & \frac{1}{7} & \frac{1}{7} & 0 & 0 \\ 0 & 3 & -1 & 0 & 1 & 0 \\ 0 & \frac{34}{7} & -\frac{11}{7} & \frac{3}{7} & 0 & 1 \end{array}\right]$$

$$\xrightarrow{\mathbf{r}_2 \mapsto \mathbf{r}_2/3} \left[\begin{array}{ccc|ccc} 1 & \frac{2}{7} & \frac{1}{7} & \frac{1}{7} & 0 & 0 \\ 0 & 1 & -\frac{1}{3} & 0 & \frac{1}{3} & 0 \\ 0 & \frac{34}{7} & -\frac{11}{7} & \frac{3}{7} & 0 & 1 \end{array}\right]$$

$$\xrightarrow{\mathbf{r}_1 \mapsto \mathbf{r}_1 - 2\mathbf{r}_2/7} \left[\begin{array}{ccc|ccc} 1 & 0 & \frac{5}{21} & \frac{1}{7} & -\frac{2}{21} & 0 \\ 0 & 1 & -\frac{1}{3} & 0 & \frac{1}{3} & 0 \\ 0 & \frac{34}{7} & -\frac{11}{7} & \frac{3}{7} & 0 & 1 \end{array}\right]$$

$$\xrightarrow{\mathbf{r}_3 \mapsto \mathbf{r}_3 - 34\mathbf{r}_2/7} \left[\begin{array}{ccc|ccc} 1 & 0 & \frac{5}{21} & \frac{1}{7} & -\frac{2}{21} & 0 \\ 0 & 1 & -\frac{1}{3} & 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{21} & \frac{3}{7} & -\frac{34}{21} & 1 \end{array}\right]$$

$$\xrightarrow{\mathbf{r}_3 \mapsto 21\mathbf{r}_3} \left[\begin{array}{ccc|ccc} 1 & 0 & \frac{5}{21} & \frac{1}{7} & -\frac{2}{21} & 0 \\ 0 & 1 & -\frac{1}{3} & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 1 & 9 & -34 & 21 \end{array}\right]$$

$$\xrightarrow{\mathbf{r}_1 \mapsto \mathbf{r}_1 - 5\mathbf{r}_3/21} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -2 & 8 & -5 \\ 0 & 1 & -\frac{1}{3} & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 1 & 9 & -34 & 21 \end{array}\right]$$

$$\xrightarrow{\mathbf{r}_2 \mapsto \mathbf{r}_2 + \mathbf{r}_3/3} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -2 & 8 & -5 \\ 0 & 1 & 0 & 3 & -11 & 7 \\ 0 & 0 & 1 & 9 & -34 & 21 \end{array}\right]$$

So $A^{-1} = \begin{bmatrix} -2 & 8 & -5 \\ 3 & -11 & 7 \\ 9 & -34 & 21 \end{bmatrix}$. You can test this result yourself by computing the products

AA^{-1} and $A^{-1}A$. Note that the inverses of most integer matrices will have non-integer rational entries; this example was chosen to look nice.

4.11 Triangular matrices

4.11.1 Defined

An *upper triangular matrix* is a matrix whose entries below the diagonal are all zero. For example, a 4×4 upper triangular matrix has the form

$$\begin{bmatrix} \star & \star & \star & \star \\ 0 & \star & \star & \star \\ 0 & 0 & \star & \star \\ 0 & 0 & 0 & \star \end{bmatrix}$$

where the entries marked with $\star$ can be anything (including, possibly, zero). A lower triangular matrix, symmetrically, has only zeros above the diagonal:

$$\begin{bmatrix} \star & 0 & 0 & 0 \\ \star & \star & 0 & 0 \\ \star & \star & \star & 0 \\ \star & \star & \star & \star \end{bmatrix}$$

Note that diagonal matrices are both upper triangular and lower triangular.

4.11.2 Properties of square triangular matrices

Square triangular matrices have a few important properties that will become useful later:

Proposition. *A square triangular matrix has full rank if and only if all of its diagonal entries are nonzero.*

Proof. We'll prove this statement for upper triangular matrices first. Proving an if-and-only-if statement requires proving two implications.

1. *Diagonal entries all nonzero implies full rank.* Suppose $\mathbf{c}_1, \ldots, \mathbf{c}_n$ are the columns of an upper triangular matrix with nonzero diagonal entries. Then in particular, $\mathbf{c}_1 \neq \mathbf{0}$, so $\{\mathbf{c}_1\}$ is a linearly independent set. And if $\{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$ is linearly independent for any integer $1 \leq k \leq n-1$, then $\{\mathbf{c}_1, \ldots, \mathbf{c}_{k+1}\}$ is also linearly independent, because $\mathbf{c}_{k+1}$ has a nonzero entry in row $k+1$ where the other column vectors $\mathbf{c}_1, \ldots, \mathbf{c}_k$ have zeros and so no linear combination of $\mathbf{c}_1, \ldots, \mathbf{c}_k$ equals $\mathbf{c}_{k+1}$. So by induction, $\{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$ is linearly independent.

2. *At least one nonzero diagonal entry implies non-full rank.* If the kth diagonal entry of an upper triangular matrix is zero, then the first k columns $\{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$ have nonzero entries only in the top $k-1$ positions, so they are all contained in the $k-1$-dimensional subspace $\mathrm{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_{k-1}\}$ of $\mathrm{Col}_n(\mathbb{F})$ and can't be linearly independent.

The argument for lower triangular matrices is symmetrical.

□

Proposition. *The product of two upper (or lower) triangular matrices is also upper (or lower) triangular, and the ith diagonal entry in the matrix product is the product of the ith diagonal entries of the two factors.*

Proof. Suppose A and B are upper triangular. Write a_{rc}, b_{rc} for the entries in row r, column c of A and B, with $a_{rc} = b_{rc} = 0$ whenever $r > c$. By definition of matrix multiplication, the entry in row r and column c of AB is $\sum_{i=1}^{n} a_{ri}b_{ic}$. The only terms in this sum that are possibly nonzero are for values of i with $r \leq i \leq c$. But for below-diagonal entries $r > c$, there are no such values of i, and for diagonal entries, then the only such value of i is r itself, for an entry of $a_{rr}b_{rr}$.

The argument for lower triangular matrices is symmetrical.

□

Proposition. *The inverse of an invertible upper (or lower) triangular matrix is also upper (or lower) triangular.*

Proof. GJE on an invertible upper triangular matrix M involves only two types of operations: row scaling operations (represented in matrix form by diagonal matrices), to make each diagonal pivot entry equal to 1; and shear operations of the form $\mathbf{r}_i \mapsto \mathbf{r}_i + \lambda\mathbf{r}_j$, where $i < j$, to clear entries in column j above the pivot on the diagonal (since the entries below the pivot are already zero and we don't need to do anything with them). These shear operations have upper triangular representations with 1s on the diagonal and another nonzero entry in row i and column j. So M^{-1} is the matrix product of upper triangular matrices, so it must also be upper triangular.

The argument for lower triangular matrices is symmetrical.

□

4.11.3 Rectangular triangular matrices

It is sometimes useful to extend the definition of triangularity to non-square matrices. The diagonal in non-square matrices descends from the top left corner down and to the right, but does not meet the bottom left corner. A 4×3 upper triangular matrix, for example, has the form

$$\begin{bmatrix} \star & \star & \star \\ 0 & \star & \star \\ 0 & 0 & \star \\ 0 & 0 & 0 \end{bmatrix}$$

and a lower triangular matrix of the same size has the form

$$\begin{bmatrix} \star & 0 & 0 \\ \star & \star & 0 \\ \star & \star & \star \\ \star & \star & \star \end{bmatrix}$$

The product of two non-square lower (or upper) triangular matrices, provided they have compatible dimensions, is also lower (or upper) triangular. You can prove this by padding non-square matrices with rows or columns of zeros to make them square: padding with zeros will keep a triangular matrix triangular.

4.11.4 Forward and back substitution

Systems of equations with triangular matrices of coefficients are especially easy to solve because instead of processing the whole system, you can simply read a value of one variable from one equation and use an iterative process of substituting known variables into other equations to find unknown variables. Consider, for example, the system

$$\begin{aligned} 2w-3x \quad +z &= -7 \\ x+2y+z &= 9 \\ y-z &= 8 \\ 2z &= -6 \end{aligned}$$

You could solve this system with GJE on the augmented matrix $\left[\begin{array}{cccc|c} 2 & -3 & 0 & 1 & -7 \\ 0 & 1 & 2 & 1 & 8 \\ 0 & 0 & 1 & -1 & 8 \\ 0 & 0 & 0 & 2 & -6 \end{array}\right]$, but it's faster to solve equations from the bottom up, substituting solutions from lower equations into higher equations, like this:

1. $2z = -6$ immediately gives $z = -3$.
2. Substituting $z = -3$ into $y - z = 8$ gives $y + 3 = 8$ or $z = 5$.
3. Substituting $y = 5$ and $z = -3$ into $x + 2y + z = 8$ gives $x + 7 = 9$ or $x = 2$.
4. Substituting $x = 2$ and $z = -3$ into $2w - 3x + z = -7$ gives $2w - 9 = -7$ or $w = 1$.

So the solution is $(w, x, y, z) = (1, 2, 5, -3)$. This procedure of working from the bottom of a system upward is called *back-substitution*, and it works because the matrix of coefficients $\begin{bmatrix} 2 & -3 & 0 & 1 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 2 \end{bmatrix}$ is upper triangular.

A symmetrical process, *forward substitution*, is possible for systems with lower triangular coefficient matrices. Consider, for example:

$$\begin{aligned} -3w \quad\quad\quad\quad &= -3 \\ 4w-x \quad\quad\quad &= 2 \\ w+x+2y \quad\quad &= 13 \\ 2x-2y-z &= -3. \end{aligned}$$

The matrix of coefficients is the lower triangular matrix

$$\begin{bmatrix} -3 & 0 & 0 & 0 \\ 4 & -1 & 0 & 0 \\ 1 & 1 & 2 & 0 \\ 0 & 2 & -2 & -1 \end{bmatrix}$$

You can solve this system by first finding $w = 1$ from the first equation, then substituting $w = 1$ into the second equation to get $4 - x = 2$ (which gives $x = 2$), then

subsituting both of these values into the third equation to get a value for y, and so on. (The solution to this system is also $(w, x, y, z) = (1, 2, 5, 3)$.)

If a matrix A can be factored into lower and upper triangular matrices as $A = LU$ (the next section is about how we can find this factorization), then we can solve $A\mathbf{x} = LU\mathbf{x} = \mathbf{b}$ by a two-step process of forward substitution followed by back substitution. Since $\mathbf{x} = U^{-1}L^{-1}\mathbf{b}$, we can first compute $L^{-1}\mathbf{b}$ (that is, the vector $\mathbf{y}$ such that $L\mathbf{y} = \mathbf{b}$) by using forward substitution. Then $\mathbf{x}$ is simply $U^{-1}\mathbf{y}$; that is, the solution to $U\mathbf{x} = \mathbf{y}$, and we can find $\mathbf{x}$ with back substitution.

4.12 LU decomposition

4.12.1 Defined; core algorithm

With a modified form of Gaussian elimination, you can bring any $r \times c$ matrix A into an upper triangular form U. The row operation matrix R for which $RA = U$ will usually be a lower triangular $r \times r$ matrix with diagonal entries all equal to 1. The inverse matrix $L \coloneqq R^{-1}$ gives a decomposition $A = LU$, where L is also lower triangular with diagonal entries all equal to 1 and and U is upper triangular. This decomposition is called the *LU decomposition* of A, and it has a practical use: if you need to solve the same linear system multiple times, with the same coefficients but different sets of equation values, LU decomposition provides a much faster alternative to solving the system from scratch each time.

In the ideal case, the modified elimination involves only shear operations with lower triangular representations, in which multiples of a higher row are added to a lower row. (In some cases, the procedure may require row swaps, which do not have triangular representation. We'll address this complication later.)

In the simplest case, the algorithm runs as follows.

1. Subtract multiples of the first row of A from every other row so that the only nonzero entry in the first column is the top left corner. Unlike in standard GJE, you do *not* have to scale the first row so that the first entry is 1. Remember that the matrix representation of the row operation $\mathbf{r}_j \mapsto \mathbf{r}_j + \lambda\mathbf{r}_1$ is the identity matrix with an additional entry of λ in position $(1, j)$, so the matrix is lower triangular.

   If the top left corner of A is zero, then we need a more complicated algorithm that involves row swaps that we'll discuss a bit later.

2. Subtract multiples of the second row from the third row and every row below it (but *not* the first row), so that the only nonzero entries in the second column are in the first and second rows. Again, if the second diagonal entry at this stage is zero, you'll need a more complicated algorithm with row swaps.

3. Subtract multiples of the third row from the fourth row and every row below it, so that the only nonzero entries of the third column are the top three. Continue in this vein until we reach the last row or the last column.

4.12.2 Example

As an example, consider the matrix

$$A = \begin{bmatrix} \mathbf{1} & 2 & -4 & 3 \\ 2 & \mathbf{3} & 7 & 0 \\ -1 & 2 & \mathbf{-4} & 5 \end{bmatrix}$$

(the bold entries mark the diagonal). We can use shear operations to turn this into an upper triangular matrix like this:

$$\begin{bmatrix} 1 & 2 & -4 & 3 \\ 2 & 3 & 7 & 0 \\ -1 & 2 & -4 & 5 \end{bmatrix} \xrightarrow{\mathbf{r}_2 \mapsto \mathbf{r}_2 - 2\mathbf{r}_1} \begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & -1 & 15 & -6 \\ -1 & 2 & -4 & 5 \end{bmatrix}$$

$$\xrightarrow{\mathbf{r}_3 \mapsto \mathbf{r}_3 + \mathbf{r}_1} \begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & -1 & 15 & -6 \\ 0 & 4 & -8 & 8 \end{bmatrix} \xrightarrow{\mathbf{r}_3 \mapsto \mathbf{r}_3 + 4\mathbf{r}_2} \begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & -1 & 15 & -6 \\ 0 & 0 & 52 & -16 \end{bmatrix}$$

Let's designate this resulting matrix U. Note we can also change U back to A with the row operations $(\mathbf{r}_3 \mapsto \mathbf{r}_3 - 4\mathbf{r}_2), (\mathbf{r}_3 \mapsto \mathbf{r}_3 - \mathbf{r}_1), (\mathbf{r}_2 \mapsto \mathbf{r}_2 + 2\mathbf{r}_1)$, each step of which reverses one of the steps that moved from A to U.

4.12.3 Methods for finding L

We can determine the matrix L by using an augmented matrix with A on the left and the identity matrix on the right, as with the matrix inversion method from Section 4.10. Alternatively, we can deduce the entries in L directly from the list of row operations, without keeping an augmented matrix. Note that the equation $LU = A$ shows that L encodes the sequence of row operations that takes U to A—that is, the reverse of the steps involved in the decomposition of A.

In this sequence from U to A, we start modifying rows at the bottom of U and work upwards, and we only modify a row by adding multiples of higher-up rows that have not yet been modified—that is, the step $\mathbf{r}_i \mapsto \mathbf{r}_i + \lambda\mathbf{r}_j$ only occurs while $\mathbf{r}_j$ has not yet been changed and still has the value that it has in U. (This is not true of working from A to U: in our worked example, for instance, we added a multiple of $\mathbf{r}_2$ to $\mathbf{r}_3$ after $\mathbf{r}_2$ had already been modified.) So if the steps $\mathbf{r}_k \mapsto \mathbf{r}_k + \lambda_1\mathbf{r}_1, \lambda_k \mapsto \mathbf{r}_k + \lambda_2\mathbf{r}_2, \ldots, \mathbf{r}_k \mapsto \mathbf{r}_k + \lambda_{k-1}\mathbf{r}_{k-1}$ all occur in the sequence from U to A, we know that the kth row of A is $\mathbf{r}_k + \lambda_1\mathbf{r}_1 + \lambda_2\mathbf{r}_2 + \cdots + \lambda_{k-1}\mathbf{r}_{k-1}$ where $\mathbf{r}_i$ is the *unmodified* ith row of U. So if $LU = A$, then the kth row of L gives the coefficients that determine the kth row of A as a linear combination of the rows of U; that is, the kth row of L is $\begin{bmatrix} \lambda_1 & \lambda_2 & \cdots & \lambda_{k-1} & 1 & 0 & \cdots & 0 \end{bmatrix}$.

In our example with the sequence of operations $(\mathbf{r}_2 \mapsto \mathbf{r}_2 - 2\mathbf{r}_1), (\mathbf{r}_3 \mapsto \mathbf{r}_3 + \mathbf{r}_1), (\mathbf{r}_3 \mapsto \mathbf{r}_3 + 4\mathbf{r}_2)$ from A to U and the reverse sequence $(\mathbf{r}_3 \mapsto \mathbf{r}_3 - 4\mathbf{r}_2), (\mathbf{r}_3 \mapsto \mathbf{r}_3 - \mathbf{r}_1), (\mathbf{r}_2 \mapsto \mathbf{r}_2 + 2\mathbf{r}_1)$ from U to A, we can read the off-diagonal entries of L from the row operations:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -4 & 1 \end{bmatrix}$$

and you can check for yourself that $LU = A$.

To summarize, the algorithm for determining an LU decomposition of an $m \times n$ matrix A is the following:

1. Conduct the modified Gaussian elimination of A to U.

2. Set up an $m \times m$ matrix L, with entries of 1 on the diagonal and 0 above the diagonal.

3. If going from A to U requires the row operation $\mathbf{r}_i \mapsto \mathbf{r}_i + \lambda \mathbf{r}_j$, then put $-\lambda$ into position (i, j) of L.

This algorithm works as long as the rows of A do not need to be reordered. We'll discuss this complication in a bit.

4.12.4 Alternate algorithm for LU decomposition

There's a way to do LU decomposition without explicit Gaussian elimination. First, we set up L and U with variables for the unknown entries. L is always square with diagonal entries 1, and U has the same dimensions as A. Then solve a set of linear systems to find the unknown entries.

To illustrate this method, let's use our example matrix from the last section, with an additional row at the bottom. (Adding a row to A means that the decomposition $A = LU$ will have additional rows at the bottom of L, but the old rows will remain unmodified, and U stays the same. To see why this is so, think about which steps in LU decomposition involve the bottom row of A.)

$$A = \begin{bmatrix} 1 & 2 & -4 & 3 \\ 2 & 3 & 7 & 0 \\ -1 & 2 & -4 & 5 \\ -6 & 4 & 10 & -2 \end{bmatrix}.$$

Now set up a template for the factors L and U:

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \ell_{21} & 1 & 0 & 0 \\ \ell_{31} & \ell_{32} & 1 & 0 \\ \ell_{41} & \ell_{42} & \ell_{43} & 1 \end{bmatrix} \qquad U = \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix}.$$

We can find the unknown quantities in each matrix by working top-down, multiplying each row in L with all the columns in U. Each row gives a system that can be solved by forward substitution:

1. Multiply the first row of L by each column of U to get the first row of U, which must equal the first row of A: $(u_{11}, u_{22}, u_{33}, u_{44}) = (1, 2, -4, 3)$. The state of the LU decomposition is now

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \ell_{21} & 1 & 0 & 0 \\ \ell_{31} & \ell_{32} & 1 & 0 \\ \ell_{41} & \ell_{42} & \ell_{43} & 1 \end{bmatrix} \qquad U = \begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix}.$$

2. Multiply the second row of L by each column of U in turn to get the system

$$\begin{aligned} \ell_{21} &= 2 \\ 2\ell_{21} + u_{22} &= 3 \\ -4\ell_{21} + u_{23} &= 7 \\ 3\ell_{21} + u_{24} &= 0 \end{aligned}$$

and solve it by substituting the top equation $\ell_{21} = 2$ into the other three equations (which is technically a simple example of forward substitution) to get the nonzero portion of the second row of U: $(u_{22}, u_{23}, u_{24}) = (-1, 15, -6)$. The state of the LU decomposition is now

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ \ell_{31} & \ell_{32} & 1 & 0 \\ \ell_{41} & \ell_{42} & \ell_{43} & 1 \end{bmatrix} \qquad U = \begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & -1 & 15 & -6 \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix}.$$

3. Multiply the third row of L by each column of U in turn to get the system

$$\begin{aligned} \ell_{31} &= -1 \\ 2\ell_{31} - \ell_{32} &= 2 \\ -4\ell_{31} + 15\ell_{32} + u_{33} &= -4 \\ 3\ell_{31} - 6\ell_{32} + u_{34} &= 5 \end{aligned}$$

We can solve this system with forward substitution. First put $\ell_{31} = -1$ into the second equation to get $\ell_{32} = -4$, then put both $\ell_{31} = -1$ and $\ell_{32} = -4$ into the next two equations to get $u_{33} = 52$ and $u_{34} = -16$. The state of the LU decomposition is now

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & -4 & 1 & 0 \\ \ell_{41} & \ell_{42} & \ell_{43} & 1 \end{bmatrix} \qquad U = \begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & -1 & 15 & -6 \\ 0 & 0 & 52 & -16 \\ 0 & 0 & 0 & u_{44} \end{bmatrix}.$$

4. Multiply the fourth row of L by each column of U to get

$$\begin{aligned} \ell_{41} &= -6 \\ 2\ell_{41} - \ell_{42} &= 4 \\ -4\ell_{41} + 15\ell_{42} + 52\ell_{43} &= 10 \\ 3\ell_{41} - 6\ell_{42} - 16\ell_{43} + u_{44} &= -2 \end{aligned}$$

This system, again, can be solved by forward substitution to get $(\ell_{41}, \ell_{41}, \ell_{43}, u_{44}) = (-6, -13, \frac{113}{26}, -\frac{136}{13})$, giving a final decomposition

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & -4 & 1 & 0 \\ -6 & -13 & \frac{113}{26} & 1 \end{bmatrix} \qquad U = \begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & -1 & 15 & -6 \\ 0 & 0 & 52 & -16 \\ 0 & 0 & 0 & -\frac{136}{13} \end{bmatrix}.$$

4.12.5 Computational advantages of LU decomposition

LU decomposition to solve a linear system with forward and back substitution can be significantly faster than Gauss–Jordan elimination. On an $n \times n$ matrix, the amount of computation required by GJE is roughly proportional to n^3 (in computer science jargon, GJE requires $O(n^3)$ time): each row operation requires a separate operation for each entry (so n operations in total), clearing the non-pivot entries in each column can take up to n row operations, and there are n columns.

Solving a triangular system with forward or back substitution, on the other hand, requires only $O(n^2)$ operations: for every integer $1 \leq k \leq n-1$, we have to substitute k known variable values into the equation with $k+1$ variables and then carry out $k-1$ multiplication and subtraction steps to isolate the unknown variable. So if you have to solve an equation of the form $A\mathbf{x} = \mathbf{b}$ repeatedly for a fixed coefficient matrix A and many different values $\mathbf{b}$, then the preprocessing time required to factor $A = LU$ (which can be done in $O(n^3)$ time) can save time overall.

LU decomposition also has advantages over directly computing A^{-1} and then $A^{-1}\mathbf{b}$. Theoretically, computing A^{-1} via GJE also takes $O(n^3)$ time, and then computing $A^{-1}\mathbf{b}$ takes $O(n^2)$ time for each vector $\mathbf{b}$, the same time requirements as LU decomposition. But there are practical drawbacks to matrix inversion. First, it's numerically unstable: rounding errors can accumulate, creating inaccuracies in floating-point computer systems that can represent most non-integer values only approximately. Matrices in scientific computing are also frequently *sparse*: most of their entries are zero. Software libraries can store sparse matrices in compact forms that don't require memory for every entry. LU decomposition on sparse matrices can create a factorization into sparse matrices, but the inverse of a sparse matrix is generally not sparse and could require prohibitive amounts of memory to store.

Take, for instance, the case of matrices with a small *bandwidth* b: every nonzero entry is located at most b places horizontally or vertically from the diagonal (that is, position (r, c) is nonzero only if $|r, c| \leq b$). On such a matrix, LU decomposition takes only $O(b^2 n)$ operations. Each of the n columns has at most b below-diagonal entries that require one shear operation each to clear. By the time we do a shear operation of the form $\mathbf{r}_i \mapsto \mathbf{r}_i + \mathbf{r}_j$, all of the left-of-diagonal entries in $\mathbf{r}_j$ have been cleared, so if we started with a matrix with bandwidth b, $\mathbf{r}_j$ can have at most $b+1$ nonzero entries and carrying out the row operation requires only $O(b)$ time. Furthermore, the necessary shear operations $\mathbf{r}_i \mapsto \mathbf{r}_i + \mathbf{r}_j$ wil always have $j+1 \leq i \leq j+b$, so the resulting matrices L and U will also only have bandwidth b.

4.12.6 LDU decomposition

LU decomposition has an annoying asymmetry: L is *unitriangular* (that is, it has diagonal entries of 1), but U is not. But we can make the factorization more symmetrical if A is invertible. In this case, the entries on the diagonal of U are all nonzero, and we can divide every row in U by its entry on the diagonal to get an upper unitriangular matrix. The diagonal entries of U go into a separate diagonal matrix notated D, and the overall decomposition is called an LDU decomposition.

For example, the LU decomposition from the last section

$$\begin{bmatrix} 1 & 2 & -4 & 3 \\ 2 & 3 & 7 & 0 \\ -1 & 2 & -4 & 5 \\ -6 & 4 & 10 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & -4 & 1 & 0 \\ -6 & -13 & \frac{113}{26} & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & -1 & 15 & -6 \\ 0 & 0 & 52 & -16 \\ 0 & 0 & 0 & -\frac{136}{13} \end{bmatrix}$$

can be recast as this LDU decomposition:

$$\begin{bmatrix} 1 & 2 & -4 & 3 \\ 2 & 3 & 7 & 0 \\ -1 & 2 & -4 & 5 \\ -6 & 4 & 10 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & -4 & 1 & 0 \\ -6 & -13 & \frac{113}{26} & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 52 & 0 \\ 0 & 0 & 0 & -\frac{136}{13} \end{bmatrix} \begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & 1 & -15 & 6 \\ 0 & 0 & 1 & -\frac{4}{13} \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The LDU decomposition of an invertible matrix, if it exists, is unique. We won't give a formal proof, but uniqueness shouldn't be surprising if you note that the algorithm of section 4.12.4 for determining the entries of L and U never creates an underdetermined system.

4.12.7 LU decomposition with row exchanges

Some invertible matrices do not have LU decompositions. Consider the LU decomposition problem

$$A = \begin{bmatrix} 0 & 2 & -2 \\ 1 & 5 & -3 \\ 4 & 7 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \ell_{21} & 1 & 0 \\ \ell_{31} & \ell_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} = LU.$$

If we try to find the unknown entries of L and U, we quickly run into trouble. The top left entry of A requires $u_{11} = 0$, for instance, but the entry in position (2, 1) requires $u_{11}\ell_{21} = 1$.

It doesn't help if we relax the requirement that L have diagonal entries of 1. If we set up the decomposition as

$$A = \begin{bmatrix} 0 & 2 & -2 \\ 1 & 5 & -3 \\ 4 & 7 & 1 \end{bmatrix} = \begin{bmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} = LU$$

instead, then the top left entry is $\ell_{11}u_{11} = 0$, so ℓ_{11} or u_{11} must be zero. We ran into problems with $u_{11} = 0$, so we could try $\ell_{11} = 0$ instead. But now the entry in position (1, 2) requires $\ell_{11}u_{12} = 2$, which is impossible if $\ell_{11} = 0$.

The basic problem is that since A has full rank, both L and U must also have full rank, so they can't have any zeros on the diagonal. If we used the method of turning A into U with row operations instead, then we would find that we can't change the zero in the top left corner of A if the only allowed row operation is adding multiples of rows to rows further down.

These problems can arise in rows below the first, as well. For example, consider the matrix

$$A = \begin{bmatrix} 1 & 2 & -4 & 3 \\ 2 & 4 & -3 & 7 \\ 0 & -6 & 3 & -10 \\ -5 & -8 & 19 & -11 \end{bmatrix}$$

Let's work out U with modified Gaussian elimination. We'll clear the first column with $\mathbf{r}_2 \mapsto \mathbf{r}_2 - 2\mathbf{r}_1$ and $\mathbf{r}_4 \mapsto \mathbf{r}_4 + 5\mathbf{r}_1$ to get

$$\begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & 0 & 5 & 1 \\ 0 & -6 & 3 & -10 \\ 0 & 2 & -1 & 4 \end{bmatrix}$$

We run into a familiar problem in the second column: there are entries that need to be cleared, but there is a zero in the pivot position on the diagonal. Let's dodge this problem by pretending that LU decompostion lets us use swaps as well as shears and fix this by swapping the second and fourth rows, $\mathbf{r}_2 \leftrightarrow \mathbf{r}_4$, getting

$$\begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & 2 & -1 & 4 \\ 0 & -6 & 3 & -10 \\ 0 & 0 & 5 & 1 \end{bmatrix}.$$

Then we can clear the second column with $\mathbf{r}_3 \mapsto \mathbf{r}_3 + 3\mathbf{r}_2$, giving

$$\begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & 2 & -1 & 4 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 5 & 1 \end{bmatrix}.$$

Again, we have a zero in the pivot position in the third row, but now we only have to swap $\mathbf{r}_3 \leftrightarrow \mathbf{r}_4$, and we're done:

$$\begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & 2 & -1 & 4 \\ 0 & 0 & 5 & 1 \\ 0 & 0 & 0 & 2 \end{bmatrix}.$$

This is upper triangular, but it isn't quite the U factor in $A = LU$, as A doesn't have an LU factorization in the first place. Instead, it's the U factor for a matrix, call it A', created by rearranging the rows of A. We could write $A' = PA$ where P is a permutation matrix—that is, an $n \times n$ matrix with n entries of 1, one in each row and each column, and all other entries zero.

How can we reconstruct the matrices P and L from the sequence of row operations that we needed to construct U? In constructing U, we mixed shear operations and row swaps, but we can rearrange any sequence of swaps and shears to get an equivalent sequence with all the swaps in the front, and glean elements of L directly from the resulting sequence. In LU decomposition with row swaps, we only use row j to modify other rows through shear operations after it has been swapped into the right place, and all subsequent swaps involve rows with numbers greater than j. That is, if we have a sequence of a shear $\mathbf{r}_i \mapsto \mathbf{r}_i + \lambda\mathbf{r}_j$ followed by a swap $\mathbf{r}_k \leftrightarrow \mathbf{r}_\ell$, then j must be strictly less than all three of i, k, ℓ.

We can interchange a shear with an adjacent swap by noting that the sequence of $\mathbf{r}_i \mapsto \mathbf{r}_i + \lambda\mathbf{r}_j$ followed by $\mathbf{r}_i \leftrightarrow \mathbf{r}_k$ is equivalent to $\mathbf{r}_i \leftrightarrow \mathbf{r}_k$ followed by $\mathbf{r}_k \mapsto \mathbf{r}_k + \lambda\mathbf{r}_j$, in which both the old and the new shear operations involve modifying a higher-numbered row $\mathbf{r}_k$ or $\mathbf{r}_i$ using a lower-numbered row $\mathbf{r}_j$.

To illustrate this, let's look at the sequence of row operations that go from A to U in our example:

$$(\mathbf{r}_2 \mapsto \mathbf{r}_2 - 2\mathbf{r}_1) \to (\mathbf{r}_4 \mapsto \mathbf{r}_4 + 5\mathbf{r}_1) \to (\mathbf{r}_2 \leftrightarrow \mathbf{r}_4) \to (\mathbf{r}_3 \mapsto \mathbf{r}_3 + 3\mathbf{r}_2) \to (\mathbf{r}_3 \leftrightarrow \mathbf{r}_4).$$

As $(\mathbf{r}_i \mapsto \mathbf{r}_i + \lambda\mathbf{r}_j) \to (\mathbf{r}_i \leftrightarrow \mathbf{r}_k)$ and $(\mathbf{r}_i \leftrightarrow \mathbf{r}_k) \to (\mathbf{r}_k \mapsto \mathbf{r}_k + \lambda\mathbf{r}_j)$ are equivalent, we can move $(\mathbf{r}_2 \leftrightarrow \mathbf{r}_4)$ to the front while changing any $\mathbf{r}_2$ in the rules that it moves past to $\mathbf{r}_4$ and vice versa, giving

$$(\mathbf{r}_2 \leftrightarrow \mathbf{r}_4) \to (\mathbf{r}_4 \mapsto \mathbf{r}_4 - 2\mathbf{r}_1) \to (\mathbf{r}_2 \mapsto \mathbf{r}_2 + 5\mathbf{r}_1) \to (\mathbf{r}_3 \mapsto \mathbf{r}_3 + 3\mathbf{r}_2) \to (\mathbf{r}_3 \leftrightarrow \mathbf{r}_4).$$

Then moving $(\mathbf{r}_3 \leftrightarrow \mathbf{r}_4)$ forward gives

$$(\mathbf{r}_2 \leftrightarrow \mathbf{r}_4) \to (\mathbf{r}_3 \leftrightarrow \mathbf{r}_4) \to (\mathbf{r}_3 \mapsto \mathbf{r}_3 - 2\mathbf{r}_1) \to (\mathbf{r}_2 \mapsto \mathbf{r}_2 + 5\mathbf{r}_1) \to (\mathbf{r}_4 \mapsto \mathbf{r}_4 + 3\mathbf{r}_2).$$

Note that once a step of the form $\mathbf{r}_i \mapsto \mathbf{r}_i + \lambda\mathbf{r}_j$ takes place—that is, once the below-diagonal entries in column j are cleared—then all subsequent row swaps involve rows strictly below j: a sequence of the form $(\mathbf{r}_i \mapsto \mathbf{r}_i + \lambda\mathbf{r}_j) \to (\mathbf{r}_j \leftrightarrow \mathbf{r}_k)$ never occurs. This means that even after rearrangement of the row operations, all steps of the form $\mathbf{r}_i \mapsto \mathbf{r}_i + \lambda\mathbf{r}_j$ will still have $j < i$, allowing us to read a lower triangular matrix off of them.

In any case, the first two steps of this new procedure from A to U accomplish a reordering of the bottom three rows of A, sending $\mathbf{r}_2 \mapsto \mathbf{r}_3 \mapsto \mathbf{r}_4 \mapsto \mathbf{r}_2$. The resulting matrix is

$$A' = \begin{bmatrix} 1 & 2 & -4 & 3 \\ -5 & -8 & 19 & -11 \\ 2 & 4 & -3 & 7 \\ 0 & -6 & 3 & -10 \end{bmatrix}$$

The remaining three steps show how to get from from A' to U. The matrix L such that $A' = LU$ can be read off this sequence of steps:

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -5 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 0 & -3 & 0 & 1 \end{bmatrix}$$

This factorization with initial row reordering is sometimes called a $PA = LU$ factorization or an *LU factorization with partial pivoting*.[3] P is the permutation matrix that induces the necessary reordering of the rows of A. In this case,

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & -4 & 3 \\ 2 & 4 & -3 & 7 \\ 0 & -6 & 4 & -10 \\ -5 & -8 & 19 & -11 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -5 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 0 & -3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & 2 & -1 & 4 \\ 0 & 0 & 5 & 1 \\ 0 & 0 & 0 & 2 \end{bmatrix}.$$

Unlike reduction to RREF, LU decomposition depends on the choice of row transpositions: different choices of P can drastically change L and U. Suppose that in the

[3] *Full pivoting* includes column rearrangements as well.

factorization of our matrix A above, once we had done $\mathbf{r}_2 \mapsto \mathbf{r}_2 - 2\mathbf{r}_1$ and $\mathbf{r}_4 \mapsto \mathbf{r}_4 + 5\mathbf{r}_1$ to get

$$\begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & 0 & 5 & 1 \\ 0 & -6 & 3 & -10 \\ 0 & 2 & -1 & 4 \end{bmatrix}$$

we swapped $\mathbf{r}_2 \leftrightarrow \mathbf{r}_3$ rather than $\mathbf{r}_2 \leftrightarrow \mathbf{r}_4$, getting

$$\begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & -6 & 3 & -10 \\ 0 & 0 & 5 & 1 \\ 0 & 2 & -1 & 4 \end{bmatrix}$$

In this case, the only remaining step to turn this matrix into upper-triangular form is $\mathbf{r}_4 \mapsto \mathbf{r}_4 + \frac{1}{3}\mathbf{r}_2$, giving

$$U = \begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & -6 & 3 & -10 \\ 0 & 0 & 5 & 1 \\ 0 & 0 & 0 & \frac{2}{3} \end{bmatrix}$$

The sequence of steps from A to U, namely

$$(\mathbf{r}_2 \mapsto \mathbf{r}_2 - 2\mathbf{r}_1) \to (\mathbf{r}_4 \mapsto \mathbf{r}_4 + 5\mathbf{r}_1) \to (\mathbf{r}_2 \leftrightarrow \mathbf{r}_3) \to (\mathbf{r}_4 \mapsto \mathbf{r}_4 + \frac{1}{3}\mathbf{r}_2)$$

can be rearranged to put all transpositions at the front:

$$(\mathbf{r}_2 \leftrightarrow \mathbf{r}_3) \to (\mathbf{r}_3 \mapsto \mathbf{r}_3 - 2\mathbf{r}_1) \to (\mathbf{r}_4 \mapsto \mathbf{r}_4 + 5\mathbf{r}_1) \to (\mathbf{r}_4 \mapsto \mathbf{r}_4 + \frac{1}{3}\mathbf{r}_2).$$

The resulting L factor can be read from this sequence of steps,

$$PA = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & -4 & 3 \\ 2 & 4 & -3 & 7 \\ 0 & -6 & 3 & -10 \\ -5 & -8 & 19 & -11 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ -5 & -\frac{1}{3} & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & -4 & 3 \\ 0 & -6 & 3 & -10 \\ 0 & 0 & 5 & 1 \\ 0 & 0 & 0 & \frac{2}{3} \end{bmatrix} = LU.$$

Finall, since row exchanges are equivalent to simply changing the order of equations in the corresponding linear system, they don't affect the solution to the system. $A\mathbf{x} = \mathbf{b}$ can be solved by choosing a permutation matrix P, decomposing $PA = LU$, and then solving $LU\mathbf{x} = P\mathbf{b}$.

Chapter 5

Subspace miscellany

Overview. This section is a grab-bag of results that will be useful in the following chapters of operator theory, or in their own right. Section 1.8.4 presents the *subspace intersection lemma*, which states that the total dimensions of two subspaces' sum and intersection must equal the total dimensions of the two subspaces themselves. This result is useful, for example, in determining whether two subspaces are sufficiently large, relative to the subspace that contains them, that they must have a nontrivial intersection.

Section 5.2 presents the key notion of a *direct sum* of two or more subspaces, along with several equivalent characterizations. Essentially, a subspace sum is direct when there are no redundant dimensions in the sum, and none of the subspaces could be made smaller without also making the sum smaller.

Section 5.3 introduces the notion of a sum of affine spaces or cosets. This leads into Section 5.4, which introduces the idea of a *quotient space*: the collection of cosets of any subspace can be made into a vector space in its own right, with arithmetic operations derived from the larger space. Understanding quotient spaces can be tricky at first, but they're worthwhile first bcause they provide the basis for other important constructions in more advanced linear algebra; second, because they allow more elegant proofs of several key results (the final section, 5.5, gives an alternate proofs of the rank–nullity theorem); and, finally, because other branches of higher mathematics use analogous quotient constructions frequently.

5.1 Dimensions of subspace intersections and sums

Key questions.

1. What is the subspace intersection lemma?

2. ($\star$) Use the subspace intersection lemma to prove that for two subspaces V, W of a larger space U, $\dim(V + W) \leq \dim V + \dim W$ and $\operatorname{codim}(V \cap W) \leq \operatorname{codim} V + \operatorname{codim} W$. (You can assume that U is finite-dimensional.) Give an example of vector spaces U, V, W for which which equality holds in both of these relationships simultaneously (that is, $\dim(V + W) = \dim V + \dim W$ and $\operatorname{codim}(V \cap W) = \operatorname{codim} V + \operatorname{codim} W$).

In section 1.6, we learned that the intersection of any two subspaces is also a subspace. This is not usually true for unions, because the sum of elements from two dif-

ferent subspaces does not have to be contained in either subspace. But the sum of two subspaces—the set of all sums of an element from one subspace and an element from another—is analogous, in certain contexts, to the union of sets. In particular, a key formula that relates the sizes of intersections and unions of sets has a closely corresponding formula that relates the dimensions of intersections and sums of subspaces.

5.1.1 Symmetry of intersection and sum; analogy between sum and set union

Let's make this analogy more precise. The intersection $S \cap T$ of two generic sets S and T is the largest subset of both S and T—"largest" in the sense that if $X \subseteq S$ and $X \subseteq T$, then $X \subseteq S \cap T$. Similarly, the union S and T is the smallest superset of both S and T: if $S \subseteq X$ and $T \subseteq X$, then $S \cup T \subseteq X$.

Similarly, the intersection $W_1 \cap W_2$ of two subspaces of a vector space V is the largest subspace contained in both W_1 and W_2: if X is a subspace of W_1 and W_2, then $X \subseteq W_1 \cap W_2$. Symmetrically, $W_1 + W_2$ is the smallest subspace that contains both W_1 and W_2: if X is a subspace that also contains both W_1 and W_2, then $W_1 \subset W_2$.

5.1.2 Subspace intersection lemma

One basic result from elementary set theory, called the *inclusion–exclusion principle*, relates the size of the intersection and union of any two sets to the size of the sets themselves: if S and T are any sets, then $|S \cup T| = |S| + |T| - |S \cap T|$. (Remember that $|S|$ means the number of elements in S.) It should be easy to convince yourself that this is true: if you calculate $|S \cup T|$ by counting the elements in S and in T separately and adding the results, then any element of $S \cap T$ is double-counted, and you can correct for this double counting by subtracting $|S \cap T|$.

There is an analogous result for subspaces: if V and W are two finite-dimensional subspaces of a larger space, then $\dim(V + W) = \dim V + \dim W - \dim(V \cap W)$—the dimensions of any two subspaces add up to the sum of the dimensions of their intersection and sum. The proof of this result simply follows from inclusion–exclusion for ordinary sets and the following two results.

Proposition. *Suppose that S and T are two linearly independent sets that do not have vectors in common. Then $S \cup T$ is linearly independent if and only if* $\operatorname{span} S \cap \operatorname{span} T = \{\mathbf{0}\}$.

Proof. Suppose $S \cup T$ is not linearly independent: that is, there is some nontrivial combination $a_1\mathbf{s}_1 + \cdots + a_m\mathbf{s}_m + b_1\mathbf{t}_1 + \cdots + b_n\mathbf{t}_n\mathbf{0}$ with $\mathbf{s}_1, \ldots, \mathbf{s}_m \in S$ and $\mathbf{t}_1, \ldots, \mathbf{t}_n \in S$. Then $a_1\mathbf{s}_1 + \cdots + a_m\mathbf{s}_m$, which is an element of $\operatorname{span} S$, equals $-b_1\mathbf{t}_1 - \cdots - b_n\mathbf{t}_n$, which is an element of $\operatorname{span} T$. These linear combinations can't equal $\mathbf{0}$ (because this would contradict the linear independence of either S or T), so there's a nonzero element of $\operatorname{span} S \cap \operatorname{span} T$.

Conversely, if $\operatorname{span} S \cap \operatorname{span} T$ contains a nonzero element, then writing this element as a linear combination of S and as a linear combination of T, and then subtracting one linear combination from the other, gives a linear combination of $S \cup T$ that equals zero. This linear combination is necessarily nontrivial: since S and T don't share vectors, subtracting the T combination from the S combination can't cancel out coefficients on the same vector. □

Remark. We need the stipulation that S and T not have vectors in common to avoid counterexamples such as if S and T are the same nonempty, linearly independent set: in this case, $\operatorname{span} S \cap \operatorname{span} T = \operatorname{span} S = \operatorname{span} T \neq \{\mathbf{0}\}$ but $S \cup T$ is linearly independent. We can drop this stipulation if we interpret S, T, and $S \cup T$ not as sets but rather as

multisets—that is, sets that can have duplicate elements (and taking the union of two sets with common elements means that the resulting set has duplicates), such that any multiset that contains a duplicate element is not linearly independent.

In general, when we talk about bases and linearly independent sets, we will typically actually mean multisets, even though we use set notation. For instance, if $\mathbf{v}_1 = \mathbf{v}_2 = (1, 0) \in \mathbb{R}^2$ and $\mathbf{v}_3 = (0, 1) \in \mathbb{R}^2$, then if we interpreted set notation pedantically, we would be forced to say that $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is a linearly independent set with only two elements: $\mathbf{v}_1$ and $\mathbf{v}_2$ are two ways to write the same elements, and sets by definition can't contain duplicates, so $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is just a redundant way of writing $\{(1, 0), (0, 1)\}$. But in practice, when we're judging the linear independence of a set, we'll interpret notation such as $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ as a multiset, not a set (thus, in this case, $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is a linearly dependent multiset with three elements). This convention will help us avoid littering theorem statements with special cases for when multiple elements in a constructed set might turn out to be the same, and shouldn't be too confusing.

Lemma. *Let V and W be two (possibly infinite-dimensional) subspaces of some larger space U. Then there exist bases of V and W whose intersection is a basis of $V \cap W$ and whose union is a basis of $V + W$.*

Proof. Let A be a basis of $V \cap W$. Extend A to bases of V and W by choosing sets $B \subset V$ and $C \subset W$, both disjoint from A, such that $A \cup B$ is a basis of V and $A \cup C$ is a basis of W. (Remember our discussion on page 48.) We're guaranteed that B and C are disjoint however we choose them, becaues any vector that they shared would also be in $V \cap W = \operatorname{span} A$, making $A \cup B$ and $A \cup C$ no longer linearly independent.

Let's prove that $A \cup B \cup C$ is a basis of $V + W$. We already know that $A \cup B \cup C$ spans $V + W$ (because $\operatorname{span} S_1 + \operatorname{span} S_2 = \operatorname{span}(S_1 \cup S_2)$), so we just have to prove that it's linearly independent. To show this (by the propsition we just established), it's enough to show that $V = \operatorname{span}(A \cup B)$ shares no vectors with $\operatorname{span} C$ besides $\mathbf{0}$.

Suppose that there's some nonzero vector $\mathbf{w} \in \operatorname{span} C$ that's also in $\operatorname{span}(A \cup B) = V$. Then since $\operatorname{span} C \subseteq W$, so $\mathbf{w} \in V \cap W = \operatorname{span} A$. So $\operatorname{span} A$ and $\operatorname{span} C$ have a nonzero common vector $\mathbf{w}$, so $A \cup C$ is not linearly independent. But $A \cup C$ was defined to be a basis for W, so it must be linearly independent, a contradiction. Thus, no nonzero vector $\mathbf{w} \in \operatorname{span}(A \cup B) \cap \operatorname{span} C$ can exist, so $A \cup B \cup C$ is linearly independent.

Thus, we have bases $A \cup B$ for U and $A \cup C$ for W, whose intersection A is a basis of $V \cap W$ and whose union $A \cup B \cup C$ is a basis of $V + W$.

□

With our preliminary results in hand, we can finally prove:

Lemma (subspace intersection lemma)**.** *If V and W are two subspaces of a common larger space, then* $\dim(V + W) = \dim V + \dim W - \dim(V \cap W)$.

Proof. Let B_1 and B_2 be bases for V and W such that $B_1 \cap B_2$ and $B_1 \cup B_2$ are bases for $V \cap W$ and $V + W$, we have

$$\begin{aligned} \dim(V + W) &= |B_1 \cup B_2| \\ &= |B_1| + |B_2| - |B_1 \cap B_2| && \text{(inclusion–exclusion for sets)} \\ &= \dim V + \dim W - \dim(V \cap W). \end{aligned}$$

□

Intuitively, you can interpret the subspace intersection lemma like this: when we compute the sum $V + W$, we're taking V and adding to it all the dimensions of W that are missing from V. That is, we're adding all the dimensions of W minus the ones that it already shares with V—that is, $\dim W - \dim(V \cap W)$.

Though the inclusion–exclusion principle has a generalization to three or more sets, the subspace intersection lemma does not. For instance, for ordinary sets,

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|.$$

But in general, for vector subspaces,

$$\begin{aligned}\dim(U + V + W) \neq \dim U + \dim V + \dim W - \dim(U \cap V) - \dim(U \cap W) \\ - \dim(V \cap W) + \dim(U \cap V \cap W).\end{aligned}$$

For instance, consider the case where U, V, W are three distinct one-dimensional subspaces of $\mathbb{R}^2$—say, $U = \operatorname{span}\{\mathbf{e}_1\}$, $V = \operatorname{span}\{\mathbf{e}_2\}$, and $W = \operatorname{span} \mathbf{e}_1 + \mathbf{e}_2$. Then $\dim(U + V + W) = \dim(\mathbb{R}^2) = 2$, but the intersection of two or more of U, V, W is $\{\mathbf{0}\}$ and has dimension 0, so the right-hand formula is 3.

5.1.3 Subspace intersection lemma for codimensions

Some slight algebraic manipulation of the subspace intersection lemma $\dim V + \dim W = \dim(V \cap W) + \dim(V + W)$ gives the consequence $(\dim U - \dim V) + (\dim U - \dim W) = (\dim U - \dim(V \cap W)) + (\dim U - \dim(V + W))$ for any space U that includes both V and W; that is, $\operatorname{codim} V + \operatorname{codim} W = \operatorname{codim}(V \cap W) + \operatorname{codim}(V + W)$. This turns out to be the same formula as before, only with codim replacing $\dim$.

5.1.4 Possible ranges of subspace dimensions

The subspace intersection lemma $\dim V + \dim W = \dim(V \cap W) + \dim(V + W)$ (and the equivalent statement for codimensions) means that if we know the dimensions of three of the spaces V, W, $V \cap W$, and $V + W$, we can find the dimension of the fourth. If we only know the dimensions of two of these spaces, we can still find possible ranges for the dimensions of the other spaces, using these facts:

1. $\dim(V \cap W)$ is at least zero and most the smaller of $\dim V$ and $\dim W$, so $\dim(V + W)$ is at most $\dim V + \dim W$ and at least the larger of $\dim V$ and $\dim W$.

2. $\dim(V + W)$ can't exceed the dimension of any larger space U that contains V and W, so $\dim(V \cap W) \geq \dim V + \dim W - \dim U$. We can rewrite this with codimensions as $\operatorname{codim}(V \cap W) \leq \operatorname{codim} V + \operatorname{codim} W$; remember that $\operatorname{codim} X = \dim U - \dim X$ if X is a subspace of U.

A few examples:

1. If $\dim V = 6$, $\dim W = 4$, and $\dim(V + W) = 9$, then what is the dimension of $V \cap W$? In this case, using the subspace intersection lemma gives an immediate answer: $\dim(V \cap W) = \dim V + \dim W - \dim(V + W) = 1$.

2. If $\dim V = 4$, $\dim(V \cap W) = 3$, and $\dim(V + W) = 7$, then what is $\dim W$? Again, the subspace intersection lemma gives an answer: $\dim W = \dim(V \cap W) + \dim(V + W) - \dim V = 6$.

3. Suppose $\dim V = 5$ and $\dim W = 3$. Furthermore, V and W are both subspaces of some space U, and $\dim U = 7$. Then $\dim(V + W)$ can range from $\max(\dim V, \dim W)$ at the low end to $\min(\dim V + \dim W, \dim U)$ at the high end. In this case, $\dim U = 7 < \dim V + \dim W = 8$, so $\dim(V + W)$ can equal 5, 6, or 7. The corresponding possible values of $\dim(V \cap W)$ are 3, 2, and 1.

4. Suppose $\dim V = 5$, $\dim(V \cap W) = 2$, and the enclosing space U has dimension 9. What are the possible dimensions of $\dim W$? We know that $\dim(V + W)$ has to be at least $\dim V = 5$ and at most $\dim U = 9$. As $\dim W = \dim(V \cap W) + \dim(V + W) - \dim V = \dim(V + W) - 3$, so $V + W$ can have any integer dimension from 5 through 9, and $\dim W$ can correspondingly have any dimension from 2 through 6.

Most problems you'll see in practice will give you $\dim V$ and $\dim W$, so you can use this pair of nicely symmetrical results, where U is any space that includes both V and W:

1. $\dim(V + W)$ is at least $\max(\dim V, \dim W)$ and at most $\min(\dim U, \dim V + \dim W)$
2. $\operatorname{codim}(V \cap W)$ is at least $\max(\operatorname{codim} V, \operatorname{codim} W)$ and at most $\min(\operatorname{codim} U, \operatorname{codim} V + \operatorname{codim} W)$.

Answers to key questions.

1. The subspace intersection lemma states that for any two subspaces V, W of a larger space U, $\dim(V + W) + \dim(V \cap W) = \dim V + \dim W$.
2. The result $\dim(V + W) \leq \dim V + \dim W$ follows immediately from the subspace intersection lemma $\dim(V + W) = \dim V + \dim W - \dim(V \cap W)$ and the fact that $\dim(V \cap W) \geq 0$.

   From the alternate formulation $\operatorname{codim}(V \cap W) = \operatorname{codim} V + \operatorname{codim} W - \operatorname{codim}(V + W)$ and the trivial fact $\operatorname{codim}(V + W) \geq 0$, we get the reuslt $\operatorname{codim}(V \cap W) = \operatorname{codim} V + \operatorname{codim} W$.

   An example for which $\dim(V + W) = \dim V + \dim W$ and $\operatorname{codim}(V \cap W) = \operatorname{codim} V + \operatorname{codim} W$ has to be one in which $\dim(V \cap W) = \operatorname{codim}(V + W) = 0$; that is, $V \cap W = \{\mathbf{0}\}$ and $V + W = U$. You can create any number of examples by taking a basis of some arbitrary vector space U, letting V be the span of some subset of this basis, and letting W be the span of the remaining basis; for instance, $U = \mathbb{R}^4, V = \operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2\}, W = \operatorname{span}\{\mathbf{e}_3, \mathbf{e}_4\}$.

5.2 Direct sums

Key questions

1. What does it mean for a sum of two subspaces to be direct? Give three equivalent criteria: one involving the number of ways to write a vector in the subspace sum as a sum of one vector from each subspace, one involving dimensions of the subspace sum, and one involving dimensions of the subspace intersection.
2. Which of the criteria in question 1 can you generalize to give a definition of a direct sum of three or more subspaces?
3. Define the subspaces $W_1 = \operatorname{span}\{(1, 0, 0)\}$, $W_2 = \operatorname{span}\{(1, 0, 0), (1, 1, 0)\}$, $W_3 = \operatorname{span}\{(0, 0, 1)\}$, and $W_4 = \operatorname{span}\{(0, 2, 1), (0, 1, 1)\}$. Is $W_1 + W_2$ a direct sum? Is $W_1 + W_3$? Is $W_1 + W_4$? Is $W_2 + W_4$?

In section 1.6.2, we defined the sum of subspaces. If $V_1, \ldots, V_n$ are multiple subspaces of some larger vector space, then $V_1 + \ldots + V_n$ is the smallest subspace that includes all of $V_1, \ldots, V_n$ as subspaces, or (equivalently) the set of all sums of one element out of each of the spaces V_n. Then in section 5.1, we noted that the sum of two spaces $V_1 + V_2$ had dimension at most $\dim V_1 + \dim V_2$—and that this maximum dimension was only possible if V_1 and V_2 had no vectors in common besides $\mathbf{0}$.

In this section, we'll look at other properties of pairs of subspaces V_1, V_2 whose sum has this maximum possible dimension $\dim V_1 + \dim V_2$, and that have the maximum possible dimension given the dimensions of the individual subspaces. Essentially, these sums have no "redundant" dimensions: every subspace expands the sum by as much as its dimension allows. These sums are called *direct sums*, and they have the additional important property that every element in the sum can be written in only one way as the sum of one vector from each of the constituent subspaces. In fact, as we'll see, these two characterizations are logically equivalent.

We'll put a definition up front. This definition may seem unintuitive for now; the rest of the chapter will be devoted to explaining its usefulness.

5.2.1 Direct sums of two spaces

Let's start with an illustrative example. We'll define three subspaces U, V, W of $\mathbb{R}^3$:

- $U \coloneqq \text{span}\{\mathbf{e}_1, \mathbf{e}_2\}$ is the set of vectors of the form $(x, y, 0)$ (that is, the xy-plane).
- $V \coloneqq \text{span}\{\mathbf{e}_2 + \mathbf{e}_3\}$ is the set of vectors of the form $(0, y, y)$ (that is, the line $x = 0, y = z$.)
- $W \coloneqq \text{span}\{\mathbf{e}_2, \mathbf{e}_3\}$ is the set of vectors of the form $(0, y, z)$ (that is, the yz-plane).

The subspace sums $U + V$ and $U + W$ both equal $\mathbb{R}^3$. Any vector $(x, y, z) \in \mathbb{R}^3$ can be decomposed into a sum of an element of U and an element of V, in exactly one way: $(x, y-z, 0)+(0, z, z)$. But the same vector can be decomposed into many different sums of an element of U and an element of W: $(x, y, 0) + (0, 0, z)$ and $(x, 0, 0) + (0, y, z)$ and $(x, \frac{3}{2}y, 0) + (0, -\frac{1}{2}y, z)$ are two of an infinite number of possibilities.

Also note these different properties of $U + V$ and $U + W$:

1. $\dim(U + V)$ equals $\dim U + \dim V$, but $\dim(U + W)$ is less than $\dim U + \dim W$.
2. U and V intersect only at $\{\mathbf{0}\}$, but the intersection of U and W is a nonzero subspace of $\mathbb{R}^3$, namely $\text{span}\{\mathbf{e}_2\}$.
3. Elements of $U + V$ can be decomposed as $\mathbf{u} + \mathbf{v}$ for some $\mathbf{u} \in U, \mathbf{v} \in V$ in an only one way. By contrast, elements of $U + W$ can be similarly decomposed into $\mathbf{u} + \mathbf{w}$ in an infinite number of ways.

Mathematicians invented the term *direct sum* to describe these differences: the sum $U + V$ is direct, but $U + W$ isn't. Direct sums can be notated with a special symbol $\oplus$: thus we can write $\mathbb{R}^3 = U \oplus V$.

Specifically, a sum $U + V$ of two subspaces is called *direct*, and can be notated $U \oplus V$, if any of these logically equivalent conditions holds:

1. $\dim(U + V) = \dim U + \dim V$.

2. $U \cap V = \{\mathbf{0}\}$.

3. *For any* vector $\mathbf{w} \in U + V$, there is only one ordered pair of elements $(\mathbf{u}, \mathbf{v})$ where $\mathbf{u} \in U$ and $\mathbf{v} \in V$ such that $\mathbf{w} = \mathbf{u} + \mathbf{v}$.

4. The only vectors $\mathbf{u} \in U, \mathbf{v} \in V$ for which $\mathbf{u} + \mathbf{v} = \mathbf{0}$ are $\mathbf{u} = \mathbf{v} = \mathbf{0}$.

5. *There exists* a vector $\mathbf{w} \in U + V$ that can be written in only one way as $\mathbf{u} + \mathbf{v}$ for $\mathbf{u} \in U, \mathbf{v} \in V$.

Proposition. *These conditions are logically equivalent.*

Proof. We'll prove the implications $1 \iff 2$, $2 \iff 5$, and $3 \implies 4 \implies 5 \implies 3$. Some portions of this argument may remind you of the proof of equivalence of the multiple notions of linear independence in section 1.7.2.

- *Equivalence of 1 and 2*: the subspace intersection lemma. (Remember that $\dim\{\mathbf{0}\} = 0$.)
- *5 implies 2*: Suppose that there are some distinct vectors $\mathbf{u}_1, \mathbf{u}_2 \in U$ and $\mathbf{v}_1, \mathbf{v}_2 \in V$ such that $\mathbf{u}_1 + \mathbf{v}_1 = \mathbf{u}_2 + \mathbf{v}_2$. Then $\mathbf{u}_1 - \mathbf{u}_2 = \mathbf{v}_2 - \mathbf{v}_1$ is a nonzero element of both U and V.
- *2 implies 5*: Suppose that there's some nonzero $\mathbf{w} \in U \cap V$. Then $(\mathbf{0}, \mathbf{w})$ and $(\mathbf{w}, \mathbf{0})$ are two distinct ordered pairs of elements from U and V that add up to $\mathbf{w}$.
- *3 implies 4 and 4 implies 5*: obvious. The progression from "for all" to "for $\mathbf{0}$" to "for any" is a logical weakening at each step.
- *5 implies 3*: We'll prove that not-3 implies not-5. Suppose we have two ways $\mathbf{w}' = \mathbf{u}_1 + \mathbf{v}_1 = \mathbf{u}_2 + \mathbf{v}_2$ of writing some particular element $\mathbf{w}'$ as a sum of an element of U and an element of V. Then $(\mathbf{u}_1 - \mathbf{u}_2) + (\mathbf{v}_1 - \mathbf{v}_2) = \mathbf{0}$, so given any other element $\mathbf{w} = \mathbf{u} + \mathbf{v} \in U + V$, we can also write $\mathbf{v} = (\mathbf{u} + \mathbf{u}_1 - \mathbf{u}_2) + (\mathbf{v} + \mathbf{v}_1 - \mathbf{v}_2)$.

□

As a final note, given a space V and a subspace W, it's always possible to find another subspace W' such that $V = W \oplus W'$. The prodcedure is simple: start with a basis of W, extend it to a basis of V (remember section 1.8.4), and define W' to be the span of these new vectors.

5.2.2 Direct sums of three or more spaces

The definition of direct sums generalizes naturally to three or more subspaces. We'll make this a definition:

Definition. *Let $V_1, \ldots, V_n$ be subspaces of the same space, with the property that the only way to choose pairs of vectors $\mathbf{v}_1, \mathbf{v}_1' \in V_1, \ldots, \mathbf{v}_n, \mathbf{v}_n' \in V_n$ from each subspace such that $\mathbf{v}_1 + \mathbf{v}_2 + \cdots + \mathbf{v}_n = \mathbf{v}_1' + \cdots + \mathbf{v}_n'$ is by choosing $\mathbf{v}_1 = \mathbf{v}_1', \cdots, \mathbf{v}_n = \mathbf{v}_n'$. Then the sum $V_1 + \cdots + V_n$ is a **direct sum**, and can be written with the special symbol $V_1 \oplus \cdots \oplus V_n$.*

Remark. This definition extends to the ability to define *infinite* subspace sums and direct sums $V_1 + V_2 + \cdots$, with the caveat that if you choose one vector from every component of an infinite subspace sum, then all but a finite number of the vectors that you choose have to be $\mathbf{0}$, because there's no notion of infinite sums in general vector spaces.

Put more intuitively: the sum $V_1+\cdots+V_n$ is direct if for every element $\mathbf{v} \in V_1+\cdots+V_n$, there is only one choice of vectors $\mathbf{v}_1 \in V_1, \cdots, \mathbf{v}_n \in V_n$ such that $\mathbf{v} = \mathbf{v}_1 + \cdots + \mathbf{v}_n$. For a sum of three or more spaces to be direct, it is necessary but not sufficient that $V_i \cap V_j = \{\mathbf{0}\}$ for any pair of indices $1 \leq i, j \leq n$. (For a counterexample, consider the subspaces $V_1 = \text{span}\{\mathbf{e}_1\}$, $V_2 = \text{span}\{\mathbf{e}_2\}$ and $V_3 = \text{span}\{\mathbf{e}_1 + \mathbf{e}_2\}$ in $\mathbb{R}^2$—geometrically: the x-axis, the y-axis, and the line $y = x$. Then $V_1 \cap V_2 = V_1 \cap V_3 = V_2 \cap V_3 = \{\mathbf{0}\}$ but the sum $V_1 + V_2 + V_3 = \mathbb{R}^2$ is not direct.)

Some of the alternate formulations, however, do hold (or can be adapted) to the sum of an arbitrary finite number of subspaces. (With slight adaptation, we could make them apply to sums of infinitely many subspaces, but again, we won't need the theory of infinite sums here.)

Proposition. *The following conditions on the subspace sum $V_1 + \cdots + V_n$ are equivalent:*

1. *$V_1 + \cdots + V_n$ is direct as defined above: no element can be written in two ways as the sum of one element from each space.*
2. *The only way to write $\mathbf{0}$ as the sum of one element from each of $V_1, \ldots, V_n$ is to choose $\mathbf{0}$ from each space.*
3. *There exists an element of $V_1 + \cdots + V_n$ that can only be written in one way as the sum of one element from each space.*
4. For any *list of bases $B_1, \ldots, B_n$ for $V_1, \ldots, V_n$, the union $B_1 \cup \cdots \cup B_n$ is a basis of $V_1 + \cdots + V_n$. (The union of bases is taken to mean a multiset: i.e. if two of the B_i contain the same vector, then the union contains two copies of the same vector and isn't a basis. In particular, this criterion means that no possible choice of bases $B_1, \ldots, B_n$ contains two bases that share an element.)*
5. There exists *a list of bases $B_1, \ldots, B_n$ of $V_1, \ldots, V_n$ such that $B_1 \cup \cdots \cup B_n$ is a basis of $V_1 + \cdots + V_n$.*

Proof. Remmeber from section 1.6.3 that if S_1 and S_2 are spanning sets of V_1 and V_2, then $S_1 \cup S_2$ is a spanning set of $V_1 \cup V_2$. The generalization to three or more sets should be relatively obvious. So to prove that $B_1 \cup \cdots \cup B_n$ is a basis of $V_1 + \cdots + V_n$ in statements 4 and 5, we only need to show that $B_1 \cup \cdots \cup B_n$ is linearly independent.

Now, on to proving the equivalence of each statement:

- *Equivalence of 1, 2, and 3*: straightforward adaptations of the equivalent portion of the proof for direct sums of two subspaces in the last subsection.
- *2 implies 4*: Suppose that not-4: that is, there exists a collection of bases $B_1, \ldots, B_n$ of $V_1, \ldots, V_n$ whose union is not linearly independent. Then if we take any nontrivial linear combination from $B_1 \cup \cdots \cup B_n$ with value $\mathbf{0}$, then we can divide it into a linear combination of n "constituent" linear combinations: one constituent is the terms from B_1 (call the total value of these terms $\mathbf{v}_1$), another is the linear terms from B_2 (call the total value of these terms $\mathbf{v}_2$), and so on. Since the complete linear combination is nontrivial, at least one of its constituents is nontrivial, and since each B_i is linearly independent, the value $\mathbf{v}_i$ of any nontrivial constituent has to be nonzero. So $\mathbf{0} = \mathbf{v}_1 + \cdots + \mathbf{v}_n$ is a sum of one element from each V_i that includes at least one nonzero term; that is, not-2.

- *4 implies 5*: Every vector space has a basis, so "for any basis ..." implies "there exists a basis such that ...".

- *5 implies 2*. Let $B_1, \ldots, B_n$ be arbitrary bases and suppose that not-2: that is, we can write $\mathbf{0} = \mathbf{v}_1 + \cdots + \mathbf{v}_n$ where $\mathbf{v}_i \in V_i$ for all indices i and at least one of the $\mathbf{v}_i$ is not zero. Then decomposing each $\mathbf{v}_i$ into a linear combination of B_i gives a nontrivial linear combination from $B_1 \cup \cdots \cup B_n$ that $\mathbf{0}$; that is, no such union of bases can be linearly independent; that is, not-5.

□

As an immediate corollary, we finally have the initial intuition about the effects of direct sums on subspace dimensions:

Corollary. $\dim(V_1 \oplus \cdots \oplus \dim V_n) = \dim V_1 + \cdots + \dim V_n$.

Proof. Taking direct sums of subspaces means taking unions of their bases, and taking unions of sets without duplicate elements means adding their sizes.

□

Answers to key questions.

1. Criteria for a sum of two subspaces to be direct: (1) every vector in the sum can be written as the sum of one vector from each subspace in exactly one way; (2) the dimension of the sum is the sum of the dimensions of the subspaces; (3) the dimension of the subspace intersection is zero.

2. Criteria (1) and (2) generalize in clear ways, but not criterion (3): it is possible to have a set of three or more subspaces whose intersection is zero (indeed, such that the intersection of every pair of subspaces is zero), but such that the other criteria for a direct sum are not satisfied.. One counterexample is three or more distinct one-dimensional subsets of $\mathbb{R}^2$.

3. $W_1 + W_3$ and $W_1 + W_4$ are direct, but $W_1 + W_2$ and $W_3 + W_4$ are not. The easiest way to tell is to compute the dimensions of the subspace sums.

5.3 Sums and intersections of affine spaces

Unlike regular subspaces, which all at least contain $\mathbf{0}$, affine subspaces don't necessarily have any elements in common. As one simple example, if $\mathbf{b}_0 \notin W + \mathbf{a}_0$, then $W + \mathbf{a}_0$ and $W + \mathbf{b}_0$ have no vectors in common: two cosets of the same subspace either are identical or don't intersect at all.

As a less trivial example, consider the one-dimensional affine subspaces of $\mathbb{R}^3$ (that is, lines) $(0, 0, 1) + \text{span}\{(1, 1, 1)\}$ and $(0, 2, 0) + \text{span}\{(1, 0, 0)\}$. The first contains all vectors of the form $(a, a, a + 1)$; the second contains all vectors of the form $(b, 2, 0)$, and no values of a and b can make these equal.

The sum of affine spaces can be defined by analogy to the sum of regular subspaces: $A + B = \{\mathbf{a} + \mathbf{b} : \mathbf{a} \in A, \mathbf{b} \in B\}$. We can also define the dimension of an affine space as the dimension of the parallel subspace.

The sums and intersections of affine spaces—as well as the conditions under which the intersection of two affine subspaces exists—can be precisely characterized. The key result is the first: the sum of cosets of two spaces V, W is a coset of $V + W$; this will be useful for the following section on quotient spaces. The result on intersections of affine spaces is not quite as useful, but included for completeness.

Proposition. *Let V and W be two subspaces of a vector space U, and let $A \coloneqq \mathbf{a}_0 + V$ and $B \coloneqq \mathbf{b}_0 + W$ be two parallel affine spaces. Then:*

1. $A + B = (\mathbf{a}_0 + \mathbf{b}_0) + (V + W)$.
2. *Define $A - B \coloneqq \{\mathbf{a} - \mathbf{b} : \mathbf{a} \in A, \mathbf{b} \in B\}$, the set of differences of an element in A and an element of B. Then either: (a) $A - B = V + W$ and $A \cap B$ is a coset of $V \cap W$, or (b) $(A - B) \cap (V + W) = \emptyset$ and $A \cap B = \emptyset$.*

Proof.

1. By definition, $A + B$ is the set of vectors of the form $(\mathbf{a}_0 + \mathbf{v}) + (\mathbf{b}_0 + \mathbf{w})$ for $\mathbf{v} \in V, \mathbf{w} \in W$, and $(\mathbf{a}_0+\mathbf{b}_0)+(V+W)$ is the set of vectors of the form $(\mathbf{a}_0+\mathbf{b}_0)+(\mathbf{v}+\mathbf{w})$. These expressions are the same sum, just parenthesized differently.
2. The set of negatives of the elements in $\mathbf{b}_0+W$ is $(-\mathbf{b}_0)+W$ (because the negative of $\mathbf{b}_0 + \mathbf{w}$ is $-\mathbf{b}_0 + (-\mathbf{w})$, and $\mathbf{w} \in W$ if and only if $-\mathbf{w} \in W$ as well). So by the formula that we established in statement 1 of this proposition, $A - B$ equals $(\mathbf{a}_0-\mathbf{b}_0)+(V+W)$; that is, it's a coset of $V+W$, so it either equals $V+W$ or doesn't intersect it at all. Let's look at each case separately, and note that $\mathbf{0} \in A-B$ if and only if A and B have an element in common.
   - Case a: $A - B = V + W$. In particular, since $\mathbf{0} \in V + W$, so $\mathbf{0} \in A - B$, so $A \cap B$ contains at least one element (call it $\mathbf{c}$). The choice of base point for a coset is arbitrary, so we can write $A = \mathbf{c}+V = \{\mathbf{c}+\mathbf{v} : \mathbf{v} \in V\}$ and, similarly, $B = \{\mathbf{c} + \mathbf{w} : \mathbf{w} \in W\}$. So the elements in $A \cap B$ are the ones that we can write both as $\mathbf{c} + \mathbf{v}$ and as $\mathbf{c} + \mathbf{w}$. That is, $A \cap B = \mathbf{c} + (V \cap W)$.
   - Case b: $A - B$ and $V + W$ are disjoint. Then $\mathbf{0} \in V + W$ (because $V + W$ is a vector subspace), so $\mathbf{0} \notin A - B$, so A and B are disjoint.

□

The result $A+B = (\mathbf{a}_0+\mathbf{b}_0)+(V+W)$ does not depend on the choice of base point. If we chose to write A and B with different base points as $A = \mathbf{a}'_0 + V$ and $B = \mathbf{b}'_0 + W$, where $\mathbf{a}'_0 - \mathbf{a}_0 \in V$ and $\mathbf{b}'_0 - \mathbf{b}_0 \in W$, then $(\mathbf{a}'_0 + \mathbf{b}'_0) - (\mathbf{a}'_0 + \mathbf{b}_0)$ is an element of $V + W$, and $\mathbf{a}'_0 + \mathbf{b}'_0 + V + W = \mathbf{a}_0 + \mathbf{b}_0 + V + W$.

Answers to key questions.

1. Geometric intuition: A is an affine space parallel to W if it's just a translated copy of W.

   A is an affine space parallel to W if for every element $\mathbf{a}_0 \in A$,

   (a) The set of differences $\{\mathbf{a} - \mathbf{a}_0 : \mathbf{a} \in A\}$ equals W.

(b) The set of sums $\{\mathbf{w} + \mathbf{a}_0 : \mathbf{w} \in W\}$ equals W.

We can get equivalent characterizations by replacing "for every element $\mathbf{a}_0 \in A$" with "there exists an element $\mathbf{a}_0 \in A$ such that ..." and using the same statements as in the list.

2. Using the definition that a coset of W is the set $\{\mathbf{w} + \mathbf{a}_0 : \mathbf{w} \in W\}$ and choosing $\mathbf{a}_0 = \mathbf{0}$ shows that W is a coset of itself.

3. Two different cosets can't have vectors in common. If A, B are two cosets of W, then they can be written in the forms $\mathbf{a}_0 + W$ and $\mathbf{b}_0 + W$ where $\mathbf{a}_0$ and $\mathbf{b}_0$ are *arbitrary* elements of A and B. So if A and B share an element, then we can choose that element as both $\mathbf{a}_0$ and $\mathbf{b}_0$ and give identical expressions for a and B.

4. The statement is true. If A is a coset of W_1 that is contained in W_2, then since W_2 is closed under addition (and therefore subtraction), it must contain every difference between two elements of A, and the set of these differences is W_1.

5. If A is a coset of some subspace W, then $\{\mathbf{a}_1 - \mathbf{a}_2 : \mathbf{a}_1 \in A\}$ must equal W for any *particular* $\mathbf{a}_2 \in A$, so the union of all such sets—that is, $\{\mathbf{a}_1 - \mathbf{a}_2 : \mathbf{a}_1, \mathbf{a}_2 \in A\}$ for arbitrary $\mathbf{a}_2$—is also W.

   For a counterexample to the converse statement: $A = \{(x, 1) : x > 0\}$ is not a subspace of $\mathbb{R}^2$, but $\{\mathbf{a}_1 - \mathbf{a}_2 : \mathbf{a}_1, \mathbf{a}_2 \in A\}$ is a subpsace (namely $\operatorname{span}\{\mathbf{e}_1\}$).

6. If $A_1 = \mathbf{a}_1 + W_1$ and $A_2 = \mathbf{a}_2 + W_2$, then $A_1 + A_2 = (\mathbf{a}_1 + \mathbf{a}_2) + (W_1 + W_2)$. So $A_1 + A_2$ could have any of the possible dimensions of W_1 and W_2: that is, any integer between $\max(\dim W_1, \dim W_2) = 7$ and $\min(\dim \mathbb{R}^{10}, \dim W_1 + \dim W_2 = 10$.

   $A_1 \cap A_2$ is either empty or a coset of $W_1 \cap W_2$. The possible codimensions of $W_1 + W_2$ range from $\max(\operatorname{codim} W_1, \operatorname{codim} W_2) = 4$ to $\min(\dim \mathbb{R}^{10}, \operatorname{codim} W_+ \operatorname{codim} W_2) = 7$, so the possible dimensions range from 3 to 6.

5.4 Quotient spaces

Key questions.

1. What is a *relation* on a set? What properties must a relation satisfy to be an *equivalence relation*? What is the relationship between equivalence relations and equivalence classes?

2. ($\star$) Define the two relations $a \sim_1 b$ and $a \sim_2 b$ on the set of integers $\mathbb{Z}$ as $a \sim_1 b$ if $a \geq b$ and $a \sim_2 b$ if $|a - b|$ is a power of 2. (Define *power of 2* to be any integer that equals 2^n for some integer $n \geq 0$.) Which equivalence relation axioms does $\sim_1$ satisfy? What about $\sim_2$?

3. What is a *quotient construction*? What's the most common kind of equivalence relation used in quotient constructions? What method do you use to extend operations on elements of the original set to elements of the quotient?

4. Why can you extend multiplication on $\mathbb{Z}$ to elements of $\mathbb{Z}/5\mathbb{Z}$ using representative elements, but you can't extend multiplication on $\mathbb{R}$ to elements of $\mathbb{R}/\mathbb{Z}$?

5. ($\star$) Consider the equivalence relation on $\mathbb{R}$ defined as $x \sim y$ if x and y are both positive, both negative, or both zero. Can you define addition on the set of equivalence classes using addition in $\mathbb{R}$ and the representative element construction? What about multiplication?

6. If V is a vector space with dimension 24 and W is a subspace of V with dimension 8, what is the dimension of V/W?

Quotient spaces are vector spaces whose elements are entire cosets of a larger vector space. To be more precise, if V is a vector space and W is a subspace of V, then the quotient space V/W is the set of cosets of W. Addition and multiplication on V/W are defined by using the corresponding operations on V and a technique called *representative element construction*, which shows up in analogous situations in many branches of algebra. It's worth discussing representative elements first as a way to construct a formalism for a slightly more familiar concept: modular arithmetic on ordinary integers.

5.4.1 Equivalence relations and modular arithmetic

Quotient spaces are analogous to modular arithmetic, which is arithmetic on integers that ignores everything except remainders relative to some fixed divisor. The idea of modular arithmetic is often introduced by an analogy with clockfaces. For instance, on a 24-hour clock numbered from 0 to 23, you could say that $9+6=15$ (because starting on the 9 and moving clockwise 6 places puts you at 15), but $20+6=2$ (because starting on 20 and moving right four places puts you back at 0, and then moving the remaining two places brings you to 2). You could notate this less confusingly than "$20+6=2$" by explicitly noting the modulus and using the congruence sign $\equiv$ instead of the equals sign, as $20+6 \equiv 2 \bmod 24$.

What makes modular arithmetic useful is that if two integers have the same remainder when divided by some "modulus" m (we can call this the remainder or "residue" "modulo m), then so do their sums or products with any other integer. That is, if a and a' have the same residue modulo m, then so do $a+b$ and $a'+b$, or ab and $a'b$, regardless of what b is. Why is this? Remember that x and y leave the same remainder modulo m iff $x-y$ is a multiple (possibly zero or negative) of m. So if $a'-a$ is a multiple of m, then so is $(a'+b)-(a+b)=a'-a$ and $a'b-ab=(a'-a)b$. This fact underlies tricks such as the digit-sum test for divisibility by 9: that is, the fact that a number is divisible by 9 if and only if the sum of its digits is divisible by 9 as well. Since $10 \equiv 1 \bmod 9$, so $10^n = 10 \times \cdots \times 10 \equiv 1 \times \cdots \times 1 \bmod 9$, so any integer $10a_n + 10^{n-1}a_{n-1} + \cdots + 10a_1 + a_0$ is congruent to the sum of its base-10 digits $a_n + a_{n-1} + \cdots + a_1 + a_0$.

Let's take one more conceptual leap: we'll define a system of modular arithmetic consisting of a set whose elements are themselves entire sets of integers, and where addition and multiplication work on entire sets at once. First, two definitions.

Definition. *A **relation** on a set S, usually denoted with a tilde as in $a \sim b$, is a statement that we define to be true or false for every ordered pair of elements $a, b \in S$. (You can also think of relations as a subset R of S^2: the relationship $a \sim b$ is true if $(a,b) \in R$.)*

Definition. *An **equivalence relation** on a set S is a relation that satisfies the following axioms:*

1. Reflexivity*: $a \sim a$ is always true for every $a \in S$.*

2. Commutativity: *$a \sim b$ and $b \sim a$ are either both true or both false for every pair of elements $a, b \in S$.*

3. Transitivity: *If $a \sim b$ and $b \sim c$, then $a \sim c$ for every triple of elements $a, b, c \in S$.*

Equivalence relations partition a set into a collection of disjoint *equivalence classes*: $a \sim b$ is true if a and b are in the same equivalence class, and false otherwise. We can firm up our understanding of this definition with a few examples.

1. Define $a \sim_1 b$ to be true if a and b have the same first digit when written in base 10, ignoring negative signs (so, for instance, $19 \sim_1 -162$ is true but $19 \sim_1 20$ is false). Then $\sim_1$ is an equivalence relation: it's reflexive ($a \sim a$ is true because every number has the same first digit as itself, it's symmetrical ($a \sim b$ implies $b \sim a$) because the definition gives a and b symmetrical roles, and it's transitive because if a and b have the same first digit, and so do b and c, then.

2. Define $a \sim_2 b$ to be true if a has at least as many digits as b when written in base 10, again ignoring a negative sign (so, for instance, $20 \sim_2 9$ and $1000 \sim_2 -1001$ are true, but $2 \sim 20$ is false). This is *not* an equivalence relation, as it's reflexive and transitive but not symmetric: for instance, $20 \sim_2 9$ is true but $9 \sim_2 20$ is false.

Now let's apply this notion to modular arithmetic. Pick some positive integer m, and let $a \sim b$ be the relation on $\mathbb{Z}$ that is true if $a - b$ is a multiple (positive, negative, or zero) of m. It's easy to show that $\sim$ is an equivalence relation:

1. $a \sim a$ is true because $a - a = 0$, and zero is a multiple of m.

2. If $a - b$ is a multiple of m, then $b - a = -(a - b)$ is also a multiple of m, so commutativity holds.

3. If $a-b$ and $b-c$ are multiples of m, then so is $a-c = (a-b)+(b-c)$, so transitivity holds.

The equivalence classes are the sets of integers with the same remainder modulo m. Let's take $m = 5$, for instance. The equivalence classes are then:

- Integers congruent to 0 modulo 5; that is, $\{\ldots, -10, -5, 0, 5, 10, \ldots\}$. We'll denote this set $\overline{0}$.

- Integers congruent to 1 modulo 5; that is, any integer of the form $5n + 1$. This set is $\{\ldots, -9, -4, 1, 6, 11\}$. We'll denote it $\overline{1}$.

- Integers congruent to 2 modulo 5: $\{\ldots, -8, -3, 2, 7, 12, \ldots\}$. We'll denote it $\overline{2}$.

- Integers congruent to 3 modulo 5: $\{\ldots, -7, -2, 3, 8, 13, \ldots\}$. We'll denote it $\overline{3}$.

- Integers congruent to 4 modulo 5: $\{\ldots, -6, -1, 4, 9, 14, \ldots\}$. We'll denote it $\overline{4}$.

Now let's write $\mathbb{Z}_5$ for the set of sets $\{\overline{0}, \overline{1}, \overline{2}, \overline{3}, \overline{4}\}$. Let's define the sum of two entire sets as follows: if A and B are equivalence classes, then choose any arbitrary integers $a \in A$ and $b \in B$. The sum $A + B$ is the equivalence class that contains $a + b$.

The word "arbitrary" should get your guard up: how do we know that no matter how we choose a and b, the sum $a+b$ will be in the same equivalence class? Fortunately, as we saw before, the properties of modular arithmetic save us. If instead of a we chose some other representative a' of the same equivalence class (which, by the way that we defined these equivalence classes, must differ from a by a multiple of m), then the resulting sum $a' + b$ also differs from $a + b$ by a multiple of m. The argument that choosing a different representative b of B doesn't change the equivalence class of $a + b$ is, of course, identical.

By this definition, then, what is $\overline{2} + \overline{4}$? If we choose representative elements $2 \in \overline{2}$ and $-6 \in \overline{4}$, for example, then the sum $\overline{2}+\overline{4}$ is the class that contains $2+(-6) = -4$: that is, $\overline{1}$. Different representative elements give the same result: if we choose, say, $7 \in \overline{2}$ and $9 \in \overline{4}$ instead, then the sum $7 + 9 = 16$ is also in $\overline{1}$.

You can check for yourself that multiplication on $\mathbb{Z}$ extends to multiplication on $\mathbb{Z}_5$ (and $\mathbb{Z}_m$ for arbitrary positive integers m more generally) by the same mechanism: choose representative integers a, b from the equivalence classes A, B, multiply them, and define the product AB to be the equivalence class containing the product ab. This equivalence class is the same no matter which representatives you choose: if you change either a or b by a multiple of m, then the product ab also changes by a multiple of m.

To sum up, the construction of $\mathbb{Z}_5$ follows a particular pattern called a *quotient construction* that occurs throughout mathematics:

1. Take a set (in this example $\mathbb{Z}$) with some arithmetic operations defined on it.

2. Define an equivalence relation on this set, often by defining two elements to be equivalent if their difference is in a particular subset. It's important to check that this relation is in fact an equivalence relation.

3. Define arithmetic operations on subclasses via representative elements. This step requires checking that every possible choice of representative elements from each pair of equivalence classes produces a result in the same equivalence class.

5.4.2 Not all equivalence relations give well-defined operations

It's crucial to check that our equivalence relation actually produces operations that don't depend on the choice of representative element: this is not guaranteed simply by the equivalence relation axioms. For instance, suppose we used the equivalence relation defined as $a \sim b$ if a and b have the same first digit (ignoring negative signs) to divide $\mathbb{Z}$ into equivalence classes: $\hat{1}$ is the set of integers whose first digit is 1, $\hat{2}$ is the set whose first digit is 2, and so on. (We'll use $\hat{1}$ to indicate that this is a different set from $\overline{1}$ in the construction of $\mathbb{Z}/5\mathbb{Z}$.) If we try to define addition of these equivalence classes using representative elements, we run into the problem that different representative elements give different results. For instance, we could have $\hat{1} + \hat{2} = \hat{1}$ with the representative elements $11 + 2 = 13$, or $\hat{1} + \hat{2} = \hat{2}$ with representatives $199 + 2 = 201$, or even $\hat{1}+\hat{2} = \hat{6}$ with $-14+20 = 6$ (and these are not the only possiblities), so addition on these equivalence classes is not well-defined. (Neither is multiplication: for instance, $7 \times 8 = 56$ and $79 \times 89 = 7031$ don't have the same first digit.)

Even constructions much closer to modular arithmetic, with equivalence classes defined as sets of elements whose differences belong to a particular subset, sometimes

don't have well-defined arithmetic operations. Consider, for instance, the set of real numbers with the equivalence relation $x \sim y$ if $x - y$ is an integer. We'll write the set of resulting equivalence classes as $\mathbb{R}/\mathbb{Z}$. Equivalence classes have the form $\{\ldots, a-2, a-1, a, a+1, a+2\}$ for real numbers a; for instance, $\{\ldots, -1.8, -0.8, 0.2, 1.2, 2.2, \ldots\}$ is one equivalence class (which we'll denote $\overline{0.2}$, choosing whatever element a lies in the range $0 \leq a < 1$ as the canonical representative), and $\{\ldots, -1.5, -0.5, 0.5, 1.5, 2.5, \ldots\}$ is another one (we'll denote this $\overline{0.5}$).

We can extend addition on $\mathbb{R}$ to addition on $\mathbb{R}/\mathbb{Z}$ without a problem: if a, a' are two representatives of the same equivalence class (that is, $a' - a$ is an integer), and ditto for b, b', then $a + b$ and $a' + b'$ must differ by an integer. But multiplication isn't well defined: different representative elements produce products that belong to different equivalence classes. For instance, if we choose $0.2 \in \overline{0.2}$ and $0.5 \in \overline{0.5}$ as representative elements, then the $\overline{0.2} \times \overline{0.5}$ would be the equivalence class that contains $0.2 \times 0.5 = 0.1$ (which is $\overline{0.1}$). But if we chose $2.2 \in \overline{0.2}$ and $3.5 \in \overline{0.5}$ as representatives instead, then the product would be whichever equivalence class contains $2.2 \times 3.5 = 7.7$, namely $\overline{0.7}$.

5.4.3 Vector space operations on cosets

Quotient spaces in linear algebra are a close analogy to modular arithmetic. Where modular arithmetic divides integers into equivalence classes based on whether their difference is in the set of multiples of some fixed modulus m, quotient spaces divide a vector space into equivalence classes based on whether their difference is in some susbpace.

To be more precise:

Definition. *let V be a vector space over some field $\mathbb{F}$, and let W be a subspace of V. Then the **quotient space** V/W is the set of all cosets of W by V/W.*

Remember: each element of V/W is itself a set of vectors in V. We can define addition and multiplication on V/W, making it into a vector space into its own right, by using arbitrary representative elements like with modular arithmetic.

To be more precise: suppose $A = \mathbf{a}_0 + W$ and $B = \mathbf{b}_0 + W$ are cosets of W. Temporarily, we'll use the symbols $\boxplus$ and $\boxtimes$ to refer to our extensions of V's vector space operations to V/W: that is,

1. $A \boxplus B$ is the coset of W containing $\mathbf{a} + \mathbf{b}$, where $\mathbf{a} \in A$ and $\mathbf{b} \in B$ are arbitrary representative elements that may or may not equal $\mathbf{a}_0$ and $\mathbf{b}_0$. (We don't know yet whether $\boxplus$ is the same operation as the more general addition of affine spaces that we defined in 5.3, and for now, we'll reserve the symbol $+$ for this latter operation.)

2. $k \boxtimes A$ is the coset of W containing $k\mathbf{a}$, where $\mathbf{a} \in A$ is an arbitrary representative element.

Let's first prove that the operations $\boxplus$ and $\boxtimes$ are well-defined: that is, every possible choice of elements from an input coset (for multiplication) or pair of cosets (for addition) produces a result in the same output coset. First, let's prove this for $\boxplus$. Take arbitrary elements $\mathbf{a}$ and $\mathbf{a}'$ of A, and $\mathbf{b}$ and $\mathbf{b}'$ are both elements of B. That is, their differences $\mathbf{w}_a \coloneqq \mathbf{a}' - \mathbf{a}$ and $\mathbf{w}_b \coloneqq \mathbf{b}' - \mathbf{b}$ are both in W. Then $(\mathbf{a}' + \mathbf{b}') - (\mathbf{a} + \mathbf{b}) = \mathbf{w}_a + \mathbf{w}_b$, which is a multiple of W, so $\mathbf{a} + \mathbf{b}$ and $\mathbf{a}' + \mathbf{b}'$ are in the same coset of W. This means

that whether you choose $\mathbf{a}$ and $\mathbf{b}$ as representatives or $\mathbf{a}'$ and $\mathbf{b}'$, you'll get the same result for $A \boxplus B$.

The proof that $\boxtimes$ is well-defined is similar. For some scalar $k \in \mathbb{F}$ and some coset A, we could define $k \boxtimes A$ as the coset containing $k\mathbf{a}$ for some representative element $\mathbf{a} \in A$, or the coset containing $k\mathbf{a}'$ for any other representative element. But $k\mathbf{a}' - k\mathbf{a} = k(\mathbf{a}' - \mathbf{a})$, and since $\mathbf{a}' - \mathbf{a} \in W$ by the definition of cosets, so is $k(\mathbf{a}' - \mathbf{a})$. So $k\mathbf{a}'$ and $k\mathbf{a}$ are in the same coset of W.

Can we write a formula for the resulting cosets? If we choose $\mathbf{a}_0$ and $\mathbf{b}_0$ as representatives for $A = \mathbf{a}_0 + W$ and $B = \mathbf{b}_0 + W$, then we get $A \boxplus B = (\mathbf{a}_0 + \mathbf{b}_0) + W$. It turns out that $A + B$, with $+$ defined as the regular sum of affine spaces as in section 5.3, is also just $(\mathbf{a}_0 + \mathbf{b}_0) + (W + W) = (\mathbf{a}_0 + \mathbf{b}_0) + W$. So $\boxplus$, which we defined for cosets of the *same* subspace, coincides in this case with the general definition for sums of cosets of *possibly different* subspaces, and we don't need to use the special symbol $\boxplus$ (which, in any case, isn't standard notation) to disambiguate them. Similarly, $k \boxtimes (\mathbf{a} + W) = k(a + W) = (k\mathbf{a}) + W$, and we can drop the symbol $\boxtimes$.

These definitions of scalar addition and multiplication satisfy the necessary vector space axioms:

1. V/W has an additive identity, namely $\mathbf{0} + W = W$. Every element $\mathbf{a}_0 + W$ has an additive inverse, namely $-\mathbf{a}_0 + W$.

2. The commutative, associative, and distributive properties all hold, because arithmetic on V/W is based on arithmetic on V, and these properties also hold in V. For instance, if A and B are two cosets of W, then $A + B$ is (by definition) the coset containing $\mathbf{a} + \mathbf{b}$ for arbitrary representatives $\mathbf{a} \in A, \mathbf{b} \in B$, and $B + A$ is the coset containing $\mathbf{b} + \mathbf{a}$. But $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$ because addition on V is commutative by definition of a vector space.

If quotient spaces still sound abstract to you, some examples might help. Let $W_1 = \text{span}\{\mathbf{e}_1, \mathbf{e}_2\} \subset \mathbb{R}^3$ be the set of vectors $(x, y, 0)$, or the plane $z = 0$. Two other vectors $\mathbf{u} = (x_1, y_1, z_1), \mathbf{v} = (x_2, y_2, z_2) \in \mathbb{R}^3$ can be in the same coset of W (that is, $\mathbf{v} - \mathbf{u} \in W$) if and only if $z_1 = z_2$; that is, W_1 is the xy-plane $z = 0$, and the quotient space $\mathbb{R}^3/W_1$ is a stack of horizontal planes $z = c$. So we can write every coset of W_1 in the form $(0, 0, z) + W_1$ where $z \in \mathbb{R}$, and, conversely, different choices of z always give different cosets. (Essentially, taking the quotient of $\mathbb{R}^3$ by a two-dimensional space W_1 "collapses" the x- and y-dimensions of $\mathbb{R}^3$ and leaves only the z-dimension. Keep this intuition in mind for the next section!)

Now let's define addition and multiplication on $\mathbb{R}^3/W_1$. Let $A = (0, 0, a) + W_1$ and $B = (0, 0, b) + W_1$ be two cosets. Then:

1. Any representatives $\mathbf{a} \in A$ and $\mathbf{b} \in B$ have the form $\mathbf{a} = (x_1, y_1, a)$, $\mathbf{b} = (x_2, y_2, b)$. Their sum is $\mathbf{a} + \mathbf{b} = (x_1 + x_2, y_1 + y_2, a + b)$, which is in $(0, 0, a + b) + W_1$.

2. Any scalar multiple $k\mathbf{a}$ has the form (x_1, y_1, ka), which is in $(0, 0, ka) + W_1$.

So $A + B = (0, 0, a + b) + W_1$ and $kA = (0, 0, ka) + W_1$ (which is easy to see just by choosing $(0, 0, a)$ and $(0, 0, b)$ as representative points and remembering that coset doesn't depend on the exact choice of representative set). This also means, incidentally, that you can model arithmetic in $\mathbb{R}^3/W_1$ by arithmetic in $\text{span}\{\mathbf{e}_3\}$, the one-dimensional subspace of $\mathbb{R}^3$ that contains all of our chosen representative points.

Now let's consider a more complex example. Let $W_2 \subset \mathbb{R}^4$ be the span of

$$\{(1, 0, -4, 2), (2, 3, -8, 3)\}$$

and let's work out a formula for the cosets of $\mathbb{R}^4/W_2$, including a convenient representative point for each. Ideally, we'd like the representative points to be a subspace of $\mathbb{R}^4$ in their own right, so that arithmetic on this subspace corresponds to arithmetic on $\mathbb{R}^4/W_2$.

The spanning set $\{(1, 0, -4, 2), (2, 3, -8, 3)\}$ of W_2 gives a general formula for its elements: $(a+2b, 3b, -4a-8b, 2a+3b)$, with $a, b \in \mathbb{R}$ freely chosen. We can get a simpler formula, though, by replacing $(2, 3, -8, 3)$ in the spanning set with a linear combination that includes it. One choice is $\frac{1}{3}[(2, 3, -8, 3) - 2(1, 0, -4, 2)] = (0, 1, 0, -\frac{1}{3})$. The new spanning set $\{(1, 0, -4, 2), (0, 1, 0, -\frac{1}{3})\}$ gives a simpler general form $(x, y, -4x, 2x - \frac{1}{3}y)$ for elements of W_2.

Two elements $\mathbf{u} = (x_1, y_1, z_1, w_1)$ and $\mathbf{v} = (x_2, y_2, z_2, w_2)$ are in the same coset of W_2 if their difference $(x_1 - x_2, y_1 - y_2, z_1 - z_2, w_1 - w_2)$ satisfies this general form: that is, if $(z_1 - z_2) = -4(x_1 - x_2)$ and $(w_1 - w_2) = 2(x_1 - x_2) - \frac{1}{3}(y_1 - y_2)$. This means (setting $z_1 = z_2 = w_1 = w_2 = 0$) that any two vectors of the form $(x_1, y_1, 0, 0)$ and $(x_2, y_2, 0, 0)$ can be in the same subspace only if $x_1 = x_2$ and $y_1 = y_2$—that is, if the vectors are identical. So $\operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2\}$ is a complement of W_2 and every coset of W_2 can be designated uniquely as $(x, y, 0, 0) + W_2$. Then if $A = (x_1, y_1, 0, 0) + W_2$ and $B = (x_2, y_2, 0, 0) + W_2$, then $A + B = (x_1 + x_2, y_1 + y_2, 0, 0) + W_2$ and $kA = (kx_1, kx_2, 0, 0) + W_2$.

5.4.4 Dimension of quotient spaces

The quotient space V/W essentially collapses every dimension of W down to a point, so you might expect the dimension of V/W to be smaller than that of V. And in the above examples, taking quotients by a two-dimensional vector space reduces the dimension of the resulting space by 2: $\mathbb{R}^3/W_1$ has dimension 1, and $\mathbb{R}^4/W_2$ has dimension 2.

These observations can be made into a general result:

Theorem. *If W is a vector subspace of V, then* $\dim V/W = \operatorname{codim} V$.

Proof. Write $n = \operatorname{codim} W$. Let B be a basis of W, and extend B to a basis of V by adding more vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ (none of which, or any linear combination of them, can be in W). Write $U := \operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$; U thus has dimension n, and $V = U \oplus W$.

We claim that every coset of W contains exactly one element of U. To see this, note that any element $\mathbf{v} \in V$ can be written uniquely as $\mathbf{v} = \mathbf{u} + \mathbf{w}$ where $\mathbf{u} \in U, \mathbf{w} \in W$, so $\mathbf{v}$ and $\mathbf{u}$ are in the same coset of $\mathbf{w}$, and every coset contains at least one element of U.

Now we need to prove that no coset of W can contain two elements of U. Suppose that $\mathbf{u}$ and $\mathbf{u}'$ were distinct elements in the same coset; that is, $\mathbf{u} - \mathbf{u}' \in W \setminus \{\mathbf{0}\}$. Then $\mathbf{u}$ could be written as a linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_n$ alone, or as $\mathbf{w}$ (a nonzero linear combination of elements in B plus the linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_n$ that gives $\mathbf{u}'$. So two linear combinations of $B \cup \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ would have to be equal. But this is impossible: $B \cup \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ is a basis of V, so it's linearly independent.

Therefore, every coset of V/W can be written as $\mathbf{u} + W$, where $\mathbf{u}$ is a unique point in U. Since we can model vector arithmetic on cosets by using vector operations on the base points, V/W must have the same structure as U, and the cosets $\mathbf{v}_1 + W, \ldots, \mathbf{v}_n + W$ give a basis for V/W.

□

This proof that $\dim V/W = \operatorname{codim} W$ holds even in the case when V and W have infinite dimension but $\operatorname{codim} W$ has finite codimension. (Codimension in this context is defined as the number of vectors we'd have to add to a basis of W to make a basis of V.) With a bit more technical set theory, we could prove that if V/W has infinite dimension, then so does $\operatorname{codim} W$ is also infinite. We won't be dealing much with infinite-dimensional vector spaces here, though.

Answers to key questions.

1. A *relation* on a set is a statement that is either true or false for every ordered pair of elements in the set. An *equivalence relation* must satisfy the three axioms of identity, reflexivity, and transitivity. An equivalence relation divides the set it's defined on into a collection of *equivalence classes*: the relation is true between elements of the same class and false between elements of different classes.

2. $\sim_1$ satisfies identity ($a \geq a$ is always true) and transitivity ($a \geq b$ and $b \geq c$ together imply $a \geq c$), but not reflexivity ($a \geq b$ doesn't imply $b \geq a$).

   $\sim_2$ satisfies reflexivity (if $|a - b|$ is a power of two, then so is $|b - a|$, because $|a - b| = |b - a|$), but not identity ($|a - a| = 0$ and 0 is not a power of 2 by our definition) or transitivity (for instance, $0 \sim 4$ and $4 \sim 6$ because $|4 - 0| = 4$ and $|6 - 4| = 2$ are both powers of 2, but $0 \not\sim 6$.

3. A *quotient construction* is a construction that defines operations on equivalence classes of some set S created by some equivalence relation, using operations defined on the original set S. The most common equivalence relations used in quotient constructions define elements to be equal if their differences are in some set of multiples of a defining element.

4. The product of integers modulo 5 is well-defined modulo 5 (that is, if $a \equiv a \bmod 5$ and $a' \equiv b' \bmod 5$ then $ab = a'b' \bmod 5$), but the product of real numbers is not well-defined modulo 1: it's possible to have real a, a' with the same fractional part, and likewise for b, b', but ab and $a'b'$ don't have the same fractional part.

5. Addition is not well-defined on these equivalence classes because if a is positive(i.e. a representative of the equivalence class of positive real numbers) and b is negative, then $a + b$ could be positive (if $|a| > |b|$), zero (if $|a| = |b|$), or negative (if $|a| < |b|$). Multiplication, though, is well-defined, because you know the sign of ab if you know the signs of a and b. Specifically, ab is always zero if either a or b is zero; otherwise, ab is positive if a and b have the same sign and negative if they have different signs.

6. The dimension of V/W is $\dim V - \dim W = 24 - 8 = 16$.

5.5 Rank–nullity proof with first isomorphism theorem

Key questions.

1. If U is a subspace of a vector space V, what is the projection map $\pi : V \to V/U$?

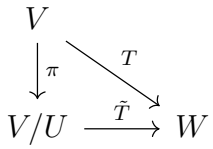
2. State the *first isomorphism theorem*. Under what circumstances is the map whose existence is guaranteed by the first isomorphism theorem a bijection?

Another method of proving the rank–nullity theorem comes from a general result called the *first isomorphism theorem*, which claims, in essence, that to every linear map on a space V whose kernel includes some subspace W, there corresponds a unique linear map on V/U. We'll use the results from section 2.4.2, but not the rank–nullity theorem itself.

Theorem (First isomorphism theorem)**.** *Let $T : V \to W$ be a linear map, and let U be a subspace of $\ker T$. Let $\pi : V \to V/U$ be the "projection map" that takes every element $\mathbf{v} \in V$ to the coset $\mathbf{v} + U$ that contains it. (It's simple to prove that π is linear, because addition and multiplication of cosets in V/U are derived from operations on representative elements in V.)*

Then there is a unique linear map $\tilde{T} : V/U \to W$ such that $T = \tilde{T} \circ \pi$. Furthermore, if $U = \ker T$, then $\tilde{T}$ is bijective and gives an isomorphism between $V/\ker T$ and $\operatorname{im} T$.

The following diagram (called a "commutative diagram") illustrates the various maps in the theorem statement; in a commutative diagram, the maps given by composing the maps on the arrow labels have to be the same for any possible path between two points:



Proof. First, note that π is surjective onto V/U, because every coset of U contains at least one vector $\mathbf{v}$ and so occurs at least once as a value of π (namely $\pi(\mathbf{v})$). So as long as $\tilde{T}$ can be defined to satisfy $T = \tilde{T} \circ \pi$, every value of T is also a value of $\tilde{T} \circ \pi$, so T and $\tilde{T}$ have the same image.

We'll define $\tilde{T}$ using arbitrary representative elements: for any coset $C \in V/U$, choose some arbitrary $\mathbf{v} \in C$, and define $\tilde{T}(C) = T\mathbf{v}$. (It should be clear that this is the only possible choice for $\tilde{T}(C)$: if we choose anything besides $T\mathbf{v}$, then $T = \tilde{T} \circ \pi$ won't hold.)

We need to check that $\tilde{T}$ is well-defined (i.e. different choices of representative give the same value of $\tilde{T}$), and $\tilde{T}$ is injective if $U = \ker T$.

1. *$\tilde{T}$ is well-defined*: By the definition of cosets, two vectors $\mathbf{v}_1, \mathbf{v}_2$ are in the same coset C of U if $\mathbf{v}_1 - \mathbf{v}_2 \in U$, and remember that $U \subseteq \ker T$. So $T\mathbf{v}_1 = T\mathbf{v}_2 + T(\mathbf{v}_1 - \mathbf{v}_2) = T\mathbf{v}_2 + \mathbf{0}_W = T\mathbf{v}_2$; that is, choosing $\mathbf{v}_1$ or $\mathbf{v}_2$ as the representative of C will give the same value of $\tilde{T}(C) = T\mathbf{v}_1 = T\mathbf{v}_2$.

2. *If $U = \ker T$, then $\tilde{T}$ is injective*: Suppose we have two representatives $\mathbf{v}_1, \mathbf{v}_2$ from different cosets C_1, C_2 of $\ker T$. Then $\tilde{T}(C_1) - \tilde{T}(C_2) = T\mathbf{v}_1 - T\mathbf{v}_2 = T(\mathbf{v}_1 - \mathbf{v}_2) \neq 0_W$, because if $\mathbf{v}_1$ and $\mathbf{v}_2$ are in different cosets of $\ker T$, then (by definition) $\mathbf{v}_1 - \mathbf{v}_2 \notin \ker T$. Therefore, $\tilde{T}(C_1) \neq \tilde{T}(C_2)$. (If U is not all of $\ker T$, then this argument fails, as two elements $\mathbf{v}_1, \mathbf{v}_2$ in different cosets of U could be in the same coset of $\ker T$.)

□

One informal way that you might hear the first isomorphism theorem described is that every linear map $T : V \to W$ can be "factored through" $V/\ker T$: that is, you can split it into the projection map to $V/\ker T$ and a bijection from $V/\ker T$ to W.

The first isomorphism theorem implies the rank–nullity theorem as a corollary, because $\dim(V/\ker T) = \operatorname{codim} \ker T$ (see page 129), and $\tilde{T}$ gives an isomorphism from $V/\ker T$ to $\operatorname{im} \tilde{T} = \operatorname{im} T$.

Answers to key questions.

1. The projection map $\pi : V \to V/U$ takes every element of V to the coset of U that contains it.

2. The first isomorphism theorem is that every map $T : V \to W$ whose kernel contains a subspace U can be written as $T = \tilde{T} \circ \pi$, where $\tilde{T} : V/U \to W$ is a map uniquely determined by T. And $\tilde{T}$ is bijective if $\ker T = U$.

Chapter 6

Operators

Overview. A linear operator (also called *endomorphism*) is a linear map from a space to itself: the domain and codomain are the same. Operators on a finite-dimensional domain, therefore, can be represented by square matrices. All of our findings on general linear maps apply to operators as well; but because operators can be composed with themselves, the theory of operators is far richer.

Section 6.1 introduces the concept of operators and the related notion of *invariant subspaces*: vector subspaces that contain their own images under a certain operator. Any operator on a larger vector space can also be considered an operator on its invariant subspaces, with the same general findings on operators applying to restricted operators on invariant subspaces.

Section 6.2 presents a few more results on invariant subspaces, of which the most important is that an invariant subspace of an operator T is also an invariant subspace of any polynomial expression $c_n T^N + c_{n-1} T^{n-1} + \cdots + c_1 T + c_0 I$.

Section 6.3 introduces the key concept of *eigenvectors*—vectors whose images under some operator are scalar multiples of the input vector (the scalar in question is called an *eigenvalue*)—and *eigenspaces*, or vector subspaces made up of eigenvectors. Knowing an operator's eigenvectors can make it much easier to reason about its properties. A set of results on the maximum possible dimension of eigenspaces—including, crucially, that no operator can have more eigenvalues than the dimension of the underlying space, and that the sum of eigenspaces is direct—is covered in Section 6.4. This latter finding lets us analyze the behavior of an operator on an entire space by decomposing it into its variant subspaces.

6.1 Operators and invariant subspaces

Key questions.

1. What is the relationship between the terms *linear map*, *operator*, and *endomorphism*?

2. What is an *invariant subspace* of an operator? What are the *trivial invariant subspaces*?

In the last chapter, we looked at general linear maps between two spaces V and W. In this chapter, we'll focus on maps $T : V \to V$ from one space to itself. These are called *operators* or *endomorphisms* on the space V. The set of endomorphisms can be

denoted $\mathrm{End}(V)$; it's also a vector space in its own right with the operations of addition and scalar multiplication of maps that we discussed in Section 2.3.

The theory of endomorphisms on one space is more interesting than the theory of general linear maps between different spaces because linear operators can be composed with themselves. In particular, if T is a linear operator on V, then the functions $T^2 := T \circ T$, $T^3 := T \circ T \circ T$, and so on are also linear operators on V, as are sums and multiples of these linear operators such as $T^3 - 5T^2 + 6T$. We can make algebraic manipulations on such polynomials of linear operators almost exactly the same way as with polynomials whose variables represent elements of ordinary fields, such as real or complex numbers—for instance, just like we can factor $x^3 - 5x^2 + 6x = x(x-2)(x-3)$ for a polynomial on a real variable x, we could also factor $T^3 - 5T^2 + 6T = T \circ (T - 2I) \circ (T - 3I)$, where I is the identity operator $I\mathbf{v} = \mathbf{v}$.

If V is finite-dimensional, then the rank–nullity theorem $\dim V = \dim \ker T + \dim \mathrm{im}\, T$ guarantees that any operator $T \in \mathrm{End}(V)$ falls in one of these two categories:[1]

1. $\ker T = \{\mathbf{0}\}$, so $\dim \mathrm{im}\, T = \dim V$ and T is a bijection from V to itself.

2. $\ker T$ has nonzero dimension and $\dim \mathrm{im}\, T < \dim V$, so $\mathrm{im}\, T$ is a proper subspace of V, so T is neither injective nor surjective.

The core of linear algebra is a classification of linear operators according to their *invariant subspaces*, defined as follows:

Definition. *An **invariant subspace** of an operator $T : V \to V$ is any subspace $W \subseteq V$ such that $T(W) \subseteq W$: that is, if $\mathbf{w} \in W$, then $T\mathbf{w} \in W$ as well.*

(Note that despite the possible implications of the term *invariant*, T doesn't have to take every element of W to itself, just to some (possibly different) element of W.

One fact that makes invariant subspaces useful is that if W is an invariant subspace of T, then the restricted map $T|_W$ is an operator on W, so every result about operators on entire vector spaces can also be applied to restricted operators on invariant subspaces.

From this definition alone, the following properties are easy to prove:

1. $\{\mathbf{0}\}$ and V are invariant subspaces of every operator. We call these *trivial* invariant spaces.

2. Every subspace is an invariant subspace of the identity operator I.

3. If W_1 and W_2 are invariant subspaces of any operator, then so are $W_1 \cap W_2$ and $W_1 + W_2$. (This claim may not be obvious: think about why it's true. Remember: every element of $W_1 + W_2$ can be written as $\mathbf{w}_1 + \mathbf{w}_2$ for some choice of $\mathbf{w}_1 \in W_1$ and $\mathbf{w}_2 \in W_2$.)

[1] Counterexamples for infinite-dimensional spaces: consider $\mathbb{F}^{\mathbb{N}}$, the set of infinite sequences of elements of $\mathbb{F}$. Then the operator $(x_1, x_2, x_3, \ldots) \mapsto (x_2, x_3, x_4, \ldots)$ is surjective but not injective (sequences that differ only in x_1 have the same values), and the map $(x_1, x_2, x_3, \ldots) \mapsto (0, x_1, x_2, \ldots)$ is injective but not surjective (the image includes only sequences that start with 0).

Answers to key questions.

1. An *operator* is a *linear map* from one space to itself. *Endomorphism* is another word for *operator*.

2. W is an *invariant subspace* of an operator $T : V \to V$ if whenever $\mathbf{w} \in W$, $T\mathbf{w} \in W$ as well. T will always have $\{\mathbf{0}\}$ and V (that is, its entire domain) as invariant subspaces, so we call these *trivial* invariant subspace.

Answers to key questions.

1. An operator is a linear map from a vector space to itself (i.e. with the same space as domain and codomain).

6.2 Results on invariant subspaces

This section will present a few small results on invariant subspaces. Throughout, V is an arbitrary (not necessarily finite-dimensional) vector space, T is an arbitrary operator on V, and I is the identity operator. We'll also adopt the notation kT to mean the map that takes $\mathbf{v}$ to $kT\mathbf{v}$ and $T_1 + T_2$ to mean the map that takes $\mathbf{v}$ to $T_1\mathbf{v} + T_2\mathbf{v}$, just as in Section 2.3.

Proposition. $\ker T$ *and* $\operatorname{im} T$ *are invariant subspaces of* T.

Proof. If $\mathbf{v} \in \ker T$, then $T\mathbf{v} = \mathbf{0} \in \ker T$, so $\ker T$ is an invariant subspace. And $T\mathbf{v} \in \operatorname{im} T$ by definition of the image for any $\mathbf{v} \in V$, so in particular, $T\mathbf{v} \in \operatorname{im} T$ for any $\mathbf{v} \in \operatorname{im} T$. So $\operatorname{im} T$ is also an invariant subspace.

□

Proposition. *If* W *is any subspace of a vector space* V*, then the set of operators that have* W *as an invariant subspace is a vector subspace of* $\operatorname{End}(V)$.

Proof. Let $S \subseteq \operatorname{End}(V)$ be the set of operators with W as an invariant subspace. We have to check that S satisfies the three subspace axioms.

1. *Non-emptiness*: Every subspace of V is invariant under the zero operator $Z\mathbf{v} = \mathbf{0}$: if W is an arbitrary subspace, then $Z\mathbf{w} = \mathbf{0} \in W$ for all $\mathbf{w} \in W$. So S at least contains Z.

2. *Closure under addition*: Suppose W is an invariant subspace of T_1 and T_2. Then for any $\mathbf{w} \in W$, we have $(T_1 + T_2)\mathbf{w} = T_1\mathbf{w} + T_2\mathbf{w}$, which is the sum of two elements of W. So W must also be invariant under $T_1 + T_2$; that is, $T_1 + T_2 \in S$.

3. *Closure under multiplication*: if k is any scalar and T is an operator with W as an invariant subspace, then $(kT)\mathbf{w} = k(T\mathbf{w})$ for any $\mathbf{w} \in W$ (by definition of scalar-by-operator multiplication). As $T\mathbf{w} \in W$, so $k(T\mathbf{w}) \in W$; thus, $kW \in S$.

□

Proposition. *If* W *is an invariant subspace of two (possibly identical) operators* $T_1, T_2 : V \to V$*, then it's also an invariant subspace of* $T_1 \circ T_2$.

Proof. For any vector $\mathbf{w} \in W$, we have $T_2\mathbf{w} \in W$ (because W is invariant under T_2) and therefore also $(T_1 \circ T_2)\mathbf{w} = T_1(T_2\mathbf{w}) \in \mathbf{w}$ (because W is invariant under T_1).

□

Corollary. *Any invariant subspace of T is also an invariant subspace of T^n for all nonnegative integers n.*

Proof. By the previous proposition, any invariant subspace of T is also an invariant subspace of $T \circ T = T^2$, $T \circ T^2 = T^3$, and so on. Furthermore, T^0 is the identity map I, and any subspace is an invariant subspace of the identity map.

□

Corollary. *Any invariant subspace of an operator T is also an invariant subspace of the polynomial $c_nT^n + c_{n-1}T^{n-1} + \cdots + c_1T + c_0I$, where $c_0, \ldots, c_n$ are arbitrary scalars and I is the identity map.*

Proof. Any invariant subspace of T is also an invariant subspace of I (because every subspace is invariant with respect to I), and we've just proved that it's invariant under $T^2, \ldots, T^n$. So it's also invariant under any linear combination of $I, T, T^2, \ldots, T^n$, because the space of operators with a certain invariant subspace is a subspace of $\mathrm{End}(V)$.

□

6.3 Eigenvectors and eigenspaces

Key questions.

1. What is an *eigenvector* of an operator T? What's the difference between an eigenvalue of a *vector* and an eigenvalue of an *operator*? (In particular, which vector's eigenvalues do we not count as eigenvalues of an operator, and why?)

2. Can any vector besides **0** have multiple eigenvalues? Why or why not?

3. Does the set of eigenvectors of an operator with eigenvalue zero form a vector subspace? What about the set of eigenvectors with a particular eigenvalue $\lambda \neq 0$? What about the set of *all* eigenvectors of an operator, regardless of eigenvalue?

4. Give a formula $T(x, y) = (ax + by, cx + dy)$ with real numbers a, b, c, d such that T has no eigenvalues as an operator on $\mathbb{R}^2$, but does have eigenvalues as an operator on $\mathbb{C}^2$.

6.3.1 Definitions

Let V be a vector space over a base field $\mathbb{F}$, and let $T : V \to V$ be a linear operator on V, and let $\mathbf{v}$ be some vector in V. Suppose that $T\mathbf{v}$ is a multiple of $\mathbf{v}$: that is, $T\mathbf{v} = \lambda\mathbf{v}$ for some scalar λ. Then we define the following terms. These definitions are absolutely crucial: make sure you learn them!

1. $\mathbf{v}$ is an *eigenvector* of T.

2. λ is an *eigenvalue* of the vector $\mathbf{v}$.

   If $\mathbf{v} = \mathbf{0}$, then $T\mathbf{v} = \lambda\mathbf{v} = \mathbf{0}$ for every possible scalar λ, so $\mathbf{v}$ is trivially an eigenvector with every scalar as an eigenvalue. But if $\mathbf{v} \neq \mathbf{0}$, then it can have at most one eigenvalue: if $T\mathbf{v} = \lambda_1\mathbf{v} = \lambda_2\mathbf{v}$ with $\lambda_1 \neq \lambda_2$, then $(\lambda_1 - \lambda_2)\mathbf{v} = \mathbf{0}$, but a nonzero scalar $\lambda_1 - \lambda_2$ times a nonzero vector $\mathbf{v}$ can't be $\mathbf{0}$ (see page 30).

3. If $\mathbf{v} \neq \mathbf{0}$, then λ is called an *eigenvalue* of the operator T. An operator can have multiple eigenvalues, as we'll see soon.

From this definition, we can prove that the set of eigenvectors S of an operator $T : V \to V$ with a particular eigenvalue λ is a vector subspace of V. To prove this, we have to check the three subspace axioms (see page 37):

1. *Non-emptiness*: S must include $\mathbf{0}$, because $T\mathbf{0} = \mathbf{0} = \lambda\mathbf{0}$ no matter what T and λ are.

2. *Closure under addition*: The sum of any two eigenvectors is an eigenvector with the same eigenvalue: if $T\mathbf{v}_1 = \lambda\mathbf{v}_1$ and $T\mathbf{v}_2 = \lambda\mathbf{v}_2$, then $T(\mathbf{v}_1 + \mathbf{v}_2) = \lambda\mathbf{v}_1 + \lambda\mathbf{v}_2 = \lambda(\mathbf{v}_1 + \mathbf{v}_2)$.

3. *Closure under multiplication*: If $T\mathbf{v} = \lambda\mathbf{v}$, then $T(k\mathbf{v}) = kT(\mathbf{v}) = k(\lambda\mathbf{v}) = \lambda k\mathbf{v}$, so $k\mathbf{v}$ is also an eigenvector with eigenvalue λ.

We'll call the vector subspace comprising all the eigenvectors with a certain eigenvalue a *maximal eigenspace*, and any subspace of a maximal eigenspace will just be an *eigenspace* with no adjective attached.

Note that eigenspaces are the sets of eigenvectors with a *particular* eigenvalue. The sum of nonzero eigenvectors with different eigenvalues, as we'll prove in the next section, can never itself be an eigenvector.

6.3.2 Examples

To make this abstract discussion more concrete, let's look at a few simple transformations and compute their eigenvalues and eigenvectors.

1. The operator $T(x, y) = (2x, 3y)$ on $\mathbb{R}^2$ sends $\mathbf{e}_1 = (1, 0)$ to $2\mathbf{e}_1 = (2, 0)$, so $\mathbf{e}_1$ is an eigenvector of $T(x, y)$ with eigenvalue 2. Every multiple of $\mathbf{e}_1$ is also an eigenvector with eigenvalue 2: $T(k\mathbf{e}_1) = T(k, 0) = (2k, 0) = 2(k\mathbf{e}_1)$. (Even more generally, if $T\mathbf{v} = \lambda\mathbf{v}$, then $T(k\mathbf{v}) = k(T\mathbf{v}) = k\lambda\mathbf{v}$ for any scalar k, so any multiple of an eigenvector is also an eigenvector with the same eigenvalue.) Similarly, $\mathbf{e}_2 = (0, 1)$, and all of its scalar multiples, are eigenvalues with eigenvector 3. So T has two maximal eigenspaces: $\mathrm{span}\{\mathbf{e}_1\}$ with eigenvalue 2, and $\mathrm{span}\{\mathbf{e}_2\}$ with eigenvalue 3.

   (Again, $\mathbf{0}$ is an eigenvector of any operator and has every value of the base field as an eigenvalue: $T\mathbf{0} = \lambda\mathbf{0} = \mathbf{0}$ for every operator T and scalar λ. When we talk about the eigenvalues *of an operator*, we mean the eigenvalues of eigenvectors other than $\mathbf{0}$.)

2. The map $T(x, y, z) = (2x, 2y, 0)$ on $\mathbb{R}^3$ has $\mathbf{e}_1 = (1, 0, 0)$ and $\mathbf{e}_2 = (0, 1, 0)$ as eigenvectors, with eigenvalue 2 in each case. Furthermore, any linear combination of $\mathbf{e}_1$ and $\mathbf{e}_2$—that is, any vector of the form $(x, y, 0)$—is also an eigenvector with eigenvalue 2, as $T(x, y, z) = (2x, 2y, 0) = 2(x, y, 0)$.

   $\mathbf{e}_3 = (0, 0, 1)$ is an eigenvector with eigenvalue zero, because $T(0, 0, 1) = (0, 0, 0) = 0(0, 0, 1)$. Any eigenvector of T with eigenvalue zero is, by definition, in $\ker T$ (and vice versa: the elements of $\ker T$ are all eigenvalues with eigenvalue zero), so T has nonzero kernel (and so can't be bijective) if and only if it has zero as an eigenvalue. This fact becomes crucial later: there is a simple formula for computing the product of the eigenvalues of an operator on a finite-dimensional vector space, and if this product is zero, the operator can't be bijective.

   So T has two maximal eigenspaces: $\operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2\}$ with eigenvalue 2, and $\operatorname{span}\{\mathbf{e}_3\}$ with eigenvalue 0.

3. The map $T(x, y) = (x + y, 3x + 3y)$ has eigenvectors $(1, 3)$ with eigenvalue 2 (because $T(1, 3) = (2, 6)$) as well as $(1, -1)$ with eigenvalue 0. The spans $\operatorname{span}\{(1, 3)\}$ and $\operatorname{span}\{(1, -1)\}$ are maximal eigenspaces of T. Again, note the link between non-bijectivity and eigenvalue zero: the image of T contains only the multiples of $(1, 3)$.

4. $T(x, y) = (-x - y, 2x - 4y)$ has eigenvectors $(1, 1)$ (with eigenvalue -2) and $(1, 2)$ (with eigenvalue -3). You may notice that it would be a bit harder to compute the eigenvalues of this map just by looking at the formula than to compute the eigenvalues of, say, $T(x, y) = (2x, 3y)$. Later, we'll discuss a method to compute the eigenvalues of a linear operator. This method requires finding roots of polynomials with degree $\dim V$, and you generally would not want to do it without a computer, but it provides the underpinning for a vast amount of scientific and statistical computing.

5. $T_{\mathbb{R}}(x, y) = (-y, x)$, as an operator on $\mathbb{R}^2$, is geometrically a rotation of the plane $\mathbb{R}^2$ about the origin 90 degrees counterclockwise. You shouldn't be surprised that this rotation changes the direction of every nonzero vector and so doesn't have any nonzero eigenvectors. But the operator $T_{\mathbb{C}} : \mathbb{C}^2 \to \mathbb{C}^2$ with the same formula $T_{\mathbb{C}}(x, y) = (-y, x)$ does have eigenvectors, namely $(1, i)$ with eigenvalue $-i$, and $(1, -i)$ with eigenvalue i. The fact that an operator over a complex vector space can have more eigenvectors than the operator with the same formula over a real vector space will prove crucial for our analysis of real operators: it's often more convenient to identify real operators with the complex operators with the same formulas, especially for performing certain calculations.

It will gradually become clearer why the theory of eigenvectors and eigenspaces is worth studying. But there are a few broad reasons that we can at least allude to in advance:

1. For operators on finite-dimensional spaces, there is an algebraic formula, the *determinant*, that gives the product of the eigenvalues of the operator (raised to an integer power that is always at least the dimension of the corresponding maximal eigenspace). Since an operator is bijective if and only if it doesn't have 0 as an eigenvalue, checking if the determinant of an operator is zero will also tell you if it's bijective.

2. Eigenvalues and eigenspaces are the foundation for a set of classification theorems for linear operators, most importantly, the result that you can always find a basis for which a linear operator on a finite-dimensional complex space has a matrix representation of a particular type called *Jordan normal form*. The matrices that result from these classification theorems, furthermore, make it easier to glean some important aspects of a linear operator's behavior at a glance, analogous to how factoring a polynomial to find its roots can also make it easier to visualize.

3. They facilitate computations, especially with iterated applications of operators. Suppose, for instance, that you know that some vector $\mathbf{u}$ can be written as $c_1\mathbf{v}_1 + c_2\mathbf{v}_2$, where $\mathbf{v}_1$ and $\mathbf{v}_2$ are eigenvectors of T with eigenvalues λ_1 and λ_2. Suppose you want to compute $T^n\mathbf{u}$: that is, the result of applying the map T to $\mathbf{u}$ n times. Finding a general formula for T^n given a formula for T can require a lot of computation, but if you notice that $T^n\mathbf{u} = T^n(c_1\mathbf{v}_1 + c_2\mathbf{v}_2) = c_1\lambda_1^n\mathbf{v}_1 + c_2\lambda_2^n\mathbf{v}_2$, you can reduce the problem to simply computing powers of the scalars λ_1 and λ_2, and this requires much less computation. (Of course, there's the problem of how to find the coefficients c_1 and c_2 in the first place—a problem that leads to the theory of matrix diagonalization and change-of-basis matrices. We'll discuss this more in the next chapter, when we introduce matrix representations of linear maps.)

Answers to key questions.

1. An eigenvector of an operator T is any vector $\mathbf{v}$ such that $T\mathbf{v}$ is a multiple of $\mathbf{v}$. If $T\mathbf{v} = \lambda\mathbf{v}$ where λ is a scalar, then λ is an eigenvalue of the vector $\mathbf{v}$. But λ only counts as an eigenvalue of T if $\mathbf{v} \neq \mathbf{0}$, because $\mathbf{0}$ has every eigenvalue.

2. No: a nonzero vector $\mathbf{v}$ can have only one eigenvalue.

3. The set of eigenvectors with any *single specified* eigenvalue (including zero) is a vector subspace. But the set of all eigenvectors regardless of eigenvalue isn't, because the sum of eigenvectors with different eigenvalues is not generally an eigenvector (In fact, it's *never* an eigenvector, as we'll prove later.)

4. One example is the rotation 90 degrees counterclockwise: $T(x, y) = (-y, x)$.

6.4 Maximum eigenspace dimensions

Key questions.

1. When is a linear combination of eigenvectors also an eigenvector?

2. Is there a linear operator $T : \mathbb{R}^4 \to \mathbb{R}^4$ that has two dimension-2 eigenspaces with different eigenvalues? What about two dimension-3 eigenspaces?

You may have noticed from the examples in the last section that most maps only have a few eigenvalues. For instance, maps such as $T(x, y) = (2x, 3y)$ and $T(x, y) = (-x - y, 2x - 4y)$ have two eigenspaces with dimension 1, and maps such as $T(x, y) = (2x, 2y)$ have a single two-dimensional eigenspace. But could we find an operator on a two-dimensional space that has three nonzero eigenvectors with different eigenvalues? It turns out that we can't: the total dimensions of the eigenspaces of an operator on V are limited by $\dim V$. We'll establish this by a series of propositions.

Proposition. *Suppose $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are nonzero eigenvectors of some operator T, with all distinct eigenvalues $\lambda_1, \ldots, \lambda_n$. Then $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is linearly independent.*

Proof. We work by induction on n. The base case $n = 1$ is trivial: every set containing a single nonzero vector is linearly independent.

Now for the induction step, assume that for $n \geq 2$ arbitrary, every set of $n - 1$ eigenvectors with distinct eigenvalues is linearly independent. Suppose there's a linearly dependent set of n eigenvectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ with distinct eigenvalues $\lambda_1, \ldots, \lambda_n$. We can write $\mathbf{v}_n$ (after renumbering the $\mathbf{v}_i$ if necessary) as a linear combination of the others:

$$\mathbf{v}_n = c_1\mathbf{v}_1 + \cdots + c_{n-1}\mathbf{v}_{n-1}$$

where at least one of the coefficients c_i is nonzero.

This equation gives us two expressions for $\lambda_n\mathbf{v}_n$. We can multiply the equation by λ_n, giving

$$\lambda_n\mathbf{v}_n = \lambda_n c_1\mathbf{v}_1 + \cdots + \lambda_n c_{n-1}\mathbf{v}_{n-1}.$$

But we can also apply T to both sides, giving

$$\lambda_n\mathbf{v}_n = \lambda_1 c_1\mathbf{v}_1 + \cdots + \lambda_{n-1} c_{n-1}\mathbf{v}_{n-1}.$$

Subtracting the first equation from the second gives

$$\mathbf{0} = (\lambda_1 - \lambda_n)c_1\mathbf{v}_1 + \cdots + (\lambda_{n-1} - \lambda_n)c_{n-1}\mathbf{v}_{n-1}.$$

This is a linear combination of $\{\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}\}$ that has to be nontrivial, because all of the expressions $\lambda_i - \lambda_n$ and at least one of the coefficients c_i are nonzero. But the existence of this linear combination contradicts the induction hypothesis that $\{\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}\}$ was linearly independent. So $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ must be linearly independent.

□

Remark. This result also applies to infinite sets. Any *infinite* set of eigenvectors with distinct eigenvalues must be linearly independent, because an infinite set is linearly independent if and only if all of its finite subsets are also linearly independent (see page 43). Such a set can only exist in an infinite-dimensional vector space.

Corollary. *No linear combination of two or more nonzero eigenvectors with different eigenvalues can be an eigenvalue itself.*

Proof. Suppose $\mathbf{u}$ is a linear combination $c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$ of eigenvectors with distinct eigenvalues. If $\mathbf{u}$ is also an eigenvector, then either it has the same eigenvalue as one of the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$, or it has a different eigenvalue from all of them. But both cases let us construct a linearly dependent set of eigenvectors with distinct eigenvalues, contradicting the previous proposition:

1. If $\mathbf{u}$ has a different eigenvalue from all of the $\mathbf{v}_1, \ldots, \mathbf{v}_n$, then $\{\mathbf{v}_1, \ldots, \mathbf{v}_n, \mathbf{u}\}$ is a linearly dependent set of eigenvectors with distinct eigenvalues.

2. If $\mathbf{u}$ has the same eigenvalue as one of the other $\mathbf{v}_i$ (let's say, $\mathbf{v}_1$ with eigenvalue λ_1), then $\mathbf{u} - c_1\mathbf{v}_1$ is the difference of two eigenvectors with eigenvalue λ_1, so it also has eigenvalue λ_1. But $\mathbf{u} - c_1\mathbf{v}_1 = c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n$, so $\{\mathbf{u} - c_1\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is a linearly dependent set of eigenvectors with distinct eigenvalues.

□

Corollary. *Suppose $W_1, \ldots, W_n \subseteq V$ are eigenspaces of T with all distinct eigenvalues. Then the subspace sum $W_1 + \cdots + W_n$ is direct.*

Proof. If this sum were not direct, then (by definition of direct sums) there would be a nontrivial linear combination $\mathbf{0} = c_1\mathbf{w}_1 + \cdots + c_n\mathbf{w}_n$ of eigenvectors $\mathbf{w}_1 \in W_1, \ldots, \mathbf{w}_n \in W_n$ with all distinct eigenvalues, but such a combination is impossible.

□

Corollary. *The sum of dimensions of maximal eigenspaces of V is at most $\dim V$.*

Proof. Call the maximal eigenspaces $W_1, \ldots, W_n$. Maximal eigenspaces must all have different eigenvalues from each other: if W and W' are distinct eigenspaces with the same eigenvalue, then $W + W'$ is also an eigenspace that contains W and W', so neither W nor W' can be maximal. The subspace sum of maximal eigenspaces is direct, so $\dim(W_1 \oplus \cdots \oplus W_n) = \dim W_1 + \cdots + \dim W_n$. But $W_1 \oplus \cdots \oplus W_n \subseteq V$, so $\dim W_1 + \cdots + \dim W_n \leq \dim V$.

□

6.5 Multiplication-like qualities of operator composition

Throughout this book, we've justified our writing operators with the multiplication-like notation $T\mathbf{v}$ rather than the more conventional functional notation $T(\mathbf{v})$ by reference to the fact that that operators satisfy a number of axioms quite similar to those of multiplication. For instance, the axiom $T(k\mathbf{v}) = kT\mathbf{v}$ looks like commutativity of multiplication, and the left distribution rule$T(\mathbf{v} + \mathbf{w}) = T\mathbf{w} + T\mathbf{w}$ looks like the analogous formula $a(b+c) = ab+ac$ from ordinary algebra. The corresponding right distribution rule $(T_1 + T_2)\mathbf{v} = T_1\mathbf{v} + T_2\mathbf{v}$ also exists; it's basically how we define the sum of maps $T_1 + T_2$.

Operators on the same space, furthermore, can be composed with each other an arbitrary number of times, and composition has many of the same properties (both by itself and in relation to addition of operators) as multiplication has in relation to addition in more basic number systems. Specifically:[2]

1. Composition of linear maps, like composition of general functions on a set (see page 13), is associative: $A \circ (B \circ C) = (A \circ B) \circ C$.

2. Scalar multiplication of a map can be freely interchanged with map composition. If $\alpha, \beta \in \mathbb{F}$ are scalars and $A, B \in \mathrm{End}(V)$ are operators, then $(\alpha A) \circ (\beta B) = (\alpha\beta)(A \circ B)$. The proof is straightforward: for any vector $\mathbf{v} \in V$, we have $(\alpha\beta)(A \circ B)(\mathbf{v}) = \alpha\beta A(B(\mathbf{v}))$ (basically by definition of map-by-scalar multiplication: kT is the map that takes $\mathbf{v}$ to $k(T\mathbf{v})$), and similarly, $(\alpha A) \circ (\beta B)\mathbf{v} = \alpha A(\beta B(\mathbf{v}))$ (again by definition of map-by-scalar multiplication), which equals $\alpha\beta A(B(\mathbf{v}))$ because A is linear.

[2]There is a general concept in abstract algebra called an *associative algebra* over a field, which is essentially a vector space whose vectors can be multiplied by each other, and where this vector-by-vector multiplication operation follows a list of axioms very similar to the properties of function composition listed below. $\mathrm{End}(V)$ is an example, with composition playing the role of the multiplication operator.

3. Composition *left-distributes* over operator addition: $A \circ (B + C) = (A \circ B) + (A \circ C)$. If we have three operators A, B, C and an arbitrary vector $\mathbf{v}$, we can expand $A \circ (B + C)\mathbf{v} = A(B\mathbf{v} + C\mathbf{v})$ by the definition of the sum of the maps B and C, and this equals $(A \circ B)\mathbf{v} + (A \circ C)\mathbf{v}$, which equals $((A \circ B) + (A \circ C))\mathbf{v}$ by definition of operator sums.

4. Composition *right-distributes* over operator addition: $(A+B)\circ C = (A\circ C)+(B\circ C)$. The proof is similar to that of left-distribution.

This analogy between between function composition and multiplication is what justifies our use of exponential notation to indicate repeated function composition, such as T^2 for $T \circ T$, and T^3 for $T \circ T \circ T$. The algebra axioms also let expand and factor operators almost as if we were expanding and factoring polynomials. For instance, if T is an operator, then we can expand the expression $(2T + 5I) \circ (3T - 4I)$ like this:

$$
\begin{aligned}
&(2T + 5I) \circ (3T - 4I) \\
&= [(2T + 5I) \circ (3T)] + [(2T + 5I) \circ (-4I)] && \text{(left-distributivity)} \\
&= [(2T) \circ (3T)] + [(5I) \circ (3T)] + [(2T) \circ (-4I)] + [(5I) \circ (-4I)] && \text{(right-distributivity)} \\
&= 6(T \circ T) + 15(I \circ T) - 8(T \circ I) - 20(I \circ I) \\
& && \text{(compatibility of scalar multiplication and operator composition)} \\
&= 6T^2 + 15T - 8T - 20I && (I \text{ is the identity of operator composition}) \\
&= 6T^2 + 7T - 20I && \text{(distributivity of scalar addition over scalar multiplication)}
\end{aligned}
$$

We could also reverse these steps and write $6T^2 + 7T - 20I$ as $(2T + 5I)(3T - 4I)$. Note the analogy to polynomial factorization $(2x + 5)(3x - 4) = 6x^2 + 7x - 20$.

The point where the analogy between polynomials and linear operators breaks down is that unlike multiplication of real or complex numbers, operator composition is not commutative: $AB \neq BA$ for generic operators on a vector space. (We'll go back to not writing explicit function composition symbols for linear maps.) So, for instance, $(A + B)^2 = (A + B)(A + B)$ could be expanded to $A(A + B) + B(A + B)$ by left-distributivity and then further to $A^2 + AB + BA + B^2$ by right-distributivity, but we can't simplify this to $A^2 + 2AB + B^2$ unless we know that A and B commute. But operators always commute at least with powers of themselves and with the identity operator; so, for instance, if B is the identity map I, we can simplify $(A + I)^2$ to $A^2 + 2A + I$.

6.6 Commutative operators

Key questions.

1. What does it mean for two operators to commute?

2. Suppose A and B are two linear operators on $\mathbb{C}^2$, and A and B both have $1 + i$ and $2 - i$ as eigenvalues. Do A and B necessarily commute?

In general, operators do not commute with each other: if $T_1, T_2 : V \to V$ are two operators, then the compositions T_2T_1 (that is, applying T_1 first, then T_2) and T_1T_2 are usually not the same. One simple example: consider the operators $T_1(x, y) = (y, x)$ and $T_2(x, y) = (2x, y)$. Then $T_2T_1(x, y) = (2y, x)$ and $T_1T_2(x, y) = (y, 2x)$.

There is, in fact, a criterion that helps you determine (and, for some types of vector space, completely determines) whether operators commute. To show this, though, we'll need a preliminary proposition:

Proposition. *For any two operators $T_1, T_2 : V \to V$, the set of vectors $\mathbf{v} \in V$ such that $T_1T_2\mathbf{v} = T_2T_1\mathbf{v}$ (sometimes called the* commutator *of T_1 and T_2) is a subspace of V.*

Proof. This set is $\ker(T_2T_1 - T_1T_2)$, and kernels are subspaces.

□

Proposition. *Let $T_1, T_2 : V \to V$ be two linear operators. If there is a spanning set of V whose elements are all eigenvectors of both T_1 and T_2, then $T_1T_2 = T_2T_1$.*

Proof. If $\mathbf{v}$ is an eigenvector of both T_1 with eigenvalue λ and T_2 with eigenvalue μ, then $T_1T_2\mathbf{v} = T_1(\mu\mathbf{v}) = \mu T_1\mathbf{v} = \lambda\mu\mathbf{v}$ and similarly $T_2T_1\mathbf{v} = T_2(\lambda\mathbf{v}) = \lambda T_2\mathbf{v} = \lambda\mu\mathbf{v}$. That is, $T_1T_2\mathbf{v} = T_2T_1\mathbf{v}$. So the commutator of T_1 and T_2 is a subspace of V that contains every common eigenvector of T_1 and T_2. So if these common eigenvectors make up a spanning set of V, then T_1 and T_2 must commute on all of V.

□

The converse of this result is not generally true: there are operators that commute even though they don't have common eigenvectors. One example on $\mathbb{R}^2$ is the set of maps $T(x, y) = (x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta)$ which produce rotations by an angle θ counterclockwise about the origin. All of these maps commute with each other, even though only the identity $\theta = 0$ and the 180-degree turn $\theta = \pi$ have any eigenvectors in $\mathbb{R}^2$ at all. The converse does, in fact, hold on finite-dimensional vector spaces over $\mathbb{C}$, because linear transformations often (to put it loosely) have more eigenvectors in $\mathbb{C}$ than in $\mathbb{R}$. But we're not ready to prove this yet.

6.7 Generalized eigenvectors

Key questions.

1. What is a *generalized eigenvector*? If $T \in \mathrm{End}(\mathbb{R}^n)$, then what operator's kernel is the set of generalized eigenvectors of order 3 and eigenvalue 2?

2. What is a *generalized eigenspace* of an operator $T \in \mathrm{End}(V)$? Are generalized eigenspaces always vector subspaces of V? Are they always invariant subspaces of T? Why or why not?

3. Suppose $\mathbf{u}, \mathbf{v}, \mathbf{w}, \mathbf{x}$ are four generalized eigenvectors of an operator T. Suppose $\mathbf{u}$ and $\mathbf{v}$ have eigenvalue -5 and order 2, $\mathbf{v}$ has eigenvalue -5 and order 4, and $\mathbf{x}$ has eigenvalue 2 and order 2. Is $\mathbf{u} + \mathbf{v}$ a always, sometimes, or never a generalized eigenvector? If it can be a generalized eigenvector, what are its possible eigenvalues and orders? Answer the same questions for $\mathbf{u} + \mathbf{w}$ and $\mathbf{u} + \mathbf{x}$.

4. Explain how the reasoning that helped you answer the previous questions also proves that any sum of generalized eigenspaces with distinct eigenvalues is direct.

5. Construct an example of an operator on $\mathbb{R}^3$ that has a generalized eigenvector of (a) order 3 and eigenvalue 0, (b) order 3 and eigenvalue 1. Can an operator on $\mathbb{R}^3$ have an eigenvalue of order 4?

6.7.1 Definitions

We can rephrase the eigenvector criterion $T\mathbf{v} = \lambda\mathbf{v}$ as $\mathbf{v} \in \ker(T - \lambda I)$, where I is the identity map $I(\mathbf{v}) = \mathbf{v}$, and subtraction and scalar multiplication on linear maps are pointwise operations. This rephrasing gives us a more general notion:

Definition. *A **generalized eigenvector of an operator** T **with order** n **and eigenvalue** λ is an element of* $\ker(T - \lambda I)^n$ *that is not in* $\ker(T - \lambda I)^{n-1}$.

Equivalently, we can give a recursive definition: $\mathbf{0}$ is a generalized eigenvector of order zero for every eigenvalue, and $\mathbf{v}$ is a generalized eigenvector of order n if $T\mathbf{v} - \lambda\mathbf{v}$ is a generalized eigenvector of order $n - 1$. (As always, a superscript integer on an operator denotes composition with itself.)

You may wonder why generalized eigenvectors are worth studying. The basic reason is that we will want a way to find, for any operator $T : V \to V$, a basis of V that consists of eigenvectors of T. This basis is useful because it lets us write T in a simple matrix form that makes many important properties clear. The problem is that such a basis doesn't always exist, but we can get a lot closer if we allow bases with generalized eigenvectors as well. In fact, if V is a finite-dimensional vector space over $\mathbb{C}$, then a basis of generalized eigenvectors of any operator on V always exists. (This is not an intuitively obvious claim and will take us most of Chapter 8 to prove, but if I've done my job well, the intuition will be relatively clear by the time we've worked through the proof.)

Henceforth, we'll also denote multiples of the identity map by simple scalars and write, for instance, $T - 2$ to mean $T - 2I$.

Before we see why generalized eigenvectors are a useful concept, we have to study a few more of their properties and give a couple more definitions closely modeled off of the corresponding definitions for regular eigenspaces. If W is a subspace of V, and every element of W is a generalized eigenvector of some operator T with the same eigenvalue λ, then we'll call W a *generalized eigenspace with eigenvalue λ and order k*, where k is the largest order of any element in W. Like with ordinary, non-generalized eigenspaces, we'll call the set of all eigenvectors with eigenvalue λ and order $\leq k$ the *maximal generalized eigenspace with eigenvalue λ and order k*. This space, by definition, is just $\ker(T - \lambda)^k$, and kernels are vector subspaces, so the maximal generalized eigenspace by this definition is also a vector subspace.

Finally, the *maximal generalized eigenspace with eigenvalue λ*, without a specified order, is the set of generalized eigenvectors with eigenvalue λ and any order. This is the infinite union $\ker(T - \lambda) \cup \ker(T - \lambda)^2 \cup \ker(T - \lambda)^3 \cup \cdots$, where each term in the union is a subspace of the next term. This is also a vector subspace of V, as an immediate corollary of this proposition:

Proposition. *Suppose V is a vector space and $W_1 \subseteq W_2 \subseteq W_3 \subseteq \cdots$ is an infinite sequence of (possibly not all distinct) nested subspaces of V. Define the infinite union $X \coloneqq W_1 \cup W_2 \cup W_3 \cup \cdots$: that is, $\mathbf{v} \in X$ if $\mathbf{v} \in W_k$ for some positive integer k (and thus is also in W_ℓ for all integers $\ell \geq k$). Then X is a subspace of V. Furthermore, if the spaces W_i are all invariant subspaces of an operator T, then X is also invariant under T.*

Proof. To prove that X is a vector subspace, we'll check the three subspace properties:

1. *Closure under addition*: Suppose $\mathbf{v}_1, \mathbf{v}_2 \in X$. Then there are integers k_1, k_2 such that $\mathbf{v}_1 \in W_{k_1}$ and $\mathbf{v}_2 \in W_{k_2}$. Define $k = \max(k_1, k_2)$. Then $\mathbf{v}_1$ and $\mathbf{v}_2$ are both in W_k because W_k itself is a vector subspace and therefore closed under addition. So $\mathbf{v}_1 + \mathbf{v}_2$ is also in W_k and thus in X.

2. *Closure under addition*: If $\mathbf{v} \in X$, then $\mathbf{v} \in W_k$ for some integer k. Thus, $c\mathbf{v} \in W_k$ for any scalar c because W_k is closed under multiplication, so $c\mathbf{v} \in X$.

3. *Non-emptiness*: The subspaces W_i all have to include $\mathbf{0}$, so X does as well.

To prove that X is invariant, note that if $\mathbf{v}$ is an arbitrary element of X, then $\mathbf{v} \in W_k$ for some k. Thus, $T\mathbf{v} \in W_k \subseteq X$ because W_k is invariant under T.

□

This result won't be that useful to us, as we're mostly interested in finite vector spaces, and it turns out that generalized eigenvectors in an n-dimensional space can't have orders greater than n (so a maximal generalized subspace of unspecified order is also a maximal generalized eigenspace of order n). But it is theoretically nice.

Maximal generalized eigenspaces (whether or not they have specified orders), as we'll see in a bit, are always invariant subspaces. Finally, note that by our definitions, $\mathbf{0}$ and $\{\mathbf{0}\}$ are a generalized eigenvector and a generalized eigenspace, both with order zero. (This is because $(T - \lambda)^0$, just like the zeroth power of any map, is the identity operator, which has kernel $\{\mathbf{0}\}$.)

6.7.2 Examples

An example should help to make this abstract discussion clearer. Consider the map $T : \mathbb{R}^4 \to \mathbb{R}^4$ given by $T(x, y, z, w) = (3x + y, 3y, 2w, 0)$. This map has two generalized eigenspaces of order 2:

1. One generalized eigenspace has basis $\{\mathbf{e}_1, \mathbf{e}_2\}$ and eigenvalue 3. One basis vector for this space is $\mathbf{e}_1 = (1, 0, 0, 0)$, which is an eigenvector because $T\mathbf{e}_1 = (3, 0, 0, 0) = 3\mathbf{e}_1$. On the other hand, $\mathbf{e}_2 = (0, 1, 0, 0)$ is a generalized eigenvector of T, because $T\mathbf{e}_2 = (1, 3, 0, 0)$ and so $(T - 3I)\mathbf{e}_2 = \mathbf{e}_1$, which is an eigenvector of T. Repeated application of the operator $T - 3I$ establishes a "chain" of generalized eigenvectors $\mathbf{e}_2 \mapsto \mathbf{e}_1 \mapsto \mathbf{0}$, each application reducing the order by 1, and the nonzero elements of this chain make up a basis for the subspace.

2. The other generalized eigenspace has basis $\{\mathbf{e}_3, \mathbf{e}_4\}$ and eigenvalue 0. In this space, $\mathbf{e}_3$ is an eigenvector because $T\mathbf{e}_3 = \mathbf{0}$, so $\mathbf{e}_3 \in \ker T = \ker(T - 0)$. Meanwhile, $\mathbf{e}_4$ is a generalized eigenvector of order 2, because $T\mathbf{e}_4 = 2\mathbf{e}_3$ and so $T^2\mathbf{e}_4 = \mathbf{0}$. Again, we can describe this subspace with a chain basis $\mathbf{e}_4 \mapsto 2\mathbf{e}_3 \mapsto \mathbf{0}$.

The phrases "generalized eigenvector" and "generalized eigenspace" are cumbersome, so from now on, we'll use the abbreviations GEV and GES.

6.7.3 Sums of generalized eigenvectors and eigenspaces

As with ordinary eigenvectors, linear combinations of GEVs with the same eigenvalue are also GEVs, but linear combinations of GEVs with different eigenvalues are not. The rest of this section presents this result, and a corollary result that the sums of generalized eigenspaces with distinct eigenvalues are always direct. The propositions here are analogous to the results in Section 6.4, and use similar proof techniques. They're a bit cumbersome to state, but we'll need them for bigger theorems later.

It's easy to see that any nonzero scalar multiple of a GEV is also a GEV with the same order: if $k \neq 0$, then $(T-\lambda)^n\mathbf{v} = \mathbf{0}$ if and only if $(T-\lambda)^n(k\mathbf{v}) = k(T-\lambda)^n\mathbf{v} = \mathbf{0}$. The sum of GEVs with the same eigenvalue is also a GEV, but it can have more than one possible order. We'll state the result below.

Proposition. *The sum of two GEVs $\mathbf{v}, \mathbf{w}$ with the same eigenvalue λ and orders m and n is a GEV with eigenvalue λ. If $m \neq n$, then $\mathbf{v}+\mathbf{w}$ has order $\max(m,n)$; if $m = n$, then $\mathbf{v}+\mathbf{w}$ can have any order between 0 and n.*

Proof. Suppose $m \leq n$ (the $m > n$ case is symmetrical). Then $(T-\lambda)^n(\mathbf{v}+\mathbf{w}) = (T-\lambda)^n\mathbf{v} + (T-\lambda)^n\mathbf{w} = \mathbf{0}+\mathbf{0} = \mathbf{0}$, so $\mathbf{v}+\mathbf{w}$ is a GEV with order at most n. If $m < n$, then $(T-\lambda)^{n-1}(\mathbf{v}+\mathbf{w}) = (T-\lambda)^{n-1}\mathbf{w} \neq \mathbf{0}$, so $\mathbf{v}+\mathbf{w}$ has order exactly n.

If $m = n$, however, then $(T-\lambda)^{n-1}(\mathbf{v}+\mathbf{w})$ is the sum of two nonzero vectors $(T-\lambda)^{n-1}\mathbf{v}$ and $(T-\lambda)^{n-1}\mathbf{w}$ that could be each other's negatives, so we can't conclude that $\mathbf{v}+\mathbf{w}$ has order n. One family of counterexamples is $\mathbf{v} = \mathbf{u}+\mathbf{x}, \mathbf{w} = \mathbf{u}-\mathbf{x}$ where $\mathbf{u}$ is a GEV of any order $h \leq n$ and $\mathbf{x}$ is a GEV of order n. In this case, $\mathbf{v}$ and $\mathbf{w}$ have order n but $\mathbf{v}+\mathbf{w}$ has order h.

□

This result has several important corollaries, the first of which we foreshadowed in the last subsection:

Corollary. *Maximal generalized eigenspaces of any order (or no order) are invariant subspaces.*

Proof. Let V be a vector space, let $T \in \mathrm{End}(V)$, and let $W = \ker($ be a GES of V containing all vectors with eigenvalue λ and order $\leq n$. For any $\mathbf{w} \in W$, define $\mathbf{v} = (T-\lambda I)\mathbf{w} = T\mathbf{w} - \lambda\mathbf{w}$. Then $\mathbf{v}$ is a GEV with eigenvalue λ and order $\leq n-1$, so $\mathbf{v} \in W$. So $T\mathbf{w} = \mathbf{v} + \lambda\mathbf{w}$ is the sum of two elements of W, so it's also in W. This proves that $T(W) \subseteq W$, so W is an invariant subspace of T.

□

Corollary. *Any set of nonzero GEVs with the same eigenvalue and distinct orders is linearly independent.*

Proof. Call the set S. Any nontrivial linear combination of vectors in S must be a GEV whose order is the maximum order of any vector in the linear combination. In particular, it can't have order zero, which is the order of the vector $\mathbf{0}$ and nothing else. So no linear combination from S can equal $\mathbf{0}$, so S is linearly independent.

□

Corollary. *No GEV of an operator on an n-dimensional space can have order greater than n.*

Proof. If $\mathbf{v}$ is a GEV with eigenvalue λ and order k, then $\{\mathbf{v}, (T-\lambda)\mathbf{v}, \ldots, (T-\lambda)^{k-1}\mathbf{v}\}$ is a set of k generalized eigenvectors with every order from 1 to k, so by the previous corollary, it must be linearly independent. And an n-dimensional vector space can't contain a set of more than n linearly independent vectors. Thus, $k \leq n$.

□

Finally, sums of GEVs with different eigenvalues can't be GEVs themselves, just as with ordinary eigenvectors.

Proposition. *Any (possibly infinite) set of nonzero GEVs with distinct eigenvalues is linearly independent.*

Proof. We'll prove this for finite sets of size n first; the logic is similar to the analogous result for ordinary eigenvectors. The $n = 1$ case is trivial: any one-vector set is linearly independent as long as the vector it contains isn't $\mathbf{0}$. For $n = 2$, any linearly dependent set of two nonzero vectors must have each vector be a scalar multiple of the other, but such vectors must have the same eigenvalue and order.

By induction for $n \geq 3$, suppose $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a linearly dependent set of nonzero GEVs with orders $m_1, \ldots, m_n$ and distinct eigenvalues $\lambda_1, \ldots, \lambda_n$. By the induction hypothesis, $\{\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}\}$ is linearly independent, so if $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is linearly dependent, then we can write $\mathbf{v}_n$ as a linear combination $\mathbf{v}_n = c_1\mathbf{v}_1 + \cdots + c_{n-1}\mathbf{v}_{n-1}$, where at least one of the scalars $c_1, \ldots, c_{n-1}$ is nonzero. Applying the operator $(T-\lambda_n)^{m_n}$ to both sides of this equation gives a linear combination $\mathbf{0} = c_1(\lambda_1 - \lambda_n)^{m_1}\mathbf{v}_1 + \cdots + c_{n-1}(\lambda_{n-1} - \lambda_n)^{m_{n-1}}\mathbf{v}_{n-1}$ that must be nontrivial because λ_n does not equal any of the other eigenvalues λ_i, contradicting the induction hypothesis.

The generaliztion to infinite sets follows because an infinite set is linearly independent if and only if all of its finite subsets are as well.

□

Corollary. *No nonzero linear combination of two or more nonzero GEVs with distinct eigenvalues can be a GEV itself.*

Proof. Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be GEVs with eigenvalues $\lambda_1, \ldots, \lambda_n$, and suppose $\mathbf{u} \coloneqq c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$ is a GEV with eigenvalue μ. Then either μ is distinct from all of the λ_i, so $\{\mathbf{v}_1, \ldots, \mathbf{v}_n, \mathbf{u}\}$ is a linearly dependent set of GEVs with distinct eigenvalues; or μ equals one of the λ_i (say, $\mu = \lambda_1$), so $\{\mathbf{u} - c_1\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is a linearly dependent set of GEVs with distinct eigenvalues. By the previous corollary, both cases are impossible, so $\mathbf{u}$ can't be a GEV.

□

Corollary. *The subspace sum of any number of generalized eigenspaces with all distinct eigenvalues is direct.*

Proof. Like the analogous results for ordinary eigenspaces on page 141.

□

This result is a building block of an important, fundamental theorem: any finite-dimensional vector space over $\mathbb{C}$ can be completely decomposed as a direct sum of the maximal generalized eigenspaces of an arbitrary operator. (This is a more abstract phrasing of a result that's usually put in matrix language: the existence of a particular matrix representation called Jordan normal form for any operator on $\mathbb{C}^n$.)

We have one more proposition that will be useful for the next section:

Proposition. *If V is the eigenspace of some operator T with eigenvalue λ, and W is a GES of order n with eigenvalue λ, then $\dim W \leq n \dim V$.*

Proof. Since $V = \ker(T - \lambda)$ and $W = \ker(T - \lambda)^n$, the result follows from the formula $\dim \ker T^n \leq n \dim \ker T$ valid for any operator T, itself an immediate corollary of the result $\dim \ker(T_2 T_1) \leq \dim \ker T_1 + \dim \ker T_2$ for arbitrary linear maps T_1, T_2 proved on page 66.

□

6.8 Jordan bases of generalized eigenspaces

This section presents one important result: we can construct a basis for a maximal GES of finite order that is made up entirely out of elements of chains. This result will be one huge piece of a major theorem in matrix theory: the existence of a *Jordan normal form* for every complex matrix.

First, some definitions. Throughout this section, V is a vector space, and T is an operator on V such that every element of V is a GEV of T with the same eigenvalue λ. (That is, V itself is a maximal generalized eigenspace of T.) For notational convenience, write $T' = T - \lambda$. Call a set $S \subset V$ *chained* if it satisfies the following properties:

1. For every element $\mathbf{v} \in S$, either $T'\mathbf{v} \in S$ or $T'\mathbf{v} = \mathbf{0}$.
2. There are no two distinct elements $\mathbf{v}_1, \mathbf{v}_2 \in S$ such that $T'\mathbf{v}_1 = T'\mathbf{v}_2 \neq \mathbf{0}$.

That is, we can organize the elements of S into a set of chains such that T' takes each element to the next element in its chain (or $\mathbf{0}$), and the chains don't merge until $\mathbf{0}$. Call $\mathbf{v} \in S$ the *predecessor* of $\mathbf{w} \in S$, and $\mathbf{w}$ the *successor* of $\mathbf{v}$, if $T'\mathbf{v} = \mathbf{w}$.

Finally, if S is a linearly independent chained set that is also a basis of V, then we'll call it a *Jordan basis* of V with respect to T.

The main result of this chapter is the following lemma, which shows with an explicit construction that a Jordan basis for V must exist.

Lemma. *If V is finite-dimensional and every element of V is a generalized eigenvector of T with the same eigenvalue λ, then there is a Jordan basis for V. Furthermore, any two Jordan bases must have matching chain lengths: that is, for every integer k, the two bases must have the same number of length-k chains.*

Proof. Again, write $T' = T - \lambda$. Let h be the maximum order of any element of V, and define the subspaces $\{\mathbf{0}\} = W_0 \subsetneq W_1 \subsetneq W_2 \subsetneq \cdots \subsetneq W_h = V$ as the maximal GESes of order $0, 1, 2, \ldots, h$ (that is, $W_k = \ker(T')^k$ for all integers $0 \leq k \leq h$). The dimensions of the spaces W_i determine the chain lengths of any possible Jordan basis: there must be $\dim W_1$ chains in total (because every chain contains exactly one element of order 1), $\dim W_2 - \dim W_1$ chains of length at least 2 (because every such chain contains exactly one element of order 2), $\dim W_3 - \dim W_2$ chains of length at least 3, and so on. So any two Jordan bases must have chains with matching lengths.

So we need to show that at least one Jordan basis exists. Remember from page 120 that the basis for a direct sum can be assembled from the union of bases for the direct sum's constituent subspaces. We'll construct one explicitly as the union of the bases $B_1, \ldots, B_h$ of a list of subspaces $U_1, \ldots, U_h$ with the following properties:

1. $W_i = W_{i-1} \oplus U_i$ for all integers $1 \leq i \leq h$. This means that $W_1 = U_1$ and $V = W_h = U_1 \oplus \cdots \oplus U_h$, so the union of bases for $U_1, \ldots, U_h$ is in fact a basis for V.

2. If $\mathbf{v} \in B_1$, then $T'\mathbf{v} = \mathbf{0}$. (This follows from the fact that B_1 is a basis of $U_1 = W_1$, the set of all ordinary (i.e. order-1 generalized) eigenvectors.)

3. If $\mathbf{v} \in B_i$ for $i \geq 2$, then $T'\mathbf{v} \in B_{i-1}$, and there is no other vector $\mathbf{v}' \in B_i$ such that $T'\mathbf{v} = T'\mathbf{v}'$. These two points establish that $B_1 \cup \cdots \cup B_h$ is in fact a chained set.

The $W_i = W_{i-1} \oplus U_i$ requirement means that $U_i \cap W_{i-1} = \{\mathbf{0}\}$; that is, every element of U_i except for $\mathbf{0}$, and in particular every element of B_i, has order exactly i. We'll define B_i and U_i iteratively as follows:

1. Take an arbitrary basis of W_{h-1}, and define B_h as any linearly independent set of vectors that we can add to extend this basis to a basis of $W_h = V$. Since the elements of B_h are all outside W_{h-1} (which contains every vector of order $h-1$ or less), they must have order h. Define $U_h = \operatorname{span} B_h$, and note that $U_h \cap W_{h-1} = \{\mathbf{0}\}$.

2. For $i = h - 1$ down to $i = 1$, assume that we already have B_{i+1} and $U_{i+1} = \operatorname{span} B_{i+1}$ defined, and that every nonzero element of U_{i+1} has order $i + 1$.

   Define $B'_i = \{T'\mathbf{v} : \mathbf{v} \in B_{i+1}\}$. So B'_i is a set of generalized eigenvectors of order i. Furthermore, if $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are arbitrary elements of B_{i+1}, then every linear combination $c_1T'\mathbf{v}_1 + \cdots + c_kT'\mathbf{v}_k = T'(c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k)$ where at least one of the coefficients $c_1, \ldots, c_n$ is nonzero must have order exactly i. To see why, note that if the linear combination $c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k$ is nontrivial, then it must be nonzero (because B_{i+1} is linearly independent) and so it must have order $i + 1$ (because the only element of $U_{i+1} = \operatorname{span} B_{i+1}$ that doesn't have order $i + 1$ is $\mathbf{0}$), which means that $T(c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k)$ has order i. In particular, if $c_1T'\mathbf{v}_1 + \cdots + c_kT'\mathbf{v}_k$ has nonzero order i, then it can't be $\mathbf{0}$, meaning that B'_i is linearly independent and no two vectors in B_{i+1} have the same successor in B'_i. And as every element of $\operatorname{span} B'_i$ has order i, furthermore, $\operatorname{span} B'_i \cap W_{i-1} = \{\mathbf{0}\}$.

   Define $U'_i = \operatorname{span} B'_i$. Extend B'_i arbitrarily to some basis C of $U'_i \oplus W_{i-1}$ by adding some basis of W_{i-1}, then let D be any set of basis vectors required to extend C to a basis of W_i, and define $B_i = B'_i \cup D$ and $U_i = \operatorname{span} B_i$. By construction, U_i and W_{i-1} are disjoint except at $\{\mathbf{0}\}$. Each vector in D begins a new chain, and each vector in B'_i continues a chain from B_{i+1}.

The union $B = B_1 \cup \cdots \cup B_h$ is thus a basis for V, and we can rearrange it to put chains together.

□

With a bit more work and notational complexity, we could generalize this theorem to prove the existence of Jordan bases for maximal generalized eigenspaces of unspecified order in infinite-dimensional vector spaces. (In this case, we may have an infinite number of chains, and individual chains may have infinite length.) But we won't need the infinite-dimensional case for this book.

6.9 Linear recurrences and differential equations

Key questions.

1. What is the space $\mathbb{F}^{\mathbb{N}}$? What is the left-shift operator $L : \mathbb{F}^{\mathbb{N}} \to \mathbb{F}^{\mathbb{N}}$? How can we translate the fact that a sequence satisfies the recurrence equation $a_{n+2} = a_n + a_{n+1}$ into a statement involving an operator constructed from L?

2. What kind of subset of $\mathbb{F}^{\mathbb{N}}$ contains solutions to the recurrence $a_{n+2} = a_n + a_{n+1} + f(n)$, where f is some function on the integers with values in $\mathbb{F}$?

3. What are the eigenvalues and corresponding eigenvectors of the left-shift operator $L : \mathbb{F}^{\mathbb{N}} \to \mathbb{F}^{\mathbb{N}}$? What about the generalized eigenvectors?

4. If A and B are two operators on the same space, then does $\ker(AB)$ always have to contain $\ker A$? Does it always have to contain $\ker B$? Explain why or give a counterexample. (Hint: consider the operators $A(x, y) = (x, 0)$ and $B(x, y) = (y, x)$ on $\mathbb{R}^2$.) When can we conclude that $\ker A \subseteq \ker(AB)$?

5. ($\star$) Suppose that T is a linear operator on a complex vector space that has 0, 1, and 2 as eigenvalues, all with one-dimensional eigenspaces. What are the possible dimensions of $\ker(T^3 - 3T^2 + 2T)$? What about $\ker(T^3 - 2T^2)$?

6. ($\star$) If A is any operator on a space and I is the identity operator, explain why $\ker(A^2 - I)$ contains both $\ker(A + I)$ and $\ker(A - I)$. Can $\ker(A^2 - I)$ contain any elements outside $\ker(A + I) \oplus \ker(A - I)$? Why or why not?

7. ($\star\star$) Explain why $\ker(A^2 - B^2)$ may not contain $\ker(A + B)$ or $\ker(A - B)$ for general operators A and B, despite the seeming identity $A^2 - B^2 = (A + B)(A - B)$. (One counterexample is $A(x, y) = (-y, x)$ [rotation of $\mathbb{R}^2$ counterclockwise by 90 degrees] and $B(x, y) = (y, x)$ [reflection about $y = x$]; then $(1, 0)$ and $(0, 1)$ are respectively in $\ker(A - B)$ and $\ker(A + B)$, but neither is in the kernel of $A^2 - B^2 = -3I$.) Are there conditions that you can impose on A and B to ensure that $\ker(A + B) \subseteq \ker(A^2 - B^2)$ and $\ker(A - B) \subseteq \ker(A^2 - B^2)$?

8. If I is a real interval, what vector space does the notation $\mathcal{C}^\infty(I)$ denote? Why is $\frac{d}{dx}$ a linear operator on this space? What are the eigenvalues and (generalized) eigenvectors of $\frac{d}{dx}$?

Our theory of linear operators may seem scant so far, but it's enough for us to solve two fundamental and closely analogous problems in discrete mathematics and differential equations: general solutions of *linear recurrences* and *linear homogeneous ordinary differential equations with constant coefficients*. Linear recurrences are sequences of integers defined by k arbitrary starting values and a recursive equation that gives all subsequent values of the sequence as a linear combination with fixed coefficients of the k preceding values. One well-known example is the Fibonacci sequence F_n, with the starting values $F_0 = 0, F_1 = 1$ and the recurrence $F_{n+2} = F_n + F_{n+1}$ for all integer $n \geq 0$: this sequence begins 0, 1, 1, 2, 3, 5, 8, 13, 21. Basic operator theory gives a method for finding closed-form equations, without recursion, for all linear recurrences.

Linear homogeneous ODEs with constant coefficients, meanwhile, are differential equations of the form $c_n \dfrac{d^n y}{dx^n} + c_{n-1} \dfrac{d^{n-1} y}{dx^{n-1}} + \cdots + c_0 y = 0$. Differential equations usually

have an infinite number of possible solutions: we can specify one solution by giving the value of $y, \frac{dy}{dx}, \ldots, \frac{d^{n-1}y}{dx^{n-y}}$ and its first $n-1$ derivatives at some specific point x_0 in the domain. A similar method to the one we use to solve linear recurrences also gives a solution for all such differential equations.

6.9.1 Solving the Fibonacci sequence

Before we develop the general theory of linear recurrences, we'll look at one specific example: finding a closed form for the Fibonacci sequence. A reminder of notation: $\mathbb{F}^{\mathbb{N}}$ is the set of infinite sequences of elements of a field $\mathbb{F}$, with vector addition defined as $(a_0, a_1, a_2, \ldots) + (b_0, b_1, b_2, \ldots) = (a_0 + b_0, a_1 + b_1, a_2 + b_2 \ldots)$ and scalar multiplication defined as $k(a_0, a_1, a_2, \ldots) = (ka_0, ka_1, ka_2, \ldots)$. (It's arbitrary whether we start sequence indices at 0, 1, or anything else.) The zero element of $\mathbb{F}^{\mathbb{N}}$ is, of course, $\mathbf{0} = (0, 0, 0, \ldots)$.[3] We'll also denote scalar multiples kI of the identity operator just as k, with I itself denoted as 1.

We'll also define the *left-shift operator* $L : \mathbb{F}^{\mathbb{N}} \to \mathbb{F}^{\mathbb{N}}$ as $L(a_0, a_1, a_2, \ldots) = (a_1, a_2, a_3, \ldots)$. The squared operator $L^2 = L \circ L$ shifts left by two places, L^3 by three, and so on. It's easy to prove that L is linear, and its eigenvectors with eigenvalue λ are sequences for which multiplying each element by λ is equivalent to replacing it with the element one place to the right: that is, the geometric sequences of the form $(k, k\lambda, k\lambda^2, k\lambda^3, \ldots)$, for some arbitrary $k \in \mathbb{F}$. This also means that the eigenvectors with eigenvalue zero are the ones with zeros in every entry but the first. (L also has generalized eigenvectors of order 2 and greater, but we'll leave those aside for now.)

Now call a sequence of complex numbers[4] $(a_0, a_1, a_2, \ldots) \in \mathbb{C}^{\mathbb{N}}$ *Fibonacci-like* if it satisfies the Fibonacci recursion $a_{n+2} - a_{n+1} - a_n = 0$ for all integers $n \geq 0$, but doesn't necessarily satisfy the Fibonacci initial condition $a_0 = 0, a_1 = 1$. We're going to find a general form for all Fibonacci-like sequences.

To start, note that we can reformulate the recurrence as a statement that the sequence $(a_2 - a_1 - a_0, a_3 - a_2 - a_1, a_4 - a_3 - a_2, \ldots, a_{n+2} - a_{n+1} - a_n, \ldots)$ is just $(0, 0, 0, \ldots) = \mathbf{0}_{\mathbb{C}^{\mathbb{N}}}$. This sequence, meanwhile, is the image of the operator $R \coloneqq L^2 - L - 1$ applied to $(a_0, a_1, a_2, \ldots)$, so the Fibonacci-like sequences are the kernel of the operator R. (We'll use the letter R to suggest "recurrence"; don't get it confused with "right-shift.")

Now the key step: since operators on $\mathbb{C}^{\mathbb{N}}$ form an associative algebra over $\mathbb{C}$ with function composition taking the role of multiplication (remember section **??**), we can factor R as either $R = (L - \phi_+)(L - \phi_-)$ or $R = (L - \phi_-)(L - \phi_+)$, where $\phi_+ = \frac{1+\sqrt{5}}{2}$ and $\phi_- = \frac{1-\sqrt{5}}{2}$ are the roots of the polynomial $x^2 - x - 1$. (Remember that L, like any operator, commutes with multiples and powers of itself and with multiples of I.) Since $\ker(T_2 T_1)$ has to contain $\ker T_1$ for any pair of operators T_1, T_2 (because $T_1\mathbf{v} = \mathbf{0}$ implies $(T_2 T_1)\mathbf{v} = T_2(T_1\mathbf{v}) = T_2\mathbf{0} = \mathbf{0}$ for any $\mathbf{v}$), so $\ker R$ must contain both $\ker(L - \phi_-)$ and $\ker(L + \phi_+)$, which are both one-dimensional eigenspaces of T. So R must also contain the *sum* of these eigenspaces, which has dimension 2 because the sum of eigenspaces with different eigenvalues is direct.

[3] Don't confuse $\mathbb{F}^{\mathbb{N}}$ with its subset $\mathbb{F}^{\infty}$, which is the set of infinite sequences with only a finite number of nonzero terms.

[4] You may wonder why we're working with the complex numbers when the Fibonacci sequence contains only integers. The answer is that polynomial factoring, which is a key part of our method, is simpler in the complex numbers thanks to the existence of the Fundamental Theorem of Algebra.

Note that $\ker(L-\phi_-)\oplus\ker(L-\phi_+)$ is the set of linear combinations of the eigenvectors $(1,\phi_-,\phi_-^2,\ldots)$ (which is a basis for the one-dimensional subspace $\ker(L-\phi_- I)$) and $(1,\phi_+,\phi_+^2,\ldots)$ (which is a basis for $\ker(L-\phi_+)$). That is, $\ker(L-\phi_-)\oplus\ker(L-\phi_+ I)$ is the set of sequences with the general form $(k_1+k_2, k_1\phi_- + k_2\phi_+, k_1\phi_-^2+k_2\phi_+^2,\ldots)$ for arbitrary constants $k_1, k_2 \in \mathbb{C}$.

We claim, finally, that $\ker(L-\phi_-)\oplus\ker(L-\phi_+)$ is in fact all of $\ker R$. This follows from a simple dimensionality argument: since R is the composition $(L-\phi_-)(L-\phi_+)$ of two maps whose kernels have dimension 1, $\dim\ker R$ can be at most 2 (see page 66).

So every Fibonacci-like sequence has the form $(k_1+k_2, k_1\phi_- + k_2\phi_+, k_1\phi_-^2+k_2\phi_+^2,\ldots)$. We can find values of k_1 and k_2 that generate the actual Fibonacci sequence by using the initial values $F_0=0, F_1=1$ and solving the system $k_1+k_2=0$, $k_1\phi_- + k_2\phi_+ = 1$ with standard linear system methods (the solution is $k_1 = -\frac{1}{\sqrt{5}}$ and $k_2 = \frac{1}{\sqrt{5}}$). So the Fibonacci sequence has the closed form

$$f(n) = \frac{1}{\sqrt{5}}\frac{(1+\sqrt{5})^n - (1-\sqrt{5})^n}{2^n}.$$

.

If you want a closed form for a sequence with different starting numbers but the same recurrence (say, the Lucas numbers $2, 1, 3, 4, 7, 11, 18, \ldots$) you would just need to use the starting numbers to compute different values of k_1, k_2.

6.9.2 Solving general linear recurrences

The method that we used for the Fibonacci sequence generalizes to all linear recurrences: write the operator $R : \mathbb{C}^{\mathbb{N}} \to \mathbb{C}^{\mathbb{N}}$ as a factored polynomial of the left-shift operator L. This polynomial is sometimes called the "characteristic polynomial" of the recurrence; note that it will never have a constant term (or, therefore, root) of zero. If L has no repeated roots (a slight complication that we'll get to later), then each factor contributes one dimension to $\ker R$, and a dimensionality argument shows that the result is the entirety of $\ker R$. This procedure gives a general form for every sequence that satisfies the recurrence, and any set of initial values of the sequence give a system that can be solved to give values for the constants in this general form.

As another example (which also shows us why complex numbers are useful for solving real sequences), let's solve the recurrence $x_{n+3} = x_{n+2} - 4x_{n+1} + 4x_n$ with initial condition $x_0 = x_2 = 1$, $x_1 = 0$. Sequences that satisfy this recurrence are the kernel of the operator $R = L^3 - L^2 + 4L^2 - 4$, which factors as $(L+2i)(L-2i)(L-1)$. Now, $\ker R$ must contain $\ker(L+2i)\oplus\ker(L-2i)\oplus\ker(L-1)$, which is the three-dimensional set of sequences with general term $a_n = k_1(2i)^n + k_2(-2i)^n + k_3$. And R is the composition of three maps with one-dimensional kernels, so $\ker R$ can have dimension at most 3, and $\ker(L+2i)\oplus\ker(L-2i)\oplus\ker(L-1)$ is the entire kernel of R. Using the initial conditions to solve for the coefficients k_1, k_2, k_3 gives the system

$$\begin{aligned} k_1 + k_2 + k_3 &= 1 \\ 2ik_1 - 2ik_2 + k_3 &= 0 \\ -4k_1 - 4k_2 + k_3 &= 1 \end{aligned}$$

with solution $(k_1, k_2, k_3) = (i/4, -i/4, 1)$. By noting that the value of $i^n - (-i)^n$ depends only on the residue of n modulo 4, you can write the general form without complex

numbers:

$$x_n = \begin{cases} 1 & n \text{ even} \\ -2^{n-1}+1 & n \equiv 1 \bmod 4 \\ 2^{n-1}+1 & n \equiv 3 \bmod 4 \end{cases}$$

(Alternatively, you could partially factor $R = (L^2+4)(L-1)$ and note directly that $\ker(L^2+4)$ is the set of sequences of the form $(a, b, -4a, -4b, 16a, 16b, -64a, -64b, \ldots)$.)

The only wrinkle in our general method comes when the characteristic equation has repeated roots, as (for example) in the recursion $x_{n+2} = 4x_{n+1} - 4x_n$, whose solutions are in the kernel of $R = L^2 - 4L + 4 = (L-2)^2$. In this case, $\ker R$ is the set of *generalized* eigenvectors of L with order 2, a subspace of $\mathbb{C}^{\mathbb{N}}$ with dimension at most 2. (Remember from page 148 that a generalized eigenspace of order n can have dimension at most n times the dimension of the ordinary maximal eigenspace with the same eigenvalue).

The dimension of $\ker(L-\lambda)^2$ is indeed 2 for any constant λ; in fact, we can find a basis for it. For $\lambda \neq 0$, we could choose the eigenvector $(1, \lambda, \lambda^2, \lambda^3, \ldots)$ and the GEV of order 2 $(0, \lambda, 2\lambda^2, 3\lambda^3, \ldots)$ to be a basis for $\ker(L-\lambda)^2$. You can check that $(L-\lambda)(0, \lambda, 2\lambda^2, 3\lambda^3, \ldots) = (\lambda, \lambda^2, \lambda^3, \ldots)$ and so $(L-\lambda)^2(0, \lambda, 2\lambda^2, 3\lambda^3, \ldots) = \mathbf{0}$. So solutions to $x_{n+2} = 4x_{n-1} - 4x_n$ have the general form $x_n = (k_1 n + k_2)2^n$. (If $\lambda = 0$, then $\ker L^2$ is the set of sequences with all zeros except possibly in the first two entries, but characteristic equations of linear recurrences never have 0 as a root.)

In general, $\ker(L-\lambda)^n$ for $\lambda \neq 0$ is an n-dimensional space whose elements have the general form

$$(p(0), p(1)\lambda, p(2)\lambda^2, p(3)\lambda^3, \ldots)$$

where p is any polynomial of degree at most $n-1$. You can prove this inductively: applying $L - \lambda$ once to $(p(0), p(1)\lambda, p(2)\lambda^2, p(3)\lambda^3, \ldots)$ gives

$$((p(1)-p(0))\lambda, (p(2)-p(1))\lambda^2, (p(3)-p(2))\lambda^3, \ldots)$$

and this sequence is in $\ker(L-\lambda)^{n-1}$ because $p(x+1) - p(x)$ is a polynomial whose degree is one below the degree of p itself.

6.9.3 Linear recurrences with a term dependent on the index

We now know how to solve sequence in which each term after the initial conditions is a linear combination of the previous terms. Now let's consider a generalization: what if each term is a linear combination of the previous terms, plus a function of the index? As an example, let's define the *Bifonacci sequence*[5] $(b_0, b_1, b_2, \ldots)$ with the initial conditions $b_0 = 0, b_1 = 1$, and then the recurrence $b_{n+2} = b_n + b_{n+1} + 2n$. The Bifonacci sequence starts $b_0 = 0, b_1 = 1, b_2 = 0 + 1 + (2 \times 0) = 1, b_3 = 1 + 1 + (2 \times 1) = 4, b_4 = 1 + 4 + (2 \times 2) = 9, b_5 = 4 + 9 + (2 \times 3) = 19, \ldots$.

The $2n$ term in the recursion prevents us from using our method for linear recurrences unmodified, but a few observations can turn much of the problem into a linear recurrence problem in disguise. First, let's call a sequence $(b_0, b_1, b_2, \ldots)$ *Bifonacci-like* if it satisfies the recurrence $b_{n+2} = b_n + b_{n+1} + 2n$ but doesn't necessarily have 0 and 1 as initial values. Note two facts about Bifonacci-like sequences.

[5] Not a standard term!

1. Suppose $b_0, b_1, \ldots$ and $b'_0, b'_1, \ldots$ are both Bifonacci-like, and define $a_n = b'_n - b_n$. Then subtracting the recurrence equation $b_{n+2} = b_n + b_{n+1} + 2n$ from the recurrence equation $b'_{n+2} = b'_n + b'_{n+1} + 2n$ gives the equation $a'_{n+2} = a'_n + a'_{n+1}$, which is the Fibonacci recurrence. That is, the difference between any two Bifonacci-like sequences is a Fibonacci-like sequence—that is, an element of $\ker(L^2 - L - I)$. This means that the Bifonacci-like sequences must all be contained in one coset of $\ker(L^2 - L - I)$.

2. In fact, the Bifonacci-like sequences are an *entire* coset of $\ker(L^2 - L - I)$, because we can reformulate the Bifonacci recurrence directly as $(L^2 - L - I)(b_0, b_1, b_2, \ldots) = (0, 2, 4, 6, \ldots, 2n, \ldots)$. That is, the Bifonacci-like sequences are the preimage of the single vector $(0, 2, 4, 6, \ldots, 2n, \ldots) \in \mathbb{C}^{\mathbb{N}}$ under $L^2 - L - I$, and in section 2.6.2, we showed that preimages of single vectors are cosets of the kernel.

So the Bifonacci-like sequences are a coset of the Fibonacci-like sequences; that is, we can write any Bifonacci-like sequence as $b_n = c_n + k_1\phi_-^n + k_2\phi_+^n$, where $k_1\phi_-^n + k_2\phi_+^n$ is a general solution for Fibonacci-like sequences, and $(c_0, c_1, c_2, \ldots)$ is an arbitrary Bifonacci-like sequence that we can use as a base point for the coset. For most linear recurrences with an additional index-dependent term, finding a base point is difficult if not impossible. But the Bifonacci recurrence equation is one of the few with a nice formula for the base point: you can check that $(-2, -4, -6, \ldots, -2n-2, \ldots)$ is Bifonacci-like, so a general formula for Bifonacci-like sequences is $b_n = k_1\phi_-^n + k_2\phi_+^n - 2n - 2$.

If you want to find the values of k_1, k_2 that give the Bifonacci sequence with initial conditions $b_0 = 0, b_1 = 1$, you can substitute $n = b_n = 0$ and $n = b_n = 1$ into the equation to get the linear system $0 = k_1 + k_2 - 2$ and $1 = k_1\phi_- + k_2\phi_+ - 4$, with solution $k_1 = 1 - \sqrt{5}$ and $k_2 = 1 + \sqrt{5}$. So the Bifonacci sequence has the closed form

$$b_n = \frac{(1-\sqrt{5})^{n+1}}{2^n} + \frac{(1+\sqrt{5})^{n+1}}{2^n} - 2n - 2.$$

We can generalize this approach to any linear recurrence of the form $b_{n+k} = c_0 b_n + \cdots + c_{k-1} b_{n+k-1} + f(n)$, where f is some arbitrary function. First, solve the corresponding linear recurrence without an index-dependent term $a_{n+k} = c_0 a_n + \cdots + c_{k-1} a_{n+k-1}$. Second, find at least one solution $(b_0, b_1, \ldots)$ to the original recurrence with the $f(n)$ term to use a base point. Not all functions f lead to a base point $(b_0, b_1, \ldots)$ with a nice formula, but if you can find only one such sequence, you know that all the other solutions can be obtained by adding $(b_0, b_1, \ldots)$ to a solution to $a_{n+k} = c_0 a_n + \cdots + c_{k-1} a_{n+k-1}$.

6.9.4 Linear homogeneous ODEs with constant coefficients

First, some notation: if I is an interval on the real line, let's denote the set of continuous complex-valued functions on I by $\mathcal{C}^0(I)$. This is a vector space over $\mathbb{C}$ if we define addition of two functions f and g as the function $x \mapsto f(x) + g(x)$ (which must be continuous if f and g are both continuous), multiplication of f by a scalar $k \in \mathbb{C}$ by the function $x \mapsto kf(x)$ (which is also continuous if k is continuous), and the zero vector as the function that takes every input to 0.

The set of complex-value functions on I whose nth derivative exists and is continuous is denoted $\mathcal{C}^n(I)$. The set of functions with derivatives of every positive order (typically called "smooth" functions), finally, is $\mathcal{C}^\infty(I)$.

Denote by D the differential map that takes every function f to its derivative f'. This is a linear map from $\mathcal{C}^n(I)$ to $\mathcal{C}^{n-1}(I)$ (linear because $\frac{d}{dx}(af(x) + bg(x)) = af'(x) + bg'(x)$), and it's a linear operator on $\mathcal{C}^\infty(I)$, the space of functions with an infinite number of complex derivatives.

The idea behind our solution method for linear homogeneous ODEs with constant coefficients is virtually identical to our idea for linear recurrences, with D taking the role of L: recast a differential equation

$$\left(\frac{d^n}{dx^n} + c_{n-1}\frac{d^{n-1}}{dx^{n-1}} + \cdots + c_1\frac{d}{dx} + c_0\right) f(x) = 0$$

as a statement about the kernel of a polynomial of D:

$$f \in \ker(D^n + c_{n-1}D^{n-1} + \cdots + c_1 D + c_0).$$

(Differential equations textbooks call this the "characteristic polynomial" of the equation.) This rewrite presupposes that f is a smooth function (because the only version of D that we can technically compose with itself is the operator on the space of smooth functions $\mathcal{C}^\infty(I)$), but it turns out that linear homogeneous ODEs with constant coefficients don't have non-smooth solutions.[6]

We can factor this operator in the form $(D-\lambda_1)^{n_1}\cdots(D-\lambda_k)^{n_k}$, where the (possibly complex) eigenvalues $\lambda_1, \ldots, \lambda_k$ are all distinct and the exponents $n_1 + \cdots + n_k = n$. This shows that the space of functions f in the kernel of this operator is the direct sum of generalized eigenspaces $\ker(D-\lambda_1)^{n_1} \oplus \cdots \oplus \ker(D-\lambda_k)^{n_k}$.

- The eigenfunctions[7] of D with value λ (i.e. the elements of $\ker(D-\lambda)$ are solutions $y = f(x)$ to $\frac{dy}{dx} = \lambda y$, which you can solve with standard calculus methods to get $f(x) = ke^{\lambda x}$. (For $\lambda = 0$, these functions are just the constant functions $f(x) = k$).

  This is a one-dimensional eigenspace; the constant k is determined by a single initial condition, such as the value of $f(x_0)$ for some special point x_0. You may wonder if there might be more exotic solutions that we've missed, but a general theorem in differential equations (which we won't cover here) that says that a first-degree differential equation with a single specified initial value can have only one solution.

- The generalized eigenfunctions of D with value λ and order n have the general form $f(x) = p(x)e^{\lambda x}$, where p is an arbitrary polynomial of degree at most $n-1$. This is an n-dimensional subspace of $\mathcal{C}^\infty(I)$. The nonexistence of any additional generalized eigenfunctions follows either from more general uniqueness

[6] The solution to $\left(\frac{d^n}{dx^n} + c_{n-1}\frac{d^{n-1}}{dx^{n-1}} + \cdots + c_1\frac{d}{dx} + c_0\right) f(x) = 0$ must have n derivatives defined, and if f has n derivatives, then the first $n-1$ derivatives must be continuous (because any differentiable function must also be continuous). We can also rearrange the differential equation as $f^{(n)}(x) = -c_{n-1}f^{(n-1)}(x) - \cdots - c_1 f'(x) - c_0 f(x)$ where $f^{(k)}(x)$ means $\frac{d^k}{dx^k}f(x)$; since the right-hand side of this equation is a sum of continuous and differentiable functions, the left-hand side $f^{(n)}(x)$ must be continuous and differentiable as well. Differentiating both sides of this equation gives $f^{(n+1)}(x) = -c_{n-1}\frac{d^{n-1}}{dx^{n-1}}f^{(n)}(x) - \cdots - c_1 f'(x) - c_0 f(x)$, which shows that $f^{(n+1)}$ is a sum of differentiable functions, so it must also exist and be differentiable. Differentiating again gives an expression for $f^{(n+2)}$ as a sum of multiples of the continuous functions $f'', \ldots, f^{(n+1)}$, so $f^{(n+2)}$ also exists and is differentiable; and so on.

[7] This is the word typically used instead of "eigenvectors" when the context is operators on function spaces such as $\mathcal{C}^\infty(I)$.

theorems for differential equation solutions, or from the bound on generalized eigenspace dimensions $\dim\ker(T-\lambda)^n \leq n\dim\ker(T-\lambda)$ from page 148.

So the general form of an element of the kernel $(D-\lambda_1)^{n_1}\cdots(D-\lambda_k)^{n_k}$ is

$$f(x) = p_1(x)e^{\lambda_1 x} + \cdots + p_k(x)e^{\lambda_k x}$$

where p_i is a polynomial of degree at most $n_i - 1$. Finding a specific function that satisfies initial conditions requires using the initial conditions to solve for the n total coefficients in the polynomials $p_1, \ldots, p_k$

Let's look at an example from physics: the damped harmonic oscillator. This is an oscillating object attached to an idealized spring that can move in one dimension, with an equilibrium position at 0. The object feels two forces: first, a restoring force from the spring pulling it back to equilibrium (and proportional to its distance from equilibrium); and second, a friction force proportional to, and in the opposite direction from, its velocity. Its acceleration is the sum of these forces divided by the object's mass; that is, if $y(t)$ is the object's position at time t, then $y''(t) = -\alpha y'(t) - \beta y(t)$, where $\alpha > 0$ gives the relative strength of the friction force compared to the object's mass, and $\beta > 0$ gives the strength of the restoring force. We can rearrange this equation as

$$y''(t) + \alpha y'(t) + \beta y(t) = 0$$

with general solution $y \in \ker(D^2 + \alpha D + \beta) = \ker(D-\lambda_-)(D-\lambda_+)$, where $\lambda_- = -\dfrac{\alpha - \sqrt{\alpha^2 - 4\beta}}{2}$ and $\lambda_+ = -\dfrac{\alpha - \sqrt{\alpha^2 - 4\beta}}{2}$ are the roots of the characteristic equation $x^2 + \alpha x + \beta = 0$.

The damped oscillator has qualitatively different behavior depending on the sign of the discriminant $\alpha^2 - 4\beta$ of the characteristic equation. There are three cases:

1. Restoring force dominates: $\alpha^2 - 4\beta < 0$. In this case, λ_+ and λ_- are complex conjugates with real part $-\alpha/2$ and complex part $\pm\omega$, where $\omega \coloneqq \sqrt{\alpha^2 - 4\beta}/2$. The general solution is $y(t) = k_1 e^{(-\alpha/2+\omega i)t} + k_2 e^{(-\alpha/2-\omega i)t} = e^{-\alpha t/2}(k_1 e^{\omega i t} + k_2 e^{-\omega i t})$. This describes an object that oscillates with angular frequency ω, with the amplitude of the oscillations decaying exponentially as $e^{-\alpha t/2}$. This is called *underdamping*.

2. Balanced friction and restoring force: $\alpha^2 - 4\beta = 0$. In this case, $\lambda_- = \lambda_+ = -\frac{\alpha}{2}$, and $\ker(D-\lambda_-)(D-\lambda_+) = \ker(D+\frac{\alpha}{2})^2$ contains generalized eigenvectors. The general solution is $y(t) = (k_1 t + k_2)e^{-\alpha t/2}$. So the object will not oscillate: if it's given an initial push in one direction (making a large value of k_1), it may overshoot equilibrium, but it will eventually decay to equilibrium roughly exponentially at a rate given by $e^{-\alpha t/2}$. This is called *critical damping*.

3. Friction force dominates: $\alpha^2 - 4\beta > 0$. The general solution is again $y(t) = k_1 e^{\lambda_- t} + k_2 e^{\lambda_+ t}$, but this time, λ_- and λ_+ are both real, with $-\frac{\alpha}{2} < \lambda_+ < 0$. The object will return without oscillating to the equilibrium position, but because friction is so strong, its return is slower than in critical damping, being given roughly by $e^{\lambda_+ t}$. This scenario is called *overdamping*.

6.9.5 Linear inhomogeneous ODEs

A linear inhomogeneous ODE with constant coefficients is a differential equation of the form

$$\left(\frac{d^n}{dx^n} + c_{n-1}\frac{d^{n-1}}{dx^{n-1}} + \cdots + c_1\frac{d}{dx} + c_0\right) f(x) = a(x)$$

where $a(x)$ is a given function of x—analogous to linear recurrences that include a term $f(n)$. These are generally much harder to solve than linear homogeneous ODEs, but one crucial observation (again, analogous to an observation that we made in developing our theory of linear recurrences) lets us solve many of them. Let's write $\Delta = D^n + c_{n-1}D^{n-1} + \cdots + c_1 D + c_0$. Then the differential equation may be written more simply $\Delta f = a$: that is, the set of solutions f is $D^{-1}(\{a\})$, the preimage of a single vector—that is, it's a *coset* of $\ker \Delta$. So we can find a single solution f_0 to $\Delta f = a$, then the general solution is just $f \in f_0 + \ker \Delta$: that is, f_0 plus any solution to the associated, mechanically solvable homogeneous equation $\Delta f = 0$. One common method for finding f_0 involves making an *Ansatz*—a guess at a form for a possible solution, with unknown parameters, that satisfies the differential equation but not necessarily the initial conditions—and then using the differential equation to solve for these parameters.

This is not a book on differential equations, so we won't go into too much detail; let's just look at one other familiar physics example: the *damped and driven* harmonic oscillator, with friction and restoration forces as well as an external driving force $F\sin(\omega t)$. The equation of motion for this oscillator is

$$y''(t) + \alpha y'(t) + y(t) = F\sin(\omega t).$$

To solve this, you might guess that the object should follow the driving force (possibly with some phase lag) and use the *Ansatz* $y_0(t) = A\sin(\omega - \phi)$.

The resulting details are boring, but in essence: substituting this equation into $y_0''(t) + \alpha y_0'(t) + y_0(t) = F\sin(\omega t)$, and using the fact that derivatives and sums of sinusoids are sinusoids with the same frequency,[8] gives a system to find (complicated) expressions for A and ϕ in terms of F, ω, α, β. Moreover, the *general* solution to the differential equation is y_0 plus a solution to $y''(t) + \alpha y'(t) + y(t) = 0$, the equation for the damped and undriven oscillator, and we've already seen that every solution for the undriven oscillator decays to zero as t increases. So in the $t \to \infty$ limit, every possible trajectory of the damped and driven oscillator approaches the special solution y_0, regardless of initial conditions.

[8] Remember that $A\sin(\omega t + \phi) = \Re(Ae^{i\omega t + i\phi})$ where $\Re$ denotes the real part, so $A_1\sin(\omega t + \phi_1) + A_2\sin(\omega_t + \phi(2) = \Re((A_1 e^{i\phi_1} + A_2 e^{i\phi_2})e^{i\omega t})$; in physics or engineering classes, you may have learned a graphical method called *phasor algebra* that turns addition of sinusoids with the same frequency, but different amplitudes A and phases ϕ, into addition of the vectors $(A\cos\phi, A\sin\phi)$.

Chapter 7

Matrix and operator determinants

7.1 Motivating intuition: determinant as volume

Suppose $\mathbb{F}$ is a field and V is a finite-dimensional vector space over $\mathbb{F}$. The *determinant* of a linear operator $T \in \mathrm{End}(V)$, or a matrix representation of T, is a single element of $\mathbb{F}$ that indicates (to put it intuitively) the factor by which T changes the volume of subsets of V.

An operator's determinant tells you several important facts about the operator: most importantly, whether it is bijective. Determinants will also let us build a theory of *eigendecompositions* of linear operators into simpler operators on invariant subspaces. We'll cover this in the next chapter.

Determinants most clearly have a geometric meaning for operators on $\mathbb{R}^n$, which can be interpreted as transformations of Euclidean space, such as dilations, compressions, and rotations, that preserve the location of the origin and map straight lines to straight lines. Since T is linear, we can divide any reasonably shaped subset of $\mathbb{R}^n$ into a set of infinitesimally small hypercubes, each of which gets scaled by the same factor.

For instance, the image of the unit hypercube $\{c_1\mathbf{e}_1 + \cdots + c_n\mathbf{e}_n : 0 \leq c_1, \ldots, c_n \leq 1\} \subset \mathbb{R}^n$ under T is generally a skewed cube or "parallelepiped" $\{c_1T\mathbf{e}_1 + \cdots + c_nT\mathbf{e}_n : 0 \leq c_1, \ldots, c_n \leq 1\}$. If we scale up the unit hypercube by a factor λ and then translate it by a vector $\mathbf{v}$, then the image of the resulting hypercube $\{\mathbf{v} + c_1\lambda\mathbf{e}_1 + \cdots + c_n\lambda\mathbf{e}_n : 0 \leq c_1, \ldots, c_n \leq 1\}$ is a correspondingly translated and resized version of the image of the unit hypercube. So there's some factor k such that for all subsets $S \subset \mathbb{R}^n$ with a finite and definable[1] volume v, the set $T(S)$ has volume kv. (If T has nonzero kernel, then it flattens any input into a space with fewer than n dimensions and volume zero, so $k = 0$.)

The determinant of T is simply this number k. At first, we'll define and prove several properties of a determinant for square *matrices*, and then define the determinant of a linear operator $T \in \mathrm{Hom}(\mathbb{F}^n)$ to be the determinant of the matrix representation of T relative to the standard basis. There's nothing special about the standard basis: all similar matrices have the same determinant. But the proof that the determinant of a linear transformation is independent of its matrix representation will have to wait until the next chapter.

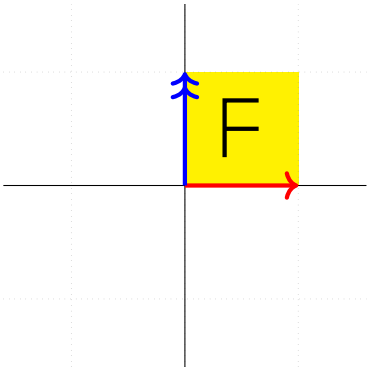
For now, let's work through the geometric intuition of determinants in a simple, easily visualized setting: linear transformations of the Cartesian plane $T : \mathbb{R}^2 \to \mathbb{R}^2$.

[1] One standard exercise in real analysis classes is constructing a subset of $\mathbb{R}^n$ with a structure so complex that it doesn't have a definable volume, either zero or nonzero.

The four points $\{(0,0),(0,1),(1,0),(1,1)\}$ form a square with area 1. If T is bijective, then $\{T(0,0),T(0,1),T(1,0),T(1,1)\}$ form the four vertices of a parallelogram (and, of course, $T(0,0)=(0,0)$). The area of this parallelogram is the factor by which T scales the volume of every input set. Furthermore, every permutation can either *preserve* orientation—that is, the output is a stretched, skewed, or rotated version of the input—or *reverse* orientation: that is, the output is a distortion of the input's mirror image.

Define the determinant of T, denoted $\det T$, to be the area of this parallelogram if T preserves orientation, or the negative area of the parallelogram if T reverses orientation. First, we'll look at several examples of linear transformations on $\mathbb{R}^2$ and figure out what their determinants have to be from geometric considerations; then we'll work out a general formula for the determinant.

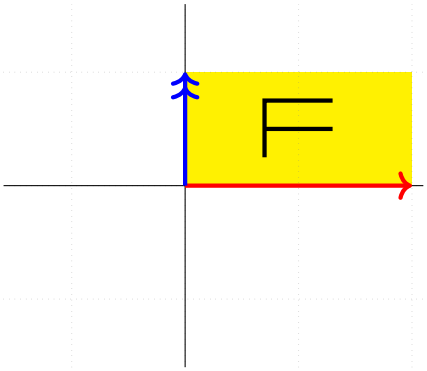
Consider this drawing of the Cartesian plane:



The unit square is highlighted, and the two standard basis vectors are also shown: $\mathbf{e}_1=(1,0)$ in red with a single arrowhead and $\mathbf{e}_2=(0,1)$ in blue with a double arrowhead. The capital F in the unit square shows orientation more clearly.

Let's consider what this picture looks like if we put it through a few different linear transformations.

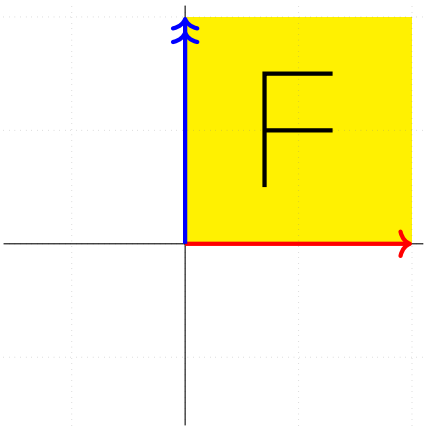
1. $T(x,y)=(2x,y)$, matrix representation with respect to the standard basis $\begin{bmatrix}2&0\\0&1\end{bmatrix}$:



   Here, T stretches the x-axis by a factor of 2 while leaving the y-axis unchanged, increasing all areas by a factor of 2. so $\det T=2$

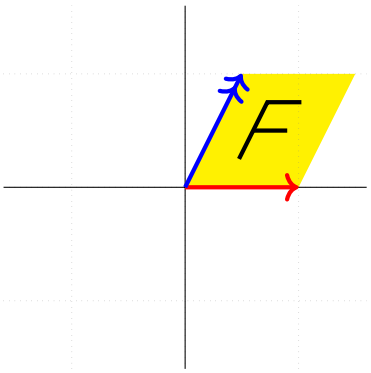
2. $T(x,y)=(2x,2y)$, matrix representation with respect to the standard basis[2] $\begin{bmatrix}2&0\\0&2\end{bmatrix}$:

[2] We'll stop writing the phrase "with respect to the standard basis" explicitly for the rest of this chapter, but remember that it's still there implicitly.



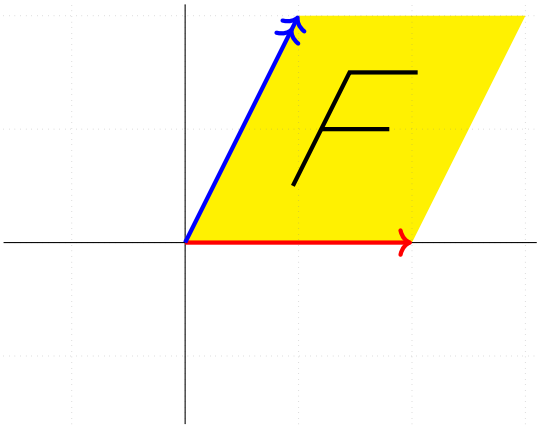
T is a uniform scaling in all directions by 2, and $\det T = 4$.

3. $T(x, y) = (x + \frac{1}{2}y, y)$, matrix representation[3] $\begin{bmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{bmatrix}$:



This is a *shear* mapping, in which points are moved parallel to one axis by a distance proportional to their coordinates on another axis. The determinant of this transformation is 1: shear transformations preserve area.

4. $T(x, y) = (2x + y, 2y)$, matrix representation $\begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$:

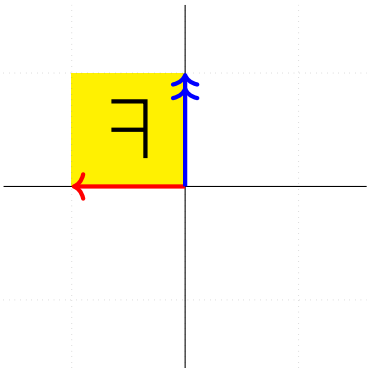


This is a uniform scaling by a factor of 2 composed with the shear transformation $(x, y) \mapsto (x + \frac{1}{2}y, y)$ from the previous example. (The composition can be done in either order: every transformation commutes with uniform scalings, just as every

[3] Again, "with respect to the standard basis" is implied for the rest of this section.

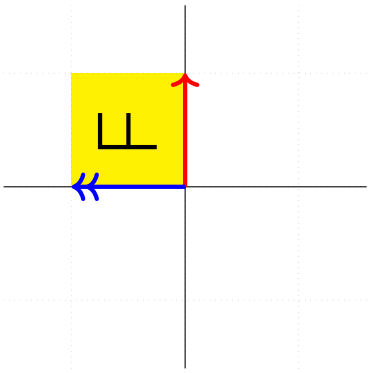
matrix commutes with a scalar multiple of the identity matrix). The determinant is 4.

5. $T(x, y) = (-x, y)$, matrix representation $\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$:



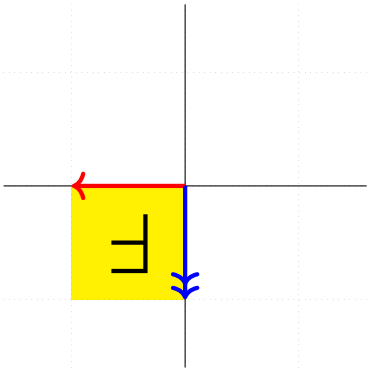
This is a reflection over the y-axis. The image of the unit square is also a square of area 1, but its orientation is reversed: note that the F appears backwards, and that the angle from $T\mathbf{e}_1$ to $T\mathbf{e}_2$ is clockwise, not counterclockwise. The determinant is -1.

6. $T(x, y) = (-y, x)$, matrix representation $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$:



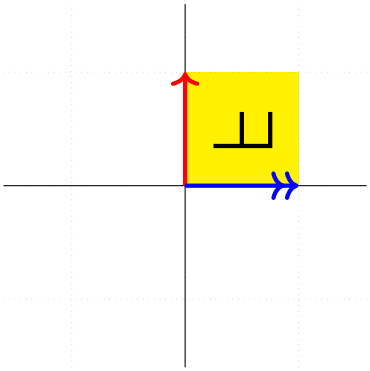
This is a rotation 90 degrees counterclockwise. The image of the unit square occupies the same area as it does under the transformation $(x, y) \mapsto (-x, y)$, but in this case, area is preserved. The determinant is 1.

7. $T(x, y) = (-x, -y)$, matrix representation $\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$:



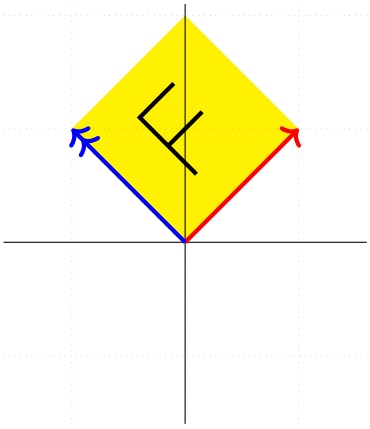
This can be interpreted either as a rotation by 180 degrees or as two reflections about perpendicular axes. The determinant is 1.

8. $T(x, y) = (y, x)$, matrix representation $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$:



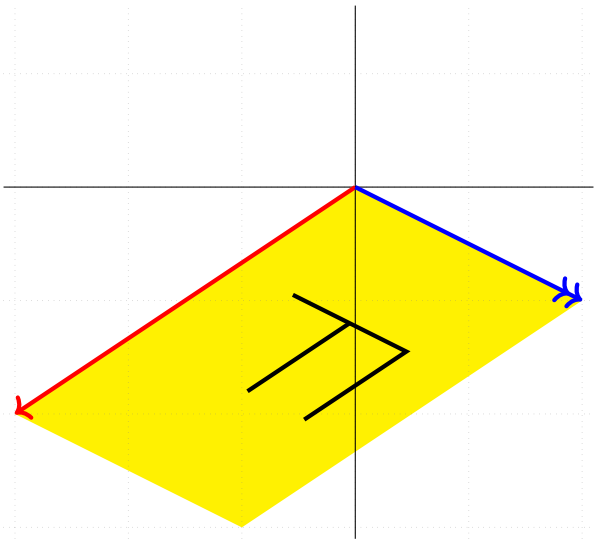
Swapping the two axes is equivalent to reflection over the line $y = x$, so this transformation has determinant -1.

9. $T(x, y) = (x - y, x + y)$, matrix representation $\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$:



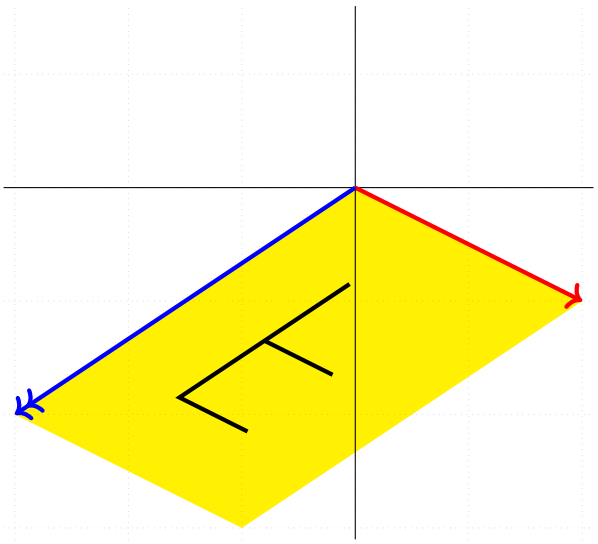
This is equivalent to rotation clockwise by 45 degrees followed by a uniform scaling by a factor of $\sqrt{2}$ in all directions, so the determinant is 2.

10. $T(x, y) = (-3x + 2y, -x)$, matrix form $\begin{bmatrix} -3 & 2 \\ -2 & -1 \end{bmatrix}$:



This transformation preserves orientation. The area of the shaded parallelogram is 7 (you can see this for yourself by considering the rectangle withe corners $(-3,-3),(-3,0),(2,0),(2,-3)$, which contains the parallelogram and four unshaded triangles, and then subtracting the areas of the triangles).

11. $T(x,y) = (2x-3y, -x-2y)$, matrix form $\begin{bmatrix} 2 & -3 \\ -1 & -2 \end{bmatrix}$:

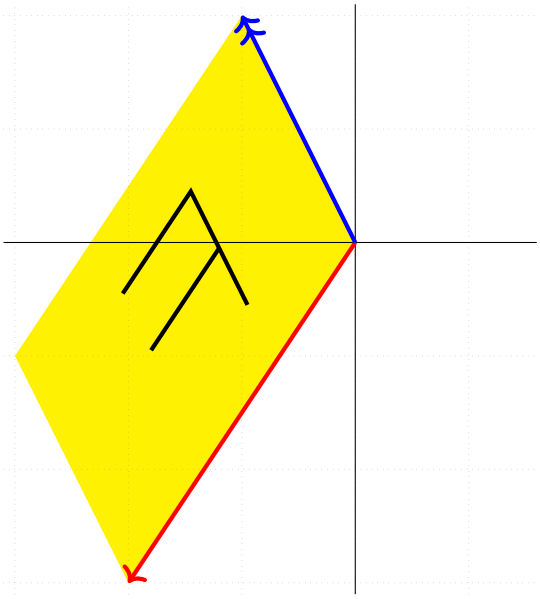


The matrix form of this transformation comes from reversing the columns of the matrix in example 10, or, equivalently, right-multiplying by $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$:

$$\begin{bmatrix} 2 & -3 \\ -1 & -2 \end{bmatrix} = \begin{bmatrix} -3 & 2 \\ -2 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

The matrix $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ represents reflection over the line $y = x$, which swaps the vectors $\mathbf{e}_1$ and $\mathbf{e}_2$, so the image of the unit square is the same as the image of the unit square in example 10, except that the two basis vector images are reversed. The orientation here is reversed, as one expects from the the composition of a reflection and an orientation-preserving transformation. The determinant is -7.

12. $T(x, y) = (-2x - y, -3x + 2y)$, matrix form $\begin{bmatrix} -2 & -1 \\ -3 & 2 \end{bmatrix}$:

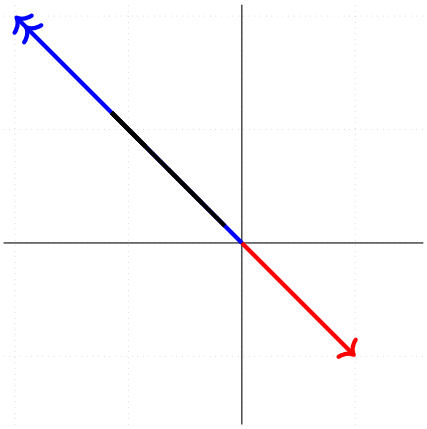


The matrix form of this example is the same as the matrix form of example 10, but with the rows swapped—or, equivalently, with the whole matrix left-multiplied by $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$:

$$\begin{bmatrix} -2 & -1 \\ 3 & -2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 3 & -2 \\ -2 & 1 \end{bmatrix}.$$

This corresponds to applying the transformation in example 10 and *then* reflecting the resulting figure over the line $y = x$. The determinant of this figure is -7.

13. $T(x, y) = (x - 2y, -x + 2y)$, matrix form $\begin{bmatrix} 1 & -2 \\ -1 & 2 \end{bmatrix}$:



This transformation collapses all of $\mathbb{R}^2$ onto a one-dimensional subspace: the line $y = -x$. Lines have no area, so the determinant is zero.

In all these cases, the determinant of the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is $ad - bc$. You can prove geometrically that the area of a parallellogram with corners $(0, 0), (a, c), (b, d), (a+b, c+$

$d)$ is, in fact, $|ad - bc|$, and that the image (b, d) of $\mathbf{e}_2$ is located less than 180 degrees counterclockwise from the image (a, c) of $\mathbf{e}_1$ (thereby preserving orientation) if and only if $ad - bc > 0$. (Hint for proving this yourself: inscribe the parallelogram in a rectangle with sides parallel to the coordinate axes, with the area inside the rectangle but outsie the parallelogram divided into four triangles, You may need to consider a few different cases depending on the signs of a, b, c, d.)

The geometric interpretation of the determinant is harder to visualize in higher dimensions and typically nonsensical for fields other than $\mathbb{R}$ or $\mathbb{Q}$, so to give a general definition of the determinant, we'll want a purely algebraic formula with properties that correspond to the geometric ones that we just explored. It turns out that there's exactly one function from $\mathrm{Mat}_{n\times n}(\mathbb{F})$ to $\mathbb{F}$ can satisfy all these properties:

1. The determinant should be multiplicative: $\det(AB) = (\det A)(\det B)$, because applying the map B first and then A should affect volumes the same way as applying AB all at once. A corollary: either every matrix has determinant 0 (and property 3 in this list rules out this possibility) or the identity matrix I has determinant 1, because $\det A = \det(AI) = \det A \det I$.

2. The determinant of any matrix with an entire row of zeros should be 0, because such matrices flatten their inputs into a smaller-dimensional (and thus zero-volume) output space that omits all the dimensions corresponding to rows of zero.

3. The matrices that represent elementary row transformations, all of which have clear geometric interpretations, should have the following determinants:

   (a) *Row swap matrices* should have determinant -1, as the linear transformations that they represent simply swap two standard basis vectors—that is, they reflect across the line (or plane, or hyperplane) that bisects the angle between them. The matrix representation for $\mathbf{r}_1 \leftrightarrow \mathbf{r}_2$ in $\mathbb{R}^3$, for instance, is $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, which simply interchanges $\mathbf{e}_1$ and $\mathbf{e}_2$—that is, it's a reflection across the plane $x = y$.

   (b) *Row scaling matrices* should have determinant equal to the scaling factor. The scaling $\mathbf{r}_1 \mapsto -2\mathbf{r}_1$ in $\mathbb{R}^3$, for instance, has matrix representation $\begin{bmatrix} -2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, which simply reflects any input set across the plane $x = 0$, thereby reversing its orientation, and then stretches it by a factor of 2. (One corollary is that the identity matrix, which is just a row scaling matrix with $\lambda = 1$, should have determinant 1.)

   (c) *Shear matrices* should have determinant 1, because they map the unit square (or cube, or hypercube) to a parallelogram (in the two-dimensional case) or, for $n \geq 3$ dimensions, to a prism or hyperprism with $n - 2$ of its dimensions all perpendicular to a parallelogram base with area 1. The shear $\mathbf{r}_1 \mapsto \mathbf{r}_1 + 2\mathbf{r}_2$ in $\mathbb{R}^3$, for instance, is represented by $\begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. This transformation preserves $\mathbf{e}_1$ and sends $\mathbf{e}_2$ to $(2, 1, 0)$, so the images of $\mathbf{e}_1$ and $\mathbf{e}_2$

form a parallelogram with vertices $(0,0,0)$, $(1,0,0)$, $(2,1,0)$, $(3,1,0)$ and area 1, and the image of the unit cube is a prism with this parallelogram as a base and height 1 along the z-axis.

It's easy to prove that if the determinant exists, it has to be unique:

Proposition. *There is at most one function from* $\mathrm{Mat}_{n\times n}(\mathbb{F})$ *to* $\mathbb{F}$ *that satisfies all these requirements.*

Proof. Every square matrix can be factored into its RREF (which, for square matrices, either equals the identity matrix or has a row of zeros) times zero or more elementary row operation matrices. Our requirements specify determinants for the identity matrix, matrices with zero rows, and elementary row operation matrices, and they also specify that the determinant of a matrix product is the product of the determinants of the individual matrix factors.

□

We haven't proved yet that a function with all these properties actually exists. But it turns out that it does exist, and we can give an explicit formula for it: it's a particular sum of products of matrix entries that does, in fact, reduce to $\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$ for the two-dimensional case. We'll two formulas for the determinant. The first uses an already familiar concept: Gauss–Jordan elimination; and this definition will be sufficient to prove many of the

The other way to define the determinant uses a set of functions called *permutations*: bijective functions from the integers $\{1, \ldots, n\}$ to themselves. We'll get to this later.

7.2 The determinant with Gauss–Jordan reduction

Let's start with a definition.

Definition. *Let* M *be a square matrix. If* $\operatorname{rref} M$ *has a row of zeros, then define the* ***determinant*** *of* M*, denoted* $\det M$*, to be zero. Otherwise, if* $\operatorname{rref} M$ *is the identity, take an arbitrary sequence of elementary row operations that reduces* M *to the identity, and define the determinant as the product of the following factors:*

1. *A factor of* λ^{-1} *for every scale operation* $\mathbf{r}_i \mapsto \lambda \mathbf{r}_i$*.*
2. *A factor of* -1 *for every swap operation* $\mathbf{r}_i \leftrightarrow \mathbf{r}_j$*.*

Shear operations do not contribute to the determinant, and if $M = I$ *(that is, there are no operations necessary to reduce* M *to RREF), then* $\det M = 0$*.*

There is one glaring problem with this definition, which should be appearance of the word *arbitrary*. There are multiple row reduction steps that could reduce any matrix M to the identity—even if we follow Gauss–Jordan elimination, there are points at which Gauss–Jordan requires us to make arbitrary choices (namely, which row to swap in to replace a row with a zero pivot).

Fortunately, this is a solvable issue, which we may solve as follows. Suppose that

It's easy to check that this definition satisfies all of the properties that we listed on page 166:

1. *The determinant is multiplicative.* If $\det A = 0$ or $\det B = 0$ (that is, either A or B does not have full rank, and so has a row of zeros in RREF), then AB also cannot have full rank.

   If A and B both have full rank, then suppose R_A is the matrix representation of a sequence of elementary row operations that reduces A to I, and likewise R_B for B (that is, $R_A = A^{-1}$ and $R_B = B^{-1}$). Then $R_B R_A AB = R_B(R_A A)B = R_B B = I$; that is, applying to AB a sequence of row operations that reduces A, and then a sequence that reduces B, will also reduce AB. Concatenating two sequences of row operations means multiplying the determinants that they yield.

   In either case, $\det AB = \det A \det B$ is true.

2. Any matrix with an entire row of zeros must also have an RREF with a row of zeros and thus have determinant zero.

3. It's easy enough for you to check that matrix representations for scale, shear, and swap operations all have the correct determinant, as it only takes another of the same operation to reduce those matrices to the identity matrix.

7.3 Elements of the theory of permutations

7.3.1 Definition and basic properties

A *permutation* on the set of integers $\{1, \ldots, n\}$ is a bijective function $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$. (The Greek letters sigma σ and tau τ are conventional for permutations, just like x and y conventionally mean real numbers, m and n mean integers, f and g mean general functions, and so on.) "Bijective," remember, means that there are no two distinct integers $i \neq j$ such that $\sigma(i) = \sigma(j)$, and (equivalently) for every $k \in \{1, \ldots, n\}$, there is exactly one i such that $\sigma(i) = k$. (Any function from a finite set to itself has to be both injective and surjective or neither: we can't cover n outputs with n inputs if two of the inputs have the same output.)

As a simple example, the function $\sigma : \{1, 2, 3, 4\} \to \{1, 2, 3, 4\}$ with values $\sigma(1) = 3, \sigma(2) = 2, \sigma(3) = 4, \sigma(4) = 1$ is a permutation: every integer from 1 to 4 occurs once as a value of σ. But $\sigma(1) = 3, \sigma(2) = 3, \sigma(3) = 4, \sigma(4) = 1$ is not a permutation, because 3 appears twice as a value, and 2 isn't a value of σ at all.

The set of all permutations on $\{1, \ldots, n\}$ is denoted S_n. The size of S_n is $n!$, the factorial of n: if you construct a permutation σ by choosing each of the values $\sigma(1), \ldots, \sigma(n)$ in turn, then you have n choices for $\sigma(1)$, then $n - 1$ possible choices for the value of $\sigma(2)$ (that is, all of $\{1, \ldots, n\}$ except for $\sigma(1)$), then $n - 2$ choices for the value of $\sigma(3)$, and so on.

Remember that the composition of two bijective functions is also bijective,[4] and that any bijective function has an inverse. This means that the functional inverse of any permutation is also a permutation. The composition of any permutation with its inverse is the identity permutation (sometimes denoted ι, Greek iota), which maps every input to itself.

[4] If $f : X \to Y$ and $g : Y \to Z$ are injective, then $x_1 \neq x_2 \in X$ implies $f(x_1) \neq f(x_2) \in Y$, which implies $(g \circ f)(x_1) \neq (g \circ f)(x_2) \in Z$, so $g \circ f$ is also injective. Similarly, the image of $g \circ f$ is the image of $g|_{\text{im } f}$ (that is, g with domain restricted to the image of f), so if f and g are both surjective, then so is $g \circ f$).

The existence of permutation inverses proves a "cancellation law" for permutation composition:

Proposition. *If σ_1 and σ_2 are different elements of S_n, then $\tau \circ \sigma_1$ and $\tau \circ \sigma_2$ are also different for any fixed τ, as are $\sigma_1 \circ \tau$ and $\sigma_2 \circ \tau$. (Equivalent phrasing: for any fixed $\tau \in S_n$, the maps $\sigma \mapsto \tau \circ \sigma$ and $\sigma \mapsto \sigma \circ \tau$ are bijections from S_n to itself.)*

Proof. If $\tau \circ \sigma_1 = \tau \circ \sigma_2$, then $\tau^{-1} \circ (\tau \circ \sigma_1)$ must equal $\tau^{-1} \circ (\tau \circ \sigma_2)$. But $\tau^{-1} \circ (\tau \circ \sigma_1) = (\tau^{-1} \circ \tau) \circ \sigma_1 = \sigma_1$ and likewise $\tau^{-1} \circ (\tau \circ \sigma_2) = \sigma_2$, because function composition is associative, so $\sigma_1 = \sigma_2$. That is, the map $\sigma \mapsto \tau \circ \sigma$ is a bijection from S_n to itself.

Similar reasoning also shows that $\sigma \mapsto \sigma \circ \tau$ is a bijective map from S_n to itself.

□

The cancellation law will be vital for proving facts about determinants because the determinant is defined as a particular sum over elements of S_n. Several proofs of properties of the determinant start with a sum of some formula with a variable σ that ranges over elements of S_n, and end with a similar formula in which σ is replaced with $\sigma \circ \tau$ or $\tau \circ \sigma$ for some fixed permutation τ. We can use the fact that if σ takes on every value of S_n in a sum, then so does $\sigma \circ \tau$, so the terms in these two sums correspond one-to-one.

7.3.2 Permutation parity

The "parity" or "sign" of a permutation σ is a somewhat strange formula whose value is either 1 or -1. It's hard to explain at first why parity is useful, so let's just define it.

If $\{i, j\}$ is a two-element subset of $\{1, \ldots, n\}$, we'll say that a permutation $\sigma \in S_n$ *inverts* $\{i, j\}$ if $i < j$ but $\sigma(i) > \sigma(j)$. We'll define a permutation that inverts an even number of two-element subsets to have *even parity* (this includes the identity permutation $\sigma(i) = i$ for all $1 \leq i \leq n$, which inverts zero pairs, because zero is even). A permutation that inverts an odd number of pairs has *odd parity*. The *sign* of a permutation is often denoted $\operatorname{sgn}(\sigma)$ or, confusingly, $(-1)^\sigma$, and has the value (again, perhaps confusingly) 1 if σ has even parity or -1 if σ has odd parity. (We'll just use the $\operatorname{sgn}(\sigma)$ notation in this book.)

The motivation for the names "even" and "odd" comes from a result that makes composition of permutations analogous to addition of even and odd numbers (or multiplication of positive and negative numbers):

Proposition. *The composition of two permutations of the same parity (i.e. both odd or both even) is even, and the composition of two permutations of opposite parity is odd. To put it more succinctly,* $\operatorname{sgn}(\sigma \circ \tau) = \operatorname{sgn}(\sigma) \operatorname{sgn}(\tau)$.

Proof. Take $\sigma, \tau \in S_n$. Denote by P_n the set of two-element subsets of $\{1, \ldots, n\}$. Divide P_n into four disjoint subsets:

- $A \subseteq P_n$ is the set of sets $\{i, j\}$ such that σ inverts $\{i, j\}$ and τ inverts $\{\sigma(i), \sigma(j)\}$.
- $B \subseteq P_n$ is the set of sets $\{i, j\}$ such that σ inverts $\{i, j\}$ and τ doesn't invert $\{\sigma(i), \sigma(j)\}$.
- $C \subseteq P_n$ is the set of sets $\{i, j\}$ such that σ doesn't invert $\{i, j\}$ and τ inverts $\{\sigma(i), \sigma(j)\}$.

- $D \subseteq P_n$ is the set of sets $\{i, j\}$ such that σ doesn't invert $\{i, j\}$ and τ doesn't invert $\{\sigma(i), \sigma(j)\}$.

The number of pairs inverted by σ is $|A| + |B|$, and the number of pairs inverted by τ is $|A| + |C|$. (Remember that σ is a bijection, so for every two-element set $\{k, \ell\}$ there is exactly one set $\{i, j\}$ such that $\{\sigma(i), \sigma(j)\} = \{k, \ell\}$: that is, the sets $\{\{i, j\} \in P_n : \tau \text{ inverts } \{\sigma(i), \sigma(j)\}\}$ and $\{\{k, \ell\} \in P_n : \tau \text{ inverts } \{k, \ell\}\}$ have the same size.)

The number of inversions of $\tau \circ \sigma$, meanwhile, is $|B| + |C|$. The sum $(|A| + |B|) + (|A| + |C|) + (|B| + |C|) = 2|A| + 2|B| + 2|C|$ is even, so $|B| + |C|$ is even (that is, $\tau \circ \sigma$ is even) if $|A| + |B|$ and $|A| + |C|$ are both even or both odd (that is, if σ and τ have the same parity) and odd otherwise.

□

Corollary. *Every transposition has the same sign as its inverse.*

Proof. The composition of any permutation and its inverse is ι, which has zero inversions and is even. (Alternatively: σ inverts $\{i, j\}$ if and only if σ^{-1} inverts $\{\sigma(i), \sigma(j)\}$.)
□

Finally, if τ is a transposition that swaps two values of $\{1, \ldots, n\}$—that is, such that $\tau(i) = i$ for every value except two values k and ℓ, for which $\tau(k) = \ell$ and $\tau(\ell) = k$—then we call τ a *transposition*.

Proposition. *Transpositions are always odd.*

Proof. If a transposition τ swaps i and j, then it inverts $\{i, j\}$ itself, but it also inverts $\{i, k\}$ and $\{k, j\}$ for each of the $|j - i| - 1$ integers k that lie between i and j. It doesn't invert any other integer pair. So τ inverts $2|j - i| - 1$ pairs, so it's odd.

□

Finally, note that if τ is an odd permutation, then $\sigma \mapsto \sigma \circ \tau$ is a bijection on S_n that takes every even permutation to an odd permutation, and vice versa. So if S_n has at least one odd permutation (that is, if $n \geq 2$), then it has the same number of even permutations as odd permutations: $n!/2$ of each, to be precise. (In abstract algebra, the set of even permutations on $\{1, \ldots, n\}$ is sometimes called the *alternating group on n elements* and denoted A_n, and the set of all permutations is the *symmetric group*.)

7.3.3 Decomposition of permutations into cycles

We can decompose any permutation σ into a set of disjoint *cycles*, in which $\sigma(i)$ is the element that immediately follows i in the cycle that contains i. Consider, for example, the element of S_{11} whose values are given by the table below:

n	1	2	3	4	5	6	7	8	9	10	11
$\sigma(n)$	4	5	1	9	7	11	2	6	3	10	8

This permutation σ sets up a cycle $1 \to 4 \to 9 \to 3 \to 1$ with four elements, two cycles $2 \to 5 \to 7 \to 2$ and $6 \to 11 \to 8 \to 6$ with three elements each, and a trivial cycle with one element sending 10 to itself. So σ could be written as a composition of cycles, using special notation in which each cycle is notated within parentheses, as $(1\,4\,9\,3)(2\,5\,7)(6\,11\,8)$, or as $(1\,4\,9\,3)(2\,5\,7)(6\,11)(10)$ explicitly including the trivial cycle.

Every length-k cycle can be split into a composition of $k-1$ transpositions. For example, $(1\,2\,3\,\cdots\,k)$ can be decomposed as $(1\,k)(1\,k-1)\cdots(1\,4)(1\,3)(1\,2)$, where composition goes from right to left: first send 1 to 2 and vice versa, then send the resulting 1 (produced from the original 2) to 3 and vice versa, and so on. So a permutation consisting solely of a length-k cycle is even if k is odd, and odd if k is even.[5]

A permutation with multiple cycles, furthermore, is the composition of its constituent cycles, and each even-length cycle changes the parity of the permutation while each odd-length cycle leaves it the same. So a permutation with an odd number of even-length cycles is odd, while a permutation with an even number of even-length cycles is even. The permutation $\sigma \in S_{11}$ above, for example, is a composition of one odd permutation (the length-4 cycle $(1\,4\,9\,3)$) and three even permutations (the length-3 cycles $(2\,5\,7)$ and $(6\,11\,8)$, and the trivial length-1 cycle (10)), so it is odd.

One final small observation: in a cycle, anything that goes up must come back down. The entries in a cycle cannot grow indefinitely: any cycle of length 2 or more contains some integer i that is smaller than its successor: that is, such that $i < \sigma(i)$. So the only element of S_n that satisfies $\sigma(i) \leq i$ for all i is the identity permutation $\sigma(i) = i$, which is a composition of n length-1 cycles. The same reasoning shows that the identity permutation is the only permutation that satisfies $(\forall i)\sigma(i) \geq i$. This point is the key to establishing a vital fact about the determinants of triangular matrices. We'll get to this in a bit.

7.4 Multilinear, symmetric, and alternating functions

Before continuing with determinants, it's worth presenting a few key concepts in a cleaner, more abstract context without messy matrix notation. The main concept is that of *multilinear* maps that take multiple inputs from V and return one output in W, and are linear maps when every argument but one is fixed to an arbitrary value and the one remaining argument is allowed to vary over elements of V. We'll see what this means more precisely soon.

7.4.1 Partial function application

First, though, we'll need a short detour into more basic function concepts and notation. Suppose X and Y are arbitrary sets, and suppose $f : X^n \to Y$ is a function that takes n inputs in X and returns one output in Y. If we choose fixed values for all the inputs to f except one, then we can get another function whose input is a single element in X and whose output is the value of f generated by inserting the new output into the slot left vacant by the outputs that we're holding fixed. In other words, we're creating a new function f through "partial application": choosing some inputs to f as fixed and letting the remaining input be an argument to a single-input function.

As a more familiar example, consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ with formula $f(x, y) = x^3 + 2xy$. We could use this formula to get a new function of x alone by choosing a fixed value for y (say, $y = 3$) and letting x vary. This new function would be the map $x \mapsto f(x, 3) = x^3 + 6y$, and you'll often see this denoted with the shorthand

[5] As with general function compositions, composition of permutations is generally not commutative: for instance, $(1\,2)(1\,3) = (1\,3\,2)$ but $(1\,3)(1\,2) = (1\,2\,3)$. Disjoint cycles, however, do commute, so the order in which we write a permutation's decomposition into disjoint cycles doesn't matter.

notation $f(\cdot, 3)$, with a dot denoting the missing input to be filled in later. Alternatively, we could get a function of y alone by choosing a fixed value for x (say, $x = -2$), creating the map $y \mapsto -4y - 8$, which we could denote $f(-2, \cdot)$.

7.4.2 Multilinear functions defined

We have a special term for functions in vector spaces in which these partial application maps are always linear:

Definition. *Suppose $f : V^n \to W$ is a function that takes n inputs in a vector space V and outputs a single element of another vector space W. Then f is **multilinear** if all of the partial application maps $f(\cdot, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_n)$, $f(\mathbf{v}_1, \cdot, \mathbf{v}_3, \ldots, \mathbf{v}_n)$, and so on up to $f(\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}, \cdot)$ are ordinary linear maps from V to W, for all values of the fixed inputs $\mathbf{v}_1, \ldots, \mathbf{v}_n$.*

For instance, if $n = 3$ and $\mathbb{F}$ is the base field for V and W, then $f : V^3 \to W$ is multilinear if all three of these axioms hold:

1. $f(\cdot, \mathbf{v}_2, \mathbf{v}_3)$ is linear for all $\mathbf{v}_2, \mathbf{v}_3 \in V$. For a map to be linear, it must satisfy the usual two axioms:
   - Respect for addition: $f(\mathbf{v}_1 + \mathbf{v}_1', \mathbf{v}_2, \mathbf{v}_3) = f(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) + f(\mathbf{v}_1', \mathbf{v}_2, \mathbf{v}_3)$ for all $\mathbf{v}_1, \mathbf{v}_1', \mathbf{v}_2, \mathbf{v}_3 \in \mathbb{F}$.
   - Respect for multiplication: $f(c\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) = cf(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ for all $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \in V$ and $c \in \mathbb{F}$.
2. $f(\mathbf{v}_1, \cdot, \mathbf{v}_3)$ is linear for all $\mathbf{v}_1, \mathbf{v}_3 \in V$. That is, $f(\mathbf{v}_1, \mathbf{v}_2 + \mathbf{v}_2', \mathbf{v}_3) = f(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) + f(\mathbf{v}_1, \mathbf{v}_2', \mathbf{v}_3)$ for all $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_2', \mathbf{v}_3 \in V$, and $f(\mathbf{v}_1, c\mathbf{v}_2, \mathbf{v}_3) = cf(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ for all $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \in V$ and $c \in \mathbb{F}$.
3. $f(\mathbf{v}_1, \mathbf{v}_2, \cdot)$ is linear for all $\mathbf{v}_1, \mathbf{v}_2 \in V$. You should now have a good idea of what this means.

One consequence:

Proposition. *The value of a multilinear function $f : V^n \to W$ is $\mathbf{0}_W$ whenever one of its inputs is $\mathbf{0}_V$.*

Proof. $f(\mathbf{0}_V, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ is the value of the partially applied map $f(\cdot, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ evalulated at $\mathbf{0}_V$. If f is multilinear, then by definition the partially applied map is linear, and any linear map from V to W takes $\mathbf{0}_V$ to $\mathbf{0}_W$. The same reasoning works when $\mathbf{0}_V$ is any other argument to f, not just the first.

□

7.4.3 Difference between multilinear and linear functions

The space V^n can also be made a vector space in its own right, with vector addition and scalar multiplication defined in terms of the corresponding operations on V: addition as $(\mathbf{v}_1, \ldots, \mathbf{v}_n) + (\mathbf{v}_1', \ldots, \mathbf{v}_n') = (\mathbf{v}_1 + \mathbf{v}_1', \ldots, \mathbf{v}_n + \mathbf{v}_n')$ and $c(\mathbf{v}_1, \ldots, \mathbf{v}_n) = (c\mathbf{v}_1, \ldots, \mathbf{v}_n)$. A function $f : V^n \to W$ could thus be a linear function, viewed as having as having one input in V^n that satisfies the usual linearity axioms. But the linearity axioms and

multilinearity axioms have very different consequences; and, in fact, only the zero function $f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \mathbf{0}_W$ can satisfy both.

Let's take $n = 2$ to give an easy-to-notate example. Let V and W are two vector spaces over a field $\mathbb{F}$, let $f : V^2 \to W$ be a function, let c be any element of $\mathbb{F}$, and let $\mathbf{v}_1, \mathbf{v}_2$ be any elements of V. If f is multilinear, then $f(c\mathbf{v}_1, c\mathbf{v}_2) = cf(\mathbf{v}_1, c\mathbf{v}_2)$ (by linearity in the first argument) $= c^2 f(\mathbf{v}_1, \mathbf{v}_2)$ (by linearity in the second argument).

But if f is linear considered as a map with a single input in V^2, then $f(c\mathbf{v}_1, c\mathbf{v}_2) = f(c(\mathbf{v}_1, \mathbf{v}_2)) = cf(\mathbf{v}_1, \mathbf{v}_2)$. So if f is both linear and multilinear, then these two expressions must be equal; that is, $(c^2 - c)f(\mathbf{v}_1, \mathbf{v}_2) = \mathbf{0}_W$. A nonzero scalar times a nonzero vector must be nonzero by the basic vector space axioms, so we can only have $(c^2 - c)f(\mathbf{v}_1, \mathbf{v}_2) = \mathbf{0}_W$ for all $c \in \mathbb{F}, \mathbf{v}_1, \mathbf{v}_2 \in V$ if f sends every input to $\mathbf{0}_W$, or every element of $\mathbb{F}$ equals its own square. (The only field with this property, as it happens, is the field with two elements.)

Likewise, suppose $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$ are arbitrary elements of V. If f is multilinear, then we can eand $f(\mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_3 + \mathbf{v}_4)$ into $f(\mathbf{v}_1, \mathbf{v}_3 + \mathbf{v}_4) + f(\mathbf{v}_2, \mathbf{v}_3 + \mathbf{v}_4)$ by linearity in the first argument, and then further into $f(\mathbf{v}_1, \mathbf{v}_3) + f(\mathbf{v}_1, \mathbf{v}_4) + f(\mathbf{v}_2, \mathbf{v}_3) + f(\mathbf{v}_2, \mathbf{v}_4)$ by linearity in the second argument. If f is also linear, though, then we also have $f(\mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_3 + \mathbf{v}_4) = f(\mathbf{v}_1, \mathbf{v}_3) + f(\mathbf{v}_2, \mathbf{v}_4)$, so $f(\mathbf{v}_1, \mathbf{v}_4) + f(\mathbf{v}_2, \mathbf{v}_3) = \mathbf{0}_W$. We could set $\mathbf{v}_2 = \mathbf{v}_3 = \mathbf{0}_V$ (which means any value of f with $\mathbf{v}_2$ or $\mathbf{v}_3$ as an argument must be $\mathbf{0}_W$) to get $f(\mathbf{v}_1, \mathbf{v}_4) = \mathbf{0}_W$ for all $\mathbf{v}_1, \mathbf{v}_4 \in V$: that is, f must be the zero map (this time, regardless of the field $\mathbb{F}$).

This result generalizes to more than two dimensions; in fact, we can prove something slightly stronger.

Proposition. *Let $n \geq 2$ be an integer, let V and W be vector spaces over the same field, and consider V^n to be a vector space with addition and multiplication derived from the same operations on V. Then the only function $f : V^n \to W$ that is both multilinear as a function on n inputs from V and linear as a function of one input from V^n is the constant map that sends all inputs to $\mathbf{0}_W$.*

Proof. Let $\mathbf{v}_1, \ldots, \mathbf{v}_n, \mathbf{x}$ be arbitrary elements of V. If f is multilinear, then it's linear in the first argument, so:

$$f(\mathbf{v}_1 + \mathbf{x}, \mathbf{v}_2, \ldots, \mathbf{v}_n) = f(\mathbf{v}_1, \ldots, \mathbf{v}_n) + f(\mathbf{x}, \mathbf{v}_2, \ldots, \mathbf{v}_n).$$

If f is also linear, then since $(\mathbf{v}_1 + \mathbf{x}, \mathbf{v}_2, \ldots, \mathbf{v}_n) = (\mathbf{v}_1, \mathbf{0}_V, \ldots, \mathbf{0}_V) + (\mathbf{x}, \mathbf{v}_2, \ldots, \mathbf{v}_n)$, we have

$$f(\mathbf{v}_1 + \mathbf{x}, \mathbf{v}_2, \ldots, \mathbf{v}_n) = f(\mathbf{v}_1, \mathbf{0}_V, \ldots, \mathbf{0}_V) + f(\mathbf{x}, \mathbf{v}_2, \ldots, \mathbf{v}_n)$$

and the first term, like the value of any multilinear function with an argument of zero, must also be zero. If these two expressions for $f(\mathbf{v}_1 + \mathbf{x}, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ are equal, therefore, then

$$f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \mathbf{0}.$$

□

Remark. Our proof didn't actually use the full hypothesis that f was multilinear—only that it was linear in the first argument, to break $f(\mathbf{v}_1 + \mathbf{x}, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ into $f(\mathbf{v}_1, \ldots, \mathbf{v}_n) + f(\mathbf{x}, \mathbf{v}_2, \ldots, \mathbf{v}_n)$, and in at least one of the other arguments (to conclude that f is zero if one of those arguments is zero). And there's nothing special about the first argument, so we've actually shown that if $f : V^n \to W$ is linear as a map on V^n as an entire vector

space, and at least two of the partial application maps $V \to W$ are linear, then f must be the zero map. As an example of a nonzero linear map on V^n whose partial application maps on exactly one argument are linear, take $f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = T(\mathbf{v}_1)$, where T is any nonzero element of $\mathrm{Hom}(V, W)$. Then the partial application maps in argument 1 are linear regardless of the values $\mathbf{v}_2, \ldots, \mathbf{v}_n$, but the partial application maps in any other argument are not linear if $\mathbf{v}_1$ is fixed to a value that is not contained in $\ker T$. (These partial application maps are constant maps with a nonzero value, which cannot be linear.)

As a final note, it's possible to make two generalizations to our definition of multilinear functions as maps from V^n to W:

1. The inputs to a function could come from different vector spaces: that is, we could have maps $V_1 \times \cdots \times V_n \to W$ whose arguments are one element from V_1 in the first position, one element from V_2 in the second position, and so on.

2. Functions could have infinite numbers of arguments.

The definition of multilinearity in either generalization is the same: a function is multilinear if the partial application maps from letting one argument vary and fixing all other arguments except one to arbitrary values are all linear. The proposition that a multilinear function has value zero whenever one argument has value zero also holds for either generalization, as do the results (when suitably modified) about the dimension of the set of linear maps as a vector space.

These generalizations are relatively straightforward, though, and we won't need them much—when we discuss tensor products in Chapter 10, we'll discuss bilinear maps that take inputs from two different vector spaces, but none of the core results really changes. And we won't need multilinear maps that take infinite numbers of inputs, either. So for now, we'll just mention that these generalizations exist, and leave it at that.

7.4.4 Multiplicative but not additive closure of multilinear kernel and image

Unlike with linear maps, the kernel (that is, preimage of $\{\mathbf{0}_W\}$) and image of a multilinear map $f : V^n \to W$ are *not* generally vector subspaces of V^n and W. The basic reason is that unlike linear maps, multilinear maps don't satisfy $f(\mathbf{u}+\mathbf{v}) = f(\mathbf{u})+f(\mathbf{v})$, and this fact generally makes it impossible to conclude that the kernel and image are closed under addition.

- *Kernel not closed under addition*: Let $V = W = \mathbb{R}$ and $n = 2$ (that is, we're treating $\mathbb{R}$ as a vector space over itself), and consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ with the formula $f(x, y) = xy$, treating $\mathbb{R}$ as a vector space over itself. You can check that this satisfies linearity in the first argument ($f(kx, y) = kf(x, y)$ and $f(x_1 + x_2, y) = f(x_1, y) + f(x_2, y)$), and likewise in the second argument. But the "kernel" $f^{-1}(\{0\})$ of this map is $\{(x, y) : x = 0 \text{ or } y = 0\}$, which is not closed under addition: for instance, it includes $(1, 0)$ and $(0, 1)$ but not their sum $(1, 1)$.

- *Image not closed under addition*: Let $V = \mathbb{R}^2$, $W = \mathbb{R}^3$, and and $n = 2$, and consider the function $f : (\mathbb{R}^2)^2 \to \mathbb{R}^3$ with the formula $f((a, b), (c, d)) = (ac, ad, bc)$. All

elements $(x, y, z) \in \mathbb{R}^3$ for $x \neq 0$ are in the image of f: for instance, $(x, y, z) = f((1, z/x), (x, y))$. But if $(x, y, z) = f((a, b), (c, d))$ and $x = 0$, then one of a or c must be zero (and thus one of y or z must be zero). So

$$\operatorname{im} f = \{(x, y, z) \in \mathbb{R}^3 : x \neq 0 \text{ or } y = 0 \text{ or } z = 0\}$$

and there are elements of $\operatorname{im} f$, for example $(0, 1, 0)$ and $(0, 1, 1)$, whose sum is not in $\operatorname{im} f$.

Images and kernels do, however, satisfy the other two subspace axioms. As a multilinear map $f : V^n \to W$ satisfies $f(\mathbf{0}_{V^n}) = f(\mathbf{0}_V, \ldots, \mathbf{0}_V) = \mathbf{0}_W$, so $\operatorname{im} f$ and $\ker f$ always contain the zero elements of their enclosing subspaces $\operatorname{im} f$ and $\ker f$. Both sets are also closed under scalar multiplication:

- If $\mathbf{w} \in \operatorname{im} f$, so $f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \mathbf{w}$ for some vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$, then $k\mathbf{w} = f(k\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) \in \operatorname{im} f$ for any scalar k.
- If $(\mathbf{v}_1, \ldots, \mathbf{v}_n) \in \ker f$, then $f(k(\mathbf{v}_1, \ldots, \mathbf{v}_n)) = f(k\mathbf{v}_1, \ldots, k\mathbf{v}_n) = k^n f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = k^n \mathbf{0}_W = \mathbf{0}_W$ for any scalar k, so $(\mathbf{v}_1, \ldots, \mathbf{v}_n) \in \ker f$ as well.

7.4.5 Dimension and basis of the space of multilinear functions

Like linear maps, multilinear functions can be added and multiplied by scalars: if $f : V^n \to W$ is a multilinear function, then kf is the function that sends $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ to $kf(\mathbf{v}_1, \ldots, \mathbf{v}_n)$, and the sum $f + g$ of f with another multilinear function $g : V^n \to W$ sends $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ to $f(\mathbf{v}_1, \ldots, \mathbf{v}_n) + g(\mathbf{v}_1, \ldots, \mathbf{v}_n)$. (Adding or multiplying multilinear functions like this also adds or multiplies the partial application maps from fixing every argument but one, and the sums and multiples of linear functions such as partial application maps are also linear, so sums and multiples of multilinear functions must also be linear. You may want to sketch this argument in more detail for yourself.)

So the set of multilinear maps from V^n to W (let's denote it $\operatorname{Multilin}(V^n, W)$) is a vector space, and we're naturally confronted with the basic questions we can ask about any vector space: what is its dimension, and can we write a basis for it? We can take a similar approach as we did for our computation of a basis for $\operatorname{Hom}(V, W)$ in section 2.3, by using from the observation that (multi)linear maps are completely determined by their values on a basis of V.

Let's take a small, easy-to-notate example: suppose that V has dimension 2, W has dimension 3, and the number of inputs n is 2. Let $\{\mathbf{b}_1^V, \mathbf{b}_2^V\}$ be a basis for V, and similarly let $\{\mathbf{b}_1^W, \mathbf{b}_2^W, \mathbf{b}_3^W\}$ be a basis for W.

If f is a multilinear function from V^2 to W, then we can use the multilinearity properties to expand any value $f(\mathbf{v}_1, \mathbf{v}_2)$, where $\mathbf{v}_1 = a\mathbf{b}_1^V + b\mathbf{b}_2^V$ and $\mathbf{v}_2 = c\mathbf{b}_1^V + d\mathbf{b}_1^V$ are two arbitrary vectors and a, b, c, d are scalars, into

$$\begin{aligned}
f(\mathbf{v}_1, \mathbf{v}_2) &= f(a\mathbf{b}_1^V + b\mathbf{b}_2^V, c\mathbf{b}_1^V + d\mathbf{b}_2^V) \\
&= f(a\mathbf{b}_1^V, c\mathbf{b}_1^V + d\mathbf{b}_2^V) + f(b\mathbf{b}_1^V, c\mathbf{b}_1^V + d\mathbf{b}_2^V) \\
&\qquad \text{(first argument respects addition)} \\
&= f(a\mathbf{b}_1^V, c\mathbf{b}_1^V) + f(a\mathbf{b}_1^V, d\mathbf{b}_2^V) + f(b\mathbf{b}_2^V, c\mathbf{b}_1^V) + f(b\mathbf{b}_2^V, d\mathbf{b}_2^V) \\
&\qquad \text{(second argument respects addition)} \\
&= ac\, f(\mathbf{b}_1^V, \mathbf{b}_1^V) + ad\, f(\mathbf{b}_1^V, \mathbf{b}_2^V) + bc\, f(\mathbf{b}_2^V, \mathbf{b}_1^V) + bd\, f(\mathbf{b}_2^V, \mathbf{b}_2^V) \\
&\qquad \text{(both arguments respect multiplication)}
\end{aligned}$$

So if we know the four values of $f(\mathbf{b}_1^V, \mathbf{b}_1^V), f(\mathbf{b}_1^V, \mathbf{b}_2^V), f(\mathbf{b}_2^V, \mathbf{b}_1^V), f(\mathbf{b}_2^V, \mathbf{b}_2^V)$, then we know all the values of f. And conversely, any choice of these four determining values gives one valid multilinear map f: specifically, if $\mathbf{w}_{11}, \mathbf{w}_{12}, \mathbf{w}_{21}, \mathbf{w}_{22} \in W$ are freely chosen vectors in W, then the one multilinear map $f : V^2 \to \mathbb{R}$ that satisfies $f(\mathbf{b}_i^V, \mathbf{b}_j^V) = \mathbf{w}_{ij}$ for $i, j \in \{1, 2\}$ is

$$f(a\mathbf{b}_1^V + b\mathbf{b}_2^V, c\mathbf{b}_1^V + d\mathbf{b}_2^V) = ac\mathbf{w}_{11} + ad\mathbf{w}_{12} + bc\mathbf{w}_{21} + bd\mathbf{w}_{22}.$$

(You can check for yourself that this is indeed a multilinear map, and it should be pretty clear that no other possible formula could give you all the right values of $f(\mathbf{b}_i^V, \mathbf{b}_j^V)$.)

Intutively, for each of these four determining values of f, we can independently choose a vector in $\mathbf{w}$, a three-dimensional vector space, so $\mathrm{Multilin}(V^2, W)$ should have dimension $4 \times 3 = 12$. A basis for this space would be the set of maps f_{ijk} for $i \in \{1, 2\}, j \in \{1, 2\}, k \in \{1, 2, 3\}$, that sends $f(\mathbf{b}_i^V, \mathbf{b}_j^V)$ to $\mathbf{b}_k^W$ and sends every other input pair of basis vectors to $\mathbf{0}_W$: there can only be one such map f_{ijk} with such properties for every choice of i, j, k. (To see that these maps can actually be defined, note that the formula for f_{123} [for example] is $f_{123}(a\mathbf{b}_1^V + b\mathbf{b}_2^V, c\mathbf{b}_1^V + d\mathbf{b}_2^V) = ad\mathbf{b}_3^W$; you can show that f_{123} is multilinear in each argument, and it's well-defined because the decomposition of any input vector into a linear combination of $\mathbf{b}_1^V$ and $\mathbf{b}_2^V$ has to be unique).

More generally, the value $f(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ of any $f \in \mathrm{Multilin}(V^n, W)$ on an arbitrary list of n inputs (where now V^n and W have arbitrary dimension) is determined by the $(\dim V)^n$ values of f when each input is chosen independently. Each of these $(\dim V)^n$ possible lists of inputs contributes $\dim W$ basis functions to $\mathrm{Multilin}(V^n, W)$, namely the functions that send that list of inputs to one basis vector of W and send the others to $\mathbf{0}_W$. So $\dim \mathrm{Multilin}(V^n, W) = (\dim V)^n \dim W$.

7.4.6 Symmetric multilinear functions

Definition. *A multilinear function is **symmetric** if swapping any two arguments preserves the value of the function: that is, if*

$$f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = f(\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, \mathbf{v}_j, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_{j-1}, \mathbf{v}_i, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n)$$

for any integers $1 \le i < j \le n$ and any inputs $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$.

Since you can make any permutation by combining enough two-element transpositions, this axiom has the immediate generalization that the value of a symmetric function stays the same if its inputs are permuted in any way: that is,

$$f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = f(\mathbf{v}_{\sigma(1)}, \ldots, \mathbf{v}_{\sigma(n)})$$

for any $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$ and any permutation $\sigma \in S_n$.

Let's denote the set of symmetric multilinear maps from V^n to W by $\mathrm{Sym}(V^n, W)$. It's straightforward to prove that $\mathrm{Sym}(V^n, W)$ is a vector subspace of $\mathrm{Multilin}(V^n, W)$ (i.e. that sums and multiples of symmetric maps are also symmetric, and that the zero map is symmetric).

Like any multilinear map, a symmetric map f is determined by its values on lists of basis elements of V, but the hypothesis that f is symmetric makes some of these lists redundant. If V is a two-dimensional vector space with basis $\{\mathbf{b}_1^V, \mathbf{b}_2^V\}$, for instance, then a general multilinear map $f : V^2 \to W$ is determined by the four values

$f(\mathbf{b}_1^V, \mathbf{b}_1^V), f(\mathbf{b}_1^V, \mathbf{b}_2^V), f(\mathbf{b}_2^V, \mathbf{b}_1^V), f(\mathbf{b}_2^V, \mathbf{b}_2^V)$. But if we also know that f is symmetric, then we don't need to know $f(\mathbf{b}_2^V, \mathbf{b}_1^V)$ if we know the other values, because it has to equal $f(\mathbf{b}_1^V, \mathbf{b}_2^V)$. So a symmetric multilinear map $f : V^2 \to R$ is determined by three inputs, not four, and the space of such maps has dimension $3 \dim W$.

Similarly, if $f : V^5 \to W$ is a symmetric multilinear map with five inputs, then the values $f(\mathbf{b}_1^V, \mathbf{b}_1^V, \mathbf{b}_1^V, \mathbf{b}_2^V, \mathbf{b}_2^V)$, $f(\mathbf{b}_1^V, \mathbf{b}_2^V, \mathbf{b}_1^V, \mathbf{b}_2^V, \mathbf{b}_1^V)$, and $f(\mathbf{b}_2^V, \mathbf{b}_2^V, \mathbf{b}_1^V, \mathbf{b}_1^V, \mathbf{b}_1^V)$ must all be equal (and equal to the values of f on all the other ordered quintuples $(\mathbf{v}_1, \dots, \mathbf{v}_5)$ where three of the $\mathbf{v}_i$ equal $\mathbf{b}_1^V$ and the other two equal $\mathbf{b}_2^V$).

If V is a general finite-dimensional vector space and $f : V^n$ is symmetric, therefore, then the non-redundant inputs that actually determine its value are $(\mathbf{v}_{i_1}^B, \mathbf{v}_{i_2}^B, \dots, \mathbf{v}_{i_n}^B)$ where $i_1, \dots, i_n$ are indices sorted in non-strictly ascending order: that is, $1 \le i_1 \le i_2 \le \cdots \le i_n \le \dim V$. (Any other value-determining input can be turned into one of these by rearranging the basis vectors to sort their indices in ascending order.) Equivalently, the number of independent value-determining inputs equals the number of n-element *multisets* of $\{1, \dots, \dim V\}$—that is, sets that (like regular sets) don't have an internal order of elements, but (unlike regular sets) can contain any number of copies of the same element. These are equivalent because sorting the elements of a multiset of integers gives a unique non-strictly ascending integer sequence; and, conversely, you can put the elements of any such sequence into a unique multiset.[6]

Counting multisets of a set with given size (or, equivalently, non-strictly ascending sequences) is a standard problem in combinatorics. Suffice it to say[7] that the number of n-element multisets in a set of size d is

$$\binom{n+d-1}{n} = \frac{(n+d-1)!}{n!(d-1)!}$$

so $\dim \mathrm{Sym}(V^n, W) = \binom{n+\dim V-1}{n} \dim W$.

7.4.7 Skew-symmetric and alternating multilinear functions

Basic definitions

Let's jump right in with definitions. There are two closely related concepts of *skew-symmetric* and *alternating* linear functions. The first is a close analogue of the concept of symmetric functions:

Definition. *A multilinear map $f : V^n \to W$ is **skew-symmetric** if swapping any two inputs flips the sign of the function: that is,*

$$f(\mathbf{v}_1, \dots, \mathbf{v}_n) = -f(\mathbf{v}_1, \dots, \mathbf{v}_{i-1}, \mathbf{v}_j, \mathbf{v}_{i+1}, \dots, \mathbf{v}_{j-1}, \mathbf{v}_i, \mathbf{v}_{j+1}, \dots, \mathbf{v}_n)$$

for arbitrary indices $1 \le i \le j \le n$ and vectors $\mathbf{v}_1, \dots, \mathbf{v}_n \in V$.

[6] One point that we're glossing over here (and we will also gloss over a similar consideration in the next section on skew-symmetric alternating functions) is whether there actually exists a symmetric multilinear function f for every possible choice of determining values. Rest assured that there is, but the proof is a bit more complicated than you might expect. We'll return to this point in section 10.6.3.

[7] If $a_1, \dots, a_n$ is a n-element non-strictly ascending sequence of values in $\{1, \dots, d\}$, then $a_1, a_2 + 1, a_3 + 2, \dots, a_n + n - 1$ is an n-element *strictly* ascending sequence of values in $\{1, \dots, n + d - 1\}$. This correspondence between non-strictly ascending sequences in $\{1, \dots, d\}$ and strictly ascending sequences in $\{1, \dots, n + d - 1\}$, furthermore, is bijective: given a strictly ascending sequence $b_1, \dots, b_n$, you can restore the original non-strictly ascending sequence $b_1, b_2 - 1, b_3 - 2, \dots, b_n - n + 1$. And there are $\binom{n+d-1}{n}$ such strictly ascending sequences of n elements in $\{1, \dots, n+d-1\}$, the same as the number of n-element subsets, because there is exactly one way to sort any such subset's entries in ascending order.

This definition immediately implies another property for arbitrary permutations $\sigma \in S_n$:

$$f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \operatorname{sgn}(\sigma) f(\mathbf{v}_{\sigma(1)}, \ldots, \mathbf{v}_{\sigma(n)n})$$

because it always takes an even number of transpositions to make an even permutation, and it always takes an odd number of transpositions to make an odd permutation.

The next core definition in this section may seem a bit stranger:

Definition. *A multilinear function $f : V^n \to W$ is **alternating** if f is zero on any input with duplicate vectors: that is, if $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a list of vectors and ther are some integers $1 \leq i < j \leq n$ such that $\mathbf{v}_i = \mathbf{v}_j$, thenw $f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \mathbf{0}_W$.*

We'll write $\operatorname{Sksym}(V^n, W)$ and $\operatorname{Alt}(V^n, W$ for the sets of skew-symmetric and alternating multilinear maps from V^n to W. It's straightforward to prove that both sets are vector subspaces of $\operatorname{Multilin}(V^n, W)$.

Alternating maps are skew-symmetric (and vice versa outside characteristic 2)

Unlike almost all of our results so far, the theory of skew-symmetric and multilinear maps varies depending on the underlying field. In particular, the theory for fields with characteristic 2 (that is, in which $1 = -1$: the multiplicative identity is its own additive inverse) is different from that of all other fields, including the familiar ones $\mathbb{R}$ and $\mathbb{C}$. (See 1.3.3 if you need a reminder of field characteristic.)

In particular, for *arbitrary fields*, we have this result:

Proposition. *Every alternating multilinear map is skew-symmetric.*

Proof. Let V, W be two vector spaces, and let $f : V^n \to W$ be an alternating multilinear map. We'll prove that $f(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_n) = f(\mathbf{v}_2, \mathbf{v}_1, \mathbf{v}_3, \ldots, \mathbf{v}_n)$ (the proof for swapping any pair of arguments other than the first and second is identical). We'll write $\mathbf{V}$ as shorthand for $\mathbf{v}_3, \ldots, \mathbf{v}_n$.

For any multilinear map $f : V^n \to W$, we can use multilinearity in the first and second arguments to get

$$f(\mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_1 + \mathbf{v}_2, \mathbf{V}) = f(\mathbf{v}_1, \mathbf{v}_1, \mathbf{V}) + f(\mathbf{v}_1, \mathbf{v}_2, \mathbf{V}) + f(\mathbf{v}_2, \mathbf{v}_1, \mathbf{V}) + f(\mathbf{v}_2, \mathbf{v}_2, \mathbf{V}).$$

If f is alternating, then all the terms with duplicate arguments become $\mathbf{0}_W$, leaving

$$\mathbf{0}_W = f(\mathbf{v}_1, \mathbf{v}_2, \mathbf{V}) + f(\mathbf{v}_2, \mathbf{v}_1, \mathbf{V})$$

which is what we needed to prove.

□

But the converse applies *only over fields that don't have characteristic 2*. This fact relies on a result from page 31: in a vector space over a field of characteristic 2, every element is its own additive inverse; in a vector space over any other field, the only vector that is its own additive inverse is $\mathbf{0}$.

Proposition. *Every skew-symmetric multilinear function on vector spaces over a field that does not have characteristic 2 is alternating.*

Proof. We'll prove that f has zero value if the first and second arguments are equal; the argument for if any other pair of arguments are equal is the same. Again, use $\mathbf{V}$ as a shorthand for $\mathbf{v}_3, \ldots, \mathbf{v}_n$.

If f is skew-symmetric, then we know that $f(\mathbf{v}_1, \mathbf{v}_2, \mathbf{V}) = -f(\mathbf{v}_2, \mathbf{v}_1, \mathbf{V})$ for arbitrary elements $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$. In particular, if we set $\mathbf{v}_1$ and $\mathbf{v}_2$ to the same element $\mathbf{x} \in V$, then we get

$$f(\mathbf{x}, \mathbf{x}, \mathbf{V}) = -f(\mathbf{x}, \mathbf{x}, \mathbf{V});$$

that is, $f(\mathbf{x}, \mathbf{x}, \mathbf{V})$ is an element of W that is its own additive inverse. If the underlying field has characteristic 2, then this tells us nothing; otherwise, it tells us that $f(\mathbf{x}, \mathbf{x}, \mathbf{V}) = \mathbf{0}_W$, the only element of W that is its own additive inverse.

□

In fact, in characteristic 2, skew-symmetric functions are the same as symmetric functions. The value of a skew-symmetric function (by definition) changes to its negative if two arguments are flipped, while the value of a symmetric function stays the same; and in a vector space over a field of characteristic 2, every vector is its own negative.

For an example in characteristic 2 of a multilinear alternating function that is not symmetric (= skew-symmetric), let $\mathbb{F}_2$ be the field of integers modulo 2, with elements $\{0, 1\}$ and operations $0+0 = 1+1 = 0, 0+1 = 1+0 = 0, 0\times 0 = 0\times 1 = 1\times 0 = 0, 1\times 1 = 1$. Define $f : (\mathbb{F}_2^2)^2 \to \mathbb{F}_2$ as

$$f((x_1, y_1), (x_2, y_2)) = x_1 x_2 + y_1 y_2.$$

This map is clearly (skew-)symmetric: if the subscript 1s and 2s in the RHS are swapped, the result is the same. But it's not alternating, as (for instance) $f((1, 0), (1, 0)) = 1$.

Results on alternating maps

The definition of alternating multilinear maps has an immediate generalization:

Corollary. *An alternating multilinear function $f : V^n \to W$ has value $\mathbf{0}_W$ whenever one of its arguments is a linear combination of the others.*

Proof. We'll prove this in the case when the first argument is a linear combination of the others (the proof for the other arguments is identical). Suppose $\mathbf{v}_1 = c_2\mathbf{v}_2 + c_3\mathbf{v}_3 + \cdots + c_n\mathbf{v}_n$. Then by linearity in the first argument, we can expand

$$f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = c_2 f(\mathbf{v}_2, \mathbf{v}_2, \ldots, \mathbf{v}_n) + c_3 f(\mathbf{v}_3, \mathbf{v}_2, \ldots, \mathbf{v}_n) + \cdots + c_n f(\mathbf{v}_n, \mathbf{v}_2, \ldots, \mathbf{v}_n).$$

Every term on the right-hand side has a duplicate argument to f and thus equals $\mathbf{0}_W$, so $f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \mathbf{0}_W$ as well.

□

Alternating multilinear functions are determined by even fewer values on basis vector inputs than symmetric multilinear functions. Let's return to our example of $f : V^2 \to W$ where V has dimension 2 and basis $\{\mathbf{b}_1^V, \mathbf{b}_2^V\}$. A fully generic multilinear function f is determined by the four values $f(\mathbf{b}_1^V, \mathbf{b}_1^V), f(\mathbf{b}_1^V, \mathbf{b}_2^V), f(\mathbf{b}_2^V, \mathbf{b}_1^V), f(\mathbf{b}_2^V, \mathbf{b}_2^V)$. If we also know that f is alternating, not only do we not need to know $f(\mathbf{b}_2^V, \mathbf{b}_1^V)$—it must equal $-f(\mathbf{b}_1^V, \mathbf{b}_2^V)$, because alternating maps are skew-symmetric—but we also

know that $f(\mathbf{b}_1^V, \mathbf{b}_1^V) = f(\mathbf{b}_2^V, \mathbf{b}_2^V) = \mathbf{0}_W$, because f must have value $\mathbf{0}_W$ if it has a duplicate input. So our choice of $f(\mathbf{b}_1^V, \mathbf{b}_2^V)$ completely determines f.

More generally, suppose $f : V^n \to W$ and V has dimension d, and $\{\mathbf{b}_1^V, \ldots, \mathbf{b}_d^V\}$ is a basis of V. Let $a_1, \ldots, a_n$ be integer indices in the range $1 \leq a_i \leq d$. Then:

1. If any of the indices $a_1, \ldots, a_n$ equals any of the others, then $f(\mathbf{b}_{a_1}^V, \ldots, \mathbf{b}_{a_n}^V) = \mathbf{0}_W$.
2. If the indices $a_1, \ldots, a_n$ are all different, then let $b_1, \ldots, b_n$ be the rearrangement of $a_1, \ldots, a_n$ in ascending order, and let σ be the permutation on $\{1, \ldots, n\}$ such that $i = \sigma(i)$. Then $f(\mathbf{b}_{a_1}^V, \ldots, \mathbf{b}_{a_n}^V) = \operatorname{sgn}(\sigma) f(\mathbf{b}_{b_1}^V, \ldots, \mathbf{b}_{b_n}^V)$.

So f is determined by its values on inputs of the form $(\mathbf{b}_{b_1}^V, \ldots, \mathbf{b}_{b_n}^V)$, where the indices are non-equal and strictly increasing: $1 \leq b_1 < \cdots < b_n \leq \dim V$. Every such sequence of indices corresponds to one n-element subset (*not* multiset) of $\{1, \ldots, \dim V\}$, and the number of such sets is

$$\binom{\dim V}{n} = \frac{(\dim V)!}{n!(\dim V - n)!}.$$

The total dimension of the vector space of alternating sets is thus $\binom{\dim V}{n} \dim W$.

7.4.8 Special properties of $\operatorname{Alt}(V^n, W)$ when $\dim V = n$

One special case of our discussion of alternating functions: if $\dim V = n$, then any alternating multilinear function $f : V^n \to W$ is determined entirely by the single value $f(\mathbf{b}_1^V, \ldots, \mathbf{b}_n^V)$. In fact, there's a bijective map $T : W \to \operatorname{Alt}(V^n, W)$: specifically, $T\mathbf{w}$ for any vector $\mathbf{w} \in W$ is the uniquely determined alternating map f such that $f(\mathbf{b}_1^V, \ldots, \mathbf{b}_n^V) = \mathbf{w}$. (I'll leave it to you to prove that T is in fact linear: the proof is straightforward.)

There is one more important pair of result and corollary that will become crucial for the definition of the determinant:

Proposition. *If V and W are two vector spaces over the same field with $\dim V = n$, and $f \in \operatorname{Alt}(V^n, W)$, then every value of f on linearly independent inputs is a scalar multiple of every other.*

Proof. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\{\mathbf{v}'_n, \ldots, \mathbf{v}'_n\}$ be two linearly independent subsets (and, therefore, bases) of V. Define $\mathbf{w} = f(\mathbf{v}_1, \ldots, \mathbf{v}_n)$.

We can write each of $\mathbf{v}'_1, \ldots, \mathbf{v}'_n$ as a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and then multilinearity of f to expand $f(\mathbf{v}'_1, \ldots, \mathbf{v}'_n)$ into a sum of scalars times values of f with all arguments drawn from $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. As f is alternating, each of these values of f is either zero (if it got a duplicate argument) or $\pm\mathbf{w}$ (if its arguments are a permutation of $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$, so the sum must be a scalar multiple of $f(\mathbf{v}_1, \ldots, \mathbf{v}_n)$.

□

Corollary. *With the same hypotheses as in the previous proposition, if f is not the zero map, then:*

1. *$\operatorname{im} f$ is a one-dimensional subspace of W.*
2. *$f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \mathbf{0}$ if and only if $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is linearly dependent. (The "if" statement is a basic result on general alternating maps; the reverse "and only if" implication is new.)*

Proof. If f has a nonzero value, then it must take that value on some list of linearly independent inputs, because the value of any alternating multilinear function on linearly dependent inputs is zero. So V has some linearly independent subset (and necessarily basis, because $\dim V = n$) $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ that satisfies $\mathbf{w} \coloneqq f(\mathbf{v}_1, \ldots, \mathbf{v}_n) \neq \mathbf{0}_W$.

Then previous results give us:

1. From section 7.4.4, the image of a multilinear map is closed under multiplication, so every multiple of $\mathbf{w}$ is in $\operatorname{im} f$.

2. From the previous proposition, every value of f on linearly independent inputs is a multiple of any other (and, in particular, a multiple of $\mathbf{w}$). And on linearly dependent inputs, f can only take the value $\mathbf{0}_W$. So $\operatorname{im} f$ contains only elements of $\operatorname{span}\{\mathbf{w}\}$.

   Statements 1 and 2 together imply $\operatorname{im} f = \operatorname{span}\{\mathbf{w}\}$, which is conclusion 1 in the corollary statement.

3. $\mathbf{w} = f(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ (which, again, by definition cannot be $\mathbf{0}_W$) must be a scalar multiple of $f(\mathbf{v}'_1, \ldots, \mathbf{v}'_n)$ for any other linearly independent set $\{\mathbf{v}'_1, \ldots, \mathbf{v}'_n\}$. This implies that $f(\mathbf{v}'_1, \ldots, \mathbf{v}'_n) \neq \mathbf{0}_W$, because the only scalar multiple of $\mathbf{0}_W$ is $\mathbf{0}_W$. This establishes conclusion 2 in the corollary statement.

□

7.5 Formal definition of determinant

We're finally ready to present the formula for the determinant of any square matrix. It's not obvious that this definition satisfies the properties that we outlined on page 166, but soon enough, we'll prove that it does.

Let A be a square $n \times n$ matrix with entries in some arbitrary field $\mathbb{F}$. Write the entry in row i and column j of A as a_{ij}. (From now on, we'll adopt the shorthand notation $A = (a_{ij})$ to indicate that A is a matrix whose entry in row i and column j is (a_{ij}). We'll also say that the slot in a matrix at row i and column j is "position (i, j).")

The determinant of A, notated $\det A$, is this:

$$\det A = \sum_{\sigma \in S_n} (-1)^\sigma a_{1,\sigma(1)} a_{2,\sigma(2)} \cdots a_{n,\sigma(n)}.$$

This definition may need some explanation. Every permutation σ of the set $\{1, \ldots, n\}$ gives a a choice of n entries of the matrix that includes one entry in each row and in each column: namely, for every row number i, pick the entry in column $\sigma(i)$. This choice of entries contributes one term to the sum for $\det A$: multiply all the entries together, and then flip the sign if σ is odd.

To see how this formula works more clearly, let's specialize it to 2×2 and 3×3 matrices. (The determinant of a 1×1 matrix is its sole entry.) Consider the general 2×2 matrix

$$\begin{bmatrix} w & x \\ y & z \end{bmatrix}.$$

S_2 contains two permutations: the identity, which is even, and the transposition $(1\,2)$, which is odd. The identity gives the term $a_{11}a_{22} = wz$ and $(1\,2)$ gives the term $-a_{12}a_{21} = -xy$, so the determinant of a 2×2 matrix is $wz - xy$.

Now consider the 3×3 matrix

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

S_3 has $3! = 6$ elements, three odd and three even. The even permutations are the identity, which gives the term aei, and the cycles $(1\,2\,3)$ (which gives the product of the entries at positions $(1,2)$, $(2,3)$, and $(3,1)$, namely bfg) and $(1\,3\,2)$ (which gives the product of the entries at positions $(1,3)$, $(2,1)$, and $(3,2)$, namely cdh).

The odd permutations are the transpositions $(1\,2)$, $(2\,3)$, and $(1\,3)$, which respectively give $-bdi$, $-afh$, and $-ceg$. Therefore,

$$\det \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = aei + bfg + cdh - afh - bdi - ceg.$$

You may want write out an expanded formula for 4×4 determinants. S_4 contains 24 elements: the identity, 6 transpositions (i.e. length-2 cycles), 8 cycles of length 3, 6 cycles of length 4, and 3 permutations with cycle structure 2+2. Remember that exactly half of the elements of S_n for $n \geq 2$ are even.

One final bit of notation: the determinant of a matrix is sometimes notated by writing the matrix entries within vertical bars rather than brackets: thus, for instance, you may see $\begin{vmatrix} a & b \\ c & d \end{vmatrix}$ to mean $\det \begin{bmatrix} a & b \\ c & d \end{bmatrix}$.

For now, we'll define the determinant of an *operator* $T : \mathbb{F}^n \to \mathbb{F}^n$ as the determinant (as defined above) of the *matrix* that represents T with respect to the standard basis. We'll eventually prove that the choice of basis doesn't matter: all matrix representations of T have the same determinant.

One useful exercise, finally, may help you solidify your understanding of transpositions and the definition of the determinant. Consider the $n \times n$ matrix with entries of 1 on the *anti-diagonal* from top right to bottom left, and 0 elsewhere: that is,

$$\begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Prove that the determinant of this matrix is 1 if n has remainder 0 or 1 when divided by 4, and -1 otherwise. (Hint: only one permutation $\sigma \in S_n$ picks out all the nonzero terms, and only nonzero terms, from this matrix. What is this permutation's sign?) This matrix is an example of a *permutation matrix*, which has entries of 1 in the positions $(i, \sigma(i))$ for some fixed permutation σ and entries of 0 everywhere else. The determinant of a permutation matrix is just $\operatorname{sgn} \sigma$.

7.6 Properties of the determinant

The definition of the determinant given in the previous section is complicated and may seem like it came out of thin air, but thankfully, most computations with determinants

can just use several general properties of determinants, not the sum-over-permutations definition. These four properties are the most important. In this section, we'll prove properties 2, 3, and 4, as well as implication 3b; property 1 and implication 3a will take a bit more work.

We'll assume for the rest of this section that the underlying field does not have characteristic 2, so we have access to all the results from sections 7.4.7 and 7.4.8.

1. The determinant of a matrix product is the product of the individual matrix determinants: $\det(AB) = (\det A)(\det B)$.

2. Every matrix has the same determinant as its transpose.

3. The determinant, viewed as an n-input function that takes each matrix row or column as a separate input, is an alternating (and therefore skew-symmetric, even in characteristic 2) multilinear function from $(\mathrm{Row}_n(\mathbb{F}))^n$ to $\mathbb{F}$ as well as from $(\mathrm{Col}_n(\mathbb{F}))^n$ to $\mathbb{F}$. This fact has two important implications:

   (a) Elementary row operations affect determinants in predictable ways, namely: scale operations multiply the determinant by the scale factor, swap operations flip the sign of the determinant, and shear operations leave the determinant unchanged.

   (b) The determinant is zero if and only the matrix has linearly dependent rows (and, therefore, columns); that is, if and only if it does not have full rank and (equivalently) creates a non-bijective multiplication operator on $\mathrm{Col}_n(\mathbb{F})$.

4. The determinant of a triangular matrix is the product of the diagonal entries.

This section is unavoidably a blizzard of small propositions, but all of them are either for proving one of these four properties or a lemma that will be useful for later sections. No proof relies on an inference that a skew-symmetric multilinear function is alternating, so they all work in characteristic 2.

Proposition. *The determinant of an upper or lower triangular matrix is the product of the diagonal entries. (This is key property 4.)*

Proof. If $A = (a_{ij})$ is an $n \times n$ upper triangular matrix, then the terms $a_{1\sigma(1)}, \ldots, a_{n\sigma(n)}$ can only be all nonzero if $i \leq \sigma(i)$ for all integers $1 \leq i \leq n$. The only element $\sigma \in S_n$ that satisfies this criterion, as we remarked on page 171, is the identity permutation, which gives the element $a_{11} \cdots a_{nn}$. The reasoning for lower triangular matrices is similar.

□

Remark. Diagonal matrices are upper and lower triangular, so the determinant of a diagonal matrix is also the product of its diagonal entries. The diagonal matrix with diagonal entries of 1—that is, the identity matrix—thus has determinant 1.

Proposition. *If $A = (a_{ij})$ is a square matrix and A^T is its transpose, then $\det A = \det A^T$. (This is key property 2.)*

Proof. Write $A^T = B = (b_{ij})$. Then $b_{ij} = a_{ji}$ for all integers $1 \leq i, j \leq n$.

The entries $a_{1\sigma(1)}, \ldots, a_{n\sigma(n)}$ in one term of $\det A$ are the same as (though in a different order from) the entries $b_{1\sigma^{-1}(1)}, \ldots, b_{n\sigma^{-1}(n)}$ in one term of $\det B$: if $j = \sigma(i)$, then $a_{i\sigma(i)}$ equals $b_{j\sigma^{-1}(j)}$, and σ gives a bijection between values of i and j. So the terms for σ in $\det A$ and σ^{-1} in $\det B$ have the same values and the same sign, because $\operatorname{sgn}\sigma = \operatorname{sgn}(\sigma^{-1})$. The sum-over-permutations expressions for $\det A$ and $\det B$ thus have all the same terms with the same sign, just ordered differently.

□

Proposition. *The determinant is a multilinear function on matrix rows. (This is part of key property 3.)*

Proof. Some notation: write $R(\mathbf{r}_1, \ldots, \mathbf{r}_n)$ for the matrix with rows $\mathbf{r}_1, \ldots, \mathbf{r}_n \in \mathrm{Row}_n(\mathbb{F})$, and write $D(\mathbf{r}_1, \ldots, \mathbf{r}_n)$ for the determinant of this matrix.

To prove that D is a multilinear function from $(\mathrm{Row}_n(\mathbb{F}))^n$ to $\mathbb{F}$, we need to prove that the partial application maps from fixing all inputs to D except one to arbitrary values are linear. We'll prove that the map $D(\cdot, \mathbf{r}_2, \ldots, \mathbf{r}_n)$, with all arguments but the first fixed, is linear for all $\mathbf{r}_2, \ldots, \mathbf{r}_n \in \mathrm{Row}_n(\mathbb{F})$. (The proof for the other partial application maps is identical.)

To prove that $D(\cdot, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ is linear, we need to check that it satisfies the two linear map axioms:

1. *Respect for addition*: $D(\mathbf{r}_1 + \mathbf{r}'_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) = D(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) + D(\mathbf{r}'_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ for all $\mathbf{r}_1, \mathbf{r}'_1, \mathbf{r}_2, \ldots, \mathbf{r}_n \in \mathrm{Row}_n(\mathbb{F})$. Proof: Define $A = (a_{ij}) \coloneqq R(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$, $B = (b_{ij}) \coloneqq R(\mathbf{r}'_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$, and $C = (c_{ij}) \coloneqq R(\mathbf{r}_1 + \mathbf{r}'_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$. Then: $c_{1j} = a_{1j} + b_{1j}$ for every column index $1 \leq j \leq n$, and $a_{ij} = b_{ij} = c_{ij}$ for every row index $2 \leq i \leq n$. Remember that every term in the formula for the determinant includes exactly one matrix element from the first row.

   So for each permutation σ, we can expand the term $c_{1\sigma(1)}c_{2\sigma(2)}$ in $\det C$ as $(a_{1\sigma(1)} + b_{1\sigma(1)})c_{2\sigma(2)} \cdots c_{n\sigma(n)} = a_{1\sigma(1)}a_{2\sigma(2)} \cdots a_{n\sigma(n)} + b_{1\sigma(1)}b_{2\sigma(2)} \cdots b_{n\sigma(n)}$: that is, the term with σ in the formula $\det C$ is the sum of the terms with σ in $\det A$ and in $\det B$. This means that $\det C = \det A + \det B$.

2. *Respect for multiplication*: $D(k\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) = kD(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ for all $\mathbf{r}_1, \ldots, \mathbf{r}_n \in \mathrm{Row}_n(\mathbb{F})$ and $k \in \mathbb{F}$. Proof: every term in the determinant includes exactly one matrix element in the first row, so multiplying everything in the first row by k means multiplying every term in the determinant (and, therefore, the total value of the determinant) by k.

□

Corollary. *The determinant is a multilinear function on matrix columns.*

Proof. Virtually identical to the proof for the preceding proposition with the words "row" and "column" interchanged where necessary. You can also argue that every matrix has the same determinant as its transpose, and vector space operations on the columns of a matrix produce the same operations on the rows of its transpose.

□

Proposition. *The determinant of a matrix with two equal rows is zero. (This establishes that the determinant is an alternating function on matrix rows, which is the remaining part of key property 3.)*

Proof. Let $M = (m_{ij})$ be an $n \times n$ matrix, and suppose rows k and ℓ of M are equal. Let $\tau \in S_n$ be the permutation (necessarily odd) that transposes k and ℓ while leaving every other element the same. Let A_n denote the set of even permutations in S_n. Note that as with all transpositions, $\tau = \tau^{-1}$.

The map $\sigma \mapsto \sigma \circ \tau$, as we mentioned on page 169, is a bijection on S_n that takes even permutations to odd permutations and vice versa, so every element $\sigma \in S_n$ can be written in exactly one way as either $\sigma = \tilde{\sigma}$ or $\sigma = \tilde{\sigma} \circ \tau$ where $\tilde{\sigma} \in A_n$.

So we can split the terms in $\det M$ into pairs, like this:

$$\begin{aligned}\det M &= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) m_{1,\sigma(1)} \cdots m_{n(\sigma n)} \\ &= \sum_{\tilde{\sigma} \in A_n} \left(\operatorname{sgn}(\tilde{\sigma}) m_{1,\tilde{\sigma}(1)} \cdots m_{n,\tilde{\sigma}(n)} + \operatorname{sgn}(\tilde{\sigma} \circ \tau) m_{1,\tilde{\sigma}\circ\tau(1)} \cdots m_{n,\tilde{\sigma}\circ\tau(n)} \right) \\ &= \sum_{\tilde{\sigma} \in A_n} \left(m_{1,\tilde{\sigma}(1)} \cdots m_{n,\tilde{\sigma}(n)} - m_{1,\tilde{\sigma}\circ\tau(1)} \cdots m_{n,\tilde{\sigma}\circ\tau(n)} \right).\end{aligned}$$

The products $m_{1,\tilde{\sigma}(1)} \cdots m_{n,\tilde{\sigma}(n)}$ and $m_{1,\tau\circ\tilde{\sigma}(1)} \cdots m_{n,\tau\circ\tilde{\sigma}(n)}$ include mostly the same matrix elements, except that while the first product includes the elements in positions $(k, \tilde{\sigma}(k))$ and $(\ell, \tilde{\sigma}(\ell))$, the second product includes the elements in positions $(k, \tilde{\sigma} \circ \tau(k)) = (k, \tilde{\sigma}(\ell))$ and $(\ell, \tilde{\sigma} \circ \tau(\ell)) = (\ell, \tilde{\sigma}(k))$. But if rows k and ℓ are equal, then these two products of matrix entries must be equal for every $\tilde{\sigma} \in A_n$, and their difference must be zero.

□

Corollary. *If two columns of a matrix are equal, then the matrix has determinant zero.*

Proof. You can make a similar argument to the previous proposition based on the fact that if columns k and ℓ are equal and τ is the transposition of k and ℓ, then $m_{1,\sigma(1)} \cdots m_{n,\sigma(n)}$ and $m_{1,\tau\circ\sigma(1)} \cdots m_{n,\tau\circ\sigma(n)}$ are equal, and $\sigma \mapsto \tau \circ \sigma$ is a bijection on S_n that takes even permutations to odd permutations and vice versa.

Alternatively, note that $\det M = \det M^T$, and M has two equal columns if and only if M^T has two equal rows.

□

We have now established key properties 2, 3, and 4. Since the determinant is an alternating multilinear function on n inputs from an n-dimensional vector space $\operatorname{Row}_n(\mathbb{F})$ or $\operatorname{Col}_n(\mathbb{F})$, we can apply of the results from sections 7.4.7 and 7.4.8.

Any such map must be uniquely determined its value on any one set of linearly independent inputs. In particular, we could choose the identity matrix I as the determining input, which gives us the following elegant characterization of the determinant:

Corollary. *The determinant is the only multilinear function on matrix rows (or columns) that satisfies* $\det I = 1$.

Finally, the corollary on page 180, plus the fact that there is at least one matrix with nonzero determinant, gives us these results:

1. Any matrix has determinant zero if and only if its rows are not linearly independent. The same goes for columns. (This is implication 3b in our list.)

2. The image of the multiplication operator on $\operatorname{Col}_n(\mathbb{F})$ created by a matrix is the span of the columns, so this operator is bijective if and only if the determinant of the matrix that creates it is not zero.

7.7 Multiplicativity of the determinant

We've established properties 2 through 4 of the list at the beginning of the last section (with the exception of 3a), but we haven't proved yet that $\det(AB) = \det A \det B$. To prove this, we're going to use the following facts:

1. Every matrix M of dimension $r \times c$ can be factored as $M = R_1 \cdots R_n E$ for some integer $n \geq 0$, where $R_1, \ldots, R_n$ are $r \times r$ matrix representations of elementary row operations and E is in RREF.

2. A square matrix in RREF either is the identity matrix or has a row of zeros.

3. An $n \times n$ matrix has nonzero determinant if and only if its rank is zero.

The missing piece of our argument is a correspondence between the determinants of elementary row operations' matrix representations, on the one hand, and the effect that the row operations have on the determinants of other matrices, on the other: in particular, if R is an elementary row operation matrix and B is any matrix, then $\det(RB) = \det R \det B$. This result will let us prove that $\det(AB) = \det A \det B$ for arbitrary matrices.

Let's look at each of the three elementary row operations in turn:

1. Row scaling operations $\mathbf{r}_i \mapsto \lambda \mathbf{r}_i$ multiply determinants by the factor λ, because every term in a determinant includes one entry from each row. The matrix representation of row scaling is a diagonal matrix with one entry of λ at position (i, i) and all other diagonal entries 1, so its determinant is the product of the diagonal entries, namely λ.

2. Row swaps $\mathbf{r}_i \leftrightarrow \mathbf{r}_j$ flip the sign of the determinant (that is, multiply it by -1), because the determinant is alternating and therefore skew-symmetric (even in characteristic 2). The matrix representation of a row swap is the permutation matrix with entries of 1 in the positions (i, j) and (j, i) and zeros elsewhere: that is, the permutation matrix that represents the transposition τ_{ij}. In the sum-over-permutations expression for the determinant of this matrix, the only permutation σ that chooses only nonzero terms is $\sigma = \tau_i$, so the determinant is $\operatorname{sgn}(\tau_{ij}) = -1$.

3. Row shears $\mathbf{r}_i \mapsto \mathbf{r}_i + \lambda \mathbf{r}_j$ leave the determinant unchanged. To see this, write $D(\mathbf{r}_1, \ldots, \mathbf{r}_n)$ for the determinant of the matrix with rows $\mathbf{r}_1, \ldots, \mathbf{r}_n \in \operatorname{Row}_n(\mathbb{F})$. Then for the operation $\mathbf{r}_1 \mapsto \mathbf{r}_1 + \lambda \mathbf{r}_2$, we have

$$\begin{aligned} D(\mathbf{r}_1 + \lambda \mathbf{r}_2, \mathbf{r}_2, \ldots, \mathbf{r}_n) &= D(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) + \lambda D(\mathbf{r}_2, \mathbf{r}_2, \ldots, \mathbf{r}_n) \\ &\qquad \text{(linearity in first argument)} \\ &= D(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) \\ &\qquad \text{(alternating function with duplicate arguments has value zero)} \end{aligned}$$

   The matrix representation of $\mathbf{r}_i \mapsto \mathbf{r}_i + \lambda \mathbf{r}_j$ has entries of 1 along the diagonal and one nonzero off-diagonal entry of λ in position (i, j). This matrix is either lower triangular if $i > j$ or upper triangular if $i < j$; in either case, it has determinant 1, because the determinant of a triangular matrix is the product of its diagonal entries.

These results can be summed up in the following lemma:

Lemma. *If $R_1, \dots, R_n$ are $n \times n$ matrix representations of elementary row operations and A is any $n \times n$ matrix, then* $\det(R_1 \cdots R_n A) = \det R_1 \cdots \det R_n \det A$.

Proof. We've basically just proved this result for the $n = 1$ case. If R is a row scaling by λ, then $\det R = \lambda$ and $\det(RA) = \lambda \det A$; if R is a row swap, then $\det R = -1$ and $\det(RA) = -\det A$; if R is a shear, then $\det R = 1$ and $\det(RA) = \det A$.

To prove the general case, first parenthesize $R_1 \cdots R_n A$ as $R_1(R_2 \cdots R_n A)$ and apply the $n = 1$ case to get $\det(R_1 \cdots R_n A) = \det R_1 \det(R_2 \cdots R_n A)$; then apply the $n = 1$ case again to get $\det R_2 \cdots R_n A = (\det R_2) \det(R_3 \cdots R_n A)$, and so forth.

□

Remember also that a matrix A has determinant zero if and only if the multiplication operator $\mathbf{v} \mapsto A\mathbf{v}$ on $\mathrm{Col}_n(\mathbb{F})$ is not bijective (and thus $\mathrm{nullsp}\, A$ contains nonzero elements). This gives us the following result:

Lemma. *If A and B are matrices, and at least one of A and B has determinant zero, then AB also has determinant zero.*

Proof. Two cases:

1. *A has determinant zero, but B doesn't.* Choose some nonzero vector $\mathbf{v} \in \mathrm{nullsp}\, A \subseteq \mathrm{Col}_n(\mathbb{F})$, and choose $\mathbf{w}$ such that $B\mathbf{w} = \mathbf{v}$. (As B doesn't have determinant zero, so its multiplication map is bijective, so such a $\mathbf{w}$ must exist.) Thus, $(AB)\mathbf{w} = A(B\mathbf{w}) = A\mathbf{v} = \mathbf{0}$. And $\mathbf{w}$ can't be zero (because $M\mathbf{0} = \mathbf{0}$ for any matrix M, but $\mathbf{v} \neq 0$). So $\mathbf{w}$ is a nonzero element of $\mathrm{nullsp}\, AB$, so $\det(AB) = 0$.

2. *B has determinant zero.* Then if $\mathbf{v}$ is any nonzero element of $\mathrm{nullsp}\, B$, then $(AB)\mathbf{v} = A(B\mathbf{v}) = A\mathbf{0} = \mathbf{0}$, so $\mathbf{v} \in \mathrm{nullsp}(AB)$.

□

Remark. The above proof didn't actually use any properties of matrices as opposed to operators besides the correspondence between bijectivity and nonzero determinant; we could have phrased it in operator language that if V is a finite-dimensional vector space and $T_1, T_2 \in \mathrm{End}(V)$, then $\dim\ker(T_1 \circ T_2) \neq 0$ if either $\dim\ker T_1 \neq 0$ or $\dim\ker T_2 \neq 0$.

Finally, we have our final result:

Theorem. *The determinant respects matrix multiplication. That is, for* any *two matrices $A, B \in \mathrm{Mat}_{n\times n}(\mathbb{F})$, we have* $\det(AB) = \det A \det B$.

Proof. Break into two cases:

1. *$\det A = 0$ or $\det B = 0$ or both.* Then the lemma we just proved shows that $\det(AB) = 0 = \det A \det B$.

2. *$\det A \neq 0$ and $\det B \neq 0$.* Then $\mathrm{rref}\, A = \mathrm{rref}\, B = I$, so A and B factor completely as products of elementary row operation matrices $A = R_1 \cdots R_m$ and $B = S_1 \cdots S_n$, so $AB = R_1 \cdots R_m S_1 \cdots S_n$. We've already proved that the determinant respects matrix multiplication when every matrix in a product except possibly the rightmost is an elementary row operation matrix, so $\det A = \det R_1 \cdots \det R_m$, $\det B = \det S_1 \cdots \det S_n$, and $\det(AB) = \det R_1 \cdots \det R_m \det S_1 \cdots \det S_m$.

□

Remark. This result immediately generalizes to three or more matrices: for instance, $\det(ABC) = \det A \det(BC) = \det A \det B \det C$.

The multiplicativity of the determinant has two important consequences:

1. We can compute an $n \times n$ matrix's determinant from its LU factorization in $O(n)$ time: if $PA = LU$, then $\det P \det A = \det L \det U$, and $\det P$ is a permutation matrix (with determinant ± 1) and L and U are triangular matrices whose determinants are the products of their diagonal entries.

2. Since $\det I = 1$ and $AA^{-1} = I$ for any invertible matrix A, so $\det A^{-1} = 1/\det A$.

7.8 Minors, cofactors, adjugate matrix

This section and the following sections discuss an alternate method of computing matrix determinants that can be useful for computations by hand. Along the way, we'll find formulas in closed form for the entries of a matrix inverse and the solutions to a linear system, not merely an algorithm for computing them (though these formulas are too complicated to be of much use).

First, some definitions. Let A be an $n \times n$ matrix. A *first minor* of A, notated M_{ij} for some particular integers i, j, is the determinant of the $(n-1) \times (n-1)$ matrix created by removing row i and column j from A. We'll call this matrix $A^{(ij)}$, so $M_{ij} = \det A^{(ij)}$. (There are also second, third, etc. minors created by removing two, three, etc. rows and columns, but we won't discuss those.)

For example, consider the matrix

$$A = \begin{bmatrix} 1 & -4 & 4 & 3 \\ 0 & 2 & 6 & 2 \\ -3 & 5 & -1 & -2 \\ 0 & 4 & 0 & -5 \end{bmatrix}$$

Let's calculate the second minor M_{32}. Removing the required row and column gives

$$\begin{bmatrix} 1 & \star & 4 & 3 \\ 0 & \star & 6 & 2 \\ \star & \star & \star & \star \\ 0 & \star & 0 & -5 \end{bmatrix} \longrightarrow A^{(32)} = \begin{bmatrix} 1 & 4 & 3 \\ 0 & 6 & 2 \\ 0 & 0 & -5 \end{bmatrix}$$

which is an upper triangular matrix with determinant $M_{32} = -30$, the product of the entries on the diagonal.

A *cofactor* of a matrix, notated C_{ij}, is the corresponding first minor M_{ij} if $i + j$ is even, or $-M_{ij}$ if $i + j$ is odd. In short, $C_{ij} = (-1)^{i+j} M_{ij}$. So, for example, $C_{11} = M_{11}$, $C_{12} = -M_{12}$, and $C_{21} = -M_{21}$. The cofactor C_{32} of the matrix A above is $-M_{32} = 30$, because $3 + 2$ is odd.

The *cofactor matrix* of A is the matrix whose entry in row i and column j is C_{ij}. The *adjugate matrix*, notated $\operatorname{adj} A$, is the transpose[8] of the cofactor matrix:

$$\operatorname{adj} A = \begin{bmatrix} C_{11} & C_{21} & C_{31} & \cdots & C_{n1} \\ C_{12} & C_{22} & C_{32} & \cdots & C_{n2} \\ C_{13} & C_{23} & C_{33} & \cdots & C_{n3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{1n} & C_{2n} & C_{3n} & \cdots & C_{nn} \end{bmatrix} = \begin{bmatrix} M_{11} & -M_{21} & M_{31} & \cdots & \pm M_{n1} \\ -M_{21} & M_{22} & -M_{23} & \cdots & \mp M_{n2} \\ M_{13} & -M_{23} & M_{33} & \cdots & \pm M_{n3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \pm M_{1n} & \mp M_{2n} & \pm M_{3n} & \cdots & M_{nn} \end{bmatrix}$$

where $\pm$ means $+$ if n is odd and $-$ if n is even, and vice versa for $\mp$.

7.9 Expansion by cofactors

The cofactor and adjugate matrices give us a recursive method of computing the determinant, called *Laplace expansion* or *expansion by cofactors*. This method is impractical for computation in the general case because it expands the determinant of an $n \times n$ matrix into a sum of n determinants of $(n-1) \times (n-1)$ matrices; expanding these determinants recursively takes $O(n!)$ computation time. But it can be useful for computations by hand on small matrices, especially if one one row or column of the matrix has a large number of zeros.

7.9.1 Restricted permutations

Laplace expansion hinges on the following result: if you take any row $a_{i1}, a_{i2}, \ldots, a_{in}$ of a matrix A and multiply each entry by the cofactor C_{i1} in the same position in the matrix of cofactors, then the sum of results $a_{i1}C_{i1} + a_{i2}C_{i2} + \cdots + a_{in}C_{in}$ is $\det A$. The same result holds for columns: $a_{1i}C_{1i} + \cdots + a_{ni}C_{ni} = \det A$ for any integer $1 \leq i \leq n$. More concisely: every diagonal entry in the matrices $A(\operatorname{adj} A)$ and $(\operatorname{adj} A)A$ is $\det A$. To prove the theorem, first we'll prove a lemma about general transpositions that lets us relate the signs of terms in the expansion of $\det A$ to the signs of corresponding terms in A's first minors. What we mean by "corresponding" should be clearer in a bit. The statement of the lemma is unavoidably complicated; I've tried to make the intuition clear.

Lemma. *Let n be an integer ≥ 2, and let $\sigma \in S_n$ be some permutation of $\{1, \ldots, n\}$. Write $[n]$ for the set of integers $\{1, \ldots, n\}$, and remember that $[n] \setminus \{k\} = \{1, \ldots, k-1, k+1, \ldots, n\}$ for every integer $1 \leq k \leq n$. For each integer $1 \leq k \leq n$, define the bijective functions $f_k : [n-1] \to [n] \setminus \{k\}$ as*

$$f_k(i) = \begin{cases} i & 1 \leq i \leq k-1 \\ i+1 & k \leq i \leq n-1 \end{cases}.$$

That is, $f_k(i)$ is the ith smallest element of $[n] \setminus \{k\}$, and for an $n \times n$ matrix A, the entry at position (i, j) of $A^{(k,\ell)}$ equals the entry at position $(f_k(i), f_\ell(j))$ of A. Note that $f_k^{-1} : [n] \setminus \{k\} \to [n-1]$ assigns every element of $[n] \setminus \{k\}$ to its rank order.

[8] Remember that the transpose of a matrix is the reflection of the matrix across the diagonal, so that row i becomes column i and vice versa.

Also define the restricted permutation $\tilde{\sigma}_k \in S_{n-1}$ *as* $\tilde{\sigma}_k = f_{\sigma(k)} \circ \sigma \circ f_k^{-1}$*; that is, if* σ *gives a bijection from* $[n] \setminus \{k\}$ *to* $[n] \setminus \{\sigma(k)\}$*, then if we identify both of these sets with* $[n-1]$ *by using* f_k^{-1} *and* $f_{\sigma(k)}$*, then* $\tilde{\sigma}_k \in S_{n-1}$ *is the resulting permutation on* $[n-1]$*. So if* σ *gives some term in* $\det A$ *that includes the entry* $a_{k,\sigma(k)}$*, then* $\tilde{\sigma}$ *gives the term in* $\det A^{(k,\sigma(k))}$ *that includes all the same entries except* $a_{k,\sigma(k)}$*.*

Then $\operatorname{sgn} \tilde{\sigma}_k = (-1)^{k+\sigma(k)} \operatorname{sgn} \sigma$*; that is,* σ *and* $\tilde{\sigma}_k$ *have the same sign if* $k + \sigma(k)$ *is even, and opposite signs if* $k + \sigma(k)$ *is odd.*

Proof. Since f_k^{-1} and $f_{\sigma(k)}$ are monotonically increasing, $\tilde{\sigma}_k$ inverts two integers $i, j \in [n-1] \setminus \{k\}$ if and only if σ inverts the corresponding integers $f_k^{-1}(i), f_k^{-1}(j)$. So σ and $\tilde{\sigma}_k$ have the same parity if and only if σ inverts k with an even number of elements of $[n] \setminus \{k\}$: these are the inversions of σ that don't have a counterpart inversion of $\tilde{\sigma}$. Partition $[n] \setminus \{k\}$ into four disjoint subsets:

1. $A = \{i \in [n] : i < k, \sigma(i) < \sigma(k)\}$
2. $B = \{i \in [n] : i < k, \sigma(i) > \sigma(k)\}$
3. $C = \{i \in [n] : i > k, \sigma(i) < \sigma(k)\}$
4. $D = \{i \in [n] : i > k, \sigma(i) > \sigma(k)\}$

Then σ inverts the set $\{k, i\}$ if and only if $i \in B$ or $i \in C$. Furthermore, $A \cup B = \{1, \dots, k-1\}$, an $A \cup C$ is the preimage of $\{1, \dots, \sigma(k)-1\}$ under σ. As σ is bijective, so $|A \cup C| = \sigma(k) - 1$. And as the sets A, B, C, D are all disjoint, the size of any union of two or more of them is equal to the sum of the constituent sets' sizes, so

$$\begin{aligned} |B \cup C| &= |A \cup B| + |A \cup C| - 2|A| \\ &= k + \sigma(k) - 2|A| - 2 \end{aligned}$$

so $|B \cup C|$ is even (and σ and $\tilde{\sigma}_k$ have the same parity) if and only if $k + \sigma(k)$ is even. □

Remark. The correspondence between σ and $\tilde{\sigma}_k$ is bijective: for any fixed integers $k, \ell \in [n]$ and any permutation $\tau \in S_{n-1}$, there is one and exactly one permutation $\sigma \in S_n$ such that $\sigma(k) = \ell$ and the reduced permutation $f_\ell \circ \sigma \circ f_k^{-1}$ equals τ.

7.9.2 Laplace expansion formula

Theorem (Laplace expansion formula)**.** *Let* A *be an* $n \times n$ *matrix, write* a_{ij} *for the entry in position* (i, j) *of* A*, and let* C_{ij} *denote the entries of the cofactor matrix of* A*. Then* $a_{r1}C_{r1} + \cdots + a_{rn}C_{rn} = a_{1c}C_{1c} + \cdots + a_{nc}C_{nc} = \det A$ *for all row and column indices* $1 \le r, c \le n$*.*

Proof. We'll first prove the result for expansion across a fixed row r: that is, $\det A = a_{r1}C_{r1} + \cdots + a_{rn}C_{rn}$. For each column number j, consider the terms in $\det A = \sum_{\sigma \in S_n} (-1)^\sigma a_{1\sigma(1)} \cdots a_{n\sigma(n)}$ that include a_{rj}; that is, the terms for permutations $\sigma \in S_n$ for which $\sigma(r) = j$. We'll prove that the sum of these terms is $a_{rj}C_{ij} = \operatorname{sgn}(\sigma) a_{rj} M_{ij}$, and the result follows from taking the sum over all values of j.

If the term of $\det A$ given by some permutation σ includes a_{rj}, then this term equals $a_{rj} \operatorname{sgn} \sigma$ times one entry per row and column of $A^{(rj)}$. This choice of entries is the same as those in the term in $\det A^{(rj)}$ given by the restricted permutation $\tilde{\sigma}_k$; recall that for

every permutation $\sigma \in S_n$ for which $\sigma(k)$ has some definite value, there is one distinct corresponding restricted permutation $\tilde{\sigma}_k$.

If $k + \sigma(k)$ is even, furthermore, then $\operatorname{sgn}\sigma = \operatorname{sgn}\tilde{\sigma}_k$, and every term with $\tilde{\sigma}_k$ in $\det A^{(rj)}$ has the same sign as the corresponding term with σ in $\det A$. In this case, the sum of terms in $\det A$ that include a_{rj} is $\operatorname{sgn}(\sigma)a_{rj}M_{ij} = \operatorname{sgn}(\sigma)a_{rj}C_{ij}$. If $k + \sigma(k)$ is odd, on the other hand, then corresponding terms in $\det A$ and $\det A^{(rj)}$ have opposite signs, so the sum of terms that include a_{rj} is $-\operatorname{sgn}(\sigma)a_{rj}M_{rj} = a_{rj}C_{rj}$.

The Laplace expansion formula for expansion across columns follows from the fact that $\det A^T = \det A$ (and transposing a matrix also transposes all the sub-matrices that determine its first minors), so expansion along a column of A is equivalent to expansion along a row of A^T.

□

Remark. Note that $a_{r1}C_{r1} + \cdots + a_{rn}C_{rn}$ is the dot product of the rth row of A and the rth row of the matrix of cofactors. Equivalently, it's the product of the rth row of A and the rth column of $\operatorname{adj} A$ (which, remember, is the transposed matrix of cofactors). We proved that this sum equals $\det A$, so the diagonal entries of $A \operatorname{adj} A$ all equal $\det A$. Similarly, $a_{1c}C_{1c} + \cdots + a_{nc}C_{nc} = \det A$ is the dot product of the cth column of the matrix of cofactors (equivalently, the cth row of $\operatorname{adj} A$) and the cth column of A, so the diagonal entris of $(\operatorname{adj} A)A$ are also $\det A$.

To illustrate Laplace expansion, let's rederive the formula for a 3×3 matrix determinant using Laplace expansion in two ways, first expanding across the first row and then across the second column. In the matrix

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

the first minors along the top row are:

$$\begin{aligned} M_{11} &= \begin{vmatrix} e & f \\ h & i \end{vmatrix} = ei - fh \\ M_{12} &= \begin{vmatrix} d & f \\ g & i \end{vmatrix} = di - fg \\ M_{13} &= \begin{vmatrix} d & e \\ g & h \end{vmatrix} = dh - eg \end{aligned}$$

The cofactors relate to the minors as $C_{11} = M_{11}$ and $C_{13} = M_{13}$ but $C_{12} = -M_{12}$, so we have an expression for the determinant:

$$\begin{aligned} \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} &= a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix} \\ &= a(ei - fh) - b(di - fg) + c(dh - eg) \\ &= aei + bfg + cdh - afh - bdi - ceg \end{aligned}$$

which matches the formula we computed directly from the sum-over-permutations definition of the determinant on page 181.

If we expand along the second column instead, then we have minors

$$M_{12} = \begin{vmatrix} d & f \\ g & i \end{vmatrix}$$
$$M_{22} = \begin{vmatrix} a & c \\ g & i \end{vmatrix}$$
$$M_{32} = \begin{vmatrix} a & c \\ d & f \end{vmatrix}$$

and the corresponding cofactors are $C_{12} = -M_{12}$, $C_{22} = M_{22}$, and $C_{32} = M_{32}$. This gives an expression for the determinant

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = -b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + e \begin{vmatrix} a & c \\ g & i \end{vmatrix} - h \begin{vmatrix} a & c \\ d & f \end{vmatrix}$$

and you can check that this formula also works.

Laplace expansion is inefficient for large matrices: most computer algebra programs use a method based on LU decomposition instead. But it is sometimes useful for computations by hand, especially when the matrix has a column or row of mostly zeros that is convenient to expand along.

7.10 Matrix inversion via adjugate matrix

We remarked in the last section that Laplace's formula shows that the diagonal entries of $A(\operatorname{adj} A)$ and $(\operatorname{adj} A)A$ are all $\det A$. The *off*-diagonal entries of both matrices, furthermore, are all zero. That is, $A(\operatorname{adj} A) = (\operatorname{adj} A)A = (\det A)I$, so $(\det A)^{-1} \operatorname{adj} A$ is the inverse matrix of A. Let's prove this.

Proposition. *The off-diagonal entries of $A \operatorname{adj} A$ and $(\operatorname{adj} A)A$ are all zero for any $n \times n$ matrix A.*

Proof. Let's compute entry (k, ℓ) (where $k \neq \ell$) of $A \operatorname{adj} A$; that is, $a_{k1}C_{\ell 1} + \cdots + a_{kn}C_{\ell n}$. Let A' be the matrix derived from A by replacing row ℓ with a copy of row k, and denote the entries and cofactors of A' by a'_{ij} and C'_{ij}.

The ℓth diagonal entry of $A' \operatorname{adj} A'$ is $a'_{\ell 1}C'_{\ell 1} + \cdots + a'_{\ell n}C'_{\ell n}$, which we proved equals $\det A'$ in the last section. And A' has two identical rows, so $\det A' = 0$. As $a'_{\ell j} = a_{kj}$ for all $1 \leq j \leq n$, so $0 = a_{k1}C'_{k1} + \cdots + a_{kn}C'_{kn} = 0$.

Furthermore, since A and A' equal each other outside row ℓ, the cofactors $C_{\ell j}$ and $C'_{\ell j}$, which are the determinants of a matrix formed by eliminating row ℓ, equal each other for every column j. So $0 = a_{k1}C'_{k1} + \cdots + a_{kn}C'_{kn} = a_{k1}C_{\ell 1} + \cdots + a_{kn}C_{kn}$, which is entry (k, ℓ) of $A \operatorname{adj} A$,.

To prove that the off-diagonal entries of $(\operatorname{adj} A)A$ are also zero, you can write a symmetrical proof that derives A' from A by replacing a column of A rather than a row, or simply note that once you know that $A \operatorname{adj} A$ is a multiple of the identity matrix (that is, $\operatorname{adj} A$ is a multiple of A^{-1}), it follows that $(\operatorname{adj} A)A$ must also be a multiple of the identity matrix. (Remember that for generic functions $f, g : X \to X$, $f \circ g$ is the identity if and only if $g \circ f$ is the identity as well.)

□

Corollary. *For any matrix A, we have $(\operatorname{adj} A)A = A \operatorname{adj} A = (\det A)I$ and so, if $\det A \neq 0$, then $A^{-1} = \frac{1}{\det A} \operatorname{adj} A$.*

7.11 Cramer's rule

Determinants can be used for another theoretically neat (though seldom practical) method of solving square systems of equations. Consider the generic 3×3 system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{32}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned}$$

or, in matrix form $A\mathbf{x} = \mathbf{b}$,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

The coefficients a_{ij} and b_i are known; the variables x_i are not.

If A is invertible, then $\mathbf{x} = A^{-1}\mathbf{b} = (\det A)^{-1}(\operatorname{adj} A)\mathbf{b}$; that is,

$$\begin{aligned} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} &= \frac{1}{\det A} \begin{bmatrix} M_{11} & -M_{21} & M_{31} \\ -M_{12} & M_{22} & -M_{32} \\ M_{13} & -M_{23} & M_{33} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \\ &= \frac{1}{\det A} \begin{bmatrix} M_{11}b_1 - M_{21}b_2 + M_{31}b_3 \\ -M_{12}b_1 + M_{22}b_2 - M_{32}b_3 \\ M_{13}b_1 - M_{23}b_2 + M_{33}b_3 \end{bmatrix} \end{aligned}$$

where M_{ij} is the first minor created by removing row i and column j from A. Every entry in this last vector is the Laplace expansion of the determinant of a matrix created by replacing one column of A with the column vector $\mathbf{b}$. To be precise,

$$M_{11}b_1 - M_{21}b_2 + M_{31}b_3 = \begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix}$$

by Laplace expansion along the first column. Similarly,

$$-M_{12}b_1 + M_{22}b_2 - M_{32}b_3 = \begin{vmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_3 & a_{33} \end{vmatrix}$$

by Laplace expansion along the second column, and

$$M_{13}b_1 - M_{23}b_2 + M_{33}b_3 = \begin{vmatrix} a_{11} & a_{12} & b_1 \\ a_{12} & a_{22} & b_2 \\ a_{13} & a_{32} & b_3 \end{vmatrix}$$

by Laplace expansion along the third column. Thus, the variables x_1, x_2, x_3 can each be written as a ratio of determinants, the denominator of each ratio being $\det A$ and the numerator being the determinant of the matrix derived from A by replacing the

coefficients for x_i with the equation values b_i:

$$x_1 = \frac{\begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}}$$

$$x_2 = \frac{\begin{vmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_3 & a_{33} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}}$$

$$x_3 = \frac{\begin{vmatrix} a_{11} & a_{12} & b_1 \\ a_{12} & a_{22} & b_2 \\ a_{13} & a_{32} & b_3 \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}}$$

This result generalizes to square systems with any number of variables.

An alternate proof: let X_k be the matrix created by replacing column k of the identity matrix with $\mathbf{x}$. Then $\det X_k = x_k$, because if we choose one entry from each row and column of X_k, then we have to choose the diagonal entry in every column other than k to avoid getting a zero, and this forces the choice of diagonal entry in k as well. Furthermore, AX_k is the matrix created by replacing column k in A with the vector of values $\mathbf{v}$ (proof: remember that $A\mathbf{x} = \mathbf{v}$ and consider how A acts column-by-column on X_k, whose columns except for k are standard basis vectors). So $\det X_k = \det(A^{-1}AX_k) = \det(AX_k)/\det A$, which is Cramer's rule.

Chapter 8

Generalized eigenspace decompositions

In previous chapters, we've talked about matrix representations of linear maps, determined by choosing bases for the domain and the codomain of the map. For linear operators $T : V \to V$, however, the theory of matrix representations is only interesting if we apply an important restriction: we have to use the same basis of V in the domain and in the codomain—that is, there must be some basis B such that whenever $\mathbf{x} \in \mathrm{Col}_n(\mathbb{F})$ represents some vector $\mathbf{v} \in V$ relative to B, $M\mathbf{x}$ also represents $T\mathbf{v}$ relative to B.

Now consider the relation on $\mathrm{Mat}_{n\times n}(\mathbb{F})$ defined as $J \sim M$ if there's some operator $T : V \to V$ (where V is an n-dimensional vector space over $\mathbb{F}$) that has both J and M as matrix representations relative to different bases. If this is true, we'll call these matrices *similar*. The main problem of this chapter is to characterize this similarity relation.

It turns out that similarity is an equivalence relation. Similarity is clearly reflexive and symmetric, and it turns out (we'll prove it later) that it is also transitive: if matrices A and B both represent one operator T_1 relative to two different bases, and matrices B and C both represent another operator T_2 relative to a potentially different pair of bases, then we can find a third operator that has both A and C as representations. The basic questions that we will ask are the following:

1. Can we find a natural set of representatives for the equivalence classes defined by the similarity relation? That is, is there some explicit formula for a set S of $n \times n$ matrices such that every $n \times n$ matrix is similar to exactly one element of S?

2. Is there an algorithm to find out *which* representative element is similar to any given matrix?

In this chapter, you'll need to think at two layers of abstraction at once: both concrete matrix manipulations and high-level considerations of vector space and linear operator axioms. Remember that these two considerations are closely related, and we'll point out frequently how findings about operators translated into matrices and vice versa.

8.1 Invariant subspaces and block diagonal matrices

Let's start by remembering a definition from section 6.1: an *invariant subspace* of an operator $T : V \to V$ is any vector subspace $W \subseteq V$ such that if $\mathbf{w} \in W$, then $T\mathbf{w} \in \mathbf{w}$ as well.

Now suppose that V is an n-dimensional vector space over $\mathbb{F}$, that $B \coloneqq \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis of V, and $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is an m-dimensional invariant subspace of T for some integer $1 \leq m < n$. (V and $\{\mathbf{0}\}$ are trivially invariant subspaces of any operator.) Let's write $A \in \operatorname{Mat}_{n\times n}(T)$ for the matrix representation of T with respect to B

What can we say about A? Remember that the kth row of A contains, in order, the coefficients $c_1, \ldots, c_n$ such that $T\mathbf{v}_k = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$. If $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is an invariant subspace, then the coefficients $c_{m+1}, \ldots, c_n$ are all zero for $1 \leq k \leq m$, so the first m columns of A are zero below the first m rows.

If, furthermore, $\operatorname{span}\{\mathbf{v}_{m+1}, \ldots, \mathbf{v}_n\}$ is also an invariant subspace, then columns $m+1$ through n must also have all of their nonzero entries in the rows $m+1$ through n. So A would have a form that called *block diagonal*, in which all the nonzero entries of A are contained in two square "blocks," one of size m and one of size $n-m$, lined up along the diagonal. For instance, if $m = 3$ and $n = 5$, then A must have the form

$$\begin{bmatrix} \star & \star & \star & 0 & 0 \\ \star & \star & \star & 0 & 0 \\ \star & \star & \star & 0 & 0 \\ 0 & 0 & 0 & \star & \star \\ 0 & 0 & 0 & \star & \star \end{bmatrix}$$

where $\star$ denotes a possibly nonzero entry. If T has a matrix representation of this form relative to some basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_5\}$, then $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ and $\operatorname{span}\{\mathbf{v}_4, \mathbf{v}_5\}$ are both invariant subspaces of T.

Knowing that a matrix A has block diagonal form—or, equivalently, that the domain of the corresponding operator T can be divided into a direct sum of invariant subspaces—simplifies many calculations that we can conduct on each block separately. For instance, if A has the block form $\begin{bmatrix} B & 0 \\ 0 & C \end{bmatrix}$, then powers of A have the block form $A^n = \begin{bmatrix} B^n & 0 \\ 0 & C^n \end{bmatrix}$. The computation required to multiply two $n \times n$ matrices scales up faster than the number of matrix entries (the straightforward algorithm that computes dot products of every row and column takes $O(n^3)$ time, and though more complicated algorithms can do better than this, there is no known $O(n^2)$ algorithm), so we can save a lot of time by operating on smaller submatrices.

Eigenspaces and generalized eigenspaces are also invariant spaces, and the blocks that such spaces contribute to matrix representations of operators have particularly simple forms:

- Suppose $W \coloneqq \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is an m-dimensional eigenspace with eigenvalue λ: that is, $T\mathbf{v}_i = \lambda\mathbf{v}_i$ for $1 \leq i \leq n$. Then the matrix representation of $T|_W$ with the basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ (and, therefore, the upper left corner of any matrix representation of T relative to a basis that has $\mathbf{v}_1, \ldots, \mathbf{v}_m$ as its first n vectors) is

$$\begin{bmatrix} \lambda & 0 & \cdots & 0 \\ 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda \end{bmatrix} = \lambda I,$$

  a diagonal matrix with all entries of λ on the diagonal.

- Suppose $W \coloneqq \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is a GES of order m, and the operator $T - \lambda$ sets up a chain $\mathbf{v}_m \mapsto \mathbf{v}_{m-1} \cdots \mapsto \mathbf{v}_1 \mapsto \mathbf{0}$: that is, $T\mathbf{v}_i = \lambda\mathbf{v}_i + \mathbf{v}_{i-1}$ for $1 \leq i \leq m-1$, and $T\mathbf{v}_1 = \lambda\mathbf{v}_1$. Then the matrix representation of $T|_W$ relative to $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is

$$\begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda. \end{bmatrix}$$

  This matrix has λ on the diagonal and 1 on the *superdiagonal*: the entries at positions $(i, i+1)$ for $1 \leq i \leq m-1$.

Detemrinants of block diagonal matrices have especially easy to compute, as a result of the following proposition:

Proposition. *The determinant of a block diagonal matrix is the product of the determinants of the individual blocks.*

Proof. It's enough to prove this for a matrix containing two blocks. (You can prove the general case for three or more blocks by combining all the blocks but one into a larger block and then applying the two-block case recursively.)

Suppose that an $n \times n$ matrix M has block diagonal form $\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$, where the blocks A and B are square and have sizes m and $n-m$. (From now on, whenever we mention a matrix in block diagonal form, it will go without saying that the diagonal blocks are square.) If either A or B has linearly dependent rows, then the corresponding rows in M must also be linearly dependent, so $\det M = \det A \det B = 0$. Otherwise, A has the $m \times m$ identity matrix as its RREF and B has the $(n-m) \times (n-m)$ as its RREF. Applying the steps of A's Gauss–Jordan reduction to the top m rows of M will reduce the top left block of M to the identity matrix without disturbing the block B or either of the rectangular zero blocks at top right and bottom left, and similarly applying B's Gauss–Jordan reduction to the bottom $n-m$ rows of M (adding m to row indices as necessary) reduces the bottom right block of M to the identity matrix.

Thus, the row reduction of M to the identity is the composition of the steps of the row reductions of A and B, and since the determinant of an invertible matrix is the product of the determinants of the elementary row-operation matrices corresponding to these steps (that is, λ for every scaling $\mathbf{r}_i \mapsto \lambda\mathbf{r}_i$, 1 for every shear, and -1 for every swap), so $\det M = \det A \det B$.

□

We can rephrase this result in operator language as:

Corollary. *Suppose $T : V \to V$ is an operator that has two invariant subspaces U and W, with $U \oplus W = V$. Then $\det T$ is the product of the determinants of the restricted maps $\det T|_U$ and $\det T|_W$, where U and W are treated as vector spaces in their own right.*

Proof. Write T in matrix form relative to a basis of V whose first several elements are a basis of U and whose remaining elements are a basis of W. The resulting matrix is block diagonal, with the upper left block giving a representation of $T|_U$ and the lower right block giving a representation of $T|_W$, and the determinants of these blocks are also the determinants of the operators themselves.

□

8.2 Translations between bases

If we have one matrix representation of an operator $T : V \to V$, it's possible that we could find a much simpler matrix representation of the same operator if we can find some eigenvectors and generalized eigenvectors of T, which would let us divide V at least partially into invariant subspaces. The question of how to find these eigenvectors will occupy us for the rest of this chapter.

Before we start looking at this question, though, let's think about a slightly more general problem: suppose that some operator $T : V \to V$ on an n-dimensional space V has matrix form M_1 relative to some basis $B_1 = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. We want to find the matrix representation M_2 of T relative to some other basis $B_2 = \{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$. Suppose also that we know the formulas for all the vectors in B_2 in terms of the vectors in B_1: that is, we know the coefficients c_{ij} for $1 \leq i, j \leq n$ such that $\mathbf{w}_i = c_{i1}\mathbf{v}_1 + \cdots + c_{in}\mathbf{v}_n$.

Now write $S_{12} : \mathrm{Col}_n(\mathbb{F}) \to \mathrm{Col}_n(\mathbb{F})$ for the function that takes the column vector representation of a vector relative to the basis B_1 and produces the column vector representation of the same vector relative to the basis B_2. That is, if $\mathbf{u} = a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n = b_1\mathbf{w}_1 + \cdots + b_n\mathbf{w}_n$, then $S_{12}\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$. Write S_{21} for the inverse function that translates B_2 representations to B_1 representations.

You should be able to convince yourself that S_{12} and S_{21} are linear maps: adding two vectors means adding their representations with respect to any basis, and ditto for multiplying one vector by a scalar. Every linear map from $\mathrm{Col}_n(\mathbb{F})$ to itself is just multiplication by an $n \times n$ matrix, so we can equate S_{12} and S_{21} with the matrices that represent them relative to the standard basis on $\mathrm{Col}_n(\mathbb{F})$.

So what are the matrices S_{12} and S_{21}? Let's first look at S_{21}. Remember that column j of any matrix M is $M\mathbf{e}_j$, where $\mathbf{e}_j$ is the jth standard basis column vector. Relative to the basis B_2, $\mathbf{e}_j$ represents $\mathbf{w}_j$, so the column vector $S_{21}\mathbf{e}_j$ (that is, the jth column of S_{21}) has to be the representation of $\mathbf{w}_j = c_{1j}\mathbf{v}_1 + \cdots + c_{nj}\mathbf{v}_n$ relative to the basis B_1; that is, the jth column of S_{21} is $\begin{bmatrix} c_{1j} \\ \vdots \\ c_{nj} \end{bmatrix}$. This means

$$S_{21} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix}.$$

That is, to translate from one basis to another, write a matrix whose columns are the coefficients of each vector of the *origin* basis relative to the *destination* basis. If $\mathbf{a} \in \mathrm{Col}_n(\mathbb{F})$ is a column vector, then we can interpret the matrix product $S_{21}\mathbf{a}$ in two ways:

1. $\mathbf{a}$ represents some element $\mathbf{u}$ of V relative to the basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, and $S_{21}\mathbf{a}$ represents, *relative to the same basis*, the *different* vector produced by giving $\mathbf{u}$ to the linear operator that sends every basis element $\mathbf{v}_i$ to the corresponding $\mathbf{w}_i$.

2. $\mathbf{a}$ represents some element $\mathbf{u}$ of V relative to the basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$, and $S_{21}\mathbf{a}$ represents, *relative to the different basis* $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, the *same* vector $\mathbf{u}$.

Now that we have S_{21}, we can compute the matrix $S_{12} = S_{21}^{-1}$ by matrix inversion.

So if M_1 represents T relative to B_1, then $M_2 := S_{12}M_1S_{21} = S_{21}^{-1}M_1S_{21}$ represents T relative to B_1. Remember that matrix products act right to left: first S_{12} translates a representation of an input vector $\mathbf{v}$ from B_2 to B_1, then M_1 gives the output $T\mathbf{v}$ relative to B_2, then S_1.

This result gives us another method of characterizing the similarity relation on matrices. Remember that A and B, by definition, are similar if they represent the same transformation T relative to different bases. We can equivalently define A and B to be similar if there is some invertible matrix S such that $SAS^{-1} = B$. (In this case, S translates from the basis used for A to the basis used for B; equivalently, its columns represent the basis vectors used for A in terms of the basis used for B.) From this, we can use the general fact that $(M_1M_2)^{-1} = M_2^{-1}M_1^{-11}$ to prove that the similarity relationship is transitive: if $A = S_1BS_1^{-1}$ and $B = S_2CS_2^{-1}$, then $A = S_1S_2CS_2^{-1}S_1^{-1} = (S_1S_2)C(S_1S_2)^{-1}$. That is, if A is similar to B via translation matrix S_1, and B is similar to C via S_2, then A is similar to C via S_1S_2.

Usually, when we use basis translations over $\mathbb{F}^n$, one basis will be the standard basis and the other basis will be clear from context, so we won't use subscripts on translation matrices. Typically, S denotes translation from an alternate basis to the standard basis (and its columns are the coefficients of the alternate basis with respect to the standard basis), and S^{-1} denotes translation out of the standard basis.

Two final notes:

1. If $M = SJS^{-1}$, then $\det M = \det S \det J \det S^{-1}$, and of course $\det S \det S^{-1} = 1$. So similar matrices have equal determinant.

2. Changes of basis and matrix multiplication are compatible operations. If $A = SBS^{-1}$ and $C = SDS^{-1}$, then $AC = SBS^{-1}SDS^{-1} = SBDS^{-1}$; that is, if A and B represent the same map relative to two different bases, and C and D represent another map relative to the same pair of bases, then AC and BD also represent the same map relative to the same pair of basis. This result is useful in computing, for example, matrix exponents: if M is a matrix and J is a diagonal (or almost-diagonal) matrix similar to J, and $M = SJS^{-1}$, then $M^n = SJ^nS^{-1}$, and finding J^n may require far less computation than finding M^n.

Now let's use our theory of basis translations to work out a few example problems in which we find the form for a linear operator $T : \mathbb{R}^2 \to \mathbb{R}^2$ based on its effects on a specific nonstandard basis of $\mathbb{R}^2$.

Example 1. Suppose that $T : \mathbb{R}^2 \to \mathbb{R}^2$ has two eigenvectors $\mathbf{v}_1 = (3, 4)$, with eigenvalue -1, and $\mathbf{v}_2 = (2, 1)$, with eigenvalue 3. What's the matrix representation for T relative to the standard basis?

Let's start by setting up the matrix factorization $M = SJS^{-1}$, where J is the representation of T relative to the basis $\{\mathbf{v}_1, \mathbf{v}_2\}$ and S translates from $\{\mathbf{v}_1, \mathbf{v}_2\}$ to the standard basis $\{\mathbf{e}_1, \mathbf{e}_2\}$. The columns of S are the representation of the origin basis vectors

[1]Proof: note that the map represented by M_1M_2 applies M_2 first and then M_1, so the reverse has to apply M_1 first and then M_2. Alternatively, note that $(M_1M_2)(M_2^{-1}M_1^{-1}) = M_1(M_2M_2^{-1})M_1^{-1} = M_1M_1^{-1} = I$ because matrix multiplication is associative. Remember that in general, $M_2^{-1}M_1^{-1} \neq M_1^{-1}M_2^{-1}$ because matrix multiplication is generally not commutative.

$\{\mathbf{v}_1, \mathbf{v}_2\}$ in terms of the destination basis vectors $\{\mathbf{e}_1, \mathbf{e}_2\}$, so

$$S = \begin{bmatrix} 3 & 2 \\ 4 & 1 \end{bmatrix}.$$

From the general 2×2 matrix inversion formula $\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$ we can compute

$$S^{-1} = -\frac{1}{5}\begin{bmatrix} 1 & -2 \\ -4 & 3 \end{bmatrix} = \begin{bmatrix} -\frac{1}{5} & \frac{2}{5} \\ \frac{4}{5} & -\frac{3}{5} \end{bmatrix}.$$

Finally, since $T\mathbf{v}_1 = -\mathbf{v}_1$ and $T\mathbf{v}_2 = 3\mathbf{v}_2$, the matrix J representing T relative to the basis $\{\mathbf{v}_1, \mathbf{v}_2\}$ is just

$$\begin{bmatrix} -1 & 0 \\ 0 & 3 \end{bmatrix},$$

so the representation of T relative to the standard basis is

$$M = SJS^{-1} = \begin{bmatrix} 3 & 2 \\ 4 & 1 \end{bmatrix}\begin{bmatrix} -1 & 0 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} -\frac{1}{5} & \frac{2}{5} \\ \frac{4}{5} & -\frac{3}{5} \end{bmatrix} = \begin{bmatrix} 3 & 2 \\ 4 & 1 \end{bmatrix}\begin{bmatrix} \frac{1}{5} & -\frac{2}{5} \\ \frac{12}{5} & -\frac{9}{5} \end{bmatrix} = \begin{bmatrix} \frac{27}{5} & -\frac{24}{5} \\ \frac{16}{5} & -\frac{17}{5} \end{bmatrix}.$$

You can check that $M\begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} -3 \\ -4 \end{bmatrix}$ and $M\begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 3 \end{bmatrix}$.

Example 2. Suppose that $T : \mathbb{R}^2 \to \mathbb{R}^2$ satisfies $T(3,4) = (2,1)$ and $T(2,1) = (3,4)$. What is the matrix representation for T relative to the standard basis?

In this case, relative to the basis vectors $\mathbf{v}_1 = (3,4)$ and $\mathbf{v}_2 = (2,1)$, T has the matrix representation

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

The basis translation matrices S and S^{-1} are the same as before, so the standard basis representation M is

$$M = SJS^{-1} = \begin{bmatrix} 3 & 2 \\ 4 & 1 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} -\frac{1}{5} & \frac{2}{5} \\ \frac{4}{5} & -\frac{3}{5} \end{bmatrix} = \begin{bmatrix} 3 & 2 \\ 4 & 1 \end{bmatrix}\begin{bmatrix} \frac{4}{5} & -\frac{3}{5} \\ -\frac{1}{5} & \frac{2}{5} \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 3 & -2 \end{bmatrix}.$$

Example 3. Suppose that $T(3,4) = (1,-1)$ and $T(2,1) = (-2,-5)$. What's the standard basis representation for T?

In this case, the provided values of T don't give us an easy matrix representation that uses $\{\mathbf{v}_1, \mathbf{v}_2\}$ as the basis for domain and codomain. But note that $\begin{bmatrix} 1 & -1 \\ -2 & -5 \end{bmatrix}$—that is, the matrix that sends $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ to $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ to $\begin{bmatrix} -2 \\ -5 \end{bmatrix}$—represents T with *input* vectors represented relative to $\{\mathbf{v}_1, \mathbf{v}_2\}$ and *output* vectors represented relative to the standard basis. That is, with our matrix decomposition $M = SJS^{-1}$, we have $\begin{bmatrix} 1 & -1 \\ -2 & -5 \end{bmatrix} = SJ$ and so

$$M = (SJ)S^{-1} = \begin{bmatrix} 1 & -1 \\ -2 & -5 \end{bmatrix}\begin{bmatrix} -\frac{1}{5} & \frac{2}{5} \\ \frac{4}{5} & -\frac{3}{5} \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ -\frac{18}{5} & -\frac{11}{5} \end{bmatrix}$$

8.3 Elements of the theory of polynomials

Now that we've seen the general theory of translating between different matrix representations of a linear operator, our goal will be finding a way to find a representation as close to a diagonal matrix as possible. Since elements in a diagonal matrix are eigenvalues, we'll also need a way to compute the eigenvalues of a matrix. It turns out that we can compute from any square matrix M a special polynomial called the *characteristic* polynomial whose roots are the eigenvalues of M.

To appreciate the characteristic polynomial fully, we'll need a small detour into the theory of polynomials. We'll need two principal results:

1. If you can prove that a set of polynomials satisfy three simple axioms that make it a special kind of set called an *ideal*, then this set must contain every multiple of a special element called the *minimal polynomial*, and nothing else.

2. Every polynomial with complex coefficients has a complex root. (This is the *fundamental theorem of algebra*, which you likely saw mentioned in high school mathematics but may not have seen proved.)

8.3.1 Polynomial ideals

First, some definitions. A *polynomial* in one variable x with degree n is an expression of the form $c_n x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0$, where the coefficients $c_0, \ldots, c_n$ are in some specified field $\mathbb{F}$ and $c_n \neq 0$. The coefficient c_n attached to the highest power of the variable x is called the *leading coefficient*, and if this equals 1, then the entire polynomial is *monic*. Note that nonzero constant polynomials have degree zero; by convention, the constant zero polynomial has degree $-\infty$.

Denote by $\mathbb{F}[x]$ the set of polynomials of any degree with coefficients in $\mathbb{F}$. We can add and multiply polynomials in the way that you're used to: for instance, $(x+1) + (2x^2 - x) = 2x^2 + 1$ and $(x+1)(2x^2 - x) = 2x^3 + x^2 - x$. We can also scale polynomials by a constant factor $c \in \mathbb{F}$ (which is equivalent to multiplying by the zero-degree constant polynomial $p(x) = c$) and subtract polynomials from each other (which is equivalent to multiplying one of the polynomials by -1 and then adding them).

Finally, an *ideal* of $\mathbb{F}[x]$ is a subset $I \subseteq \mathbb{F}[x]$ that satisfies these three axioms:

1. *Non-emptiness*: I has at least one element. (As a consequence of axiom 3, this means I must contain the zero polynomial.)

2. *Closure under addition by other ideal elements*: If $p(x)$ and $q(x)$ are two (possibly identical) elements of I, then their sum $p(x) + q(x)$ is also in I.

3. *Closure under multiplication by arbitrary elements*: If $p(x) \in I$, then $p(x)q(x) \in I$ for any other polynomial $q(x) \in \mathbb{F}[x]$ (even if $q(x) \notin I$). Note that this includes the case when q is constant or zero.

This axiomatic definition may seem complicated, but any set of polynomials that we can prove satisfies these three axioms turns out to have a simple structure:

Theorem. *Every ideal $I \subseteq \mathbb{F}[x]$ is the set of multiples of some element $p(x) \in \mathbb{F}[x]$, called a* primitive element *of I. That is, $I = \{p(x)q(x) : q(x) \in \mathbb{F}[x]\}$.*

Proof. If I contains any polynomial p, then it must also contain the monic polynomial that comes from dividing all coefficients of p by the leading coefficient. Now let's consider two cases: either I contains a nonzero constant polynomial, or it doesn't.

If I contains a nonzero constant polynomial $p(x) = c$, then (by axiom 3) it also contains $c^{-1}p(x)$, which is the constant polynomial with value 1. Also by axiom 3, it must contain all multiples of 1 by any element of $\mathbb{F}[x]$. This set equals $\mathbb{F}[x]$, and I can't be any larger than $\mathbb{F}[x]$, so $I = \mathbb{F}[x]$ and also equals the set of multiples of the constant polynomial 1.

The other case is when the smallest degree of any nonzero polynomial in I is $k \geq 1$. We'll proceed in three steps, the last of which establishes the claim that every polynomial $p(x) \in I$ has to be a multiple of some specific degree-k polynomial by induction on the degree of p:

1. *I contains exactly one monic polynomial of degree k.* Proof: if I has at least one polynomial $p(x) = c_k x^k + \cdots + c_1 x + c_0$ of degree k, then dividing that polynomial by its leading coefficient (i.e. multiplying it by the constant polynomial $q(x) = c_k^{-1}$) creates a monic polynomial that must also be in I by axiom 3. If there were two or more monic polynomials of degree k in I, though, then subtracting one polynomial from the other would cancel the x^k terms and leave a nonzero polynomial with degree strictly less than k. But this polynomial would also be in I by axiom 2—a contradiction, as we assumed k was the smallest degree of any nonzero polynomial in I.

2. *Every degree-k polynomial in I is a scalar multiple of every other.* Proof: if any two degree-k polynomials in I were not scalar multiples of each other, then dividing both polynomials by their leading coefficients would give distinct monic degree-k polynomials, but there can only be one monic polynomial of degree k in I.

3. *Let $m(x)$ be the monic polynomial of minimal degree in an ideal I, and let k be the degree of m. Suppose that for some integer $n \geq k$, every polynomial in I of degree at most n is a scalar multiple of $m(x)$. Then so is every polynomial in I of degree $n+1$.* Proof: let $p(x)$ be a degree-$n+1$ element of I, and let c be its leading coefficient. Then $p(x)$ and $cx^{n+1-k}m(x)$ both have leading terms cx^{n+1}, so $q(x) \coloneqq p(x) - cx^{n+1}m(x)$ has degree at most n. So by the induction hypothesis, $q(x)$ is a multiple of $m(x)$. Thus, $p(x) = q(x) + cx^{n+1-k}m(x)$ is the sum of two multiples of $m(x)$, so it has to be a multiple of $m(x)$ itself.

□

In the vocabulary of abstract algebra, this is a proof that $\mathbb{F}[x]$ is a *principal ideal domain*.

8.3.2 Algebraically complete fields

An *algebraically complete field* is one in which every polynomial has a root: if p is a nonconstant polynomial with coefficients in an algebraically complete field $\mathbb{F}$, then there's guaranteed to be some $x \in \mathbb{F}$ such that $p(x) = 0$. As a corollary, any such polynomial can be factored completely into the form $(x - c_1)(x - c_2)\cdots(x - c_n)$, where $c_1, \ldots, c_n$ are constants in $\mathbb{F}$.

The most important algebraically complete field is $\mathbb{C}$, the complex numbers. The result that $\mathbb{C}$ is algebraically complete is often called the *fundamental theorem of algebra*. In high school algebra, you likely saw it stated, but not quite proved. Here is the formal statement:

Theorem. *If $p(x) \in \mathbb{C}[x]$ is a nonconstant polynomial, then there's at least one complex value $z \in \mathbb{C}$ such that $p(z) = 0$.*

Proof. A completely rigorous proof would require introducing formal topology, but the fundamental geometric idea is quite intuitive.. First, remember the standard way to visualize the complex plane $\mathbb{C}$, with the x-axis representing the real component of a number and the y-axis representing the imaginary component.

Now suppose that p is a monic polynomial. (The theorem for general p immediately reduces to the case for monic p if you divide p by its leading coefficient, because factoring a constant out of a polynomial doesn't change the roots.) Write $p(z) = z^n + c_{n-1}z^{n-1} + \cdots + c_1 z + c_0$. If $c_0 = 0$, then $p(0) = 0$, and we're done. So let's consider only the case where $c_0 \neq 0$.

Now imagine tracing the value of $z = re^{i\theta}$ in the complex plane as $r \leq 0$ stays fixed and θ increases from 0 to 2π. What does the graph of $f(z)$ looks like for these values of z? It's hard to say in general, but if r is very large or very small, we can draw a couple of conclusions:

1. Suppose r and thus z are extremely small, so that $c_n z^n + c_{n-1}z^{n-1} + \cdots + c_1 z$ is much smaller than c_0. Then the image of $f(re^{i\theta})$ for varying θ will be a tiny squiggle around c_0 that doesn't go anywhere near the origin.

2. Suppose r and thus z are extremely large, so that $c_n z^n$ has a much greater absolute value than all of the terms $c_{n-1}z^{n-1} + \cdots + c_1 z + c_0$. Remember that the map $z \mapsto cz^n$ maps $re^{i\theta}$ to $cr^n e^{in\theta}$, so as θ increases from 0 to 2π (that is, z goes around the origin once), z^n goes around the origin n times. So if $c_n z^n$ is much larger than all of the other terms in $p(z)$, then as z goes around the origin once, $p(z)$ will trace a very large, nearly circular loop that wraps n times around the origin. The non-leading terms in $p(z)$ will mean that this loop won't be a perfect circle, but just by making r larger, we can make these deviations can be made as small a fraction of the distance separating the loop from the origin as we want, and the graph of $p(re^{i\theta})$ can be made to look as close to a circle when zoomed out as we want.

So as r goes from very small to very large, the graph of $p(re^{i\theta})$ has to go from a small squiggle around c_0 that has the origin outside to a looping path that has the origin inside. It follows (more or less—defining "inside" and "outside" turns out to be harder than you might think!) that at some intermediate value r, the graph of $p(re^{i\theta})$ has to pass through the origin; that is, there is some value $z_0 = re^{i\theta}$ such that $p(z_0) = 0$. □

Corollary. *Every degree-n polynomial $p \in \mathbb{C}[x]$ can be written as $p(z) = c(z - r_1)(z - r_2) \cdots (z - r_n)$ where $r_1, \ldots, r_n \in \mathbb{C}$ are (possibly not all distinct) zeros of n.*

Proof. Obvious if $n = 1$. Otherwise, we know that p has at least one root r_1, and $p(z)$ must be a multiple of $z - r_1$. Apply the theorem again to $p(z)/(z - r_1)$, which is a degree-$n - 1$ polynomial and must have another root r_2, and so on.

□

Polynomials that split entirely into a product of terms of the form $x - c_i$ are much easier to deal with than general polynomials, which is why many results about polynomials with real coefficients are most easily proved by treating them as polynomials with complex coefficients instead.

8.4 Characteristic polynomials

8.4.1 Defined

Let's now recall a few definitions and theorems from previous sections. Throughout, $T : V \to V$ is a linear operator on a vector space over a field $\mathbb{F}$.

1. A scalar $\lambda \in \mathbb{F}$ is an *eigenvalue* of T if the operator $T - \lambda I$, or equivalently its negative $\lambda I - T$, is not injective. The corresponding *eigenvectors* of T are all the elements in the kernel of $\lambda I - T$.

2. If V is a finite-dimensional space, then we can define the *determinant* of any matrix representation of T. Every matrix representation has the same determinant (recall $\det SJS^{-1} = \det J$), so we can define $\det T$ to be the determinant of any matrix that represents T.

3. The determinant of any operator T is zero if and only if T has nonzero kernel. By the same token, $\det(xI - T) = 0$ if and only if x is an eigenvalue of T.

The good news is that $\det(xI - T)$ turns out to be a monic degree-n polynomial that we call the *characteristic polynomial* of T and denote $\chi_T(x)$. We've shown a result that's important enough to call out as a proposition:

Definition. *The **characteristic polynomial** of an operator T or matrix M is the expression $\det(xI - T)$ or $\det(xI - M)$ as a function of a scalar variable x (where I is the identity operator or matrix, respectively).*

Proposition. *λ is an eigenvalue of T if and only if it is a root of the characteristic polynomial χ_T: that is, if $\chi_T(\lambda) = 0$.*

8.4.2 Computing characteristic polynomials

To compute the determinant of an operator $xI - T$ for some arbitrary scalar x, we must find a matrix representation of this operator and calculate its determinant. It turns out, intuitively enough, that if M represents T relative to some basis, then the matrix $xI - M$ represents the operator $xI - T$ relative to the same basis. And since determinants of matrices that represent the same operator are equal, it doesn't matter which matrix M we choose: if J and M are two representations of the same operator with $M = SJS^{-1}$, then $xI - M = S(xI - J)S^{-1}$ as well (because $SIS^{-1} = I$), and so $\det(xI - M) = \det S \det(xI - J) \det S^{-1} = \det(xI - J)$.

Let's look at an example. Consider the operator $T : \mathbb{R}^2 \to \mathbb{R}^2$ with the formula $T(a, b) = (a + 2b, 5a - 2b)$. With respect to the standard basis, T has the matrix representation

$$M = \begin{bmatrix} 1 & 2 \\ 5 & -2 \end{bmatrix}$$

so $xI - T$ has the matrix representation

$$x\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} - \begin{bmatrix}1 & 3\\ 4 & -2\end{bmatrix} = \begin{bmatrix}x-1 & -2\\ -5 & x+2\end{bmatrix}.$$

The determinant of this matrix is the characteristic polynomial $\chi_T(x)$. From the generic 2×2 determinant formula $\begin{vmatrix}a & b\\ c & d\end{vmatrix} = ad - bc$, therefore,

$$\chi_T(x) = \begin{vmatrix}x-1 & -2\\ -5 & x+2\end{vmatrix} = (x-1)(x+2) - 10 = x^2 + x - 12$$

We chose T's matrix representation with respect to the standard basis to choose χ_T, but we could have chosen a matrix representation with respect to any basis.

From this example, furthermore, it should be clear why the characteristic polynomial is a monic polynomial of degree n: every term in the sum over permutations for $\det(xI - M)$ selects n terms from the matrix $xI - M$, each of which is either a constant or a degree-1 polynomial; and the only term that includes n degree-1 polynomials is the term that selects all the diagonal entries.

8.4.3 The characteristic polynomial as an aid to matrix diagonalization

The characteristic polynomial is an essential tool for finding diagonal matrices similar to a given matrix because it reduces the problem of computing matrix eigenvalues to computing the roots of a polynomial. Consider the same example $T(a, b) = (a+2b, 5a-2b)$ from the previous section, with characteristic polynomial $\chi_T(x) = x^2 + x - 12$.

We can factor this polynomial as $(x+4)(x-3)$, so its roots are -4 and 3; these must be the eigenvalues of T.

Now let's try to find the corresponding eigenvectors. We can use the general RREF-based method for finding matrix nullspaces that we outlined on page 92: substitute the necessary value for x in the matrix representation of $xI - T$, then reduce to RREF, insert and remove rows of zero to line up all pivots on the diagonal, and take all non-pivot columns, with diagonal entries changed from 0 to -1, as a basis for the nullspace.

For $x = 3$, for instance, we need to find the nullspace of

$$\begin{bmatrix}2 & -2\\ -5 & 5\end{bmatrix},$$

which the elementary row operations $\mathbf{r}_1 \mapsto \frac{1}{2}\mathbf{r}_1$ followed by $\mathbf{r}_2 \mapsto \mathbf{r}_2 + 5\mathbf{r}_1$ bring to the RREF

$$\begin{bmatrix}1 & -1\\ 0 & 0\end{bmatrix}.$$

Every pivot in this matrix (that is to say, the sole pivot in the first row) is already on the diagonal, so the pivotless second column gives a basis column vector $\begin{bmatrix}-1\\ -1\end{bmatrix} \in \mathrm{Col}_2(\mathbb{R})$ of the eigenspace of matrix M with eigenvalue 3. As M represents the original operator T relative to the standard basis, this column vector corresponds to an eigenvector $(-1, -1) \in \mathbb{R}^2$ for T (though $(1, 1)$ is a nicer choice).

Likewise, the eigenvectors of T with eigenvalue -4 are represented by the nullspace of

$$\begin{bmatrix} -5 & -2 \\ -5 & -2 \end{bmatrix},$$

which has RREF

$$\begin{bmatrix} 1 & \frac{2}{5} \\ 0 & 0 \end{bmatrix}$$

and thus has $\begin{bmatrix} \frac{2}{5} \\ -1 \end{bmatrix}$ as a basis column vector for its nullspace, corresponding to $(\frac{2}{5}, -1) \in \mathbb{R}^2$ (though, again, $(2, -5)$ is just as valid a choice). You can confirm for yourself that $T(1,1) = (3,3)$ and $T(2,-5) = (-8,20)$.

Note that we've found a basis $\mathbf{v}_1 = (1,1), \mathbf{v}_2 = (2,-5)$ for $\mathbb{R}^2$ that contains only eigenvectors of T. With respect to this basis, T has the diagonal matrix representation $\begin{bmatrix} 3 & 0 \\ 0 & -4 \end{bmatrix}$, and we can even write a full $M = SJS^{-1}$ matrix diagonalization

$$\begin{bmatrix} 1 & 3 \\ 4 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & -5 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & -4 \end{bmatrix} \begin{bmatrix} -\frac{5}{7} & -\frac{2}{7} \\ -\frac{1}{7} & \frac{1}{7} \end{bmatrix}$$

where on the right-hand side, the rightmost matrix translates column vector representations relative to the standard basis to the column vector representations of the same element of $\mathbb{R}^2$ relative to $\{\mathbf{v}_1, \mathbf{v}_2\}$, the middle matrix expresses T on column vector representations relative to $\{\mathbf{v}_1, \mathbf{v}_2\}$, and the left matrix translates back to column vectors relative to the standard basis.

One final point that will be worth remembering later. Remember that if T has a matrix representation with all real numbers, then:

1. $xI - T$ has a matrix representation with all real numbers whenever x is real.
2. We can extract eigenvectors with eigenvalue x from the RREF of this matrix with the usual nullspace calculation algorithm discussed in Section 4.6.
3. Gauss–Jordan elimination doesn't introduce complex numbers into a real matrix.

The following proposition is an immediate consequence:

Proposition. *If a matrix M with all real entries has a real eigenvalue with multiplicity k (meaning $kI - M$ has nullspace dimension n), then we can find k linearly independent eigenvectors with all real entries.*

Proof. See above.

□

8.4.4 Generalized eigenspace dimensions

We can make an even stronger statement connecting characteristic polynomials to eigenvectors than that the roots of the characteristic polynomial are also eigenvalues of the operator. In fact, the *multiplicity* of λ as a root of χ_T equals the dimension of the largest generalized eigenspace of T with eigenvalue λ. That is, if $(x-\lambda)^2$ (but not $(x-\lambda)^3$) is a factor of χ_T, then the largest GES with eigenvalue λ has dimension 2; if $(x-\lambda)^3$ but not $(x-\lambda)^4$ is a factor, then GES has dimension 3; and so on.

We'll need a few preliminary propositions to prove this:

Proposition. *Suppose $T : V \to V$ is an operator on a finite-dimensional vector space V, and suppose further that V can be decomposed as a direct sum $V = U \oplus W$ where U and W are invariant subspaces of T. Write χ_V for the characteristic polynomial of T on all of V, and χ_U and χ_W for the characteristic polynomials of the restricted maps of T on U and W. Then $\chi_V(x) = \chi_U(x)\chi_W(x)$.*

Proof. Remember from page 136 that if U and W are invariant spaces of T, then they're also invariant spaces of any polynomial of T, such as $T - x$ for any scalar x. So the result comes just from noting that $\chi_V(x) = \det(T - x)$ (and similarly for the restricted maps) and applying the result on page 197 to get $\det(T-x) = \det(T|_U - x)\det(T|_W - x)$. □

Translated into matrix language:

Corollary. *The characteristic polynomial of a block diagonal matrix is the product of the polynomials of the blocks.*

Proof. Each block of a block diagonal matrix represents a restricted map on an invariant subspace.

□

Proposition. *If $T : V \to V$ is an operator on a vector space such that $\ker T = \ker T^2$, then the sum $\ker T + \operatorname{im} T$ is direct. If V is finite-dimensional, furthermore, then $\ker T \oplus \operatorname{im} T = V$.*

Proof. If the sum $\ker T + \operatorname{im} T$ isn't direct, then there's some vector $\mathbf{w}$ that's in both $\ker T$ and $\operatorname{im} T$. Choose any $\mathbf{v} \in V$ such that $T\mathbf{w} = \mathbf{v}$. Then $T\mathbf{v} = \mathbf{w}$ and $T^2\mathbf{v} = T\mathbf{w} = \mathbf{0}$, so $\mathbf{v}$ is an element of $\ker T^2$ that's not in $\ker T$, a contradiction. So $\ker T + \operatorname{im} T$ is a direct sum.

By rank–nullity, $\dim \ker T + \dim \operatorname{im} T = \dim V$, and $\dim(\ker T \oplus \operatorname{im} T) = \dim \ker T + \dim \operatorname{im} V$ because the dimension of a direct sum is the sum of the dimensions of its constituents. So $\ker T \oplus \operatorname{im} T$ is a $(\dim V)$-dimensional subspace of T; i.e. it equals V. □

Proposition. *Let V be any vector space, let T be any operator on V, let λ be any scalar, and let k be any positive integer. Then $\ker(T - \lambda)^k$ and $\operatorname{im}(T - \lambda)^k$ are both invariant subspaces of T.*

Proof. $\ker(T - \lambda)^k$ is the set of GEVs of T with eigenvalue λ and order $\leq k$, so if $\mathbf{v} \in \ker(T - \lambda)^k$, then $T\mathbf{v} = \lambda\mathbf{v} + \mathbf{w}$, where $\mathbf{w}$ is also a GEV with eigenvalue λ and order $\leq k - 1$. So $\lambda\mathbf{v} + \mathbf{w}$ is the sum of two GEVs with eigenvalue λ and order $\leq k$, so it also has order $\leq k$. So we've established that $\ker(T - \lambda)^k$ is an invariant subspace of T.

To see that $\operatorname{im}(T - \lambda)^k$ is invariant, take $\mathbf{v} \in \operatorname{im}(T - \lambda)^k$ arbitrary and choose $\mathbf{u}$ such that $(T - \lambda)^k\mathbf{u} = \mathbf{v}$. Then $T\mathbf{v} = T(T - \lambda)^k\mathbf{u}$. Since T and $T - \lambda$ commute, we can can rewrite $T\mathbf{v}$ as $(T - \lambda)^k(T\mathbf{u})$, which is clearly an element of $\operatorname{im}(T - \lambda)^k$. This establishes that $\operatorname{im}(T - \lambda)^k$ is an invariant subspace of T.

□

Proposition. *Suppose V is an n-dimensional vector space, and T is an operator on V such that every element of V is a GEV with the same eigenvalue λ. Then $\chi_T(x) = (x - \lambda)^n$.*

Proof. Let B be a Jordan basis for V. (See section 6.8 if you need a reminder of what a Jordan basis is, and a proof of their existence.) Number the elements of B as $\mathbf{v}_1, \ldots, \mathbf{v}_n$ so that the chains created by application of $T-\lambda$ run over a contiguous range of indices from highest to lowest: that is, $(T-\lambda)\mathbf{v}_1 = \mathbf{0}$ and, for every integer $2 \leq i \leq n$, either $(T-\lambda)\mathbf{v}_i = \mathbf{v}_{i-1}$ or $(T-\lambda)\mathbf{v}_i = \mathbf{0}$).

Relative to this basis, T has the matrix representation

$$\begin{bmatrix} \lambda & c_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & c_2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \lambda & c_3 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \lambda & c_{n-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix}$$

where $c_i = 0$ if $(T-\lambda)\mathbf{v}_{i+1} = \mathbf{0}$, and $c_i = 1$ if $(T-\lambda)\mathbf{v}_{i+1} = \mathbf{v}_i$. Regardless of the values of $c_1, \ldots, c_{n-1}$, the characteristic polynomial

$$\chi_T(x) = \det \begin{bmatrix} x-\lambda & c_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & x-\lambda & c_2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & x-\lambda & c_3 & \cdots & 0 & 0 \\ 0 & 0 & 0 & x-\lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & x-\lambda & c_{n-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & x-\lambda \end{bmatrix}$$

is the determinant of an upper triangular $n \times n$ matrix in which every diagonal element is $x-\lambda$, so it equals $(x-\lambda)^n$.

□

With all these results in hand, our main theorem is quick to establish:

Theorem. *Let V be a finite-dimensional vector space, $T : V \to V$ a linear operator, and λ any scalar. Let U be the subspace of V consisting of every GEV with eigenvalue λ. Then $\dim U$ is the largest integer k such that $(x-\lambda)^k$ divides the characteristic polynomial $\chi(x)$ of T.*

Proof. Remember that $\ker(T-\lambda)^h$ is the set of GEVs with eigenvalue λ and order $\leq h$. So if we let h be any integer larger than the maximum order of any GEVs with eigenvalue λ (which must be finite: remember from page 146 that an n-dimensional space can't have GEVs of order greater than n), then $\ker(T-\lambda)^h = \ker(T-\lambda)^{2h}$ and (by the previous proposition) $\ker(T-\lambda)^h \oplus \mathrm{im}(T-\lambda)^h = 0$.

Denote $U = \ker(T-\lambda)^h$ (this subspace is every GEV with eigenvalue λ; i.e. the same space as the U in the theorem statement) and $W = \mathrm{im}(T-\lambda)^h$. Both U and W, as we've just proved, are invariant subspaces of T. So $\chi(x) = \chi_U(x)\chi_W(x)$ where χ_U, χ_W are the characteristic polynomials of the restricted operators $T|_U, T|_W$.

Every element of U is a GEV of $T|_U$ with eigenvalue λ, so $\chi_U(x) = (x-\lambda)^k$ where $k = \dim U$. Furthermore, W can't contain any nonzero GEVs with eigenvalue λ, because we defined U to contain all of them and $U \cap W = \{\mathbf{0}\}$. In particular, W can't contain any nonzero ordinary eigenvectors with eigenvalue λ, so λ is not a root of χ_W. So the exponent of $x-\lambda$ in $\chi(x)$, when fully factored, equals the exponent of $x-\lambda$ in χ_U, i.e. k.

□

Note that all the results in this section hold even for vector spaces over an algebraically incomplete field: we didn't assume anywhere that characteristic polynomials have to factor completely into monomials $(x - \lambda_1) \cdots (x - \lambda_n)$.

8.5 The trace

The *trace* of a square matrix A, commonly denoted $\operatorname{tr} A$, is the sum of its diagonal elements; for example,

$$\operatorname{tr} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = 1 + 4 + 9 = 14.$$

8.6 Matrix triangularization

If V is a finite-dimensional vector space over the complex numbers $\mathbb{C}$ (from now on we'll say "complex vector space" rather than "vector space over the complex numbers"), then every linear operator $T : V \to V$ has a characteristic polynomial with complex coefficients whose roots are the eigenvalues of T. The Fundamental Theorem of Algebra states that every polynomial with complex coefficients has a complex root. We've therefore proved the following result:

Lemma. *Every operator T over a finite-dimensional complex vector space has an eigenvalue.*

Proof. Just given.

Alternate proof without characteristic polynomials: if $\mathbf{v}$ is some arbitrary nonzero vector in an n-dimensional space, then the $n + 1$ vectors $\mathbf{v}, T\mathbf{v}, \ldots, T^n\mathbf{v}$ are linearly dependent: that is, there are some constants $c_0, \ldots, c_{n-1}$ such that $T^n\mathbf{v} + c_{n-1}T^{n-1}\mathbf{v} + \cdots + c_1T\mathbf{v} + c_0\mathbf{v} = \mathbf{0}$. Write $p(x)$ for the polynomial $x^n + c_{n-1}x^{n-1} + \cdots + c_1x + c_0$ and factor it as $(x-h_1)(x-h_2)\cdots(x-h_n)$ (we can do this factoring because $\mathbb{C}$ is algebraically complete). Then $p(T)\mathbf{v} = (T - \lambda_1)\cdots(T - h_n)\mathbf{v} = \mathbf{0}$.

If you apply the maps $T - h, T - hn - 1, \ldots, T - h$ one at a time to $\mathbf{v}$, there must be some point at which the result becomes $\mathbf{0}$: that is, $(T - hi + 1)\cdots(T - h_n)\mathbf{v} \neq \mathbf{0}$, but $(T - h_i)\cdots(T - h_n)\mathbf{v} = \mathbf{0}$ This means that $(T - h_{i+1})\cdots(T - h)\mathbf{v}$ is an eigenvector of T, with eigenvalue h_i. (Or, possibly, $(T - \lambda_n)\mathbf{v} = \mathbf{0}$, in which case h_n is an eigenvalue.) □

Remark. It's essential that V be finite-dimensional. For an example of an operator on an infinite-dimensional space that has no eigenvalues, take $\mathbb{C}^{\mathbb{N}}$, the space of infinite sequences of complex numbers, and the right-shift operator $R(a_1, a_2, a_3, \ldots) = (0, a_1, a_2, a_3, \ldots)$. For any sequence $\mathbf{v} \in \mathbb{C}^{\mathbb{N}}$ with at least one nonzero entry, $R\mathbf{v}$ also has a nonzero entry (so $\mathbf{v}$ can't have eigenvalue 0) but always starts with one more entry of zero than $\mathbf{v}$ does (so $R\mathbf{v}$ can't be a nonzero multiple of $\mathbf{v}$).

This is an important building block for another result: the fact that every linear operator over a finite-dimensional complex vector space has a triangular representation.

Theorem. *Let V be an n-dimensional complex vector space, and let $T : V \to V$ be a linear operator. There exists a set of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ such that $T\mathbf{v}_k \in \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ for every integer $1 \leq k \leq n$.*

Proof. Let V be an n-dimensional complex vector space, let $\mathbf{v}_1$ be an eigenvector of $T : V \to V$, and let $W_1 \subset V$ be a complement of $\text{span}\{\mathbf{v}_1\}$, so every vector $\mathbf{v} \in V$ can be written in one unique way as $\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{w}$ for some $\mathbf{w} \in W_1$. (Remember: we can always extend a linearly independent set, such as $\{\mathbf{v}_1\}$, to a full basis of V, and then take the vectors that we added as a basis of W.) Let P_1 be the projection map $P(c_1\mathbf{v}_1 + c_2\mathbf{w}) = c_2\mathbf{w}$ from V onto W_1.

The map $P_1 \circ T|_{W_1}$, the composition of T and P_1 with the domain restricted to W_1, maps the $(n-1)$-dimensional complex vector space W_1 to itself. This map must have an eigenvector $\mathbf{v}_2 \in W_1$, and if $\mathbf{v}_2$ is an eigenvector of $P_1 \circ T$, then $T\mathbf{v}_2 \in \text{span}\{\mathbf{v}_1, \mathbf{v}_2\}$.

We can continue in the same way. Pick some subspace $W_2 \subset V$ that is the complement of $\text{span}\{\mathbf{v}_1, \mathbf{v}_2\}$; that is, such that every $\mathbf{v} \in V$ can be written uniquely as $\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{w}$ for some $\mathbf{w} \in W_2$. Let P_2 be the projection map $P_2(c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{w}) = c_1\mathbf{v}_1 + c_2\mathbf{v}_2$ and choose some eigenvector $\mathbf{v}_3 \in W_2$ of $P_2 \circ T|_{W_2}$.

□

Corollary. *Every linear transformation $T : V \to V$ has an upper triangular and a lower triangular matrix representation.*

Proof. The basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ defined in the previous theorem gives an upper triangular matrix representation for T, and the reversed basis $\{\mathbf{v}_n, \ldots, \mathbf{v}_1\}$ gives a lower triangular matrix representation.

□

8.7 Minimal polynomials

8.7.1 Minimal polynomials of matrices

If p is a polynomial with coefficients in a field $\mathbb{F}$, then we can assign a value to $p(x)$ not just when x is a member of the field $\mathbb{F}$, but when x is a square matrix with elements in $\mathbb{F}$. For instance, if $p(x) = x^2 + 2x - 3$ and M is the matrix $\begin{bmatrix} 4 & 2 \\ 0 & 1 \end{bmatrix}$, then

$$\begin{aligned} p(A) &= \begin{bmatrix} 4 & -2 \\ 0 & 3 \end{bmatrix}^2 + 2\begin{bmatrix} 4 & -2 \\ 0 & 3 \end{bmatrix} - 3\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 16 & -14 \\ 0 & 9 \end{bmatrix} + \begin{bmatrix} 8 & -4 \\ 0 & 6 \end{bmatrix} - \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 24 & -18 \\ 0 & 12 \end{bmatrix} \end{aligned}$$

where we've identified the constant term -3 in the polynomial with the matrix $-3I$.

Let S be the set of of polynomials $p \in \mathbb{F}[x]$ that satisfy $p(A) = 0$ for some specified matrix $A \in \text{Mat}_{n\times n}(\mathbb{F})$. It turns out that S satisfies all the properties for being an ideal of $\mathbb{F}[x]$:

1. *Non-emptiness*: $p(A) = 0$ is true if p is the zero polynomial, so S contains the zero polynomial at the very least.

2. *Closure under addition by other elements of the ideal*: If $p \in S$ and $q \in S$, then $(p + q)(A) = p(A) + q(A)$ (by definition of polynomial addition) $= 0 + 0 = 0$, so $p + q \in I$ as well.

3. *Closure under multiplication by arbitrary polynomials*: If $p \in S$, then $(pq)(A) = p(A)q(A) = 0q(A) = 0$, so $pq \in I$. (Remember that here, the notation pq means polynomial multiplication, not composition.)

So either S contains only the zero polynomial, or it equals the set of multiples of some monic polynomial of minimal degree which we'll call the *minimal polynomial* of A. It turns out that S must contain at least one nonzero element. One way to see this: remember that $\mathrm{Mat}_{n\times n}(\mathbb{F})$ is a vector space with dimension n^2. so for any matrix A, the n^2+1 matrices $I, A, A^2, \cdots, A^{n^2}$ have to be linearly dependent, and whatever linear combination of them gives the zero matrix also specifies a polynomial with degree at most n^2. (As we'll see in the next section, we can get a much better upper bound on the degree than n^2.)

8.7.2 Minimal polynomials of operators

We can also apply p to linear operators on a vector space over $\mathbb{F}$, not just to matrix representations. For instance, if $T \in \mathrm{End}(V)$ is a linear operator over a real vector space V and $p(x) = x^2 + 2x - 3 \in \mathbb{R}[x]$, then $p(T) = T^2 + 2T - 3I$; that is, $p(T)$ sends $\mathbf{v}$ to $T(T\mathbf{v}) + 2T\mathbf{v} - 3\mathbf{v}$. We can define the minimal polynomial of T either as the minimial polynomial of any matrix representation of T (though we'll have to prove that all representations give the same polynomial), or the minimal element of the ideal $\{p \in \mathbb{F}[x] : p(T) \text{ is the zero operator}\}$.

To prove that these definitions coincide, remember that if a linear operator T has matrix representations J and M related by the basis translation matrix S (that is, $M = SJS^{-1}$), then $M^n = (SJS^{-1})^n = SJ^nS^{-1}$, so J^n and M^n represent T^n with respect to the same two bases. Similarly, if some other operator T' has matrix representations J', M' with $M' = SJS^{-1}$ (where the change of basis matrix S is the same as before), then $M + M' = SJS^{-1} + SJ'S^{-1} = S(J + J')S^{-1}$, so $T + T'$ has matrix representations M and J' with respect to the same two bases.

So since changes of basis respect both operator powers and operator sums, the operator $p(T)$ for any polynomial p can be represented in matrix form either as $p(J)$ or as $p(M) = p(SJS^{-1}) = Sp(J)S^{-1}$. The only matrix representation of the map that sends everything to zero is the zero matrix. So $p(T)$ is the zero map if and only if $p(M)$ and $p(J)$ are the zero matrix; and if $p(M)$ is the zero matrix, then $p(J)$ is the zero matrix for any matrix J similar to M.

It bears reiteration that this theory makes sense only for finite-dimensional vector spaces: operators on infinite-dimensional spaces may not have a minimal polynomial. One example is the right-shift operator $R(x_1, x_2, x_3, \ldots) = (0, x_1, x_2, x_3, \ldots)$ on the space of infinite sequences $\mathbb{F}^{\mathbb{N}}$. If the nonzero polynomial $a_nR^n + \cdots + a_1R + a_0I$ is applied to the sequence $(1, 0, 0, 0, \ldots)$ with a 1 in the first slot and zeros elsewhere, then it produces the sequence $(a_0, a_1, \ldots, a_n, 0, 0, 0, \ldots)$. Thus, no nonzero polynomial of R can map every element of $\mathbb{F}^{\mathbb{N}}$ to zero, so R does not have a minimal polynomial.

8.7.3 Minimal polynomials and invariant subspace decompositions

In our discussion of characteristic polynomials, we noted that if a space V can be decomposed as $V = U \oplus W$ where U and V are invariant subspaces of some operator T, then the characteristic polynomial T is the product of the characteristic polynomials of

its restrictions to U and W. A natural question is whether a similar finding exists for minimal polynomials. The answer is that it does, but instead of *multiplying* minimal polynomials, we have to compute the *least common multiple*.

The least common multiple of two polynomials p, q is the monic polynomial r of smallest degree such that both p and q divide r. In practical terms, we can calculate the LCM by factoring p and q as far as we can, then taking the largest exponent of each factor: for instance, if $p(x) = a(x)b(x)^2$ and $q(x) = b(x)^3c(x)$ where a, b, c are polynomials that can't be factored any further (the more proper term is *irreducible*), then the least common multiple is $a(x)b(x)^3c(x)$.

Proposition. *Suppose $V = U \oplus W$ where both U and W are invariant subspaces of some operator $T \in \mathrm{End}(V)$, and let m_U, m_W be the minimal polynomials of the restricted maps $T|_U, T|_W$. Let p be any other polynomial. Then $p(T)$ is the zero operator if and only if m_U and m_W both divide p.*

Proof. First, suppose p is divisible by both m_U and m_W: that is, we can write $p(x) = q_U(x)m_U(x) = q_W(x)m_W(x)$ where q_U, q_W are two other polynomials. For any $\mathbf{u} \in U$, therefore, $p(T)\mathbf{u} = q_U(T)m_U(T)\mathbf{u} = q_U(T)\mathbf{0} = \mathbf{0}$, so $p(T)$ is the zero map on U. By identical logic, $p(T)$ is the zero map on W. Since $V = U \oplus W$, any element $\mathbf{v} \in V$ can be written $\mathbf{v} = \mathbf{u} + \mathbf{w}$, so $p(T)\mathbf{v} = p(T)\mathbf{u} + p(T)\mathbf{w} = \mathbf{0} + \mathbf{0} = \mathbf{0}$, so $p(T)$ is the zero operator.

Conversely, suppose that $p(T)$ is the zero operator on V. Then in particular, it must be the zero operator on U and on V separately. The set of polynomials q such that $q(T|_U) = \mathbf{0} \in \mathrm{End}(U)$, as we've discussed earlier in this section, is an ideal, so it's the set of all multiples of some minimal polynomial that, by definition, is m_U. The same logic goes for W. So both m_U and m_W must divide p.

□

Corollary. *With the same setup as in the previous proposition, the minimal polynomial of T is the least common multiple of m_U and m_W.*

And there's a natural translation into matrix language:

Corollary. *If a matrix M has the block diagonal structure $\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$, then the minimal polynomial of M is the LCM of the minimal polynomials of A and B.*

8.7.4 Maximum generalized eigenvector order

In section 8.4.4, we proved that the dimension of the subspace containing all generalized eigenvectors of some operator T with a fixed eigenvalue λ was the multiplicity of λ as a root of T's characteristic polynomial. You may wonder if T's minimal polynomial tells us something else useful about generalized eigenspaces. It turns out that the answer is yes: it tells you the maximum order of any generalized eigenvector with eigenvalue λ.

Throughout, V is a finite-dimensional vector space, T is a linear operator on V, λ is a scalar, and m is the minimal polynomial of T.

Proposition. *Suppose that every element of V is a GEV of T with eigenvalue λ, and the maximum order of these GEVs is h. Then $m(x) = (x - \lambda)^h$.*

Proof. If $\mathbf{v} \in V$ is a GEV with eigenvalue λ and order $\leq h$, then $(T-\lambda)^h\mathbf{v} = \mathbf{0}$. If this is true for every $\mathbf{v} \in V$, then $p(x) \coloneqq (x-\lambda)^h$ satisfies $p(T) = \mathbf{0}_{\mathrm{End}(V)}$, so p is a multiple of m; that is, $m(x) = (x-\lambda)^k$ for some integer $k \leq h$. But if $\mathbf{v}$ is a GEV of order h, then $(T-\lambda)^k$ for $k < h$ is a GEV of order $h-k$, and in particular can't be $\mathbf{0}$. So $k = h$ and $m(x) = (x-\lambda)^h$.

□

Theorem. *If V is a finite-dimensional vector space and T is an arbitrary operator on V, then the the largest order of any GEV of T with eigenvalue λ is also the largest exponent k such that $(x-\lambda)^k$ divides $m(x)$.*

Proof. Write h for the maximal order of a GEV with eigenvalue λ. By repeating the logic from section 8.4.4, we can decompose V into a direct sum of T-invariant subspaces $U \oplus W$, where $U = \ker(T-\lambda)^h$ contains all the GEVs with eigenvalue λ, and $V = \mathrm{im}(T-\lambda)^h$ contains no GEVs with eigenvalue λ except $\mathbf{0}$. If m_U, m_W are the minimal polynomials of the restricted maps $T|_U, T|_W$, then m must be the LCM of m_U and m_W.

We know from the previous proposition that $m_U(x) = (x-\lambda)^h$, so $m(x)$ is guaranteed to have a factor of $(x-\lambda)^h$ (but no higher factor of $x-\lambda$) if $x-\lambda$ does not divide $m_W(x)$.

We claim that $x-\lambda$ does not, in fact, divide $m_W(x)$. To prove this, suppose the contrary: that m_W can be factored as $m_W(x) = (x-\lambda)p(x)$ where p is some polynomial of degree less than that of m_W. Then $p(T|_W)$ can't be the zero operator (if it were, then p would be the minimal polynomial of $T|_W$ or a multiple thereof, not m_W). So there's some vector $\mathbf{w} \in W$ such that $p(T|_W)\mathbf{w} \neq \mathbf{0}$ but $m_W(T|_W) = (T|_W - \lambda)p(T|_W)\mathbf{w} = \mathbf{0}$. But this means that $p(T|_W)\mathbf{w}$ would be an nonzero eigenvector of T with eigenvalue λ—a contradiction, as U and W were constructed to put all such eigenvectors in U and none of them in W.

Therefore, $x-\lambda$ doesn't divide m_W, so $m(x)$ is divisible by $(x-\lambda)^h$ but no higher power of $x-\lambda$.

□

Note that this result doesn't require us to assume that V is a complex vector space, or that m can be factored completely into monomials. An alternate proof that $x-\lambda$ can't divide m_W is a consequence of the Cayley–Hamilton theorem proved in the next section: if $x-\lambda$ divided the minimal polynomial of $T|_W$, then it would have to divide its characteristic polynomial; that is, $T|_W$ would need to have λ as an eigenvalue.

8.8 Cayley–Hamilton theorem

It turns out that the minimal polynomial of any $n \times n$ matrix (equivalently, operator on an n-dimensional vector space) has degree at most n; in particular, the characteristic polynomial will always be a multiple of the minimal polynomial.

Theorem (Cayley–Hamilton theorem)**.** *If M is an $n \times n$ matrix with complex entries (possibly all real) and χ is its characteristic polynomial, then $\chi(M)$ is the zero matrix.*

Proof. Write $M = SJS^{-1}$ where J is upper triangular. (Note that J may have complex entries even if M has all real entries.) Then M and J have the same characteristic polynomial. Furthermore, $p(M) = Sp(J)S^{-1}$ for any polynomial p, and the only matrix similar to the zero matrix is itself, so $\chi(M)$ is zero if and only if $\chi(J)$ is zero as well.

So we just need to prove that $\chi(J) = 0$. Let $\lambda_1, \ldots, \lambda_n$ be the diagonal entries of J. Since $xI - J$ is also upper triangular, its determinant—that is, the characteristic polynomial χ of J—is the product $(x - \lambda_1) \cdots (x - \lambda_n)$ of the diagonal entries of J. Let $\mathbf{e}_1, \ldots, \mathbf{e}_n \in \mathrm{Col}_n(\mathbb{C})$ be the standard basis column vectors, and note that $J\mathbf{e}_k$ equals $\lambda_k \mathbf{e}_k$ plus a linear combination of $\mathbf{e}_1, \ldots, \mathbf{e}_{k-1}$.

So the column space of $J - \lambda_n$ (remember that we identify constants such as λ_n with multiples $\lambda_n I$ of the identity map) is contained in $\mathrm{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}\}$. Similarly, the column space of $(J - \lambda_{n-1})(J - \lambda_n)$ is contained in the image of $\mathrm{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}\}$ under $J - \lambda_{n-1}$, which itself is contained in $\mathrm{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_{n-2}\}$. Continuing likewise, we get that $(J - \lambda_2) \cdots (J - \lambda_n)$ sends every column vector to a multiple of $\mathbf{e}_1$, which is an eigenvector of J, so $(J - \lambda_1) \cdots (J - \lambda_n)$ sends every column vector to $\mathbf{0}_{\mathrm{Col}_n(\mathbb{C})}$. That is, $\chi(J)$ is the zero matrix.

□

Even though we've used complex numbers in its proof, the Cayley–Hamilton theorem establishes a statement about matrices that is equally valid even if the entries in M are all in a smaller field, such as the real or rational numbers.

Corollary. *For any linear operator $T : V \to V$ on a vector space whose base field is a subfield of $\mathbb{C}$ (such as $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$ itself), $\chi_T(T) = \mathbf{0}_{\mathrm{End}(V)}$.*

Proof. χ_T equals χ_M for any matrix representation M of T, and $\chi_M(M)$ is the zero matrix and a representation of $\chi_T(T)$, and the zero matrix can only represent, and is the only representation of, the zero map.

□

Corollary. *The characteristic polynomial of any matrix is a multiple of its minimal polynomial.*

Proof. The set I of polynomials that satisfy $p(A) = 0$ for some specified matrix A is an ideal of $\mathbb{F}[x]$, so it equals the set of all multiples of its primitive element. This primitive element is (by definition) the minimal polynomial, and the characteristic polynomial is in I by the Cayley–Hamilton theorem.

□

Remark. These results are true even for more exotic fields not contained in $\mathbb{C}$. The key result from abstract algebra is that given any field $\mathbb{F}$, we can construct an algebraically complete field (that is, a field in which every polynomial has a root, like $\mathbb{C}$) that contains $\mathbb{F}$ as a subfield. But the existence of algebraically complete field extensions would require a detour into abstract algebra, so we won't talk about it more here.

8.9 Jordan normal form

8.9.1 Existence and essential uniqueness

We've now laid the groundwork for the main result of this chapter: every matrix is similar to exactly one matrix with a specified block diagonal form made out of *Jordan blocks*. A Jordan block is an upper triangular matrix with identical entries λ on the

diagonal, entries of 1 on every entry above a diagonal entry (called the *superdiagonal*), and zeros elsewhere. Jordan blocks of size 1 through 5, for example, have the forms

$$\begin{bmatrix}\lambda\end{bmatrix} \quad \begin{bmatrix}\lambda & 1\\ 0 & \lambda\end{bmatrix} \quad \begin{bmatrix}\lambda & 1 & 0\\ 0 & \lambda & 1\\ 0 & 0 & \lambda\end{bmatrix} \quad \begin{bmatrix}\lambda & 1 & 0 & 0\\ 0 & \lambda & 1 & 0\\ 0 & 0 & \lambda & 1\\ 0 & 0 & 0 & \lambda\end{bmatrix} \quad \begin{bmatrix}\lambda & 1 & 0 & 0 & 0\\ 0 & \lambda & 1 & 0 & 0\\ 0 & 0 & \lambda & 1 & 0\\ 0 & 0 & 0 & \lambda & 1\\ 0 & 0 & 0 & 0 & \lambda\end{bmatrix}$$

The Jordan block of size n and eigenvalue λ has characteristic and minimal polynomial $(x-\lambda)^n$. The characteristic polynomial is easy to compute from the general result on determinants of triangular matrices. To find the minimal polynomial, note that if M is a Jordan block with diagonal entry λ, then $(M-\lambda I)$ is a matrix with 1s on the superdiagonal and 0s everywhere else, and every power of this matrix moves the 1s another step up and to the right. For instance, for a 4×4 block:

$$\begin{bmatrix}0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0\end{bmatrix}^2 = \begin{bmatrix}0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$

$$\begin{bmatrix}0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0\end{bmatrix}^3 = \begin{bmatrix}0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$

$$\begin{bmatrix}0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0\end{bmatrix}^4 = \begin{bmatrix}0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$

If an operator $T : V \to V$ has a matrix representation as a single Jordan block with diagonal entries λ relative to the basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, then all of V is a generalized eigenspace with order n and eigenvalue λ, and repeated application of the operator $T-\lambda I$ produces the chain $\mathbf{v}_n \mapsto \mathbf{v}_{n-1} \mapsto \cdots \mapsto \mathbf{v}_1 \mapsto \mathbf{0}$.

A matrix in *Jordan normal form* (JNF) is a block diagonal matrix in which every block is a Jordan block. We can formulate the main theorem on JNF in two ways:

1. In *operator language*: let V be a finite-dimensional complex vector space and T be a linear operator on V. Then V has a basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ consisting entirely of GEVs of T with respective eigenvalues $\lambda_1, \ldots, \lambda_n$ organized into chains: that is, $(T-\lambda_1)\mathbf{v}_1 = \mathbf{0}$, and for all integers $i \geq 2$, either $(T-\lambda_i)\mathbf{v}_i = \mathbf{0}$, or $(T-\lambda_i)\mathbf{v}_i = \mathbf{v}_{i-1}$ and $\lambda_{i-1} = \lambda_i$. We'll call this a *Jordan basis* of V.

   Furthermore, the lengths and eigenvalues of the chains are determined by T: any two bases must have the same number of chains of length ℓ and eigenvalue λ for any positive integer λ and scalar λ.

2. In *matrix language*: Every matrix M with complex entries is similar to a matrix in Jordan normal form. Furthermore, if M is similar to multiple matrices in Jordan normal form, then these matrices can be turned into each other by rearranging the Jordan blocks along the diagonal without changing their sizes.

Why are these formulations equivalent? If $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a Jordan basis, then column i of the matrix representation of T with respect to this basis either has a diagonal entry of λ_k and zeros elsewhere (if $(T - \lambda)\mathbf{v}_k = \mathbf{0}$), or it has an entry of λ_k on the diagonal, an entry of 1 just above the diagonal in row $k - 1$, and zeros elsewhere (if $(T - \lambda)\mathbf{v}_k = \mathbf{v}_{k-1}$). This means that every chain $(T - \lambda)\mathbf{v}_j = \mathbf{v}_{j-1}, (T - \lambda)\mathbf{v}_{j-1} = \mathbf{v}_{j-2}, \ldots, (T - \lambda)\mathbf{v}_i = \mathbf{0}$ creates a Jordan block in the matrix representation with eigenvalue λ, extending from row and column i to row and column j.

If the chain lengths and eigenvalues in any Jordan basis are uniquely determined, therefore, then the block sizes and eigenvalues in the matrix representation must also be uniquely determined. Rearranging Jordan blocks in the matrix representation, meanwhile, corresponds to rearranging a Jordan basis while keeping chains together.

Theorem. *A Jordan basis for any finite-dimensional complex vector space V with respect to an arbitrary operator $T \in \mathrm{End}(V)$, as spelled out above, exists, and the eigenvalues and lengths of its chains are uniquely determined by T.*

Proof. We've essentially already proved this! It just requires putting together three building blocks from previous sections.

First, some notation: let n be the dimension of V, let $\lambda_1, \ldots, \lambda_k$ be the (distinct) eigenvalues of T, and let W_i be be the subspace of V consisting of all the generalized eigenvectors of T, of any order, with eigenvalue λ_i. (We'll call this the "maximal" GES with eigenvalue λ_i.) Remember that W_i is an invariant subspace of T, so any results on general operators and spaces also apply to the restricted map $T|_{W_i}$. Let χ be the characteristic polynomial of T. Since V is a complex vector space and $\mathbb{C}$ is algebraically complete, χ must factor completely as $\chi(x) = (x - \lambda_1)^{a_1} \cdots (x - \lambda_k)^{a_k}$, with $a_1 + \cdots + a_k = n$.

The conclusion follows from the following easy inferences:

1. The sum $W_1 + \cdots + W_k$ is a sum of GESes with all distinct eigenvalues, so it is direct (page 147), and $\dim(W_1 \oplus \cdots \oplus W_k) = \dim W_1 + \cdots + \dim W_k$.

2. The dimension $\dim W_i$ of any of the maximal GESes is a_i, the exponent of $x - \lambda_i$ in $\chi(x)$ (page 206). Therefore, $\dim W_1 + \cdots + \dim W_k = a_1 + \cdots + a_k = n = \dim V$, so $V = W_1 \oplus \cdots \oplus W_k$. That is, V can be completely decomposed into its maximal GESes.

3. Each of the spaces W_i has a Jordan basis relative to the restricted operator $T|_{W_i}$, with uniquely determined chain lengths. This was the main finding of Section 6.8, page 148.

Taking the union of the Jordan bases for each W_i gives a Jordan basis for V.

□

8.9.2 Characteristic and minimal polynomials

From the Jordan normal form of a matrix or operator, it's possible to discern, at a glance, all that operator's eigenvalues, as well as the sizes of the associated (generalized) eigenspaces: every Jordan block with size k and diagonal entry λ adds 1 to the dimension of the subspace of eigenvectors with eigenvalue λ, adds 2 to the dimension of the subspace of GEVs with order at most 2, and so on up to k.

It's also possible to find the characteristic and minimal polynomials at a glance by putting together three observations that we've already made:

1. The Jordan block with size k and diagonal λ has minimal polynomial and characteristic $(x-\lambda)^k$ (page 215).
2. The characteristic polynomial of a block diagonal matrix is the product of the characteristic polynomials of the blocks (page 207).
3. The minimal polynomial of a block diagonal matrix is the least common multiple of the minimal polynomials of the blocks (page 212).

These three points together mean that if T is an operator on a finite-dimensional complex vector space and has distinct eigenvalues $\lambda_1, \dots, \lambda_k$, then:

1. The characteristic polynomial of T is $\chi_T(x) = (x-\lambda_1)^{a_1} \cdots (x-\lambda_k)^{a_k}$, where a_i is the *sum of block sizes* with diagonal entry λ_i (or equivalently, the *count* of diagonal entries λ_i)
2. The minimal polynomial of T is $m_T(x) = (x-\lambda_1)^{b_1} \cdots (x-\lambda_k)^{b_k}$, where b_i is the *size of the largest block* with diagonal entry λ_i.

This result gives an alternate proof of the Cayley–Hamilton theorem: obviously the size of the largest block with a certain eigenvalue must be less than the total sum of all the blocks.

The process for actually computing a Jordan basis is a bit complicated, and we won't go into it too much here. Computing a (not necessarily Jordan) basis of a generalized eigenspace is easy, though—just use the typical RREF method to compute $\ker(M-\lambda_i)^h$—and you may enjoy thinking about how the proof of existence of Jordan bases in Section 6.8 could be adapted into a practical algorithm.

8.10 Real matrices and conjugate eigenspaces

In most of this chapter, we've been dealing with complex matrices, and some results for complex matrices don't generalize in the obvious way to real matrices because polynomials with real coefficients don't necessarily have real roots. For instance, any real matrix M is conjugate to a matrix in Jordan normal form with complex coefficients by means of a change-of-basis matrix with complex entries. But the only real matrices M whose Jordan normal forms are real are the ones whose characteristic polynomials have all real roots.

The existence of Jordan normal form, though, does give us a way to determine if two matrices are similar in $\mathbb{R}$. If A and B are real matrices that both have complex Jordan normal form J, then since matrix similarity is a transitive relation, A and B are similar to each other as elements of $\mathrm{Mat}_{n\times n}(\mathbb{C})$; that is, there is some complex invertible matrix S such that $A = SBS^{-1}$. One natural question: are A and B also similar as elements of $\mathrm{Mat}_{n\times n}(\mathbb{R})$—that is, can we choose S to have all real entries? The answer turns out to be yes.

Proposition. *Suppose A and B are $n \times n$ matrices with real entries, and suppose there's some invertible matrix $S \in \mathrm{Mat}_{n\times n}(\mathbb{C})$ such that $A = SBS^{-1}$. Then there's also a matrix S' with all real entries such that $A = S'BS'^{-1}$.*

Proof. Note first that if S is invertible, then the equations $A = SBS^{-1}$ and $AS = SB$ are equivalent by right-multiplying both sides by S or S^{-1}.

Separate the real and imaginary parts of S into two matrices: $S = S_r + iS_i$ where S_r, S_i are both real. If $A = SBS^{-1}$ (that is, $AS = SB$), then $AS_r + iAS_i = S_rB + iS_iB$. As AS_r and S_rB are real and iAS_i and iS_iB are purely imaginary, so $AS_r = S_rB$ and $AS_i = S_iB$. So any linear combination $S' \coloneqq S_r + cS_i$ also satisfies $AS' = S'B$, so if S' is invertible and c is real, then S' is a real change-of-basis matrix from A to B.

Consider the function $p(x) = \det(S_r + xS_i)$. This function is a polynomial with degree at most n and real coefficients (because every entry in $S_r + xS_i$ has the form $ax + b$, and each term in the determinant is the product of n entries). Furthermore, p can't be the zero polynomial, because $p(i) = \det S \neq 0$. So p has only a finite number of roots, and in particular, there's sone real number c such that $S_r + cS_i$ is invertible and gives a change-of-basis matrix from A to B.

□

Another question: how close can we get to Jordan normal form while staying in the real numbers? The answer turns out to be: pretty close. The complex generalized eigenspaces for complex conjugate eigenvectors turn out to have identical chain structures, and we can merge these eigenspaces together into a real almost-but-not-quite generalized eigenspace that gives the operator an almost-but-not-quite triangular matrix representation.

The starting point is this basic result in the theory of polynomials:

Proposition. *The non-real roots of any polynomial p with real coefficients occur in conjugate pairs. That is, if z is a root of p with multiplicity k, then $\bar{z}$ is also a root with multiplicity k.*

Proof. Remember that complex conjugation respects addition and multiplication: that is, $\overline{z + w} = \bar{z} + \bar{w}$ and $\overline{zw} = \bar{z}\bar{w}$. Therefore, if $p(z) = c_n z^n + \cdots + c_1 z + c_0$ and the coefficients c_n are all real (that is, they equal their own complex conjugates), then $p(z)$ and $p(\bar{z}) = c_n \bar{z}^n + \cdots + c_1 \bar{z} + c_0$ are complex conjugates. In particular, if $p(z) = 0$, then $p(\overline{z}) = 0$.

To see that z_0 and $\bar{z}_0$ have the same multiplicity, note that $(z - z_0)(z - \bar{z}_0) = z^2 - (z_0 + \bar{z}_0)z + z_0\bar{z}_0$ is a polynomial with real coefficients. If z_0 has multiplicity k as a root of p, but $\bar{z}_0$ has multiplicity greater than k, then $\dfrac{p(z)}{(z - z_0)^k (z - \bar{z}_0)^k}$ is also a polynomial with real coefficients (because the quotient of polynomials with real coefficients must have real coefficients itself) that has $\bar{z}_0$ but not z_0 as a root, a contradiction of what we just proved. (Similar logic shows that $\bar{z}_0$ can't have multiplicity less than that of z_0, either.)

□

Corollary. *The non-real eigenvalues of any real matrix (or operator on a finite-dimensional real vector space) occur in conjugate pairs: that is, if λ is an eigenvalue, then so is $\bar{\lambda}$. Furthermore, the dimensions of the maximal generalized eigenspaces of λ and $\bar{\lambda}$ (that is, the exponents of $x - \lambda$ and $x - \bar{\lambda}$ in the characteristic polynomial), and the maximal order of generalized eigenvectors with eigenvalue λ and $\bar{\lambda}$ (that is, the exponents of $x - \lambda$ and $x - \bar{\lambda}$ in the minimal polynomial), are equal.*

We can even say more than this: not only does every non-real eigenvalue of a real matrix occur in complex conjugate pairs, but the corresponding generalized eigenvectors also occur in conjugate pairs.

Proposition. *Let $T : \mathbb{C}^n \to \mathbb{C}^n$ be a linear transformation whose matrix representation relative to the standard basis has all real entries. For every $\mathbf{v} = (a_1, \ldots, a_n) \in \mathbb{C}^n$, write $\bar{\mathbf{v}}$ for the vector $(\bar{a}_1, \ldots, \bar{a}_n)$ whose components are the complex conjugates of corresponding components of $\mathbf{v}$. Suppose that λ is a non-real eigenvalue of T, and let $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ be a basis for the maximal generalized eigenspace with eigenvalue λ that consists of a set of Jordan chains: that is, $(T - \lambda)\mathbf{v}_1 = \mathbf{0}$ and $(T - \lambda)\mathbf{v}_i$ is either $\mathbf{0}$ or $\mathbf{v}_{i-1}$ for $2 \leq i \leq k$. Then $\bar{B} \coloneqq \{\bar{\mathbf{v}}_1, \ldots, \bar{\mathbf{v}}_k\}$ is a basis for the maximal generalized eigenspace with eigenvalue $\overline{\lambda}$, and $\bar{B}$ has the same chain structure with respect to $T - \bar{\lambda}$ that B has with respect to $T - \lambda$.*

Proof. Remember that every component in $T\mathbf{v}$ is the sum of products of entries in $\mathbf{v}$ with entries in the matrix representation of T with respect to the standard basis. If all of these matrix entries are real, then $T\mathbf{v}$ and $T\bar{\mathbf{v}}$ are complex conjugates (because if x is real then xz and $x\bar{z}$ are complex conjugates), so $(T-\lambda)\mathbf{v}$ and $(T-\overline{\lambda})\overline{\mathbf{v}}$ are also complex conjugates, so the conclusion follows.

□

We can use this result to get a sort of "real JNF" that recasts corresponding complex Jordan blocks for λ and $\overline{\lambda}$ into a single Jordan block. How can we do this? Suppose that $W \subset \mathbb{C}^n$ is a maximal primitive generalized eigenspace of eigenvalue λ: that is, its basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is a single Jordan chain $\mathbf{v}_k \mapsto \cdots \mapsto \mathbf{v}_1 \mapsto \mathbf{0}$ with respect to $T - \lambda$. Denote by $\bar{W}$ the space of complex conjugates of every element in W; this (by the proposition that we just proved) is also a maximal generalized eigenspace of T with eigenvalue $\bar{\lambda}$.

Now, let the symbols $\Re$ and $\Im$ denote the real and imaginary parts of a complex number, and remember the general formulas $z + \bar{z} = 2\Re(z)$ and $z - \overline{z} = 2i\Im(z)$. By analogy, define the vectors $\mathbf{u}_i = \frac{1}{2}(\mathbf{v}_i + \bar{\mathbf{v}}_i)$ and $\mathbf{w}_i = \frac{1}{2i}(\mathbf{v}_i - \bar{\mathbf{v}}_i)$ for $1 \leq i \leq k$: these are the real and imaginary parts of $\mathbf{v}_i$. The vectors $\mathbf{u}_i$ and $\mathbf{w}_i$ must have real coefficients, and $\operatorname{span}\{\mathbf{u}_i, \mathbf{w}_i\} = \operatorname{span}\{\mathbf{v}_i, \bar{\mathbf{v}}_i\}$ for all integers $1 \leq i \leq k$. To see that $\{\mathbf{u}_i, \mathbf{w}_i\}$ does in fact span all of $\operatorname{span}\{\mathbf{v}_i, \bar{\mathbf{v}}_i\}$, note that the matrix $S = \begin{bmatrix} 1/2 & -i/2 \\ 1/2 & i/2 \end{bmatrix}$, which translates column-vector representations relative to the $\{\mathbf{u}_i, \mathbf{w}_i\}$ basis to representations relative to the $\{\mathbf{v}_i, \bar{\mathbf{v}}_i\}$ basis, has nonzero determinant $i/2$. The inverse of S is $S^{-1} = \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}$, which gives the formulas $\mathbf{v}_i = \mathbf{u}_i + i\mathbf{w}_i$ and $\bar{\mathbf{v}}_i = \mathbf{u}_i - i\mathbf{w}_i$.

So $\{\mathbf{u}_1, \mathbf{w}_1, \mathbf{u}_2, \mathbf{w}_2, \ldots, \mathbf{u}_k, \mathbf{w}_k\}$ is a basis for $W \oplus \bar{W}$. To figure out the matrix representation of $T|_{W \oplus \bar{W}}$ relative to this basis, we need to compute $T\mathbf{u}_i$ and $T\mathbf{w}_i$ for every integer $1 \leq i \leq k$. First, let's compute $T\mathbf{u}_1$ and $T\mathbf{w}_1$:

$$\begin{aligned} T\mathbf{u}_1 &= \frac{1}{2}T(\mathbf{v}_1 + \bar{\mathbf{v}}_1) \\ &= \frac{1}{2}(\lambda\mathbf{v}_1 + \bar{\lambda}\bar{\mathbf{v}}_1) \\ &= \frac{1}{2}(\lambda(\mathbf{u}_1 + i\mathbf{w}_1) + \bar{\lambda}(\mathbf{u}_1 - i\mathbf{w}_1)) \\ &= \frac{\lambda + \bar{\lambda}}{2}\mathbf{u}_1 + \frac{i(\lambda - \bar{\lambda})}{2}\mathbf{w}_1 \\ &= \Re(\lambda)\mathbf{u}_1 - \Im(\lambda)\mathbf{w}_1 \end{aligned}$$

Similarly,

$$\begin{aligned} T\mathbf{w}_1 &= \frac{1}{2i}T(\mathbf{v}_1 - \bar{\mathbf{v}}_1) \\ &= \frac{1}{2i}(\lambda\mathbf{v}_1 - \bar{\lambda}\bar{\mathbf{v}}_1) \\ &= \frac{1}{2i}(\lambda(\mathbf{u}_1 + i\mathbf{w}_1) - \bar{\lambda}(\mathbf{u}_1 + i\mathbf{w}_1)) \\ &= \frac{1}{2i}((\lambda - \bar{\lambda})\mathbf{u}_1 + i(\lambda + \bar{\lambda})\mathbf{w}_1) \\ &= \Im(\lambda)\mathbf{u}_1 + \Re(\lambda)\mathbf{w}_i. \end{aligned}$$

For the vectors $\mathbf{u}_i, \mathbf{w}_i$ with $i \geq 2$, we can compute

$$\begin{aligned} T\mathbf{u}_i &= \frac{1}{2}T(\mathbf{v}_i + \bar{\mathbf{v}}_i) \\ &= \frac{1}{2}(\mathbf{v}_{i-1} + \bar{\mathbf{v}}_{i-1} + \lambda\mathbf{v}_i + \bar{\lambda}\bar{\mathbf{v}}_i) \\ &= \mathbf{v}_{i-1} + \Re(\lambda)\mathbf{u}_i - \Im(\lambda)\mathbf{w}_i \\ T\mathbf{w}_i &= \frac{1}{2i}T(\mathbf{v}_i - \bar{\mathbf{v}}_i) \\ &= \frac{1}{2i}(\mathbf{v}_{i-1} - \bar{\mathbf{v}}_{i-1} + \lambda\mathbf{v}_i - \bar{\lambda}\bar{\mathbf{v}}_i) \\ &= \mathbf{w}_{i-1} + \Im(\lambda)\mathbf{u}_i + \Re(\lambda)\mathbf{w}_i. \end{aligned}$$

So with respect to the basis $\{\mathbf{u}_1, \mathbf{w}_1, \ldots, \mathbf{u}_k, \mathbf{w}_k\}$, $T|_{W \oplus \bar{W}}$ has the following almost-but-not-quite-triangular form. Define $r = \Re(\lambda)$ and $c = \Im(\lambda)$ for clarity:

$$\begin{bmatrix} r & c & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ -c & r & 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & r & c & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -c & r & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & r & c \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & -c & r \end{bmatrix}$$

Note the relationship to Jordan form: we've replaced every occurrence of λ in the Jordan block for W with the 2×2 block $\begin{bmatrix} r & c \\ -c & r \end{bmatrix}$ and every occurrence of 1 with $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. The matrix with the same block form but with $\begin{bmatrix} r & -c \\ c & r \end{bmatrix}$ as a block instead of $\begin{bmatrix} r & c \\ -c & r \end{bmatrix}$ would also be valid; this corresponds to designating $\bar{\lambda}$ as the "original" eigenvalue and λ as the conjugate instead of vice versa. And we can assemble a "real JNF" for the entire transformation on T by lining up blocks of this form on the diagonal for non-real eigenvalues and using ordinary Jordan blocks for real eigenvalues.

Chapter 9

Inner products and vector space geometry

Most of this chapter looks at vector spaces that have an extra structure called an *inner product*, which gives a notion of how large and close together vectors are. The theory of inner products in $\mathbb{R}^2$ and $\mathbb{R}^3$, in particular, is essential for using linear algebra to model real-world geometry and physics. These notions will get clearer as we go on.

First, a few matrix definitions:

1. The *transpose* of an $r \times c$ matrix M is the $c \times r$ matrix produced by reflecting M across the diagonal. Row number i of M is column number i of M^T, and vice versa. For instance, if $M = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$, then $M^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$. (We've seen transposed matrices a few times before, most extensively with the adjugate and cofactor matrices in section 7.8, but now we'll start talking about the properties of matrix transposes in general.)

2. An $n \times n$ matrix is *symmetric* if it equals its transpose (that is, entry (i, j) equals entry (j, i) for all integers $1 \leq i, j \leq n$) and *antisymmetric* if it equals the negative of its transpose. Diagonal entries on an antisymmetric matrix must all be zero.

3. The *Hermitian conjugate* or *conjugate transpose* of M, which we'll denote M^H (some books use the notation $M^\dagger$ or M^*), is the matrix created from M by transposing it and then taking the complex conjugate of all of its entries. For instance, if $M = \begin{bmatrix} i & 4 \\ 1-2i & 2+5i \end{bmatrix}$, then $M^H = \begin{bmatrix} -i & 1+2i \\ 4 & 2-5i \end{bmatrix}$.

4. An $n \times n$ matrix is *Hermitian* if it equals its own conjugate transpose, and *anti-Hermitian* if it equals the negative of its own conjugate transpose.

   If you break a complex matrix M into its real and imaginary parts $M = A + iB$ where A and B are real, then M is Hermitian if and only if A is symmetric and B is antisymmetric, and M is anti-Hermitian if and only if A is antisymmetric and B is symmetric. So real symmetric and antisymmetric matrices are automatically (respectively) Hermitian and anti-Hermitian; furthermore, the diagonal entries of a Hermitian matrix are real, and the diagonal entries of an anti-Hermitian matrix are purely imaginary. (Remember that zero counts as purely imaginary.)

And one important property of the transpose of matrix products: $(AB)^T = B^T A^T$ and, similarly, $(AB)^H = B^H A^H$. (Note that entry (i, j) of $(AB)^T$ and of $B^T A^T$ is produced from row j of A and column i of B.)

9.1 Bilinear and sesquilinear forms

Let V be a vector space over a field $\mathbb{F}$. A *bilinear form* on V is a function $B : V^2 \to \mathbb{F}$ that takes ordered pairs of vectors in V and returns a value in the base field, such that if you hold either argument of B constant, the map from V to $\mathbb{F}$ that you get from varying the other argument is linear. That is:

1. For any three vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{w} \in V$ and constants $c_1, c_2 \in \mathbb{F}$, $B(c_1\mathbf{v}_1 + c_2\mathbf{v}_2, \mathbf{w}) = c_1 B(\mathbf{v}_1, \mathbf{w}) + c_2 B(\mathbf{v}_2, \mathbf{w})$. (That is, the map $\mathbf{v} \mapsto B(\mathbf{v}, \mathbf{w})$ is linear for any constant vector $\mathbf{w}$.)

2. For any three vectors $\mathbf{v}, \mathbf{w}_1, \mathbf{w}_2 \in V$ and constants $c_1, c_2 \in \mathbb{F}$, $B(\mathbf{v}, c_1\mathbf{w}_1 + c_2\mathbf{w}_2) = c_1 B(\mathbf{v}, \mathbf{w}_1) + c_2 B(\mathbf{v}, \mathbf{w}_2)$. (That is, $\mathbf{w} \mapsto B(\mathbf{v}, \mathbf{w})$ is linear for any constant vector $\mathbf{v}$.)

One immediate consequence is that $B(\mathbf{v}, \mathbf{w}) = 0$ if (but not only if) either $\mathbf{v}$ or $\mathbf{w}$ equals $\mathbf{0}$. And as with linear maps with one argument, these axioms imply their own generalizations to sums of three or more vectors: that is,

$$B(a_1\mathbf{v}_1 + \cdots + a_m\mathbf{v}_m, b_1\mathbf{w}_1 + \cdots + b_n\mathbf{w}_m) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j B(\mathbf{v}_i, \mathbf{w}_j).$$

So B is completely determined by its values on every ordered pair of elements from a basis of V.

9.2 Matrix representations of bilinear forms

9.2.1 Definitions

Like linear maps, bilinear forms $B : V^2 \to \mathbb{F}$ have matrix representations if V is finite-dimensional. But to use the matrix representation, you have to multiply it representation by *two* vector representations: a row vector to represent the first input, and a column vector to represent the second.

Suppose that $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis of V, and write $b_{ij} = B(\mathbf{v}_i, \mathbf{v}_j)$. Then the matrix representation of B (sometimes called the *Gram matrix*) is

$$\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix}$$

To use this matrix to compute the value of a bilinear product on two vectors $B(\mathbf{u}, \mathbf{w})$, first find the coefficients $\mathbf{u} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$ and $\mathbf{w} = d_1\mathbf{v}_1 + \cdots + d_n\mathbf{v}_n$ of the arguments

to B, then represent $\mathbf{u}$ as a row vector and $\mathbf{w}$ as a column vector. The matrix product

$$\begin{bmatrix} c_1 & c_2 & \cdots & c_n \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{bmatrix}$$

is a 1×1 matrix whose sole entry is the value of $B(\mathbf{u}, \mathbf{w})$. (We won't be pedantic about distinguishing the space $\mathrm{Mat}_{1\times 1}(\mathbb{F})$ of 1×1 matrices from the underlying field $\mathbb{F}$.) You may want to prove for yourself that if $\mathbf{u} = \mathbf{v}_i$ and $\mathbf{w} = \mathbf{v}_j$ (that is, the row vector representing $\mathbf{u}$ has a 1 in position i and 0 everywhere else, and similar for $\mathbf{w}$), then the matrix product is b_{ij}.

We can ask the same questions about matrix representations of bilinear forms as about matrix representations of linear maps. We'll look at two core questions:

1. If we have a representation relative to one basis, how can we find a representation relative to another basis?

2. Can we define a set of canonical matrices with simple structures, analogous to matrices in Jordan normal form for linear operators, such that every bilinear form is equivalent to exactly one of those canonical matrices?

3. In particular, can we find such a set of canonical matrices if we also put some restrictions on the bases that we're allowed to use?

Finding a fully general set of canonical matrices is very difficult, but for the most useful subset of bilinear forms, these canonical matrices do exist. For example, for symmetric bilinear forms on $\mathbb{R}^n$—that is, for those that satisfy $B(\mathbf{v}, \mathbf{w}) = B(\mathbf{w}, \mathbf{v})$—a result called *Sylvester's law of inertia* proves that there is exactly one representation of B as a diagonal matrix whose diagonal entries are a string of entries of 1, followed by a string of entries of 0, followed by a string of entries of -1.

9.2.2 Matrix congruence and changes of basis

To answer our second and third questions, we'll first have to look at the first one: how to change the basis for a matrix representation of a bilinear form. Suppose that we have a matrix M that represents a bilinear form B relative to the basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. That is, if two vectors $\mathbf{u}_1, \mathbf{u}_2 \in M$ are represented relative to this basis by the column vectors $\mathbf{a}, \mathbf{b} \in \mathrm{Col}_n(\mathbb{F})$, then $B(\mathbf{u}_1, \mathbf{u}_2) = \mathbf{a}^T M \mathbf{b}$.

Suppose we want to find the matrix representation M' of B relative to some other basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$—that is, if $\mathbf{a}', \mathbf{b}'$ are the column vector representations of $\mathbf{u}_1, \mathbf{u}_2$ relative to the basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$, then $\mathbf{a}'^T M' \mathbf{b}' = \mathbf{a}^T M \mathbf{b}$. We know, of course, how to translate between column vector representations: if S is the matrix whose column i represents $\mathbf{v}_i$ relative to $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$, then S translates column-vector representations relative to $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ to column-vector representations relative to $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$. That is, $\mathbf{b} S \mathbf{b}'$ and $\mathbf{a} = S\mathbf{a}'$, so $\mathbf{a}^T = \mathbf{a}'^T S^T$ (remember the formula $(M_1 M_2)^T = M_2^T M_1^T$ for generic matrices M_1, M_2). Substituting these expressions into $\mathbf{a}'^T M' \mathbf{b}' = \mathbf{a}^T M \mathbf{b}$ gives $\mathbf{a}'^T M' \mathbf{b}' = \mathbf{a}'^T S^T M S \mathbf{b}$, which has to be true for every possible pair of column vectors $\mathbf{a}', \mathbf{b}'$—in particular, it must be true if $\mathbf{a}'$ and $\mathbf{b}'$ are both standard basis vectors, in which

case the products $\mathbf{a}'^T M'\mathbf{b}'$ and $\mathbf{a}'^T S^T MS\mathbf{b}$ simply extract one entry from the matrices M' and $S^T MS$. So $M' = S^T MS$.

We call two $n \times n$ matrices M', M *congruent*—that is, they represent the same bilinear form relative to possibly different bases—if there is some invertible matrix S such that $M' = S^T MS$. Note the resemblance to the concept of *similarity*: M and M' represent the same linear operator if there's some matrix S such that $M' = S^{-1}MS$. The definition of similarity used both S and S^{-1} because they express a two-way conversion from one basis to another and then back: S translates from the basis for M' into the basis for M, and then S^{-1} translates back into the basis for M'. In the product $M' = S^T MS$, however, both S^T and S translate vector representations in the same direction, for M to the basis for M': but left-multiplication by S translates column vector representations, and right-multiplication by S^T translates row vector representations.

It's easy to prove that congruence, like similarity, satisfies the three axioms of an equivalence relation:

1. Reflexivity: We have $M = S^T MS$ if $S = I$, so M is congruent to itself.

2. Commutativity: Suppose that M is congruent to M': that is, $M' = S^T MS$ for some matrix S. Then $(S^T)^{-1}MS^{-1} = S$, and it's easy to prove that $(S^T)^{-1} = (S^{-1})^T$ (because $I = SS^{-1}$, so $I = I^T = (SS^{-1})^T = (S^{-1})^T S^T$), so $(S^{-1})^T$ and S^T are inverses). So S^{-1} gives a change of basis from M' to M, and M' is congruent to M.

3. Transitivity: If $M_2 = S_1^T MS_1$ and $M_3 = S_2^T M_2 S_2$, then $M_3 = S_2^T(S_1^T M_1 S_1)S_2 = (S_1S_2)^T M(S_1S_2)$, so S_1S_2 gives a change-of-basis matrix from M_1 to M_3.

In general, matrices can be similar without being congruent, or congruent without being similar. An example of each possibility:

1. Every matrix M is congruent to its scalar multiple $4M$ (because choosing $S = 2I$ gives $4M = S^T MS$), but M and $4M$ are not in general similar. For instance, if M has nonzero determinant, then M and $4M$ have different determinants and so can't be similar.

2. Two similar but noncongruent real matrices are $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$. These matrices both have characteristic and minimal polynomial $x(x-1)$ and so must have the same JNF; in fact, A is the JNF of B.[1] But if A and B were congruent, there would have to be $a, b, c, d \in \mathbb{R}$ such that

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

   The product on the left is $\begin{bmatrix} a^2 & ab \\ ab & b^2 \end{bmatrix}$, which obviously can't equal $\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$, as we can't have both $ab = 1$ and $ab = 0$.

[1] The explicit change of basis $B = SAS^{-1}$ has $S = S^{-1} = \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix}$.

For symmetric matrices, which represent the most important class of bilinear forms on real vector spaces, it turns out that similarity does in fact imply congruence, but this result is not obvious and will take some effort to prove.

We'll mostly discuss bilinear forms in the context of vector spaces over $\mathbb{R}$: the theory of bilinear forms for vector spaces over $\mathbb{C}$ turns out not to be practically useful. Vector spaces over $\mathbb{C}$ do have a closely related, and much more useful, concept of *sesquilinear* forms, which we'll discuss a bit later.

9.3 Dot products, orthogonality, and geometry of $\mathbb{R}^n$

The simplest bilinear form on $\mathbb{R}^n$ is the form whose Gram matrix (relative to the standard basis) is the identity matrix: that is, the *dot product* defined as

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + \cdots + u_n v_n$$

where $\mathbf{u} = (u_1, \ldots, u_n)$ and $\mathbf{v} = (v_1, \ldots, v_n)$.

The dot product satisfies a property called *positive definiteness*: the product $\mathbf{v} \cdot \mathbf{v} = \sqrt{v_1^2 + \cdots + v_n^2}$ of a vector with itself is a positive real number, and it's zero if and only if $\mathbf{v}$ itself is zero. We'll denote the square root of $\mathbf{v} \cdot \mathbf{v}$ as $||\mathbf{v}||$ and call this the *norm* of $\mathbf{v}$; the norm of a vector is essentially a measure of its size or length.

The norm satisfies these three important properties:

1. It *respects scalar multiplication in absolute value*: $||k\mathbf{u}|| = |k|\,||\mathbf{u}||$ for any $k \in \mathbb{R}$. Proof: $||k\mathbf{u}|| = \sqrt{(k\mathbf{u}) \cdot (k\mathbf{u})} = \sqrt{k^2||\mathbf{u}||^2} = |k|\,||\mathbf{u}||$.

2. It satisfies the *Cauchy–Schwartz inequality*: $|\mathbf{u} \cdot \mathbf{v}| \leq ||\mathbf{u}||\,||\mathbf{v}||$ for all vectors $\mathbf{u}$ and $\mathbf{v}$, with equality if and only if $\mathbf{u}$ and $\mathbf{v}$ are scalar multiples of each other. Proof: note (or remember from high school algebra) the following facts:

   - The quadratic polynomial $ax^2 + bx + c$ has zero, one or two real roots according as its discriminant $b^2 - 4ac$ is negative, zero, or positive.
   - $||\mathbf{u} + x\mathbf{v}||^2 \geq 0$ for all real numbers x, because the square of any real number can't be negative.
   - The norm is positive definite, so $||\mathbf{u} + x\mathbf{v}||^2$ must be strictly greater than zero for all values of x unless $\mathbf{u} + x_0\mathbf{v} = \mathbf{0}$ for some x_0; that is, if $\mathbf{u}$ is a multiple of $\mathbf{v}$. In this case, $||\mathbf{u} + x_0\mathbf{v}||^2 = 0$, and $||\mathbf{u} + x\mathbf{v}||^2 > 0$ for $x \neq x_0$.

   We can expand $||\mathbf{u} + x\mathbf{v}||^2 = (\mathbf{u} + x\mathbf{v}) \cdot (\mathbf{u} + x\mathbf{v}) = ||\mathbf{v}||^2 x^2 + 2(\mathbf{u} \cdot \mathbf{v})x + ||\mathbf{u}||^2$, which is a quadratic polynomial in x with discriminant $4(\mathbf{u} \cdot \mathbf{v})^2 - 4||\mathbf{u}||\,||\mathbf{v}||$. So if $\mathbf{u}$ and $\mathbf{v}$ are not scalar multiples of each other, this quadratic polynomial has no real roots, so its discriminant is negative: that is, $(\mathbf{u} \cdot \mathbf{v})^2 \leq ||\mathbf{u}||\,||\mathbf{v}||$. If $\mathbf{u} = -x_0\mathbf{v}$, then this polynomial has exactly one real root at $x = x_0$, so its discriminant is zero.

3. It satisfies the *triangle inequality* $||\mathbf{u} + \mathbf{v}|| \leq ||\mathbf{u}|| + ||\mathbf{v}||$. Proof: $||\mathbf{u} + \mathbf{v}||^2 = (\mathbf{u} + \mathbf{v}) \cdot (\mathbf{u} + \mathbf{v}) = ||\mathbf{u}||^2 + ||\mathbf{v}||^2 + 2(\mathbf{u} \cdot \mathbf{v})$ and $(||\mathbf{u}|| + ||\mathbf{v}||)^2 = ||\mathbf{u}||^2 + ||\mathbf{v}||^2 + 2||\mathbf{u}||\,||\mathbf{v}||$, and $\mathbf{u} \cdot \mathbf{v} \leq ||\mathbf{u}||\,||\mathbf{v}||$ by Cauchy–Schwartz.

In high school algebra or physics, you may have learned, but may not have seen a proof, that the dot product is a measure of the angle between two vectors: if $\mathbf{u}, \mathbf{v}$ are two vectors and θ is the angle between them, then $\mathbf{u} \cdot \mathbf{v} = ||\mathbf{u}|| \, ||\mathbf{v}|| \cos\theta$. This fact is not obvious from the definition of dot product; it would be good to prove it.

To prove it, we'll first need some results about a general class of operators on $\mathbb{R}^n$ called *orthogonal* operators. An operator $T \in \mathrm{End}(\mathbb{R}^n)$ is defined as *orthogonal* if it preserves the dot product: that is, if $(T\mathbf{u}) \cdot (T\mathbf{v}) = \mathbf{u} \cdot \mathbf{v}$ for all pairs of vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$.

Proposition. *Orthogonal operators are bijective.*

Proof. If T is not bijective, then $(T\mathbf{u}) \cdot (T\mathbf{u}) = 0$ but $\mathbf{u} \cdot \mathbf{u} > 0$ for any nonzero vector $\mathbf{u} \in \ker T$, so T is not orthogonal.

□

Proposition. *If T is orthogonal, then so is T^{-1}.*

Proof. Take arbitrary vectors $\mathbf{u}, \mathbf{v} \in V$. Then since T is orthogonal, so $T(T^{-1}\mathbf{u}) \cdot T(T^{-1}\mathbf{v}) = T^{-1}\mathbf{u} \cdot T^{-1}\mathbf{v}$. But since $T \circ T^{-1}$ is the identity, so $T(T^{-1}\mathbf{u}) \cdot T(T^{-1}\mathbf{v}) = \mathbf{u} \cdot \mathbf{v}$. So $T^{-1}\mathbf{u} \cdot T^{-1}\mathbf{v} = \mathbf{u} \cdot \mathbf{v}$, so T^{-1} is orthogonal.

□

Lemma. *Suppose that T is a linear operator that preserves dot products of the standard basis vectors: that is, $(T\mathbf{e}_i) \cdot (T\mathbf{e}_j)$ is 0 if $i \neq j$ and 1 if $i = j$. Then the matrix representation M of T with respect to the standard basis is the inverse of its own transpose (that is, $M^T = M^{-1}$), and T is an orthogonal operator (that is, it preserves dot products of all vectors, not just the standard basis).*

Proof. If $\mathbf{a}, \mathbf{b} \in \mathrm{Col}_n(\mathbb{R})$ are column vector representations of arbitrary vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ with respect to the standard basis, then $\mathbf{u} \cdot \mathbf{v} = \mathbf{a}^T\mathbf{b}$ and $(T\mathbf{u}) \cdot (T\mathbf{v}) = (M\mathbf{a})^T(M\mathbf{b}) = \mathbf{a}^T M^T M\mathbf{b}$. By hypothesis, the equation $\mathbf{a}^T\mathbf{b} = \mathbf{a}^T M^T M\mathbf{v}$ must hold when $\mathbf{a}$ and $\mathbf{b}$ are the standard basis vectors i and j, in which case $\mathbf{a}^T\mathbf{b}$ is 1 if $i = j$ and 0 if $i \neq j$, and $\mathbf{a}^T M^T M\mathbf{b}$ is the entry at position (i, j) of $M^T M$. So $M^T M$ is the identity matrix I (that is, $M^T = M^{-1}$), and $(T\mathbf{u}) \cdot (T\mathbf{v}) = (M\mathbf{a})^T M\mathbf{b} = \mathbf{a}^T\mathbf{b} = \mathbf{u} \cdot \mathbf{v}$, so T is an orthogonal operator.

□

We'll call a matrix M *orthogonal* if it represents an orthogonal operator with respect to the standard basis. We have a way to characterize orthogonal matrices purely in terms of their entries. Remember that the columns of M are representations of the images $T\mathbf{e}_1, \ldots, T\mathbf{e}_n$ of the standard basis vectors, and we just showed that T is orthogonal if and only if thse vectors $T\mathbf{e}_1, \ldots, T\mathbf{e}_n$ all have norm 1 and dot products 0 with each other. Thus, M is an orthogonal matrix if and only if its columns all have norm 1 and the dot product of any column with any other is zero. (We're extending the dot product from $\mathbb{R}^n$ to column vectors $\mathrm{Col}_n(\mathbb{R})$ in the obvious way: just identify $\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathrm{Col}_n(\mathbb{R})$ with $(x_1, \ldots, x_n) \in \mathbb{R}^n$.)

Finally, some more vocabulary: a set of vectors in $\mathbb{R}^n$ (or similar spaces such as $\mathrm{Col}_n(\mathbb{R})$) is *orthogonal* if the dot product of any two distinct vectors is zero. The set is *orthonormal* if it's orthogonal and every element also has norm 1. There's some unfortunate terminological skew here: a matrix is ortho*gonal* if its columns are ortho*normal*.

Matrices with merely orthogonal columns don't have a special name, and don't have special properties. For instance, if M is a matrix with orthogonal but not orthonormal columns, then not only does M^T not generally equal M^{-1}, but M and M^T don't necessarily even commute. For one example, consider $M = \begin{bmatrix} 2 & 1 \\ 2 & -1 \end{bmatrix}$; in this case, $MM^T = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$ and $M^TM = \begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix}$.

One more simple but important result:

Corollary. *The columns of a matrix M are orthonormal if and only if its rows are.*

Proof. If M's columns are orthonormal, then M represents an orthogonal operator, so $M^T = M^{-1}$ represents the inverse operator, which must also be orthogonal, and the columns of M^T are the rows of M.

Conversely, if M's rows are orthonormal, then M^T's columns are orthonormal, so M^T represents an orthogonal operator and $(M^T)^T = M$ represents the inverse operator, so M's columns are orthonormal.

□

One vital subclass of orthogonal operators is the set of what we'll call *primitive rotations* (this is not a standard term). These are the generalizations of the linear operator $T : \mathbb{R}^2 \to \mathbb{R}^2$ that rotates the Cartesian plane by an angle θ counterclockwise around the origin, sending $(1, 0)$ to $(\cos\theta, \sin\theta)$ and $(0, 1)$ to $(-\sin\theta, \cos\theta)$ (relative to the standard basis, T has matrix representation $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$).

The higher-dimensional analogues, which we'll denote $R_{ij}(\theta)$ for some implicit dimension n, rotates the plane spanned by $\mathbf{e}_i$ and $\mathbf{e}_j$ while leaving other dimensions fixed. The matrix representations of these operators have entries of 1 along the diagonal except at positions (i, i) and (j, j), where they have entry $\cos\theta$; they further have $-\sin\theta$ at position (i, j) and $\sin\theta$ at position (j, i).

For $n = 5$, for example, the operator $R_{24}(\theta)$ has form

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \cos\theta & 0 & -\sin\theta & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & \sin\theta & 0 & \cos\theta & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

You should be able to convince yourself by taking dot products of the columns of this matrix (which, recall, represent the images $R_{24}(\theta)\mathbf{e}_1, \ldots, R_{24}(\theta)\mathbf{e}_n$ of the standard basis vectors) that $R_{24}(\theta)$ is an orthonormal operator, and so are all the primitive rotation operators. I'll also ask you to accept that rotation operators also preserve the angle between any two vectors.[2] Now to the main result:

Theorem. *If $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ are nonzero vectors separated by an angle θ, then $\mathbf{u}\cdot\mathbf{v} = ||\mathbf{u}||\,||\mathbf{v}||\cos\theta$.*

[2]It turns out that every orthonormal operator can be decomposed as a product of primitive rotation operators times, optionally, the reflection operator $T(a_1, a_2, a_3, \ldots, a_n) = (-a_1, a_2, a_3, \ldots, a_n)$, but this result would take a bit of effort and we won't really need it.

Proof. If we replace $\mathbf{u}$ and $\mathbf{v}$ by any nonzero scalar multiples $a\mathbf{u}, b\mathbf{v}$, then the truth of the equation $\mathbf{u} \cdot \mathbf{v} = ||\mathbf{u}|| \, ||\mathbf{v}|| \cos\theta$ doesn't change, because $(a\mathbf{u}) \cdot (b\mathbf{v}) = ab(\mathbf{u} \cdot \mathbf{v})$ and $||a\mathbf{u}|| \, ||b\mathbf{v}|| \cos\theta = ab||\mathbf{u}|| \, ||\mathbf{v}|| \cos\theta$. So it's enough to prove that if $||\mathbf{u}|| = ||\mathbf{v}|| = 1$, then $\mathbf{u} \cdot \mathbf{v} = \cos\theta$. (We call vectors with norm 1 *unit vectors*.)

We'll prove this case, in which $\mathbf{u}$ and $\mathbf{v}$ are unit vectors, in three steps:

1. Prove the special case $n = 2, \mathbf{u} = \mathbf{e}_1, ||\mathbf{v}|| = 1$.

2. Prove the case $n = 2$ and $||\mathbf{u}|| = ||\mathbf{v}|| = 1$. We'll do this by finding a primitive rotation operator $R : \mathbb{R}^2 \to \mathbb{R}^2$ such that $R\mathbf{u} = \mathbf{e}_1$. As R preserves both dot products and angles, we can thus reduce this case to that of step 1.

3. Prove the general case $||\mathbf{u}|| = ||\mathbf{v}|| = 1$ for dimension $n \geq 3$. We'll do this inductively by finding a product Q of primitive rotation operators such that the last components of $Q\mathbf{u}$ and $Q\mathbf{v}$ are zero: that is, $Q\mathbf{u}$ and $Q\mathbf{v}$ are in $\mathrm{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}\}$. This reduces the case for dimension n to the case for dimension $n-1$ and inductively establishes the result for all dimensions ≥ 2.

Step 1. Suppose $\mathbf{u} = \mathbf{e}_1$. Since $||\mathbf{v}|| = 1$, there has to be some ϕ in the interval $0 \leq \pi < 2\pi$ such that $\mathbf{v} = (\cos\phi, \sin\phi)$, so $\mathbf{u} \cdot \mathbf{v} = \cos\phi$. The angle θ from $\mathbf{u}$ to $\mathbf{v}$ is either ϕ counterclockwise (if $0 \leq \phi \leq \pi$) or $\theta := 2\pi - \phi$ clockwise (if $\pi < \phi < 2\pi$), and ϕ and θ have the same cosine, so $\cos\theta = \cos\phi = \mathbf{u} \cdot \mathbf{v}$.

Step 2. Choose ϕ such that $\mathbf{u} = (\cos\phi, \sin\phi)$, and choose $R = R_{12}(-\phi)$.

Step 3. Our operator Q will be the composition $Q_3Q_2Q_1$ of three primitive rotation operators:

1. Let u_1 and u_n be the first and last components of $\mathbf{u}$. If $u_n = 0$, then let Q_1 be the identity. Otherwise, define $k_1 = \sqrt{u_1^2 + u_n^2}$ and choose ϕ_1 such that $(u_1, u_n) = (k_1 \cos\phi_1, k_1 \sin\phi_1)$, and let $Q_1 = R_{1n}(-\phi_1)$.

2. Note that $Q_1\mathbf{u}$ has all the same components as $\mathbf{u}$ except possibly the first (which is k_1 instead of $k_1 \cos\phi_1$) and the last (which is 0 instead of $\sin\phi_1$). Let u_{n-1} be the second-to-last component of $\mathbf{u}$, let $k_2 = \sqrt{k_1^2 + u_{n-1}^2}$, and let ϕ_2 be such that $(k_1, u_{n-1}) = (k_2 \cos\phi_2, k_2 \sin\phi_2)$, and define $Q_2 = R_{1,n-1}(-\phi_2)$. (If $k_2 = 0$, then ϕ_2 is arbitrary.)

3. Define $\mathbf{w} = Q_2Q_1\mathbf{v}$ and let w_{n-1}, w_n be the last two components of $\mathbf{w}$. Define $k_3 = \sqrt{w_{n-1}^2 + w_n^2}$. and define ϕ_3 such that $(w_{n-1}, w_n) = (k_3 \cos\phi_3, k_3 \sin\phi_3)$, and define $Q_3 = R_{n-1,n}(-\phi_3)$ (if $k_3 = 0$, then ϕ_3 is arbitrary). The last two components of $Q_2Q_1\mathbf{u}$ are zero, so $R_{n-1,n}(\phi)$, as a rotation of dimensions $n-1$ and n of $\mathbb{R}^n$ that leaves other dimensions alone, must leave $Q_2Q_1\mathbf{u}$ unchanged.

So $Q = Q_3Q_2Q_1$, as the composition of three primitive rotation operators, must preserve the dot product, and $Q\mathbf{u}$ and $Q\mathbf{v}$ all have zero in their last entry, effectively reducing the dimension of the problem by 1.

□

In particular, if the dot product of two nonzero vectors is zero, then the angle θ between them is $90°$ (or $\pi/2$ radians). We'll call vectors with a nonzero dot product "orthogonal." This term will also apply even to vector spaces other than $\mathbb{R}^n$ that are harder to give a geometric interpretation.

One final possible terminological confusion that's important to head off early on: we call a matrix *orthogonal* if the columns are what we'll call *orthonormal*—that is, not only are the columns orthogonal to each other, but each column has norm 1. Matrices whose columns are merely orthogonal—that is, the dot product of any pair of distinct columns is zero, but the columns may not have norm 1—don't have any special name.

9.4 Sesquilinear forms and unitary matrices

Bilinear forms on complex vector spaces aren't usually very useful. The reason for this is that we we want a form for which $B(\mathbf{v}, \mathbf{v})$ is a measure of the size of $\mathbf{v}$—in particular, it should be a positive real number. So if we multiply $\mathbf{v}$ by any complex number with magnitude 1—for instance, i—then the value of $B(\mathbf{v}, \mathbf{v})$ should remain unchanged. But this is impossible if B is bilinear, as $B(i\mathbf{v}, i\mathbf{v}) = i^2 B(\mathbf{v}, \mathbf{v}) = -B(\mathbf{v}, \mathbf{v})$.

The definition that we'll use instead is *sesquilinear* forms. To be precise, a function $S : V^2 \to F$ on a complex vector space V is sesquilinear if it satisfies these requirements:

1. It is linear in the second argument: $S(\mathbf{u}, a\mathbf{v}_1 + b\mathbf{v}_2) = aS(\mathbf{v}_u, \mathbf{v}_1) + bS(\mathbf{v}_u, \mathbf{v}_2)$ for all $\mathbf{u}, \mathbf{v}_1, \mathbf{v}_2 \in V$ and $a, b \in \mathbb{C}$.

2. It is *anti*-linear in the first argument: $S(a\mathbf{u}_1 + b\mathbf{u}_2, \mathbf{v}) = \bar{a}S(\mathbf{u}_1, \mathbf{v}) + \bar{b}S(\mathbf{u}_2, \mathbf{v})$. That is, scalar factors in the first argument of S multiply the value of S by their *complex conjugates*.

Some books reverse these axioms and make the first argument linear rather than the second, but when we're working with matrix and column vector representations, having the second argument be linear turns out to be more convenient. (Linearity in the second argument is also the convention in quantum mechanics.) The prefix "sesqui-" comes from the Latin word for "one-and-a-half" and refers to the fact that sesquilinear forms are linear in the second argument and in the real part of the first argument, but not the complex part. The sesquilinear forms that we're most interested in are positive-definite forms—that is, forms for which $S(\mathbf{v}, \mathbf{v})$ is a positive real number for every nonzero vector $\mathbf{v} \in V$—but positive-definiteness isn't one of the axioms.

Like real bilinear forms, sesquilinear forms have matrix representations. Specifically, if $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis for V, then the Gram matrix of some sesquilinear form S has an entry of $B(\mathbf{v}_i, \mathbf{v}_j)$ at position (i, j). If $\mathbf{u}, \mathbf{w} \in V$ are two vectors with column vector representations $\mathbf{a}$ and $\mathbf{b}$, and M is the Gram matrix of some form S, then $S(\mathbf{u}, \mathbf{w})$ can be computed with the matrix product $\mathbf{a}^H M\mathbf{b}$. Note that we need to use the conjugate transpose $\mathbf{a}^H$ to make sure that the product $\mathbf{a}^H M\mathbf{b}$ is antilinear in $\mathbf{a}$: multiplying $\mathbf{a}$ by k means multiplying $\mathbf{a}^H M\mathbf{b}$ by $\bar{k}$.

The sesquilinear equivalent of the dot product (which we'll often notate using ordinary dot product notation) is $(z_1, \cdots, z_n) \cdot (w_1, \cdots, w_n) = \bar{z}_1 w_1 + \cdots + \bar{z}_n w_n$. Note that any complex vector dotted with itself produces a positive real number: we can consider this number to be the squared vector norm, just as with real spaces.

We'll call a matrix (or a linear operator in $\mathbb{C}^n$ represented by that matrix) *unitary* if this complex dot product equals zero for any two different columns or 1 for a column dotted with itself. All of the results that we proved about real orthogonal operators have analogues for unitary operators: for instance, an operator on $\mathbb{C}^n$ is unitary if and only if the columns of its matrix representation relative to the standard basis have norm 1 and the complex dot product of any two different columns is zero, and the inverse of a unitary matrix is also its *conjugate* transpose. We won't prove these here, but it may be a useful exercise for you to go over the proofs in the last section and make the changes required to have them apply to complex matrices.

Finally, the equivalent concept to congruence for sesquilinear forms is called **-congruence* or *star-congruence*. Remember, two real matrices J, M are congruent if they represent the same bilinear form relative to different bases, or equivalently if there's some invertible real matrix S such that $M = SJS^T$. Two complex matrices are star-congruent if there's some invertible complex matrix S such that $M = SJS^H$; this guarantees that the matrices are two representations of the same sesquilinear form.

9.5 Orthogonalization and orthogonal complements

9.5.1 The orthogonal projection operator

Throughout this section, let V be a vector field over a base field $\mathbb{F}$ (which is either $\mathbb{R}$ or $\mathbb{C}$. We'll impose a geometry on V by choosing some basis B of V and defining a bilinear (if $\mathbb{F} = \mathbb{R}$) or sesquilinear (if $\mathbb{F} = \mathbb{C}$) form whose Gram matrix with respect to B is the identity matrix. For this matrix, we'll use the special notation $\langle \mathbf{u}, \mathbf{v} \rangle$. For any two vectors $\mathbf{u}, \mathbf{v} \in B$, we have $\langle \mathbf{u}, \mathbf{v} \rangle = 1$ if $\mathbf{u} = \mathbf{v}$, and $\langle \mathbf{u}, \mathbf{v} \rangle = 0$ otherwise. We'll call this form the *inner product* on V; it's essentially a generalization of the dot product in $\mathbb{R}^n$. (If V is a finite-dimensional space, in fact, then the Gram matrix of this form with respect to to B is the identity matrix.)

Even if V is infinite-dimensional, all vectors can be written as finite sums of elements of B. So every quantity $\langle \mathbf{u}, \mathbf{v} \rangle$, for arbitrary vectors $\mathbf{u}, \mathbf{v} \in V$, can be written as

$$\langle a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n, b_1\mathbf{v}_1 + \cdots + b_n\mathbf{v}_n \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i b_j \langle \mathbf{v}_i, \mathbf{v}_j \rangle$$

where $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are elements of B. This sum equals $a_1b_1 + \cdots + a_nb_n$ if $\mathbb{F} = \mathbb{R}$, or $\bar{a}_1b_1 + \cdots + \bar{a}_nb_n$ if $\mathbb{F} = \mathbb{C}$. This means, that if $\mathbb{F} = \mathbb{R}$, then inner product is *symmetric* (that is, $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$ for all pairs of vectors $\mathbf{u}, \mathbf{v} \in V$); and if $\mathbb{F} = \mathbb{C}$, then $\langle \mathbf{u}, \mathbf{v} \rangle$ and $\langle \mathbf{v}, \mathbf{u} \rangle$ are always complex conjugates. Sesquilinear inner products whose values change to their complex conjugates when the arguments are swapped are called *Hermitian*; this is the closest that a sesquilinear form can get to being symmetrical.)

The basic problem in this section is this: suppose that we have a set of linearly independent vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$. We'd like to orthogonalize this set: that is, define a set of linearly independent and *orthogonal* vectors $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ with the same span.

It turns out, in fact, we can do this and even more: we can guarantee that the set of vectors $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ satisfies $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_j\} = \operatorname{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_j\}$ for every integer $1 \leq j \leq m$. The way we can do this is to define an algorithm called *Gram–Schmidt*

orthogonalization that uses the following *projection* operator:

$$\text{proj}_{\mathbf{u}}\,\mathbf{v} = \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{||\mathbf{u}||^2}\mathbf{u}$$

Note that $\frac{\langle \mathbf{u}, \mathbf{v}\rangle}{||\mathbf{u}||^2}$ is just a scalar, so $\text{proj}_{\mathbf{u}}\,\mathbf{v}$ is a multiple of $\mathbf{u}$. Essentially, we're breaking $\mathbf{v}$ into two components, a component $\text{proj}_{\mathbf{u}}\,\mathbf{v}$ parallel to $\mathbf{u}$ and another component $\mathbf{v} - \text{proj}_{\mathbf{u}}\,\mathbf{v}$ orthogonal to $\mathbf{u}$ (we haven't yet proved that $\mathbf{u}$ and $\mathbf{v} - \text{proj}_{\mathbf{v}}\,\mathbf{u}$ are orthogonal, but we will), and extracting the component parallel to $\mathbf{u}$. For a clear example of this, consider the case when $\mathbf{u}$ is one of the standard basis vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ of $\mathbb{F}^n$ and the inner product is either the dot product (if $\mathbb{F} = \mathbb{R}$) or its sesquilinear equivalent $\langle (u_1, \ldots, u_n), (v_1, \ldots, v_n)\rangle = \bar{u}_1 v_1 + \cdots + \bar{u}_n v_n$ (if $\mathbb{F} = \mathbb{C}$). In this case, $\langle \mathbf{e}_i, \mathbf{e}_j\rangle$ is 1 if $i = j$ and 0 if $i \neq j$. So if $\mathbf{v} = (v_1, \ldots, v_n) = v_1\mathbf{e}_1 + \cdots + v_n\mathbf{e}_n$, then $\langle \mathbf{e}_i, \mathbf{v}\rangle = v_1 \langle \mathbf{e}_i, \mathbf{e}_1\rangle + \cdots + v_n \langle \mathbf{e}_i, \mathbf{e}_n\rangle = v_i$ and so $\text{proj}_{\mathbf{e}_i}\,\mathbf{v} = \frac{v_i}{||\mathbf{e}_i||^2}\mathbf{e}_i = \mathbf{e}_i$.

Let's prove some properties of the projection operator:

Proposition. *The coefficient* $\frac{\langle \mathbf{u}, \mathbf{v}\rangle}{||\mathbf{u}||^2}$ *attached to* $\mathbf{u}$ *in* $\text{proj}_{\mathbf{u}}\,\mathbf{v}$ *has the following properties:*

1. *It is the only scalar* $k \in \mathbb{F}$ *for which* $\mathbf{u}$ *and* $\mathbf{v} - k\mathbf{u}$ *are orthogonal.*

2. *It is the unique value* $k \in \mathbb{F}$ *that minimizes* $||\mathbf{v} - k\mathbf{u}||$.

Proof. To prove the first statement: if $\mathbf{u}$ and $\mathbf{v} - k\mathbf{u}$ are orthogonal, then we can expand $0 = \langle \mathbf{u}, \mathbf{v} - k\mathbf{u}\rangle = \langle \mathbf{u}, \mathbf{v}\rangle - k \langle \mathbf{u}, \mathbf{u}\rangle$, so $k = \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle} = \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{||\mathbf{u}||^2}$.

To prove the second statement, we'll start with the following expansion (remember that the absolute value of a complex number z is notated $|z| = \sqrt{z\bar{z}}$:

$$\begin{aligned}
||\mathbf{v} - k\mathbf{u}||^2 &= \langle \mathbf{v} - k\mathbf{u}, \mathbf{v} - k\mathbf{u}\rangle \\
&= \langle \mathbf{v}, \mathbf{v}\rangle - \langle \mathbf{v}, k\mathbf{u}\rangle - \langle k\mathbf{u}, \mathbf{v}\rangle + \langle k\mathbf{u}, k\mathbf{u}\rangle \\
&= ||\mathbf{v}||^2 - k \langle \mathbf{v}, \mathbf{u}\rangle - \bar{k} \langle \mathbf{u}, \mathbf{v}\rangle + k\bar{k}||\mathbf{u}||^2 \\
&= ||\mathbf{v}||^2 - k\overline{\langle \mathbf{u}, \mathbf{v}\rangle} - \bar{k} \langle \mathbf{u}, \mathbf{v}\rangle + k\bar{k}||\mathbf{u}||^2 \\
&= \left(k||\mathbf{u}|| - \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{||\mathbf{u}||}\right)\left(\bar{k}||\mathbf{u}|| - \frac{\overline{\langle \mathbf{u}, \mathbf{v}\rangle}}{||\mathbf{u}||}\right) + ||\mathbf{v}||^2 - \frac{|\langle \mathbf{u}, \mathbf{v}\rangle|^2}{||\mathbf{u}||^2} \\
&= \left|k||\mathbf{u}|| - \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{||\mathbf{u}||}\right|^2 + ||\mathbf{v}||^2 - \frac{|\langle \mathbf{u}, \mathbf{v}\rangle|^2}{||\mathbf{u}||^2}
\end{aligned}$$

(We're essentially writing $||\mathbf{v} - k\mathbf{u}||^2$ as a sort of quadratic equation in k and then completing the square.) Minimizing this expression means minimizing $\left|k||\mathbf{u}|| - \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{||\mathbf{u}||}\right|$, the only term in it that depends on k, and the minimum value of this term is zero, achieved when $k = \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{||\mathbf{u}||^2}$.

□

9.5.2 Gram–Schmidt orthogonalization

We can use the projection operator in a procedure that, given any linearly independent set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, produces an orthogonal set $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ such that $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\} = \operatorname{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ for every integer $1 \leq k \leq n$.

The procedure, called *Gram–Schmidt orthogonalization*, works like this:

1. Define $\mathbf{w}_1 = \mathbf{v}_1$.
2. Define $\mathbf{w}_2$ by removing the component of $\mathbf{v}_2$ parallel to $\mathbf{w}_1$: that is, $\mathbf{w}_2 = \mathbf{v}_2 - \operatorname{proj}_{\mathbf{w}_1} \mathbf{v}_2$. Note that $\operatorname{proj}_{\mathbf{w}_1} \mathbf{v}_2$ is a scalar multiple of $\mathbf{w}_1$, so $\mathbf{w}_2 \in \operatorname{span}\{\mathbf{w}_1, \mathbf{v}_2\}$.
3. Define $\mathbf{w}_3$ removing the components of $\mathbf{v}_3$ parallel to $\mathbf{w}_1$ and $\mathbf{w}_2$: that is, $\mathbf{w}_3 = \mathbf{v}_3 - \operatorname{proj}_{\mathbf{w}_1} \mathbf{v}_1 - \operatorname{proj}_{\mathbf{w}_2} \mathbf{v}_2$. As before, $\mathbf{w}_3 \in \operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2, \mathbf{v}_3\}$.
4. Continue in the same vein, recursively defining $\mathbf{w}_k = \mathbf{v}_k - \operatorname{proj}_{\mathbf{w}_1} \mathbf{v}_k - \cdots - \operatorname{proj}_{\mathbf{w}_{k-1}} \mathbf{v}_k$.

Once we have an orthogonal basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$, we can define an orthonormal basis $\{\mathbf{n}_1, \ldots, \mathbf{n}_n\}$ simply by rescaling: $\mathbf{n}_k = \dfrac{\mathbf{w}_k}{||\mathbf{w}_k||}$. The unitary matrix U whose columns are the coefficients of $\mathbf{n}_1, \ldots, \mathbf{n}_n$ relative to the basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ can be interpreted two ways: first, as a rotation matrix that sends $\mathbf{v}_i$ to $\mathbf{n}_i$; second, as a change-of-basis matrix that translates from representations relative to $\{\mathbf{n}_1, \ldots, \mathbf{n}_n\}$ to $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$.

9.5.3 Orthogonal complements

One final note. Suppose we have a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of a space V of which the first k vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are a basis of some subspace W. By using Gram–Schmidt orthogonalization, we can get another basis of orthogonal vectors $\mathbf{v}'_1 = \mathbf{v}_1, \mathbf{v}'_2 = \mathbf{v}_2 - \operatorname{proj}_{\mathbf{v}'_1} \mathbf{v}_2, \mathbf{v}'_3 = \mathbf{v}_3 - \operatorname{proj}_{\mathbf{v}'_1} \mathbf{v}_3 - \operatorname{proj}_{\mathbf{v}'_2} \mathbf{v}_3, \ldots$ such that $\mathbf{v}'_1, \ldots, \mathbf{v}'_k$ is still a basis of W. The remaining vectors $\mathbf{v}'_{k+1}, \ldots, \mathbf{v}'_n$ form the basis for a space all of whose elements are orthogonal to every element of W—remember that we can expand the inner product

$$\left\langle c_1\mathbf{v}'_1 + \cdots + c_k\mathbf{v}'_k, c_{k+1}\mathbf{v}'_{k+1} + \cdots + c_n\mathbf{v}'_n \right\rangle$$

into a sum of terms $\bar{c}_i c_j \langle \mathbf{v}_i, \mathbf{v}_j \rangle$ where $i \leq k$ and $j > k$, so $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$. We'll denote the span of $\{\mathbf{v}'_{k+1}, \ldots, \mathbf{v}'_n\}$ by $W^\perp$, and note that its dimension is $\dim V - \dim W$. In particular, if W is a proper subspace of V, then $W^\perp$ must contain at least one nonzero vector.

The space $W^\perp$, in fact, must contain *every* vector in V that is orthogonal to every vector in W: if some vector $\mathbf{u} = c_1\mathbf{v}'_1 + \cdots + c_n\mathbf{v}'_n$ is orthogonal to every element in W, then it must in particular be orthogonal to $\mathbf{v}'_i$ for $1 \leq i \leq k$, so $\langle \mathbf{v}'_i, \mathbf{u} \rangle = c_i = 0$. This space $W^\perp$ is called the *orthogonal complement* of W, and it's also the kernel of the *orthogonal projection* operator $\operatorname{proj}_W(c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n) = c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k$ that projects a vector onto its "shadow" in a multidimensional subspace. Analogously to projections onto single-dimensional subspaces, you can prove that $\operatorname{proj}_W \mathbf{v}$ is the unique element $\mathbf{w} \in W$ such that $\mathbf{w}$ and $\mathbf{v} - \mathbf{w}$ are perpendicular, as well as the element that minimizes $||\mathbf{v} - \mathbf{w}||$.

9.6 Unitary triangularization

We showed back in 8.6 that every operator $T : V \to V$ on a complex finite-dimensional vector space V can be given an upper triangular matrix form—that is, we can find some basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of V such that $T\mathbf{v}_i \in \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_i\}$ for $1 \leq i \leq n$.

Now let's suppose that V is $\mathbb{C}^n$, so we have the sesquilinear dot product as an inner product on V. Let T be some operator on $\mathbb{C}^n$ and let M be its matrix representation with respect to to the standard basis. Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be a basis of $\mathbb{C}^n$ with respect to which T has an upper triangular form, and let $\mathbf{w}_1, \ldots, \mathbf{w}_n$ be the basis of V derived from applying Gram–Schmidt orthogonalization to $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and then rescaling every vector to have norm 1; that is, $\mathbf{w}_i \cdot \mathbf{w}_j$ is either 1 if $i = j$ or 0 otherwise.

The basis $\mathbf{w}_1, \ldots, \mathbf{w}_n$ also has the property that $T\mathbf{w}_i \in \operatorname{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_i\}$, because the Gram–Schmidt process guarantees that $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_i\} = \operatorname{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_i\}$. Furthermore, the change-of-basis matrix U that translates from column vector representations relative to $\mathbf{w}_1, \ldots, \mathbf{w}_n$ to column vector representations relative to the standard basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ is the matrix with $\mathbf{w}_1, \ldots, \mathbf{w}_n$ written as columns. That is, U has orthonormal columns, so it is a unitary matrix. The inverse matrix U^{-1} translates the other way, and for unitary matrices, $U^{-1} = U^H$, giving us a matrix factorization $M = U\Gamma U^H$ where Γ is upper triangular and M is unitary. That is, *every square complex matrix can be triangularized via a unitary change-of-basis matrix*.

This result may be interesting by itself, but it's also a key building block of the culminating result of this chapter: a set of *spectral theorems* proving that several large classes of matrices are always diagonalizable.

9.7 Symmetric forms and self-adjoint operators

First, a bit of vocabulary. Let V be a vector space over a field $\mathbb{F}$, and let $B : V^2 \to \mathbb{F}$ be a bilinear form.

Definition. *B is **symmetric** if $B(\mathbf{u}, \mathbf{v}) = B(\mathbf{v}, \mathbf{u})$ for all vectors $\mathbf{u}, \mathbf{v} \in V$.*

Remark. If B is symmetric, then the matrix representation of B relative to any basis also has to be symmetric: remember that $B(\mathbf{v}_i, \mathbf{v}_j)$ gives entry (i, j) of the Gram matrix relative to the base $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$.

Definition. *A symmetric bilinear form B is **degenerate** if there's some fixed vector $\mathbf{u} \in V$ whose product with every other vector is zero: that is, $B(\mathbf{u}, \mathbf{v}) = 0$ for all $\mathbf{v} \in V$, or (equivalently) $T(\mathbf{v}) = B(\mathbf{u}, \mathbf{v})$ is the zero map from V to $\mathbb{F}$. Since B is symmetric, taking the second argument fixed and the first variable gives us an equivalent criterion: B is degenerate if there's some fixed $\mathbf{v}$ such that $\mathbf{u} \mapsto B(\mathbf{u}, \mathbf{v})$ is the zero map. A symmetric bilinear form that is not degenerate is called, naturally enough, **nondegenerate**.*

You can tell whether B is degenerate or nondegenerate simply by computing the rank of its Gram matrix, thanks to the following proposition:

Proposition. *If $B : V^2 \to \mathbb{R}$ is a symmetric bilinear form on a finite-dimensional real vector space V, then its Gram matrix relative to an arbitrary basis of V is invertible if and only if B is nondegenerate.*

Proof. Let S be an arbitrary basis of V, let $n = \dim V$, and let M_B be the Gram matrix of B relative to S. We will prove two claims:

1. *If B is degenerate, then M_B is noninvertible.* Suppose $\mathbf{v} \in V$ is some fixed nonzero vector such that $B(\mathbf{u}, \mathbf{v}) = 0$ for all $\mathbf{u} \in V$. Let $\mathbf{a}, \mathbf{b} \in \mathrm{Col}_n(\mathbb{R})$ be representations relative to S of the arbitrary vector $\mathbf{u}$ and the fixed vector $\mathbf{v}$. Then $\mathbf{a}^T M_B \mathbf{b} = 0$ for all $\mathbf{a} \in \mathrm{Col}_n(\mathbb{R})$. In particular, choosing $\mathbf{a} = M_B\mathbf{b}$ gives $(M_B\mathbf{b})^T M_B\mathbf{b} = 0$. But this expression is the squared norm of $M_B\mathbf{b}$; i.e. if $M_B\mathbf{b} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, then $(M_B\mathbf{b})^T M_B\mathbf{b} = x_1^2 + \cdots + x_n^2$. If this expression is zero, then $M_B\mathbf{b} = \mathbf{0}_{\mathrm{Col}_n(\mathbb{R})}$; that is, $\mathbf{b}$ is a nonzero element of $\mathrm{nullsp}\, M_B$, so M_B can't be invertible.

2. *If M_B is noninvertible, then B is degenerate.* Let $\mathbf{b}$ be some nonzero element of the nullspace of M_B: that is, $M_B\mathbf{b} = \mathbf{0}_{\mathrm{Col}_n(\mathbb{R})}$ and so $\mathbf{a}^T M_B\mathbf{b} = 0$ for all $\mathbf{a} \in \mathrm{Col}_n(\mathbb{R})$. Every element $\mathbf{u} \in V$ corresponds to some $\mathbf{a} \in \mathrm{Col}_n(\mathbb{R})$, so if we let $\mathbf{v} \in V$ be the vector with representation $\mathbf{b}$, then $\mathbf{a}^T M_B\mathbf{b} = 0$ for all $\mathbf{a} \in \mathrm{Col}_n(\mathbb{R})$, then $B(\mathbf{u}, \mathbf{v}) = 0$ for all $\mathbf{u} \in V$, and B is degenerate.

□

Now suppose $B : V^2 \to \mathbb{R}$ is a symmetric bilinear form on a real finite-dimensional vector space. (We'll later define an analogous notion for complex vector spaces.) Let $T : V \to V$ and $T^\dagger : V \to V$ be two linear operators. We'll say that $T^\dagger$ is an *adjoint* of T if the equation $B(\mathbf{u}, T\mathbf{v}) = B(T^\dagger\mathbf{u}, \mathbf{v})$ for all pairs of vectors $\mathbf{u}, \mathbf{v} \in V$.

If B is the dot product with respect to some basis S—that is, its matrix representation relative to S is the identity—then you can get an adjoint of T by writing its Gram matrix relative to S, transposing it, and interpreting the result as a Gram matrix also relative to S. How come? If $M_B, M_T, M_T^\dagger$ are the matrix representations relative to S of $B, T, T^\dagger$ respectively, then the equation $B(\mathbf{u}, T\mathbf{v}) = B(T^\dagger\mathbf{u}, \mathbf{v})$ for all elements $\mathbf{u}, \mathbf{v} \in V$ translates into the matrix equation $\mathbf{a}^T M_B M_T \mathbf{b} = (M_T^\dagger\mathbf{a})^T M_B\mathbf{b}$ for all column vectors $\mathbf{a}, \mathbf{b} \in \mathrm{Col}_n(\mathbb{R})$. That is, $M_T^\dagger$ must satisfy $M_B M_T = (M_T^\dagger)^T M_B$, so if $M_B = I$, then $M_T^\dagger = M_T^T$.

We can also use matrices to get one adjoint if B is not the dot product (i.e. if M_B is not the identity), but M_B is still invertible. In this case, the equation $\mathbf{a}^T M_B M_T \mathbf{b} = (M_T^\dagger\mathbf{a})^T M_B\mathbf{b}$ is still satisfied if $M_T^\dagger = (M_B M_T M_B^{-1})^T$. This gives us a construction for an adjoint relative to any bilinear form with an invertible Gram matrix—that is, relative to any nondegenerate bilinear form.

We've referred to $T^\dagger$ as *an* adjoint of T, not *the* adjoint, but it turns out that if B is nondegenerate, this caution is unnecessary: every operator has one and only one adjoint relative to B.

Proposition. *If B is a nondegenerate symmetric bilinear form on a finite-dimensional real vector space V, then every linear operator $T : V \to V$ has exactly one adjoint relative to B.*

Proof. We know that transposing T's Gram matrix gives one adjoint. To prove that the adjoint is unique, suppose that $T_1^\dagger$ and $T_2^\dagger$ both satisfy $B(T_1^\dagger\mathbf{u}, \mathbf{v}) = B(T_2^\dagger\mathbf{u}, \mathbf{v}) = B(\mathbf{u}, T\mathbf{v})$ for all vector pairs $\mathbf{u}, \mathbf{v} \in V$. If $T_1^\dagger$ and $T_2^\dagger$ are distinct maps, then there's some $\mathbf{u}$ such that $T_1^\dagger\mathbf{u} \neq T_2^\dagger\mathbf{u}$. Then $0 = B(T_1^\dagger\mathbf{u}, \mathbf{v}) - B(T_2^\dagger\mathbf{u}, \mathbf{v}) = B(T_1^\dagger\mathbf{u} - T_2^\dagger\mathbf{u}, \mathbf{v})$ for every $\mathbf{v} \in V$ by linearity of B in the first argument. But then B is degenerate, a contradiction. □

If B is degenerate, on the other hand, then operators can have multiple adjoints relative to B. In the simplest case, B is the zero form $B(\mathbf{u}, \mathbf{v}) = 0$ for all $\mathbf{u}, \mathbf{v} \in V$. In this case, $B(T^\dagger \mathbf{u}, \mathbf{v}) = B(\mathbf{u}, T\mathbf{v})$ for any pair of operators $T, T^\dagger \in \mathrm{End}(V)$, so every operator is an adjoint of every operator.

We'll usually be interested in bilinear forms B that can be represented relative to some basis as the identity matrix. Relative to such a form B, an operator is *self-adjoint*—that is, it is its own adjoint—if its matrix representation with respect to the same basis is symmetric.

For complex vector spaces, the analogous concepts are *Hermitian* sesquilinear products; that is, those for which $S(\mathbf{u}, \mathbf{v})$ and $S(\mathbf{v}, \mathbf{u})$ are always complex conjugates of each other. Gram matrices of sesquilinear products are Hermitian, and an operator and its adjoint with respect to a sesquilinear form have matrix representations that are conjugate transposes of each other.

One final note that may be useful in higher mathematics and in quantum physics: an adjoint in the sense that we've defined it may not exist for infinite-dimensional vector spaces. Consider, for example, the space $\mathbb{R}^\infty$ of infinite sequences of real numbers that have only a finite number of nonzero entries, and let B be the dot product $(x_1, x_2, \ldots) \cdot (y_1, y_2, \ldots) = x_1 y_1 + x_2 y_2 + \cdots$ (this sum only has a finite number of nonzero terms, so it always converges). Let $T \in \mathrm{End}(\mathbb{R}^\infty)$ be the operator $T(x_1, x_2, x_3, \ldots) = (x_1 + x_2 + x_3 + \cdots, 0, 0, \ldots)$. Then $(x_1, x_2, \ldots) \cdot T(y_1, y_2, \ldots) = x_1(y_1 + y_2 + \cdots)$, so the adjoint $T^\dagger$ with respect to B could only have the formula $T^\dagger(x_1, x_2, x_3, \ldots) = (x_1, x_1, x_1, \ldots)$. But $(x_1, x_1, x_1, \ldots)$ has an infinite number of nonzero entries if $x_1 \neq 0$, so it is not an element of $\mathbb{R}^\infty$.

In real analysis, adjoint operators are defined as operators on the *dual space* V^* of V, which is the set of linear functions from V to its base field $\mathbb{F}$. For instance, the map $T : \mathbb{R}^\infty \to \mathbb{F}$ with the formula $f(y_1, y_2, y_3, \ldots) = x_1 y_1 + x_2 y_2 + x_3 y_3 + \cdots$, where $x_1, x_2, x_3, \ldots$ are fixed coefficients that could all be infinite, is an element of the dual space of $\mathbb{R}^\infty$, and we can identify this map with the infinite sequence $(x_1, x_2, x_3, \ldots) \in \mathbb{R}^\mathbb{N}$. (Remember: $\mathbb{R}^\mathbb{N}$ contains infinite sequences with a potentially infinite number of nonzero entries.) The adjoint of an operator $T : V \to V$ is the operator $T^\dagger : V^* \to V^*$ defined as $T^\dagger(f) = f \circ T$: that is, $(T^\dagger(f))(\mathbf{v}) = f(T\mathbf{v})$ for every vector $\mathbf{v} \in V$ and element $f : V \to \mathbb{F}$ of the dual space. In finite-dimensional vector spaces, we can use the bilinear form B to establish an isomorphism between V and V^* that identifies every vector $\mathbf{u} \in V$ with the map $\mathbf{v} \mapsto B(\mathbf{u}, \mathbf{v})$, and this identification lets us define "adjoints" as operators on V, not V^*.

9.8 Normal matrices and the finite-dimensional spectral theorem

Spectral theorems (sometimes referred to in the singular as *the spectral theorem*, because spectral theorems in different contexts are quite similar) are results guaranteeing the existence of eigenvectors for certain special matrices or linear operators. The results on matrices also guarantee some important results for the structure of bilinear forms.

For finite-dimensional vector spaces, the most important spectral theorem concerns one special class of matrices:

Definition. *A square matrix M with complex (possibly all real) entries is **normal** if it commutes with its conjugate transpose: that is, $MM^H = M^H M$.*

Remark. As $(M^H)^H = M$, so if M is normal, then so is M^H.

The word "normal" is a bit misleading, because most matrices aren't normal. Several important classes of matrices, however, are normal:

1. Diagonal matrices: if M is diagonal, then MM^H and $M^H M$ are also diagonal, and their diagonal entries are the squared absolute values of the corresponding entries of M.

2. Hermitian matrices. If $M = M^H$, then $M^H M = MM^H = M^2$.

3. Real symmetric matrices, as a subclass of Hermitian matrices.

4. Skew-Hermitian matrices (i.e. matrices that equal their negative conjugate transposes). If $-M = M^H$, then $M^H M = MM^H = -M^2$.

5. Real skew-symmetric matrices, as a subclass of skew-Hermitian matrices.

6. Unitary matrices, as $U^H U = UU^H = I$.

7. Real orthogonal matrices, as a subclass of unitary matrices.

One crucial class of provably *non*-normal matrices, though, are the triangular matrices with at least one off-diagonal element..

Proposition. *No triangular non-diagonal matrix with real or complex entries is normal.*

Proof. We'll prove that any normal upper triangular matrix must be diagonal. The proof for lower triangular matrices is symmetrical, but it also follows as a corollary from the result for upper triangular matrices: M is normal if and only if M^H is normal, and every lower triangular matrix has an upper triangular conjugate transpose.

Let $R \in M_{n\times n}(\mathbb{C})$ be upper triangular, and write r_{ij} for the entry in row i and column j of R. Let α_i be the ith diagonal entry of $R^H R$ (which is also the squared norm of the ith column of R; that is, $\alpha_i = |r_{1i}|^2 + \cdots + |r_{ii}|^2$). Let β_i be the ith diagonal entry of RR^H (which is also the squared norm of the ith row of R; that is, $\beta_i = |r_{ii}| + \cdots + |r_{in}|^2$).

Suppose that R is normal: that is, $R^H R = RR^H$. Then $\alpha_i = \beta_i$ for all i. Thus:

1. $\alpha_1 = |r_{11}|^2$ and $\beta_1 = |r_{11}|^2 + |r_{12}|^2 + \cdots + |a_{1n}|^2$. So the off-diagonal first-row entries r_{12} through r_{1n} must be all zero.

2. $\alpha_2 = |r_{12}|^2 + |r_{22}|^2$ and $\beta_2 = |r_{22}|^2 + |r_{23}|^2 + \cdots + |r_{2n}|^2$. But we have already shown $r_{12} = 0$, so the off-diagonal second-row entries r_{23} through r_{2n} must also be zero. (r_{21} is also zero, of course, because it's a below-diagonal element and R is upper triangular.)

3. $\alpha_3 = |r_{13}|^2 + |r_{23}|^2 + |r_{33}|^2$ and $\beta_3 = |r_{33}|^2 + |r_{34}|^2 + \cdots + |r_{3n}|^2$. But we have already shown that $r_{13} = r_{23} = 0$, so the third row must also be all zeros except the diagonal entry r_{33}.

By continuing, we can prove that $r_{ij} = 0$ whenever $i \neq j$, so R is diagonal.

□

This lemma gives us the most important result of the section:

Theorem (Finite-dimensional complex spectral theorem)**.** *If M is a normal matrix, then M can be factored as $M = U^H \Lambda U$, where U is a unitary matrix and Λ is diagonal.*

Proof. We know that every matrix is similar to an upper triangular matrix via a unitary change-of-basis matrix (Section 9.6), so write $M = U^H R U$ where R is upper triangular and U is unitary. We'll prove that R is normal, so it has to be diagonal.

Remember that $(ABC)^H = C^H B^H A^H$ for any matrices A, B, C, and also that $U^H = U^{-1}$. So $MM^H = (U^H RU)(U^H RU)^H = (U^H RU)(U^H R^H U) = U^H RR^H U$ and $M^H M =$

$(U^H RU)^H (U^H RU) = (U^H R^H U)(U^H RU) = U^H R^H RU$. If M is normal (that is, $MM^H = M^H M$), then $U^H RR^H U = U^H R^H RU$. Multiplying both sides of this equation by U on the left and U^H on the right leaves $RR^H = R^H R$, so R must be normal as well. But the only normal triangular matrices are diagonal, so R is diagonal.

□

Translated from matrix language into operator language, this result is:

Corollary. *If $T : \mathbb{C}^n \to \mathbb{C}^n$ is an operator whose matrix representation relative to the standard basis is a normal matrix, then there is a basis of $\mathbb{C}^n$ made up entirely of eigenvectors of T that are orthogonal to one another relative to the standard sesquilinear dot product.*

A few more important corollaries:

Corollary. *For normal matrices, similarity implies star-congruence.*

Proof. First, suppose M_1, M_2 are similar normal matrices. Any pair of similar matrices, whether or not they're normal, must have the same Jordan normal form. By the spectral theorem, the Jordan normal form for a normal matrix is a diagonal matrix Λ, and we can choose the change-of-basis matrices S_1, S_2 giving $M_1 = S_1 \Lambda S_1^{-1}$ and $M_2 = S_2 \Lambda S_2^{-1}$ such that S_1 and S_2 are unitary, and $M_1 = S_1 \Lambda S_1^H$ and $M_2 = S_2 \Lambda S_2^H$. So M_1 and M_2 are star-congruent to Λ. And star-congruence is transitive, so M_1 and M_2 are star-congruent to each other.

□

Corollary. *Eigenvectors of a normal matrix with complex entries and distinct eigenvalues must be orthogonal to each other.*

Proof. Let M be a normal matrix, and diagonalize it as $M = U^H \Lambda U$ where $\mathbf{u}_1, \ldots, \mathbf{u}_n \in \mathrm{Col}_n(\mathbb{C})$ are the columns of U. These column vectors, of course, are all eigenvalues of M and orthonormal to each other, and every element of $\mathrm{Col}_n(\mathbb{C})$ can be written as a linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_n$.

Remember from page 140 that a linear combination of eigenvectors can be an eigenvector itself if and only if all the eigenvectors in the linear combination have the same eigenvalue. So if $\mathbf{v}_1, \mathbf{v}_2 \in \mathrm{Col}_n(\mathbb{C})$ are eigenvectors of M with distinct eigenvalues μ_1, μ_2, then $\mathbf{v}_1$ must be a linear combination of the subset of $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ consisting of vectors with eigenvalue μ_1, and $\mathbf{v}_2$ is a linear combination of the *necessarily disjoint* subset of $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ consisting of vectors with eigenvalue μ_2. So if we expand $\mathbf{v}_1^H \mathbf{v}_2$ into a sum of terms of the form $\mathbf{u}_i^H \mathbf{u}_j$, then none of these terms will ever have $i = j$, and $\mathbf{u}_i^H \mathbf{u}_j = 0$ if $i \neq j$. So $\mathbf{v}_1^H \mathbf{v}_2 = 0$.

□

9.9 Eigenvalues and eigenvectors of some normal matrices

In this section, we'll present a grab-bag of results about the eigenvectors of certain important classes of matrices that the preceding sections' theorems have let us establish easily. These results prove to be very important in other fields, especially many fields of physics—such as quantum mechanics and the study of rotating rigid bodies—that involve symmetrical or Hermitian matrices.

9.9.1 Hermitian and real symmetric matrices

The two most important facts about Hermitian matrices, beyond the general results of the spectral theorem, are:

1. Their eigenvalues are real.

2. Real symmetric matrices can be diagonalized by real orthogonal, not merely unitary, matrices—that is, they have an orthogonal eigenbasis of vectors with only real entries.

These results have wide application in other fields of mathematics and physics, so they're worth knowing! Now we'll present proofs.

Proposition. *The eigenvalues of a Hermitian matrix are real.*

Proof. Let H be a Hermitian (possibly real symmetric) $n \times n$ matrix, and let $\mathbf{v} \in \mathrm{Col}_n(\mathbb{C})$ be a nonzero vector such that $H\mathbf{v} = \lambda\mathbf{v}$ and, therefore, $\mathbf{v}^H H = (H^H\mathbf{v})^H = (H\mathbf{v})^H = (\lambda\mathbf{v})^H = \overline{\lambda}\mathbf{v}^H$. Then $\mathbf{v}^H(H\mathbf{v}) = \mathbf{v}^H(\lambda\mathbf{v}) = \lambda\,||\mathbf{v}||^2$ and $(\mathbf{v}^H H)\mathbf{v} = (\overline{\lambda}\mathbf{v}^H)\mathbf{v} = \overline{\lambda}\,||\mathbf{v}||^2$. But, of course, $\mathbf{v}^H(H\mathbf{v}) = (\mathbf{v}^H H)\mathbf{v}$, so $\lambda = \overline{\lambda}$: that is, λ is real.

□

Proposition. *Any real symmetric matrix can be diagonalized by a real orthogonal matrix (i.e. a unitary matrix with all real entries).*

Proof. Recall from page 206 that if a matrix M with all real entries has a real eigenvalue λ with multiplicity k, then we can choose k column vectors with all real entries as a basis for the eigenspace. Then the Gram–Schmidt process creates an orthonormal basis with equal span.

Finally, remember that that eigenvectors of any normal matrix with distinct eigenvalues must be orthogonal. So if we have orthonormal bases for each eigenspace of a real symmetric matrix M, then every element of one basis must also be orthogonal to every element of every other basis, and we can take the union of these bases to get an orthonormal basis for all of $\mathrm{Col}_n(\mathbb{R})$. Combining these column vectors into a change-of-basis matrix gives a real orthogonal matrix that diagonalizes M.

□

One important consequence of this finding is the existence of principal axes for quadratic forms. Consider the function $f : \mathbb{R}^3 \to \mathbb{R}$ given by $f(x, y, z) = \alpha x^2 + \beta y^2 + \gamma z^2 + \delta xy + \epsilon xz + \zeta yz$. A sum of products of two variables like this is called a *quadratic form*. f can be expressed in matrix form as

$$f(x, y, z) = \begin{bmatrix} x & y & z \end{bmatrix} \begin{bmatrix} \alpha & \delta/2 & \epsilon/2 \\ \delta/2 & \beta & \zeta/2 \\ \epsilon/2 & \zeta/2 & \gamma \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

The central matrix is symmetrical, so it can be diagonalized by a real orthogonal matrix that gives new orthonormal coordinates u, v, w in terms of x, y, z. The resulting quadratic form has the form $\lambda_1 u^2 + \lambda_2 v^2 + \lambda_3 w^2$; the new coordinates u, v, w are called the *principal axes* of f. The quantities $\lambda_1, \lambda_2, \lambda_3$ are simply the eigenvalues of the matrix

representation of f, and this matrix is evidently positive definite if and only if all of the quantities $\lambda_1, \ldots, \lambda_n$ are positive.[3]

The equivalent findings for quadratic forms in two variables, or in four or more, also hold, and can have interesting geometric interpretations. For instance, the generic quadratic form in two variables is $f(x,y) = \alpha x^2 + \beta xy + \gamma y^2 = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} \alpha & \beta/2 \\ \beta/2 & \gamma \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$, and equations of the form $f(x,y) = k$ for some constant k define *conic sections*. The value of the matrix determinant $\alpha\gamma - \beta^2/4$ determines the type of this conic section: if the determinant is positive, then $f(x,y) = k$ determines an ellipse (for example, $x^2 + 2y^2 = 1$, whereas if it is negative, then $f(x,y) = k$ determines a hyperbola (for example, $xy = 1$ or $x^2 - y^2 = 2$).

Finally, we have one vital corollary: a Hermitian matrix is positive definite if and only if it has all positive eigenvalues. Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be an orthonormal eigenbasis of H with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$: that is, $\mathbf{v}_i^H \mathbf{v}_j$ is 1 if $i = j$ and 0 otherwise. Then if $\mathbf{v} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$, then $\mathbf{v}^H H\mathbf{v} = |c_1|^2\lambda_1 + \cdots + |c_n|^2\lambda_n$. If all of the λ_i are positive, then this expression is also positive as long as $\mathbf{v} \neq \mathbf{0}$; if $\lambda_i \leq 0$, however, then $\mathbf{v}_i^H H\mathbf{v}_i \leq 0$ as well.

9.9.2 Skew-Hermitian and real skew-symmetric matrices

The most important finding is this:

Proposition. *The eigenvectors of any skew-Hermitian (including real skew-symmetric) matrix are purely imaginary.*

Proof. Let S be skew-Hermitian (possibly real skew-symmetric). If $S\mathbf{v} = \lambda\mathbf{v}$, then $\mathbf{v}^H S = (S^H\mathbf{v})^H = (-S\mathbf{v})^H = (-\lambda\mathbf{v})^H = -\overline{\lambda}\mathbf{v}^H$. So by a similar argument to our proof that Hermitian matrices have all real eigenvalues, $\lambda = -\overline{\lambda}$; that is, λ is purely imaginary.

□

One corollary: since non-real eigenvalues of a real matrix occur in conjugate pairs (because they're roots of the characteristic polynomial, which has all real coefficients), every real skew-symmetric matrix of odd dimension has 0 as an eigenvalue, so it does not have full rank.

9.9.3 Unitary and real orthogonal matrices

The most important result is this:

Proposition. *The eigenvalues of a unitary matrix have absolute value 1 (i.e. they're located on the complex unit circle.*

[3] In physics, this finding helps us analyze the motion of rigid objects in zero gravity or free fall. The rotational momentum of a rigid object is the product of a real symmetrical 3×3 matrix I, the object's *moment-of-inertia tensor*, with the angular rotation vector $\boldsymbol{\omega}$, which points along the object's axis of rotation and has a magnitude equal to the speed of rotation. If the rotational momentum and rotational velocity are not aligned (that is, if $\boldsymbol{\omega}$ is not an eigenvector of I), then the object's axis of rotation precesses, as in the wobble of a poorly thrown football. The spectral theorem, though, shows that any rigid body in three dimensions has three perpendicular axes about which it will rotate without precession.

Proof. If U is a unitary (possibly real orthogonal) matrix, then U has to preserve the norm of any column vector: $U\mathbf{a}$ has the same norm as $\mathbf{a}$ for any $\mathbf{a} \in \mathrm{Col}_n(\mathbb{C})$. In particular, if $\mathbf{a}$ is an eigenvector of U with eigenvalue λ, then $\mathbf{a}$ and $\lambda\mathbf{a}$ must have the same norm, so $|\lambda| = 1$.

□

9.10 Sylvester's law of inertia

Sylvester's law of inertia is a result that completely classifies symmetric bilinear (or Hermitian sesquilinear) forms up to change of basis; you can think of this as the equivalent of Jordan normal form for such bilinear or sesquilinear forms. We'll prove it for real spaces and bilinear forms; the statement and proof for complex spaces and sesquilinear forms is rather more complicated. First, a preparatory lemma.

Lemma. *Let V be a finite-dimensional real vector space and $B : V^2 \to \mathbb{R}$ a symmetric bilinear form. Call a subspace $W \subseteq V$ a* positive subspace with respect to B *if $B(\mathbf{w}, \mathbf{w}) > 0$ for every nonzero $\mathbf{w} \in W$, and a* negative subspace *if $B(\mathbf{w}, \mathbf{w}) < 0$ for every nonzero $\mathbf{w} \in W$. (By this definition, $\{\mathbf{0}\}$ is both a positive subspace and a negative subspace.) Call a positive (or negative) subspace* maximal *if no larger positive (or negative) subspace contains it.*

Then all maximal positive subspaces of V with respect to B have the same dimension, as do all maximal negative subspaces.

Proof. We'll prove that all maximal positive subspaces have the same dimension; the proof that all negative subspaces have the same dimension is practically identical.[4] Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a basis of V that gives B a diagonal Gram matrix with diagonal entries $\lambda_1, \ldots, \lambda_n$: that is, $B(\mathbf{v}_i, \mathbf{v}_i) = \lambda_i$, and $B(\mathbf{v}_i, \mathbf{v}_j) = 0$ if $i \neq j$. (By the spectral theorem, such a matrix must exist.) We can order $\mathbf{v}_1, \ldots, \mathbf{v}_n$ in descending order of eigenvalues, so there's some integer k such that $\lambda_1, \ldots, \lambda_k > 0$ and $\lambda_{k+1}, \ldots, \lambda_n \leq 0$. (If every eigenvalue is nonpositive, then $k = 0$.)

Define $W \coloneqq \mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ and $W^\perp \coloneqq \mathrm{span}\{\mathbf{v}_{k+1}, \ldots, \mathbf{v}_n\}$. We claim that W is a maximal positive subspace. Proof:

1. *W is positive*: if $\mathbf{w} = c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k \in W$, then $B(\mathbf{w}, \mathbf{w}) = c_1^2\lambda_1 + \cdots + c_k^2\lambda_k$ is a sum of non-negative terms at least one of which must be positive if $c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k \neq \mathbf{0}$.

2. *W is maximal*: if W' is a larger subspace that includes W and some additional vector $\mathbf{w}' = c_1\mathbf{v}_n + \cdots + c_n\mathbf{v}_n$ that is not in W (meaning one of the coefficients $c_{k+1}, \ldots, c_n$ is nonzero), then $\mathbf{x} \coloneqq \mathbf{w}' - c_1\mathbf{v}_n - \cdots - c_k\mathbf{v}_k = c_{k+1}\mathbf{v}_{k+1} + \cdots + c_n\mathbf{v}_n$ is a nonzero linear combination of elements of W', so it must also be in W'. But $\mathbf{x} \in W^\perp$ and $B(\mathbf{x}, \mathbf{x}) = c_{k+1}^2\lambda_{k+1} + \cdots + c_n^2\lambda_n \leq 0$, so W' is not positive.

So W is a maximal positive subspace with dimension k. To prove that all maximal subspaces have dimension k, we need two more results:

[4] It may be a useful exercise to go through this proof and make the necessary changes to apply it to negative subspaces.

Any subspace with dimension greater than k cannot be positive. Suppose U is a subspace with dimension $k+1$ or greater, and recall the subspace dimension lemma $\dim(U+W^\perp)+\dim(U\cap W^\perp) = \dim U + \dim W^\perp$ from section 5.1. Since $\dim W^\perp = n-k$ and $\dim U > k$, but $\dim(U+W^\perp) \leq \dim V = n$, it follows that $\dim(U\cap W^\perp) > 0$: that is, U shares at least one nonzero vector (call it $\mathbf{u}$) with $W^\perp$. So $B(\mathbf{u},\mathbf{u}) \leq 0$, so U cannot be positive.

Any positive subspace with dimension less than k cannot be maximal. Call two vectors $\mathbf{u},\mathbf{v}$ "orthogonal" if $B(\mathbf{u},\mathbf{v}) = 0$. Also define the operator $\text{proj}_W \mathbf{v}$ to be the result of writing $\mathbf{v}$ as a linear combination of $\mathbf{v}_1,\ldots,\mathbf{v}_n$ and changing the coefficients of $\mathbf{v}_{k+1},\ldots,\mathbf{v}_n$ to zero. Note that if $\mathbf{v} = a_1\mathbf{v}_1+\cdots+a_n\mathbf{v}_n$ is an arbitrary element of V and $\mathbf{w} = b_1\mathbf{w}_1+\cdots+b_k\mathbf{w}_k$ is an arbitrary element of W, then $B(\mathbf{v},\mathbf{w}) = a_1b_1+\cdots+a_kb_k$, and changing $a_{k+1},\ldots,a_n$ has no effect on $B(\mathbf{v},\mathbf{w})$. In particular, $B(\text{proj}_W \mathbf{v},\mathbf{w}) = B(\mathbf{v},\mathbf{w})$.

Now suppose that U is a positive subspace with dimension less than k. Let $U' := \{\text{proj}_W \mathbf{u} : \mathbf{u}\in U\} \subset W$ be the orthogonal projection of all elements of U into W. As $\dim U' \leq \dim U < \dim W$, the orthogonal complement of U' in W must have positive dimension: that is, there's some nonzero vector $\mathbf{w}\in W$ orthogonal to all of U'. As $B(\mathbf{u},\mathbf{w}) = B(\text{proj}_W \mathbf{u},\mathbf{w})$ for any $\mathbf{u}\in U, \mathbf{w}\in W$, so $\mathbf{w}$ must also be orthogonal to all of U.

So every vector in $U\oplus\text{span}\{\mathbf{w}\}$ can be written in the form $\mathbf{u}+k\mathbf{w}$ where $\mathbf{u}\in U$ and $k\in\mathbb{R}$. Thus, $B(\mathbf{u}+k\mathbf{w},\mathbf{u}+k\mathbf{w}) = B(\mathbf{u},\mathbf{u}) + 2kB(\mathbf{u},\mathbf{w}) + k^2B(\mathbf{w},\mathbf{w})$. As $B(\mathbf{u},\mathbf{u})$ and $B(\mathbf{w},\mathbf{w})$ are positive and $B(\mathbf{u},\mathbf{w}) = 0$, it follows that $U\oplus\text{span}\{\mathbf{w}\}$ is a positive subspace that strictly contains U, so U cannot be maximal.

□

Theorem (Sylvester's law of inertia for real matrices)**.** *Let J and M be two symmetric matrices with all real entries (and, necessarily, all real eigenvalues with real orthogonal eigenvectors). Then J and M are congruent if and only if they have the same number of positive, negative, and zero eigenvalues (counted up to multiplicity). That is, if J has characteristic polynomial $(x-\lambda_1)\cdots(x-\lambda_n)$ and M has characteristic polynomial $(x-\mu_1)\cdots(x-\mu_n)$, then we can order the roots λ_i and μ_i such that λ_i and μ_i are both positive, both zero, or both negative for all indices $1\leq i\leq n$.*

Proof. This theorem states an if-and-only-if result, so we need to prove two implications.

Eigenvalue condition implies congruence. First, suppose that J and M follow the stated condition on the eigenvalues, and suppose the eigenvalues λ_i and μ_i are ordered so that λ_i and μ_i are both positive, both zero, or both negative for every index i. Let Λ_J be the diagonal matrix with entries $\lambda_1,\ldots,\lambda_n$ and let Λ_M be a diagonal matrix with entries $\mu_1,\ldots,\mu_n$. Then J and Λ_J are real symmetric (and therefore normal) matrices that are similar to each other, so they are also congruent to each other (see page 237). Likewise, M and its diagonalization Λ_M are also congruent.

Finally, Λ_J and Λ_M are congruent as $\Lambda_M = S\Lambda_J S^T$, where S is a diagonal matrix whose ith diagonal entry is $\sqrt{\mu_i/\lambda_i}$ if λ_i and μ_i are either both positive or both negative and arbitrary if $\lambda_i = \mu_i = 0$. Matrix congruence is transitive, so J and M are congruent.

Congruence implies eigenvalue condition. Suppose again that J and M are congruent real symmetric matrices. Then if Λ_J and Λ_M are diagonal matrices similar (and thus congruent) to J and M respectively, then Λ_J and Λ_M must be congruent to each other. So we only need to prove that congruence implies the eigenvalue condition for diagonal matrices—whose eigenvalues, of course, are also their diagonal entries.

Suppose that $B : (\mathbb{R}^n)^2 \to \mathbb{R}$ is a bilinear form represented by Λ_J relative to the basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ (with eigenvalues $\lambda_1, \ldots, \lambda_n$) and by Λ_M relative to the basis $\mathbf{w}_1, \ldots, \mathbf{w}_n$ (with eigenvalues $\mu_1, \ldots, \mu_n$). The set of vectors $\mathbf{v}_i$ such that the corresponding eigenvalues λ_i are positive gives a basis for a maximal positive subspace of $\mathbb{R}^n$ with respect to B. (Adding any other vector to this space would mean adding a linear combination of the vectors $\mathbf{v}_i$ for which $\lambda_i \leq 0$.) Similarly, the set of vectors $\mathbf{w}_i$ such that $\mu_i > 0$ is also a basis of a maximal positive subspace. But these two maximal positive spaces must have equal dimension, so Λ_J and Λ_M must have an equal number of positive entries. Similarly, Λ_J and Λ_M must have an equal number of negative entries, and all the remaining entries must be zero.

□

As a final note, Sylvester's law of inertia also applies to complex Hermitian matrices (which also must have real eigenvalues). There is an even futher generalization for all normal matrices: the star-congruence of two normal matrices is determined by the complex *arguments* of their eigenvalues. (Remember that the argument of a complex number z is the angle $0 \leq \theta < 2\pi$ for which $z = r(\cos\theta + i\sin\theta)$ for some positive real number r.) Specifically, two normal matrices J and M are star-congruent if and only if they have the same number of zero eigenvalues and, for each angle θ, the same number of eigenvalues with argument θ. (Since positive reals have argument zero and negative reals have argument π, and real symmetric matrices are also Hermitian matrices, this means that star-congruence and regular congruence for real symmetric matrices are equivalent.)

Chapter 10

Tensor products

In section 7.4, we gave an introduction to the theory of multilinear, symmetric, and alternating linear maps. We noted in particular that multilinear maps from V^n to W are generally different from linear maps that treat V^n as a single vector space with operations extended from V.

There is, however, a way to construct another vector space X such that multilinear maps $V^n \to W$ always correspond to linear maps $X \to V$. Building this space requires a new concept called the *tensor product* which, though it requires a couple of leaps of abstraction, doesn't introduce any fundamentally new mathematics. We can also represent symmetric and alternating maps on V^n as linear maps on a certain quotient space of X.

The tensor product and the other ideas that we build in this chapter may seem like another tower of abstractions without much payoff. There are, however, a few benefits that can at least be alluded to up front:

1. We can find cleaner expressions of some definitions and results that we have already proved, such as the concept of a matrix determinant.

2. We can derive several important results about the *trace* of a matrix, or the sum of its diagonal entries (equivalently, the sum of its eigenvalues), without having to deal with matrix.

3. In quantum mechanics, tensor, symmetric, and alternating products are the natural way to represent the possible states of multi-particle systems.

With that out of the way, let's begin.

10.1 Free vector spaces

Suppose that S is a set of some elements—say, $S = \{X, Y, Z\}$. We don't know anything about X, Y, and Z: they could be any mathematical objects. Nevertheless, if we have some field $\mathbb{F}$, we can write "linear combinations" of the elements of S with coefficients taken from $\mathbb{F}$, like $aX + bY + cZ$ with $a, b, c \in \mathbb{F}$. (We'll allow any number of the coefficients to be zero: in particular, $a = b = c$ gets us the zero element.)

These linear combinations are just abstract expressions that don't have any values (mathematicians have a special name for expressions that aren't necessarily evaluable: *formal* expressions): for the purpose of constructing a free vector space, we have no idea

what aX is, nor how we could add it to bY. Nevertheless, we can turn this set of linear combinations into a vector space in a natural way: multiplication is $k(aX+bY+cZ) = (ka)X + (kb)Y + (kc)Z$, and addition is $(a_1X + b_1Y + c_1Z) + (a_2X + b_2Y + c_2Z) = (a_1 + a_2)X + (b_1 + b_2)Y + (c_1 + c_2)Z$: you're allowed to simplify a linear combination by combining terms that have the same element of S.

The set of finite linear combinations like this is called the *free vector space* on S. (Another way to think of the free vector space is as the space of functions from S to $\mathbb{F}$ with a finite number of nonzero values: the value of a function on any $x \in S$ in the function representation of the free vector space corresponds to the coefficient of x in the formal linear combination representation. For instance, $aX + bY + cZ$ corresponds to the function $f : S \to \mathbb{F}$ with $f(X) = a, f(Y) = b, f(Z) = c$. Vector space operations in the function representation of the free vector space are just pointwise operations on corresponding function values.)

Again, the elements of free vector spaces are just abstract linear combinations without reference to any values they might have: the objects of the underlying set are opaque. For instance, if S is the set of three vectors $\{(1,0),(0,1),(1,1)\}$ from $\mathbb{R}^2$, then two *distinct* elements of the free vector space on S are $2(1,0) + 3(1,1)$ and $6(1,0) + 4(0,1) - (1,1)$. Both of these linear combinations evaluate to $(5,3)$, of course, but we consider them different elements of the free vector space because they have different coefficients on corresponding elements: we're ignoring any operations that might be defined on the set S.

Free vector spaces by themselves are pretty boring. Their interest comes when we can use properties of the underlying set S to define ways in which two linear combinations can be regarded as equivalent—usually not in a completely straightforward way. We can then define a new vector space that is the *quotient* of the free vector space by the subspace of all linear combinations that are equivalent to the trivial linear combination with every coefficient zero, and this space can have more interesting properties.

10.2 Tensor product of two spaces

Key questions.

1. ($\star$) Find the flaw in the following argument: "$\mathbb{R}^2 \otimes \mathbb{R}^2$ has dimension 4, and every pure tensor has the form $(a,b)\otimes(c,d)$, where a,b,c,d are freely chosen elements of $\mathbb{R}$. So by dimensional considerations, every element of $\mathbb{R}^2 \otimes \mathbb{R}^2$ is a pure tensor."

2. ($\star\star$) Let $f : (\mathbb{R}^2)^2 \to \mathbb{R}$ be the bilinear map with formula $f((x,y),(z,w)) = xz+yw$ (this is the ordinary dot product). Give an explicit general form for elements of the kernel of the corresponding linear map $\tilde{f} : \mathbb{R}^2 \otimes \mathbb{R}^2 \to \mathbb{R}$.

10.2.1 Defined

First, a reminder of some notation. Suppose U, V, W are three vector spaces over the same field $\mathbb{F}$. The space $U \times V$ is the set of ordered pairs containing one element of U and one element of V (in that order). To express that a function f takes one input from each of U and V and returns an output in W, we write $f : U \times V \to W$. This function f is *bilinear* if:

1. f is linear in the first argument: $f(k_1\mathbf{u}_1 + k_2\mathbf{u}_2, \mathbf{v}) = k_1 f(\mathbf{u}_1, \mathbf{v}) + k_2 f(\mathbf{u}_2, v)$ for all $\mathbf{u}_1, \mathbf{u}_2 \in U$, $\mathbf{v} \in V$, and $k_1, k_2 \in \mathbb{F}$.

2. f is linear in the second argument: $f(\mathbf{u}, k_1\mathbf{v}_1 + k_2\mathbf{v}_2) = k_1 f(\mathbf{u}, \mathbf{v}_1) + k_2 f(\mathbf{u}, \mathbf{v}_2)$ for all $\mathbf{u} \in U$, $\mathbf{v}_1, \mathbf{v}_2 \in V$, and $k_1, k_2 \in \mathbb{F}$.

(This is a natural generalization of our definition in 7.4, which only considered multilinear functions in which every input came from the same vector space.)

Now let X be the free vector space on $U \times V$: its elements are formal linear combinations $k_1(\mathbf{u}_1, \mathbf{v}_1) + \cdots + k_n(\mathbf{u}_n, \mathbf{v}_n)$, where $\mathbf{u}_1, \ldots, \mathbf{u}_n, \mathbf{v}_1, \ldots, \mathbf{v}_n$ are arbitrary elements of U or V (subject only to the restriction that the same ordered pair $(\mathbf{u}, \mathbf{v})$ can't occur twice in the same linear combination). As with all free vector spaces, we are keeping these purely as linear combinations: we're not trying to evaluate them to an element of X.

We'll also adopt a bit of slightly nonstandard notation to make things a bit easier on your eyes: instead of writing $(\mathbf{u}, \mathbf{v})$ for an ordered pair from $U \times V$ (i.e. a basis element of X), write $[\mathbf{u}, \mathbf{v}]$ instead. Besides avoiding pile-ups of nested parentheses, this notation should also remind you that since we're using these ordered pairs as basis elements of a free vector space, we can't make any manipulations on them that aren't specifically allowed—in particular, we don't have standard ordered-pair operations such as addition of corresponding components $(\mathbf{u}_1, \mathbf{v}_1) + (\mathbf{u}_2, \mathbf{v}_2) = (\mathbf{u}_1 + \mathbf{u}_2, \mathbf{v}_1 + \mathbf{v}_2)$.

For an easy-to-notate example, let's take $U = \mathbb{R}^2$ and $V = \mathbb{R}^3$. Then some elements of X are $3[(1, -2), (-2, 10, \pi)] - 5[(4, 2), (-2, 10\pi)]$ (a linear combination of two ordered pairs) and $6[(0, 0), (\frac{1}{6}, \frac{3}{6}, \frac{5}{6})]$ (a linear combination of one ordered pair).

You may be instinctively looking for ways to simplify these expressions, by somehow "factoring out" the $(-2, 10, \pi)$ from $3[(1, -2), (-2, 10, \pi)] - 5[(4, 2), (-2, 10\pi)]$, or trying to distribute the 6 in $6[(0, 0), (\frac{1}{6}, \frac{3}{6}, \frac{5}{6})]$. This is a good impulse, and the purpose of the construction that we're about to define is to give a space derived from the free vector space that allows these simplifications! But in the free vector space itself, the elements of $U \times V$ are particles without any discernible internal structure: the only simplification you're allowed is combining terms that use the same ordered pair.

Now define a few subspaces of the free vector space X:

1. X_{LM} (the LM is a mnemonic for *left multiplication*) is the span of all linear combinations of the form
$$k[\mathbf{u}, \mathbf{v}] - [k\mathbf{u}, \mathbf{v}].$$
In our example space with $U = \mathbb{R}^2$ and $V = \mathbb{R}^3$, this would include elements like $2[(1, 3), (1, 0, 7)] - [(2, 6), (1, 0, -7)]$ (and its multiples such as $-8[(1, 3), (1, 0, 7)] + 4[(2, 6), (1, 0, -7)]$).

2. X_{RM} (RM = *right multiplication*) is the span of all linear combinations of the form
$$k[\mathbf{u}, \mathbf{v}] - [\mathbf{u}, k\mathbf{v}].$$

3. X_{LA} (LA = *left addition*) is the span of all linear combinations of the form
$$[\mathbf{u}_1 + \mathbf{u}_2, \mathbf{v}] - [\mathbf{u}_1, \mathbf{v}] - [\mathbf{u}_2, \mathbf{v}].$$
Again, in our example space with $U = \mathbb{R}^2$ and $V = \mathbb{R}^3$, this would include elements like $[(1, 3), (\pi, \sqrt{2}, e)] - [(0, 2), (\pi, \sqrt{2}, e)] - [(1, 1), (\pi, \sqrt{2}, e)]$.

4. X_{RA} (RA = *right addition*) is the span of all linear combinations of the form

$$[\mathbf{u}, \mathbf{v}_1 + \mathbf{v}_2] - [\mathbf{u}, \mathbf{v}_1] - [\mathbf{u}, \mathbf{v}_2].$$

•

We finally have:

Definition. *The **tensor product** $U \otimes V$ of two spaces U, V over the same field is the quotient space $X/(X_{LM} + X_{RM} + X_{LA} + X_{RA})$, with $X, X_{LM}, X_{RM}, X_{LA}, X_{RA}$ defined as above.*

Remember that the elements of a quotient space are cosets of the space in the denominator of the quotient. The point of this construction is that each of the spaces $X_{LM}, X_{RM}, X_{LA}, X_{RA}$ gives us a way to simplify elements of X to get equivalent elements in the same coset. For instance, if we have an element of X that includes a term $k[\mathbf{u}, \mathbf{v}]$, then we can replace it with $[k\mathbf{u}, \mathbf{v}]$—or vice versa. The difference between these two elements of X is $k[\mathbf{u}, \mathbf{v}] - [k\mathbf{u}, \mathbf{v}]$, which is in X_{LM} and thus in $X_{LM} + X_{RM} + X_{LA} + X_{RA}$, so both elements are part of the same coset of X—that is, they're different ways to write the same element of $U \otimes V$. Likewise, we can swap $[\mathbf{u}_1 + \mathbf{u}_2, \mathbf{v}]$ with $[\mathbf{u}_1, \mathbf{v}] + [\mathbf{u}_2, \mathbf{v}]$. These same procedures are also valid for changing the right side of an ordered pair, not the first.

We'll notate the elements of $U \otimes V$ as linear combinations $(\mathbf{u}, \mathbf{v})$ as an element of the tensor product as opposed to the free vector space, we'll use the special notation $\mathbf{u} \otimes \mathbf{v}$. This notation, remember, really represents an *affine subspace* of the free vector space, namely $[\mathbf{u}, \mathbf{v}] + (X_{LM} + X_{RM} + X_{LA} + X_{LM})$. (For shorthand, let's write $Y = X_{LM} + X_{RM} + X_{LA} + X_{RA}$.) Addition and multiplication of cosets works like in any quotient space: for instance, $(\mathbf{u}_1 \otimes \mathbf{v}_1) + (\mathbf{u}_2 \otimes \mathbf{v}_2)$ denotes the coset $[\mathbf{u}_1, \mathbf{v}_1] + [\mathbf{u}_2, \mathbf{v}_2] + Y$ (which is also $([\mathbf{u}_1, \mathbf{v}_1] + X) + ([\mathbf{u}_2, \mathbf{v}_2] + X)$ as a sum of affine spaces; see section 5.4.3 if you don't remember this), and $k(\mathbf{u} \otimes \mathbf{v})$ denotes the coset $k[\mathbf{u}, \mathbf{v}] + Y$.

Unlike the free group X, the tensor product $U \otimes V$ has operations that let us change sums of tensors to equivalent sums of other tensors. (Two expressions with tensors are equivalent if when you change every constituent pure tensor $\mathbf{u} \otimes \mathbf{v}$ back to $[\mathbf{u}, \mathbf{v}]$, you get two expressions in X that are in the same coset of Y.) Specifically, we have the relations $(k\mathbf{u}) \otimes \mathbf{v} = k(\mathbf{u} \otimes \mathbf{v})$ (that is, $[k\mathbf{u}, \mathbf{v}]$ and $k[\mathbf{u}, \mathbf{v}]$ are in the same coset of Y) as well as $(\mathbf{u}_1 + \mathbf{u}_2) \otimes \mathbf{v} = \mathbf{u}_1 \otimes \mathbf{v} + \mathbf{u}_2 \otimes \mathbf{v}$ (that is, $[\mathbf{u}_1 + \mathbf{u}_2, \mathbf{v}]$ and $[\mathbf{u}_1, \mathbf{v}] + [\mathbf{u}_2, \mathbf{v}]$ are in the same coset of Y). Similar relations hold for the right-hand sides of tensors.

But be warned that there are some tempting operations on tensors that don't actually give valid results. Component-by-component addition of tensors, for instance, doesn't work: in general, $\mathbf{u}_1 \otimes \mathbf{v}_1 + \mathbf{u}_2 \otimes \mathbf{v}_2 \neq (\mathbf{u}_1 + \mathbf{u}_2) \otimes (\mathbf{v}_1 + \mathbf{v}_2)$. (A correct formula is $(\mathbf{u}_1 + \mathbf{u}_2) \otimes (\mathbf{v}_1 + \mathbf{v}_2) = \mathbf{u}_1 \otimes \mathbf{v}_1 + \mathbf{u}_1 \otimes \mathbf{v}_2 + \mathbf{u}_2 \otimes \mathbf{v}_1 + \mathbf{u}_2 \otimes \mathbf{v}_2$.) You also can't distribute a coefficient into both sides of a tensor once: in general, $k(\mathbf{u} \otimes \mathbf{v}) \neq (k\mathbf{u}) \otimes (k\mathbf{v})$ (a correct formula would be $k^2(\mathbf{u} \otimes \mathbf{v}) = (k\mathbf{u}) \otimes (k\mathbf{v})$).

10.2.2 Simplifying tensor sums

For instance, suppose we wanted to simplify the tensor sum $\mathbf{x} = (2, 4) \otimes (1, 0, -3) + 2(0, 2) \otimes (2, 0, -6) + 3(1, 2) \otimes (1, 2, 5)$. There are a few ways we could try to do this. One way would be:

1. Factor a 2 out from the right of $2(0,2) \otimes (2,0,-6)$ to get $4(0,2) \otimes (1,0,-3)$ (the difference between these expressions is in X_{RM}). The entire expression for $\mathbf{x}$ is now $(2,4) \otimes (1,0,-3) + 4(0,2) \otimes (1,0,-3) + 3(1,2) \otimes (1,2,5)$

2. Factor the coefficient of 4 in $4(0,2) \otimes (1,0,-3)$ back into the first vector to get $(0,8) \otimes (1,0,-3)$ (the difference between these expressions is in subspace X_{LM}). We now have $\mathbf{x} = (2,4) \otimes (1,0,3) + (0,8) \otimes (1,0,-3) + 3(1,2) \otimes (1,2,5)$.

3. Combine $(2,4) \otimes (1,0,-3) + (0,8) \otimes (1,0,-3)$ into $(2,12) \otimes (1,0,-3)$ (the difference between these expressions is in X_{LA}). We now have $\mathbf{x} = (2,12) \otimes (1,0,-3) + 3(1,2) \otimes (1,2,5)$.

It's not obvious how to simplify this expression any further, and indeed, most elements of $U \otimes V$ are not so-called "pure tensors" that can be written as a single element $\mathbf{u} \otimes \mathbf{v}$. But this isn't the only way to write $\mathbf{x}$ as the sum of two pure tensors: for instance, we could have combined the first and third terms in the original expression, instead of the first and second, to get $\mathbf{x} = (1,2) \otimes (5,6,-21) + 2(0,2) \otimes (2,0,-6)$.

We can, however, prove two useful results from pure tensor manipulation:

Proposition. *The map $\iota_{U \otimes V} : U \times V \to U \otimes V$ defined as $\iota_{U \otimes V}(\mathbf{u}, \mathbf{v}) = \mathbf{u} \otimes \mathbf{v}$ is bilinear.*

Proof. Straightforward. Linearity in the first argument is $\iota_{U \otimes V}(a\mathbf{u}_1 + b\mathbf{u}_2, \mathbf{v}) = (a\mathbf{u}_1 + b\mathbf{u}_2) \otimes \mathbf{v} = a(\mathbf{u}_1 \otimes \mathbf{v}) + b(\mathbf{u}_2 \otimes \mathbf{v})$ by the left-addition and left-multiplication properties of tensors, and the same logic works for linearity in the second argument.

□

Proposition. *All pure tensors $\mathbf{u} \otimes \mathbf{v}$ where either $\mathbf{u} = \mathbf{0}_U$ or $\mathbf{v} = \mathbf{0}_V$ are equal to the zero element (i.e. additive identity) of $U \otimes V$.*

Proof. By the right-multiplication property we have $\mathbf{u} \otimes \mathbf{0}_V = \mathbf{u} \otimes (0\mathbf{v}) = 0(\mathbf{u} \otimes \mathbf{v})$ for arbitrary $\mathbf{u}, \mathbf{v}$ (and likewise $\mathbf{0}_U \otimes \mathbf{v} = 0(\mathbf{u} \otimes \mathbf{v})$), and the basic vector space axioms imply that 0 times any vector is $\mathbf{0}$.

□

Remark. This is in fact an if-and-only-if result: $\mathbf{u} \otimes \mathbf{v} = \mathbf{0}_{U \otimes V}$ if and only if $\mathbf{u} = \mathbf{0}_U$ of $\mathbf{v} = \mathbf{0}_V$. But we can't prove the only-if implication just yet.

10.2.3 Universal property of the tensor product

Motivation and definition

A natural question when we encounter any new vector space is to find its basis and dimension. It's easy enough to find a spanning set:

Proposition (Pure tensors constructed from bases of constituent spaces span the tensor product)**.** *Suppose U and V are vector spaces over the same field, with respective bases B_U and B_V. Then a spanning set for $U \otimes V$ is the set $S = \{\mathbf{u} \otimes \mathbf{v} : \mathbf{u} \in B_U, \mathbf{v} \in B_V\}$ of pure tensors constructed from an element of B_U and an element of B_V.*

Proof. If $\mathbf{u} \in U$ and $\mathbf{v} \in V$ are two tensors that can be expanded as $\mathbf{u} = a_1\mathbf{u}_1 + \cdots + a_m\mathbf{u}_m$ and $\mathbf{v} = b_1\mathbf{v}_1 + \cdots + b_n\mathbf{v}_n$, then the pure tensor $\mathbf{u} \otimes \mathbf{v}$ can be expanded as $\sum_{i=1}^{m}\sum_{j=1}^{n} a_m b_n \mathbf{u}_i \otimes \mathbf{v}$, which is a linear combination of elements in S. Every element in $U \otimes V$ is also a finite linear combination of pure tensors and, therefore, a finite linear combination of elements of S.

□

It's natural to believe that our set S should be linearly independent, as well. It turns out that S is, in fact, linearly independent, but this is harder to prove than you might think with our tools so far. We would have to prove that no possible sequence of our four allowable tensor operations could turn one of its elements into another. Instead, we'll prove that S is linearly independent as a corollary of an important theorem characterizing the tensor product that gives a close relationship between tensors and bilinear maps.

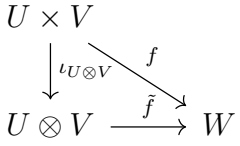
You may have noticed a suspicious resemblance between our rules for manipulating tensors and the axioms defining bilinear maps:

- The tensor manipulation rules $(k\mathbf{u} \otimes \mathbf{v}) = \mathbf{u} \otimes (k\mathbf{v}) = k(\mathbf{u} \otimes \mathbf{v}) = \mathbf{u}$ looks a lot like the axioms $f(k\mathbf{u}, \mathbf{v}) = f(\mathbf{u}, k\mathbf{v}) = k(\mathbf{u}, \mathbf{v})$ for bilinear maps $f : U \times V \to W$.
- The rules $(\mathbf{u}_1 + \mathbf{u}_2) \otimes \mathbf{v} = \mathbf{u}_1 \otimes \mathbf{v} + \mathbf{u}_2 \otimes \mathbf{v}$ and $\mathbf{u} \otimes (\mathbf{v}_1 + \mathbf{v}_2) = \mathbf{u} \otimes \mathbf{v}_1 + \mathbf{u} \otimes \mathbf{v}_2$ look a lot like the axioms $f(\mathbf{u}_1 + \mathbf{u}_2, \mathbf{v}) = f(\mathbf{u}_1, \mathbf{v}) + f(\mathbf{u}_2, \mathbf{v})$ and

In other words, operations on components of the pure tensors in a sum change the tensor's value in the same way that operations on the arguments of a bilinear function change the value of the function. (Also note that the map $(\mathbf{u}, \mathbf{v}) \mapsto \mathbf{u} \otimes \mathbf{v}$ from $U \times V$ to $U \otimes V$ is itself bilinear.)

We can formalize this observation as a precise property of the tensor product: any bilinear map on $U \times V$ can be turned into an equivalent linear map on $U \otimes V$ that gives the same value on the pure tensor $\mathbf{u} \otimes \mathbf{v}$ that the original function gives on the ordered pair $(\mathbf{u}, \mathbf{v})$. This is called a *universal property* because, it turns out, $U \otimes V$ is essentially the only space that satisfies it: any other space with the same property is isomorphic to $U \otimes V$.

Proposition (Universal property of the tensor product)**.** *Let U, V, W be vector spaces over the same field $\mathbb{F}$, and let $f : U \times V \to W$ be a bilinear map. Then there is exactly one* linear *map $\tilde{f} : U \otimes V \to W$ such that $f(\mathbf{u}, \mathbf{v}) = \tilde{f}(\mathbf{u} \otimes \mathbf{w})$, i.e. that gives the following commutative diagram with the map $\iota_{U\otimes V}(\mathbf{u}, \mathbf{v}) = \mathbf{u} \otimes \mathbf{v}$:*



Proof. This is an existence-and-uniqueness statement, so we have two tasks: prove that $\tilde{f}$ exists, and that it is unique.

Uniqueness of $\tilde{f}$ is the easy part. The pure tensors $\mathbf{u} \otimes \mathbf{v}$ are a spanning set of $U \otimes V$, and any linear map is determined by its values on a spanning set. So if $\tilde{f}_1(\mathbf{u} \otimes \mathbf{v}) = \tilde{f}_2(\mathbf{u} \otimes \mathbf{v}) = f(\mathbf{u}, \mathbf{v})$ for all $\mathbf{u} \in U, \mathbf{v} \in V$, then $\tilde{f}_1 = \tilde{f}_2$ for every input in $U \otimes V$.

To prove existence, let X be the free vector space on $U \times V$, and let X_{LM}, X_{RM}, X_{LA}, X_{RA} be the subspaces of X defined on page 245. Write $Y = X_{LM} + X_{RM} + X_{LA} + X_{RA}$,

and remember that $U \otimes V = X/Y$. Let $\pi : X \to U \otimes V$ be the projection map that takes every element of X to the coset of Y that contains it.

Remember that the general form for an element of X is $c_1[\mathbf{u}_1, \mathbf{v}_1] + \cdots + c_n[\mathbf{u}_n, \mathbf{v}_n]$ for some variable integer $n \geq 0$ and freely chosen $c_i \in \mathbb{F}, \mathbf{u}_i \in U, \mathbf{v}_i \in V$ for all $1 \leq i \leq n$. Define the map $F : X \to W$ as $F(c_1[\mathbf{u}_1, \mathbf{v}_1] + \cdots + c_n[\mathbf{u}_n, \mathbf{v}_n]) = c_1 f(\mathbf{u}_1, \mathbf{v}_1) + \cdots + c_n f(\mathbf{u}_n, \mathbf{v}_n)$. It's routine to check that F is linear, and it also satisfies $F([\mathbf{u}, \mathbf{v}]) = f(\mathbf{u}, \mathbf{v})$.

We further claim that $Y \subseteq \ker F$. To prove this, we'll prove that each of the four subspaces $X_{LM}, X_{RM}, X_{LA}, X_{RA}$ is in $\ker F$. Remember that any space that contains multiple subspaces must also contain their sum, and that to prove that a subspace is in the kernel of some linear map, it's enough to prove that every element of a spanning set of that subspace is in the kernel.

1. $X_{LM} \subseteq \ker F$: remember that X_{LM} is spanned by elements of the form $k[\mathbf{u}, \mathbf{v}] - [k\mathbf{u}, \mathbf{v}]$. On these elements, $F(k[\mathbf{u}, \mathbf{v}] - [k\mathbf{u}, \mathbf{v}]) = kf(\mathbf{u}, \mathbf{v}) - f(k\mathbf{u}, \mathbf{v})$. But this equals $\mathbf{0}_W$, because f is bilinear (and thus linear in the first argument: i.e. the partial application map $f(\cdot, \mathbf{v})$ is linear from U to W).

2. $X_{RM} \subseteq \ker F$: identical argument as for X_{LM}: f is also linear in the second argument.

3. $X_{LA} \subseteq \ker F$: on a generic element $[\mathbf{u}_1 + \mathbf{u}_2, \mathbf{v}] - [\mathbf{u}_1, \mathbf{v}] - [\mathbf{u}_2, \mathbf{v}]$ of the spanning set of X_{LA}, we have $F([\mathbf{u}_1 + \mathbf{u}_2, \mathbf{v}] - [\mathbf{u}_1, \mathbf{v}] - [\mathbf{u}_2, \mathbf{v}]) = f(\mathbf{u}_1 + \mathbf{u}_2, \mathbf{v}) - f(\mathbf{u}_1, \mathbf{v}) - f(\mathbf{u}_2, \mathbf{v})$, which must equal zero because f is linear in the first argument.

4. $X_{RA} \subseteq \ker F$: identical argument as for X_{LA}.

So $Y \subseteq \ker F$. By the first isomorphism theorem (page 131), there is some other map $\tilde{f} : U \otimes V \to W$ such that $F = \tilde{f} \circ \pi$. As π takes $[\mathbf{u}, \mathbf{v}] \in X$ to $\mathbf{u} \otimes \mathbf{v} \in U \otimes V$, it follows that $\tilde{f}(\mathbf{u} \otimes \mathbf{v}) = F([\mathbf{u}, \mathbf{v}]) = f(\mathbf{u}, \mathbf{v})$.

□

Applications

The power of the universal property of the tensor product is out of proportion to how easy it is to prove: clever use of it lets us prove results about tensor spaces that bypass tensor manipulation. We'll present two fundamental results here; the second will let us show that our spanning set S for $U \otimes V$ is, in fact, linearly independent (and thus a basis of $U \otimes V$).

Proposition (Pure tensors are zero if and only if one component is zero)**.** .

Let U, V be two vector spaces over the same field, and let $\mathbf{u} \in U, \mathbf{v} \in V$ be arbitrary. Then $\mathbf{u} \otimes \mathbf{v} = \mathbf{0}_{U \otimes V}$ if and only if $\mathbf{u} = \mathbf{0}_U$ or $\mathbf{v} = \mathbf{0}_V$.

Proof. We've already proved that if $\mathbf{u} = \mathbf{0}_U$ or $\mathbf{v} = \mathbf{0}_V$, then $\mathbf{u} \otimes \mathbf{v} = \mathbf{0}_{U \otimes V}$. So we just need to show the converse result: if both $\mathbf{u}$ and $\mathbf{v}$ are nonzero, then so is $\mathbf{u} \otimes \mathbf{v}$.

In the statement of the previous proposition, let W be the free vector space on $U \times V$, and define $f : U \times V \to W$ as follows:

1. Decompose $\mathbf{u}$ and $\mathbf{v}$ into linear combinations from B_U and B_V as $\mathbf{u} = a_1\mathbf{u}_1 + \cdots + a_m\mathbf{u}_m$ and $\mathbf{v} = b_1\mathbf{v}_1 + \cdots + b_n\mathbf{v}_n$, where the coefficients a_i, b_j are all nonzero. (If either $\mathbf{u} = \mathbf{0}_U$ or $\mathbf{v} = \mathbf{0}_V$, then correspondingly $m = 0$ or $n = 0$: B_U and B_V can have elements that don't get used in a particular linear combination.)

2. Define $f(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j [\mathbf{u}_i, \mathbf{v}_j]$: this expression has mn terms, each using a different ordered pair from $U \times V$. The crucial observation about f is that $f(\mathbf{u}, \mathbf{v})$ is the trivial, zero-term linear combination—that is, the zero element of W—if and only if either $\mathbf{u} = \mathbf{0}_U$ or $\mathbf{v} = \mathbf{0}_V$.

It's easy to check that f is bilinear (remember that we're allowed to combine terms of the form $c[\mathbf{u}_i, \mathbf{v}_j]$ in a free vector space if they use the same $\mathbf{u}_i$ and $\mathbf{v}_j$), and the universal property guarantees us some unique map $\tilde{f} : U \otimes V \to W$ such that $f(\mathbf{u}, \mathbf{v}) = \tilde{f}(\mathbf{u} \otimes \mathbf{v})$.

But if $\mathbf{u} \neq \mathbf{0}_U$ and $\mathbf{v} \neq \mathbf{0}_V$, then $f(\mathbf{u}, \mathbf{v})$ will always have at least one term, so it's not the zero element of the free vector space W. And $\tilde{f}$ is linear, so it can't take $\mathbf{0}_{U \otimes V}$ to a nonzero element of W. But $f(\mathbf{u}, \mathbf{v}) = \tilde{f}(\mathbf{u} \otimes \mathbf{v})$, so $\mathbf{u} \otimes \mathbf{v}$ can't be zero if both $\mathbf{u} \neq \mathbf{0}_U$ and $\mathbf{v} \neq \mathbf{0}_V$.

□

Proposition (Basis of the tensor space)**.** *If B_U, B_V are the bases of two vector spaces U, V over the same field, then the set*

$$S \coloneqq \{\mathbf{u} \otimes \mathbf{v} : \mathbf{u} \in U, \mathbf{v} \in V\}$$

is a basis of $U \otimes V$.

Proof. We've already proved that S spans $U \otimes V$, so we only need to prove that it's linearly independent. Let $\mathbb{F}$ be the base field of U and V and define the map $f : U \times V \to \mathbb{F}$ as follows:

1. Decompose $\mathbf{u}$ and $\mathbf{v}$ into linear combinations from B_U and B_V as $\mathbf{u} = a_1\mathbf{u}_1 + \cdots + a_m\mathbf{u}_m$ and $\mathbf{v} = b_1\mathbf{v}_1 + \cdots + b_n\mathbf{v}_n$, as in the previous proposition.
2. Define $f(\mathbf{u}, \mathbf{v}) = a_1 b_1$.

The corresponding map $\tilde{f} : U \otimes V \to \mathbb{F}$ satisfies $\tilde{f}(\mathbf{u}_1 \otimes \mathbf{v}_1) = 1$ but $\tilde{f}(\mathbf{u} \otimes \mathbf{v}) = 0$ for all other elements $\mathbf{u} \otimes \mathbf{v} \in S$. But $\ker \tilde{f}$ is a subspace of $U \otimes V$, so $\mathbf{u}_1 \otimes \mathbf{v}_1$ cannot be written as a linear combination of the other elements of S. We can repeat this logic for all other elements of S to prove that S is linearly independent. (Note that this argument also works if one or both of U and V is infinite-dimensional.)

□

Remark. We'll reuse the core idea of this proof—that is, proving that a set of pure tensors is linearly independent by constructing a map that has a nonzero value only on one of them—to prove a basis for the symmetric and alternating products, which have the same relations to the spaces of symmetric and alternating maps that the tensor space has on the space of all bilinear maps.

Uniqueness

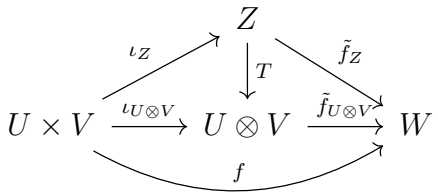
In fact, the tensor product $U \otimes V$ is essentially the only space that has this universal property of allowing arbitrary bilinear maps to be factored into unique linear maps: any other space that also has this property must have the same structure as the tensor product. What we mean by "same structure" requires careful definition an unavoidably complex theorem statement:

Theorem (Universal property uniquely defines tensor product)**.** *Let U, V, Z be vector spaces over the same field $\mathbb{F}$, let $\iota_{U\otimes V} : U \times V \to U \otimes V$ be the map $(\mathbf{u}, \mathbf{v}) \mapsto \mathbf{u} \otimes \mathbf{v}$, and let $\iota_Z : U \times V \to Z$ be a bilinear map. Suppose that for every vector space W over $\mathbb{F}$ and every bilinear map $f : U \times V \to W$, there is a unique linear map $\tilde{f}_Z : Z \to W$ such that $f = \tilde{f}_Z \circ \iota_Z$.*

Then there is a unique linear map $T : Z \to U \otimes W$ with the following properties:

1. *T is bijective.*
2. *$T \circ \iota_Z = \iota_{U\otimes V}$ (that is, $T \circ \iota_Z(\mathbf{u}, \mathbf{v}) = \mathbf{u} \otimes \mathbf{v}$ for all $\mathbf{u} \in U, \mathbf{v} \in V$).*
3. *For an arbitrary bilinear map $f : U \times V \to W$, let $\tilde{f}_{U\otimes V} : U \otimes V \to W$ be the unique linear map such that $\tilde{f}_{U\otimes V} \circ \iota_{U\otimes V} = f$. Then $\tilde{f}_Z = \tilde{f}_{U\otimes V} \circ T$.*

To put this in less formal language: Z has the same structure as $U \otimes V$, with $\iota_Z(\mathbf{u}, \mathbf{v})$ being the equivalent of the pure tensor $\mathbf{u} \otimes \mathbf{v}$, and T gives a correspondence between the structures of Z and $U \otimes V$ that also matches the pure tensor equivalents of Z to the pure tensors of $U \otimes V$. The maps defined in the theorem statement are summarized in this commutative diagram:



Proof. Two preliminary observations:

1. The image of ι_Z must span Z. Proof: otherwise, for any nontrivial vector space W, you could define distinct maps $\tilde{f}_Z, \tilde{f}'_Z : Z \to W$ that satisfy $\tilde{f} \circ \iota_Z = \tilde{f}' \circ \iota_Z$ in the following way: let B be a basis of $\operatorname{span} \operatorname{im} \iota_Z$, let C be a nonempty set of vectors that extends B to a basis of Z, and define $\tilde{f}, \tilde{f}'$ to have equal values to each other on all of B and arbitrary different values on elements of C. So the map $f : U \times V \to W$ could factor into multiple maps $\tilde{f} : Z \to W$, contradicting the hypothesis that $\tilde{f}$ is uniquely determined.

2. Property 2 of T in the theorem statement implies property 3. Proof: $\operatorname{im} \iota_Z$ spans Z, so to show that $\tilde{f}_Z = \tilde{f}_{U\otimes V} \circ T$, it's enough to show that $\tilde{f}_Z(\mathbf{z}) = \tilde{f}_{U\otimes V} \circ T(\mathbf{z})$ for every element $\mathbf{z} \in \operatorname{im} \iota_Z$.

   Choose $\mathbf{z} \in \operatorname{im} \iota_Z$ arbitrary, and let $\mathbf{u} \in U, \mathbf{v} \in V$ be such that $\iota_Z(\mathbf{u}, \mathbf{v}) = \mathbf{z}$. If $T \circ \iota_Z = \iota_{U\otimes W}$, then $T(\mathbf{z}) = \mathbf{u} \otimes \mathbf{w}$. By definition, $f = \tilde{f}_Z \circ \iota_Z = \tilde{f}_{U\otimes V} \circ \iota_{U\otimes Z}$. So $\tilde{f}_Z(\mathbf{z})$ and $\tilde{f}_{U\otimes W}(\mathbf{u} \otimes \mathbf{v}) = \tilde{f}_{U\otimes W} \circ T(\mathbf{z})$ both equal $f(\mathbf{u}, \mathbf{v})$ so they also both equal each other for arbitrary elements of $\operatorname{im} \iota_Z$, so they are equal on all of Z.

Now we'll prove that a map T exists. In the theorem statement, choose $W = U \otimes V$ and $f = \iota_{U\otimes V}$. The (necessarily unique) map $\tilde{f}_{U\otimes V} : U \otimes V \to U \otimes V$ such that $\tilde{f}_{U\otimes V} \circ \iota_{U\otimes V} = f$ must be the identity on $U \otimes V$, and there's a map $\tilde{f}_Z : Z \to U \otimes V$ such that $f = \tilde{f}_Z \circ \iota_Z$.

This map $\tilde{f}_Z$ will be our T. As $f = \iota_{U\otimes V}$, we just proved that T satisfies $T \circ \iota_Z = \iota_{U\otimes V}$, which is property 2, so T also must satisfy property 3.

If we choose $W = Z$ and $f = \iota_Z$ in the theorem statement, then by similar logic, we'll get some linear map $\tilde{f}_{U\otimes V} : U \otimes V \to Z$ such that $\iota_Z = \tilde{f}_{U\otimes V} \circ \iota_{U\otimes V}$. Call this map T'.

It remains to prove property 1: T is bijective—that is, both surjective and injective. We'll prove each of these separately:

1. *T is surjective*: In general, if f, g, h are any three functions and $f = g \circ h$, then the image of f equals the image of the restricted function $g|_{\text{im}\, h}$. Restricting a function's domain can only make its image smaller, so $\text{im}\, f \subseteq \text{im}\, g$.

   In this case, since $T \circ \iota_Z = \iota_{U\otimes V}$, so $\text{im}\, \iota_{U\otimes V} \subseteq \text{im}\, T$. And $\text{im}\, \iota_{U\otimes V}$ is the set of pure tensors, which is a spanning set of $U \otimes V$, so if $\text{im}\, T$ includes this spanning set, then it must be all of $U \otimes V$.

2. *T is injective*: We'll prove that $T'T$ is the identity map on Z. Any element of $\ker T$ must also be in $\ker T'T$, so if $\ker T'T = \{\mathbf{0}_Z\}$, then $\ker T = \{\mathbf{0}_Z\}$.

   Remember the core properties of T and T' are that $T \circ \iota_Z = \iota_{U\otimes V}$ and $T' \circ \iota_{U\otimes V} = \iota_Z$. Let $\mathbf{z}$ be any element of $\text{im}\, \iota_Z$ and choose $(\mathbf{u}, \mathbf{v}) \in U \times V$ such that $\iota_Z(\mathbf{u}, \mathbf{v}) = \mathbf{z}$. As $T \circ \iota_Z = \iota_{U\times V}$, so $T(\mathbf{z}) = T \circ \iota_Z(\mathbf{u}, \mathbf{v}) = \iota_{U\otimes V}(\mathbf{u}, \mathbf{v}) = \mathbf{u} \otimes \mathbf{v}$. Similarly, since $T' \circ \iota_{U\otimes V} = \iota_Z$, so $T' \circ \iota_{U\otimes V}(\mathbf{u}, \mathbf{v}) = T'(\mathbf{u} \otimes \mathbf{v}) = \iota_Z(\mathbf{u}, \mathbf{v}) = \mathbf{z}$.

   So $T'T\mathbf{z} = \mathbf{z}$ for every $\mathbf{z} \in \text{im}\, \iota_Z$. And $\text{im}\, \iota_Z$ spans Z, so $T'T\mathbf{z} = \mathbf{z}$ for every $\mathbf{z} \in Z$ as well.

To prove that T is unique, note that T must satisfy $T(\mathbf{z}) = \mathbf{u} \otimes \mathbf{v}$ for any $\mathbf{z} = \iota_Z(\mathbf{u}, \mathbf{v}) \in \text{im}\, \iota_Z$. So if T is uniquely determined on $\text{im}\, \iota_Z$, which spans Z, then T is uniquely determined on all of Z as well.

□

Answers to key questions.

1. Different choices of a, b, c, d can give equivalent pure tensors $(a, b) \otimes (c, d)$: for example, $(1, 2) \otimes (30, 40) = (10, 20) \otimes (3, 4)$, and $(a, b) \otimes (0, 0)$ has the same value (namely, $\mathbf{0}_{\mathbb{R}^2\otimes\mathbb{R}^2}$) no matter what a and b are.

2. The map f satisfies $f(\mathbf{e}_1, \mathbf{e}_1) = f(\mathbf{e}_2, \mathbf{e}_2) = 0$ and $f(\mathbf{e}_1, \mathbf{e}_2) = f(\mathbf{e}_2, \mathbf{e}_1) = 0$ (where as always $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$ are the standard basis vectors), so $\tilde{f}$ must satisfy $\tilde{f}(\mathbf{e}_1 \otimes \mathbf{e}_1) = \tilde{f}(\mathbf{e}_2 \otimes \mathbf{e}_2) = 1$ and $\tilde{f}(\mathbf{e}_1 \otimes \mathbf{e}_2) = \tilde{(}\mathbf{e}_1 \otimes \mathbf{e}_2) = 0$. So a general form for $\tilde{f}$ is

   $$\tilde{(}f)(a\mathbf{e}_1 \otimes \mathbf{e}_1 + b\mathbf{e}_1 \otimes \mathbf{e}_2 + c\mathbf{e}_2 \otimes \mathbf{e}_1 + d\mathbf{e}_2 \otimes \mathbf{e}_2) = a + d$$

   (remember that every element of $\mathbb{R}^2 \otimes \mathbb{R}^2$ has exactly one expression in the form $a\mathbf{e}_1 \otimes \mathbf{e}_1 + b\mathbf{e}_1 \otimes \mathbf{e}_2 + c\mathbf{e}_2 \otimes \mathbf{e}_1 + d\mathbf{e}_1 \otimes \mathbf{e}_2$), so $\ker \tilde{f}'$ is the set of elements in this form for which $d = -a$.

10.3 Tensor product of three or more spaces

The tensor product can be defined on an arbitrary number of finite vector spaces (and even for an infinite number, though this would require more complication than we'll

need). The basic procedures are the same: given vector spaces $V_1, \ldots, V_n$ over the same field, start with the free group on the set $V_1 \times \cdots \times V_n$ of ordered n-tuples with one element from each space. Then define subspaces to allow factoring out a constant from any single position in the tuple, or distributing a sum of vectors in one position into a sum of n-tuples, and take the quotient.

All the results that we proved for a tensor product of two spaces hold for arbitrarily large products:

1. The pure tensor $\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_n$ is nonzero if and only if all the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are nonzero.

2. A basis for the tensor space is given by the pure tensors constructed from taking one element from a basis of each constituent space. This means that $\dim(V_1 \otimes \cdots \otimes V_n) = \dim V_1 \cdots \dim V_n$.

3. Any multilinear map $V_1 \times \cdots \times V_n \to W$ has a unique corresponding linear map $V_1 \otimes \cdots \otimes V_n \to W$.

4. Any space Z that also allows factoring multilinear maps into linear maps is isomorphic to $V_1 \otimes \cdots \otimes V_n$ by a unique isomorphism that maps the image of the map $\iota_Z : V_1 \times \cdots \times V_n \to Z$ to the corresponding pure tensors.

The proofs are all virtually identical to the two-space case, only with uglier notation, and we won't provide them here.

10.4 Linear maps as tensors

10.4.1 Preliminary notions

A reminder of notation and some definitions. Here, V and W are vector spaces over the same field $\mathbb{F}$.

1. $\mathrm{Hom}(V, W)$ is the vector space of linear maps from V to W.

2. The *dual space* of V, which we'll denote V^*, is the space $\mathrm{Hom}(V, \mathbb{F})$ (remember that $\mathbb{F}$ is a one-dimensional vector space over itself).

In in section 2.3, we discussed that $\mathrm{Hom}(V, W)$ was a a vector space, and at least in the case that V and W are finite-dimensional, it has a readily describable basis in terms of bases of V and W. In particular, if V has basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ and W has basis $\{\mathbf{w}_1, \ldots, \mathbf{v}_n\}$, then $\mathrm{Hom}(V, W)$ has mn basis vectors.

One specialization: if W is the one-dimensional field $\mathbb{F}$, then $\mathrm{Hom}(V, \mathbb{F}) = V^*$ has a basis of the maps that take $\mathbf{v}_i$ to 1 and other basis vectors to zero. We'll call this map $\mathbf{v}_i^*$. In this case, V^* has dimension m, the same as V.[1]

[1]There is not always an obvious isomorphism betwen V and V^* if V is finite. One example: R^∞, the set of infinite sequences with a finite number of nonzero entries, has the standard basis vectors $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots\}$ as a basis, where $\mathbf{e}_i$ has an entry of 1 in position i and 0 everywhere else. Each of these has a corresponding dual element in $(R^\infty)^*$; for instance, $\mathbf{e}_1^*$ is the map that takes the sequence $(a_1, a_2, a_3, \ldots)$ to a_1. But there are other elements of $(R^\infty)^*$ that can't be expressed as a linear combination of the $\mathbf{e}_i^*$; for instance, the map that takes $(a_1, a_2, a_3, \ldots)$ to $a_1 + a_2 + a_3 + \cdots$ (this is really a finite sum, as the terms of the sequence a_i must become all zeros past a certain point). We would have to write this map as $\mathbf{e}_1^* + \mathbf{e}_2^* + \mathbf{e}_3^* + \cdots$, but of course we don't have a concept of infinite sums in standard vector spaces.

The fact that $\mathrm{Hom}(V, W)$ and $V \otimes W$ has the same dimension $mn = \dim V \dim W$ (at least in the finite-dimensional case) may have struck you. In fact, there's a way that we can represent elements of $\mathrm{Hom}(V, W)$ as elements of the tensor space $V^* \otimes W$. This may seem like another needless brick in a tower of abstraction, but it will actually be quite useful for getting a number of useful results.

How, exactly, does this representation work? Suppose that $f : V \to \mathbb{F}$ is some element of V^*, and $\mathbf{w}$ is some element of W. We can take the tensor $f \otimes \mathbf{w} \in V^* \otimes W$, therefore, to correspond to the map in $\mathrm{Hom}(V, W)$ that takes an input vector $\mathbf{v}$ to the output $f(\mathbf{v})\mathbf{w}$. This means that all the maps representable by pure tensors have images of dimension 0 (if $\mathbf{w} = \mathbf{0}_W$ or $f = \mathbf{0}_{W^*}$) or 1 (otherwise).

If we have a general linear map $T : V \to W$ where W is finite-dimensional, then we can represent T as a sum of pure tensors in $V^* \otimes W$ as follows:

1. Take a basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ of W.
2. For each basis vector $\mathbf{w}_i \in W$, define the coefficient extraction map $\mathbf{w}_i^* : W \to \mathbb{F}$ as $\mathbf{w}_i^*(c_1\mathbf{w}_1 + \cdots + c_n\mathbf{w}_n) = c_i$. It's easy to show that this is linear.
3. Represent T as $(\mathbf{w}_1^* \circ T) \otimes \mathbf{w}_1 + \cdots + (\mathbf{w}_n^* \circ T) \otimes \mathbf{w}_n$, where the maps $T \circ \mathbf{w}_i^*$ go from V to $\mathbb{F}$ (i.e. are elements of V^*).

This strategy works with slight modification even if V is finite-dimensional but W is not. In this case, just choose a basis B_W of W with a finite subset S whose span includes the (necessarily finite dimensional) space $\mathrm{im}\, T$. Relative to this basis, T has the representation $\sum_{\mathbf{w} \in B_W} (\mathbf{w}^* T) \otimes \mathbf{w}$. This may look like an infinite sum, but it's actually a finite one, because $\mathbf{w}^* T$ is the zero map from V to $\mathbf{0}$ (and thus $(\mathbf{w}^* T) \otimes \mathbf{w}$ is the zero tensor) whenever $\mathbf{w} \notin S$.

10.4.2 Example

If the discussion above seemed a bit intimidatingly abstract, a concrete example might help. Let's consider the map $T : \mathbb{R}^2 \to \mathbb{R}^2$ with formula $T(x, y) = (y, 3x + 4y)$, with matrix representation relative to the standard basis $\begin{bmatrix} 0 & 1 \\ 3 & 4 \end{bmatrix}$. As always, we'll use $\mathbf{e}_1, \mathbf{e}_2 \in \mathbb{R}^2$ to denote the standard basis vectors, and $\mathbf{e}_1^*, \mathbf{e}_2^* \in (\mathbb{R}^2)^*$ to denote the corresponding coefficient extraction functions $\mathbf{e}_1^*(x, y) = x$ and $\mathbf{e}_2^*(x, y) = y$.

How can we represent this map T as an element of $(\mathbb{R}^2)^* \otimes \mathbb{R}^2$? One basis for $(\mathbb{R}^2)^* \otimes \mathbb{R}^2$, of course, is the set $\{\mathbf{e}_1^* \otimes \mathbf{e}_1, \mathbf{e}_1^* \otimes \mathbf{e}_2, \mathbf{e}_2^* \otimes \mathbf{e}_1, \mathbf{e}_2^* \otimes \mathbf{e}_2\}$, so a general form for elements of $(\mathbb{R}^2)^* \otimes \mathbb{R}^2$ would be

$$a\mathbf{e}_1^* \otimes \mathbf{e}_1 + b\mathbf{e}_1^* \otimes \mathbf{e}_2 + c\mathbf{e}_2^* \otimes \mathbf{e}_1 + d\mathbf{e}_2^* \otimes \mathbf{e}_2.$$

The map that represents any $T : \mathbb{R}^2 \to \mathbb{R}^2$ would have the coefficients a, b, c, d chosen such that

$$a(\mathbf{e}_1^*(x, y))\mathbf{e}_1 + b(\mathbf{e}_1^*(x, y))\mathbf{e}_2 + c(\mathbf{e}_2^* T(x, y))\mathbf{e}_1 + d(\mathbf{e}_2^* T(x, y))\mathbf{e}_2 = T(x, y)$$

or, more simply,

$$(ax + cy, bx + dy) = (y, 3x + 4y)$$

that is, $(a, b, c, d) = (0, 3, 1, 4)$.

Each of these solutions is a matrix entry: in particular, the coefficient for $\mathbf{e}_i^* \otimes \mathbf{e}_j$ is the entry in position (j, i) of the matrix. This correspondence between entries of a matrix representation and coefficients on the pure tensors of a tensor representation will be very useful later.

10.4.3 Basis-independence of tensor representations of maps

Even though we used an explicit basis of W to define representations of elements of $\mathrm{Hom}(V, W)$ in $V^* \otimes W$, the resulting representation is independent of the basis. If we choose a different basis, we will get a sum of a different set of pure tensors, but this sum is guaranteed to equal the same actual element of $V^* \otimes W$ as the original sum. If both V and W are finite-dimensional, you can prove this purely from dimensional considerations: every sum of pure tensors in $V^* \otimes W$ can be interpreted as a map in $\mathrm{Hom}(V, W)$ (that is, the map $\iota : \mathrm{Hom}(V, W) \to V^* \otimes W$ that takes a map to its tensor product representation is surjective), and $\mathrm{Hom}(V, W)$ and $V^* \otimes W$ have equal finite dimension $\dim V \dim W$, so no element of $\mathrm{Hom}(V, W)$ can have more than one representation in $V^* \otimes W$.

In the infinite-dimensional case, we can slightly generalize this:

Proposition. *Let V, W be possibly infinite-dimensional vector spaces over the same field, and let F be the vector subspace of $\mathrm{Hom}(V, W)$ consisting of all maps with a finite-dimensional image.*[2] *Then:*

1. *There is a unique function $\iota : F \to V^* \otimes W$ with the property that for all $T \in F$, if $\iota T = \sum_{i=1}^n f_i \otimes \mathbf{w}_i$ (where $f_1, \ldots, f_n \in V^*$ and $\mathbf{w}_1, \ldots, \mathbf{w}_n \in W$), then $T\mathbf{v} = \sum_{i=1}^n f_i(\mathbf{v})\mathbf{w}_i$ for all $\mathbf{v} \in V$.*

2. *The function ι defined above is linear and bijective.*

3. *If $T \in \mathrm{Hom}(V, W)$ has an infinite-dimensional image, then T is unrepresentable as an element of $V^* \otimes W$: there is no element $\sum_{i=1}^n f_i \otimes \mathbf{w}_i$ of $V^* \otimes W$ such that $T\mathbf{v} = \sum_{i=1}^n f_i(\mathbf{v})\mathbf{w}_i$. (That is: we can't extend ι to a map on any subspace of $\mathrm{Hom}(V, W)$ that contains elements outside of F.)*

Proof. A preliminary note: if B_W is a basis of W, then for any element of $V^* \otimes W$, we can find an equivalent element in which the right-hand sides of every constituent pure tensor come from B_W and no two terms use the same element of B_W: write out the right-hand sides of every original tensor in terms of B_W, expand the results and then collect terms that use the same element of B_W, and add their left-hand sides together to get (at most) one pure tensor with $\mathbf{w}$ on the right-hand side for each element $\mathbf{w} \in B_W$. Any resulting coefficients can also be subsumed into the left-hand side of each pure tensor. So we can assume throughout that any two elements $\mathbf{x}, \mathbf{x}' \in V^* \otimes W$ can be written simultaneously as $\mathbf{x} = \sum_{i=1}^n f_i \otimes \mathbf{w}_i$ and $\mathbf{x}' = \sum_{i=1}^n f_i' \otimes \mathbf{w}_i$, where $\mathbf{w}_1, \ldots, \mathbf{w}_n$ are distinct elements of an arbitrary basis B_W.

[2] Proof that F is a vector subspace: $\mathrm{im}(k_1 T_1 + k_2 T_2) \subseteq \mathrm{im}\, T_1 + \mathrm{im}\, T_2$ for any maps T_1, T_2 and scalars k_1, k_2, so if $\dim \mathrm{im}\, T_1$ and $\dim \mathrm{im}\, T_2$ are finite, then so is $\dim \mathrm{im}(k_1 T_1 + k_2 T_2)$.

Proof of statement 1. This is an existence-and-uniqueness statement. We've covered existence in the preceding discussion: for any map element T, choose a basis B_W for W such that some finite subset S spans a space that includes $\operatorname{im} T$, and then choose $\iota(T) = \sum_{i=1}^{n} (\mathbf{w}_i^* T) \otimes \mathbf{w}_i$ where $\mathbf{w}_1, \ldots, \mathbf{w}_n \in S$, and the dual element $\mathbf{w}_i^*$ extracts the coefficient of $\mathbf{w}_i$ when a vector is written in the basis B_W.

To see that ι is unique, suppose that $\sum_{i=1}^{n} f_i \otimes \mathbf{w}_i$ and $\sum_{i=1}^{n} f_i' \otimes \mathbf{w}_i = 0$ are two possible values of $\iota(T)$ for some map T: that is, they both satisfy $\sum_{i=1}^{n} f_i(\mathbf{v})\mathbf{w}_i = \sum_{i=1}^{n} f_i'(\mathbf{v})\mathbf{w}_i = T\mathbf{v}$ for all $\mathbf{v} \in V$ (and where, again, $\mathbf{w}_1, \ldots, \mathbf{v}_n$ are distinct elements of some basis of W). Then if we subtract these two representations define $g_i := f_i - f_i'$, then $\sum_{i=1}^{n} g_i(\mathbf{v})\mathbf{w}_i = T\mathbf{v} - T\mathbf{v} = \mathbf{0}_W$ (that is, $\sum_{i=1}^{n} g_i \otimes \mathbf{w}_i$ is a representation of the zero map). But this is only possible if the maps $g_1, \ldots, g_n$ are all uniformly zero (i.e. $f_i = f_i'$): if $g_i(\mathbf{v}) \neq 0$, then $\sum_{i=1}^{n} g_i(\mathbf{v})\mathbf{w}_i$ would have a nonzero coefficient on $\mathbf{w}_i$ and thus could not be $\mathbf{0}_W$, because $\mathbf{w}_1, \ldots, \mathbf{w}_n$ are linearly independent.

Proof of statement 2. This statement is really three sub-statements:

1. ι *is linear*: straightforward. If $\iota(T_1) = \mathbf{x}_1 = \sum_{i=1}^{n} f_i \otimes \mathbf{w}_i$ and $\iota(T_2) = \mathbf{x}_2 = \sum_{i=1}^{n} f_i' \otimes \mathbf{w}_i$, then $k_1\mathbf{x}_1 + k_2\mathbf{x}_2 = \sum_{i=1}^{n} (k_1 f_i + k_2 f_i') \otimes \mathbf{w}_i$ satisfies the necessary property for $\iota(k_1 T_1 + k_2 T_2)$.

2. ι *is injective*: the zero element of $V^* \otimes W$ is $\sum_{i=1}^{n} f_i \otimes \mathbf{w}_i$ where all the maps f_i are the zero map; and this could only represent the map that takes $\mathbf{v}$ to $\sum_{i=1}^{n} f_i(\mathbf{v})\mathbf{w}_i = \sum_{i=1}^{n} 0\mathbf{w}_i = \mathbf{0}_W$. So if $T \neq \mathbf{0}_{\operatorname{Hom}(V,W)}$, then $\iota(T) \neq \mathbf{0}_{V^* \otimes W}$.

3. ι *is surjective*: every tensor $\sum_{i=1}^{n} f_i \otimes \mathbf{w}_i$ is the image under ι of some element $T \in F$, namely $T\mathbf{v} = f_i(\mathbf{v})\mathbf{w}_i$.

Proof of statement 3. If $\sum_{i=1}^{n} f_i \otimes \mathbf{w}_i$ represents some map $T\mathbf{v} = \sum_{i=1}^{n} f_i(\mathbf{v})\mathbf{w}_i$, then $\operatorname{im} T \subseteq \operatorname{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$. So if $\operatorname{im} T$ doesn't have a finite spanning set, then we would need an infinite sum of pure tensors to represent it, but we don't have infinite sums in vector spaces.

□

10.5 The trace

If V is finite-dimensional with basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, then the space $\operatorname{End}(V) = \operatorname{Hom}(V, V)$ of operators on V can be represented in $V^* \otimes V$ with a basis $\mathbf{v}_i^* \otimes \mathbf{v}_j$, where $\mathbf{v}_i^*$ is the coordinate extraction function defined previously. We can further define an "evaluation map" $E : V^* \otimes V \to k$ that just applies the element of V^* on the left of every pure tensor to the element of V on the right: that is, $E(\sum_{i=1}^{k} f_i \otimes \mathbf{v}_i) = \sum_{i=1}^{k} f_i(\mathbf{v}_i)$. The values of E on the basis of $V^* \otimes V$ constructed from a basis of V and its corresponding coordinate extraction functions are

$$E(\mathbf{v}_i^* \otimes \mathbf{v}_j) = \mathbf{v}_i^*(\mathbf{v}_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

We can interpret the action of $E \circ \iota$ in matrix terms. Suppose that $T \in \mathrm{End}(V)$ has a matrix representation $A = (a_{ij})$: that is, if $\mathbf{v} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$, then

$$\begin{aligned} T(c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n) &= (a_{11}c_1 + \cdots + a_{1n}c_n)\mathbf{v}_1 + \cdots + (a_{n1}c_1 + \cdots + a_{nn}c_n)\mathbf{v}_n \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\mathbf{v}_j^*(\mathbf{v})\mathbf{v}_i. \qquad (c_i = \mathbf{v}_i^*(\mathbf{v})) \end{aligned}$$

Then T has a tensor representation

$$\iota(T) = \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}\mathbf{v}_j^* \otimes \mathbf{v}_i$$

and if we apply E to this map, then it extracts adds the coefficients of terms $\mathbf{v}_j^* \otimes \mathbf{v}_i$ for which $i = j$ and ignores the rest, so

$$E \circ \iota(T) = a_{11} + \cdots + a_{nn}.$$

But the map ι is independent of basis, so this value has to be the same no matter what basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ we used to get the matrix representation A of our original operator T. So we just proved an important result that would have been far more cumbersome to get with matrix algebra:

Proposition. *All similar matrices have the same sum of diagonal entries.*

Proof. Just given.

□

The sum of the elements on the diagonal of a square matrix is called the *trace* and denoted with the abbreviation tr. Since every square matrix in $\mathbb{C}$ is also similar to a matrix in Jordan normal form, and the diagonal entries of a matrix in JNF are its eigenvalues (counted up to multiplicity), we have this result:

Proposition. *The trace of a matrix A equals the sum of its eigenvalues counted up to multiplicity: if A has eigenvalues $\lambda_1, \ldots, \lambda_k$ and the corresponding maximal generalized eigenspaces have dimensions $d_1, \ldots, d_k$, then $\mathrm{tr}\, A = d_1\lambda_1 + \cdots + d_k\lambda_k$.*

Proof. Just given.

□

There's one other core result on traces that, this time, is easier to get just by working with matrices.

Proposition. *Suppose A and B are matrices of respective dimensions $m \times n$ and $n \times m$. Then $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. (Note that AB and BA may have different dimensions, but they are always both square.)*

Proof. Write $A = (a_{ij})$ and $B = (b_{ij})$. The ith diagonal entry of AB is $\sum_{j=1}^{n} a_{ij}b_{ji}$, so the sum of all diagonal entries is $\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}b_{ji}$. Likewise, the sum of the jth diagonal entry of BA is $\sum_{i=1}^{m} b_{ji}a_{ij}$, and the sum of all diagonal entries is $\sum_{j=1}^{n}\sum_{i=1}^{m} b_{ji}a_{ij}$. These sums are clearly equal.

□

Remark. This result generalizes to products three or more matrices, but not quite as far as you might expect. You can rearrange the terms in a matrix product cyclically while preserving the trace: for instance, $\operatorname{tr}(ABCD) = \operatorname{tr}(A(BCD)) = \operatorname{tr}(BCDA)$, and likewise $\operatorname{tr}(BCDA) = \operatorname{tr}(CDAB) = \operatorname{tr}(DABC)$. But non-cyclic rearrangements won't generally preserve the trace: in general, for instance, $\operatorname{tr}(ABCD) \neq \operatorname{tr}(CBDA)$. The reason for this is that unlike determinants—where the rule $\det(AB) = \det A \det B$ allows you to break the determinant of a matrix product completely down into the determinants of the individual entries, which are field elements whose products all commute—in general $\operatorname{tr}(AB) \neq \operatorname{tr} A \operatorname{tr} B$.

10.6 Symmetric and alternating tensors

10.6.1 Defined

Some new shorthand notation: to express the tensor product of n copies of V (that is, $\underbrace{V \otimes \cdots \otimes V}_{n \text{ times}}$), we'll write $\bigotimes^n V$.

You may recall from section 7.4 that there are two important subspaces of the set $\operatorname{Multilin}(V^n, W)$ of multilinear maps that take all their inputs from the same vector space. These are *symmetric* maps (in which swapping any two entries preserves the value) and *alternating* maps (that take value $\mathbf{0}_W$ whenever two of their arguments are equal and, therefore, flip the sign of their value whenever two arguments are interchanged).

We can ask the same question about alternating and symmetric maps that motivated our construction of the tensor product for multilinear maps: is there a space that we can build from V that produces a unique linear representation of an alternating or symmetric map on V^n? There are fewer alternating and symmetric maps than there are multilinear maps, so you might expect that the required spaces are smaller.

In fact, there is: these spaces, called the *alternating product* and *symmetric product*, are quotients of the tensor product. The idea behind the construction of these spaces from the tensor product is similar to the idea behind the construction of the tensor product itself from the free vector space on V^n: choose some operations that you want to be able to do on tensors that correspond to the defining axioms for alternating or symmetric maps, define subspaces that include the differences between the tensors that you want to consider equivalent, and take the quotient of the tensor space by the subspace.

Specifically, define the following subspaces:

1. X_S is the subspace of $\bigotimes^n V$ spanned by the differences between pairs of pure tensors that have the same components in different orders: that is, $\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_n - \mathbf{v}_{\sigma(1)} \otimes \cdots \otimes \mathbf{v}_{\sigma(n)}$ for all vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$ and permutations $\sigma \in S_n$.

   (It's possible and relatively straightforward to prove that this space has an even smaller spanning set: the set of differences between pure tensors in which the positions of two components in the first pure tensor are swapped. The core observation is that every permutation is a composition of transpositions.)

2. X_A is the subspace spanned by all pure tensors with at least two equal components. For instance, for $n = 3$ again, X_A is the span of the set of elements of the form $\mathbf{v} \otimes \mathbf{x} \otimes \mathbf{x}$, $\mathbf{x} \otimes \mathbf{v} \otimes \mathbf{x}$, or $\mathbf{x} \otimes \mathbf{x} \otimes \mathbf{v}$.

We can then define the *symmetric product*[3] $\bigodot^n V = \left(\bigotimes^n V\right)/X_S$ and the *alternating product* $\bigwedge^n V = \left(\bigotimes^n V\right)/X_A$.

The elements of the symmetric product $\bigodot^n V$ are linear combinations of *pure symmetric tensors* of the form $\mathbf{v}_1 \odot \cdots \odot \mathbf{v}_n$ and, likewise, elements of the alternating product $\bigwedge^n V$ are linear combinations of *pure alternating tensors* that we'll write in the form $\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n$. To be more precise, $\mathbf{v}_1 \odot \cdots \odot \mathbf{v}_n$ represents the coset of X_S that contains $\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_n$, and more generally, some linear combination of symmetric tensors $c_1(\mathbf{v}_1^{(1)} \odot \cdots \odot \mathbf{v}_n^1) + \cdots + c_k(\mathbf{v}_1^{(k)} \odot \cdots \odot \mathbf{v}_n^{(k)})$ represents the coset of X_S that contains $c_1(\mathbf{v}_1^{(1)} \otimes \cdots \otimes \mathbf{v}_n^1) + \cdots + c_k(\mathbf{v}_1^{(k)} \otimes \cdots \otimes \mathbf{v}_n^{(k)})$. The analogous statement for pure alternating tensors and X_A is also true.

These constructions mean:

1. Constituent vectors of pure symmetric tensors in $\bigodot^n V$ may be freely rearranged without changing the value: that is, $\mathbf{v}_1 \odot \cdots \odot \mathbf{v}_n = \mathbf{v}_{\sigma(1)} \odot \cdots \odot \mathbf{v}_{\sigma(n)}$ for any permutation $\sigma \in S_n$, because $\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_n$ and $\mathbf{v}_{\sigma(1)} \otimes \cdots \otimes \mathbf{v}_{\sigma(n)}$ are in the same coset of X_S. All the other operations possible in a tensor product remain allowed in a symmetric product: for instance, coefficients on a symmetric tensor can be merged into one of the components, and vector sums in one component of a symmetric tensor can be distributed into sums of multiple tensors.

2. Pure alternating tensors with two identical components can be eliminated from any sum, and (as a consequence) the components of any alternating tensor can be permuted arbitrarily with a sign flip if the permutation is odd: $\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n = \operatorname{sgn}(\sigma)\mathbf{v}_{\sigma(1)} \wedge \cdots \wedge \mathbf{v}_{\sigma(n)}$.

   The inference from "alternating tensors with two identical components are zero" to "elements of alternating tensors can be permuted with a sign flip if the permutation is odd" is the result of two facts: the map $(\mathbf{v}_1, \ldots, \mathbf{v}_n) \mapsto \mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n)$ is an alternating multilinear map from V^n to $\bigwedge^n V$, and alternating maps are skew-symmetric (as we proved on page 178). You can also see this by expanding the right-hand side of identities such as $\mathbf{0}_{\bigwedge^n V} = (\mathbf{v}_1 + \mathbf{v}_2) \wedge (\mathbf{v}_1 + \mathbf{v}_2) \wedge \mathbf{v}_3 \wedge \cdots \wedge \mathbf{v}_n$ to note that reversing two components of an alternating tensor flips the sign.

Careful readers may have noticed one slight problem with the above discussion: we have defined $c_1(\mathbf{v}_1^{(1)} \odot \cdots \odot \mathbf{v}_n^1) + \cdots + c_k(\mathbf{v}_1^{(k)} \odot \cdots \odot \mathbf{v}_n^{(k)}) \in \bigodot^n V$ to be the coset of X_S that contains $c_1(\mathbf{v}_1^{(1)} \otimes \cdots \otimes \mathbf{v}_n^1) + \cdots + c_k(\mathbf{v}_1^{(k)} \otimes \cdots \otimes \mathbf{v}_n^{(k)}) \in \bigotimes^n V$. But, of course, elements of $\bigotimes^n V$ do not necessarily have unique representations: there are many ways to write any element of $\bigotimes^n V$ as a sum of pure tensors. How can we be sure that all of these representations, when we replace the $\otimes$ symbols with $\odot$, give different representations of the same coset of X_S? (The same questions, of course, are valid for $\bigwedge^n V$ as well as $\bigodot^n V$.)

The answer is to remember that any valid manipulations to tensor expressions in $\bigotimes^n V$ remain valid in $\bigodot^n V$ and $\bigwedge^n V$; the latter two spaces simply add new manipulations (that is: freely rearranging symmetric tensor constituents in $\bigodot^n V$, and removing tensors with duplicate constituents and rearranging constituents with a possible sign

[3]There's no universal notation for the symmetric product: you may also see the notation $\operatorname{Sym}^n V$ for this in some other books. The notation $\bigwedge^n V$ for the alternating product, though, is relatively standard (though you may also see the term *exterior product* instead, relating to a particular use of alternating products in calculus on manifolds).

flip in $\bigwedge^n V$). Therefore, if any two expressions with ordinary tensors are equivalent in $\bigotimes^n V$ (i.e. one expression can be turned into another with a sequence of valid tensor operations), then the same sequence of operations converts between these expressions when the $\otimes$ operator replaced by $\odot$ or $\wedge$ are also equivalent in $\bigodot^n V$ and $\bigwedge^n V$. So the projection maps from $\bigotimes^n V$ to $\bigodot^n V$ and $\bigwedge^n V$ given by $\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_n \mapsto \mathbf{v}_1 \odot \cdots \odot \mathbf{v}_n$ are in fact well defined.

10.6.2 Universal properties

The worth of the tensor product came from its ability to turn multilinear functions on $U \times V$ to linear functions on $U \otimes V$. The symmetric and alternating products have similar properties that shouldn't be too surprising: rather than giving linear equivalents of all multilinear functions, they give us linear equivalents only of symmetric and alternating functions.

We'll sketch out a proof for this result:

Proposition (Universal properties of symmetric and alternating products)**.** *Let $f : V^n \to W$ be a multilinear function. Then:*

1. *If f is symmetric, then there is a unique linear map $\tilde{f} : \bigodot^n V \to W$ such that $f(\mathbf{v}_1, \cdots, \mathbf{v}_n) = \tilde{f}(\mathbf{v}_1 \odot \cdots \odot \mathbf{v}_n)$ for all $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$.*
2. *If f is alternating, then there is a unique linear map $\tilde{f} : \bigwedge^n V \to W$ such that $f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \tilde{f}(\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n$ for all $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$.*

Proof. The proofs of these two statements are quite similar to each other, as well as to our technique for proving the universal property of the tensor product; we'll provide a slightly abbreviated proof sketch here.

The basic steps for proving the theorem statement are:

1. Given the multilinear function $f : V^n \to W$, let $f' : \bigotimes^n V \to W$ be the (necessarily unique) linear function that satisfies $f'(\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_n) = f(\mathbf{v}_1, \ldots, \mathbf{v}_n)$.
2. If f is symmetrical and $\mathbf{x}$ is an element of the spanning set of X_S (that is, $\mathbf{x} = \mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_n - \mathbf{v}_{\sigma(1)} \otimes \cdots \otimes \mathbf{v}_{\sigma(n)}$), then $f'(\mathbf{x}) = f(\mathbf{v}_1, \ldots, \mathbf{v}_n) - f(\mathbf{v}_{\sigma(1)}, \ldots, \mathbf{v}_{\sigma(n)}) = \mathbf{0}_V$. Thus, $f'(\mathbf{x}) = \mathbf{0}_W$ for any element $\mathbf{x} \in X_S$.

   Similarly, if f is alternating, then $f'(\mathbf{x}) = \mathbf{0}_W$ whenever $\mathbf{x} \in X_A$.
3. Let $\pi_S : \bigotimes^n V \to \bigoplus^n V$ and $\pi_A : \bigotimes^n V \to \bigwedge^n V$ be the projection maps defined on pure tensors as $\pi_S(\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_n) = \mathbf{v}_1 \odot \cdots \odot \mathbf{v}_n$ and $\pi_A(\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_n) = \mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n$. (We've just discussed why these are well-defined maps that have the same values no matter what form their inputs in $\bigotimes^n V$ are written in.) Then $X_S = \ker \pi_S$ and $X_A = \ker \pi_A$.

   So if f is symmetric, then $X_S \subseteq \ker f'$, and the first isomorphism theorem guarantees the existence of a unique linear map $\tilde{f} : \bigodot^n V \to W$ such that $f' = \tilde{f} \circ \pi_S$. Similarly, if f is alternating, then $X_S \subseteq \ker f'$, and there's a unique linear map $\tilde{f} : \bigwedge^n V \to W$ such that $f' = \tilde{F} \circ \pi_A$. In either case, this map $\tilde{f}$ has the properties required in the theorem statement.

□

It's also possible to prove (though we won't prove it here) that these universal properties uniquely determine the alternating and symmetric products in the same way as with the universal property: if there's some other space Z with a multilinear symmetric (or alternating) map $\iota : V^n \to Z$ such that any symmetric (or alternating) map $f : V^n \to W$ can be written as $f = \tilde{f} \circ \iota$ where $\tilde{f} : Z \to W$ is a unique linear map, then there's some linear bijection from Z to $\bigodot^n V$ or $\bigwedge^n V$ that also takes vectors in the image of ι to the corresponding pure symmetric or alternating tensors.

10.6.3 Bases of symmetric and alternating products

It should be relatively intuitive that if B is a basis of V, then the sets $\{\mathbf{v}_1 \odot \cdots \odot \mathbf{v}_n : \mathbf{v}_1, \ldots, \mathbf{v}_n \in V\}$ and $\{\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n : \mathbf{v}_1, \ldots, \mathbf{v}_n \in B\}$ are spanning sets of $\bigodot^n V$ and $\bigwedge^n V$—after all, these sets are images under the projection maps π_S and π_A of sets of pure tensors that span $\bigotimes^n V$. These sets, though, aren't linearly independent: they include tensors with duplicate elements (which are zero in the alternating product), and tensors that have the same components as each other in different orders (which equal each other in the symmetric product and are each other's negatives in the alternating product).

We can eliminate these redundancies in the following way. Impose an arbitrary total order on the elements of B: that is, a relation[4] that we'll denote with the sign $\leq$ that satisfies the following axioms:

1. *Totality*: for any two vectors $\mathbf{v}_1, \mathbf{v}_2 \in B$, either $\mathbf{v}_1 \leq \mathbf{v}_2$ or $\mathbf{v}_2 \leq \mathbf{v}_1$.

2. *Antisymmetry*: if $\mathbf{v}_1 \leq \mathbf{v}_2$ and $\mathbf{v}_2 \leq \mathbf{v}_1$, then $\mathbf{v}_1 = \mathbf{v}_2$.

3. *Transitivity*: if $\mathbf{v}_1 \leq \mathbf{v}_2$ and $\mathbf{v}_2 \leq \mathbf{v}_3$, then $\mathbf{v}_1 \leq \mathbf{v}_3$.

There's a finding in set theory, which we won't get into here, that establishes that we can always find a total order for any set. (This relation $\leq$ doesn't have to correspond to any useful properties of a vector: it's purely notional.) This relation $\leq$ gives another relation $<$, defined, naturally enough, as $\mathbf{v}_1 < \mathbf{v}_2$ if $\mathbf{v}_1 \leq \mathbf{v}_2$ and $\mathbf{v}_1 \neq \mathbf{v}_2$.

Now define the following sets:

1. $B_S = \{\mathbf{v}_1 \odot \cdots \odot \mathbf{v}_n : \mathbf{v}_1, \ldots, \mathbf{v}_n \in B, \mathbf{v}_1 \leq \cdots \leq \mathbf{v}_n\}$ is the set of pure symmetric tensors constructed from elements of B in *non-strictly ascending* order. Every pure symmetric tensor constructed from elements of B can be made into exactly one equal element of B_S by sorting its elements with respect to our arbitrary relation $\leq$.

2. $B_A = \{\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n : \mathbf{v}_1, \ldots, \mathbf{v}_n \in B, \mathbf{v}_1 < \cdots < \mathbf{v}_n\}$ is the set of pure alternating tensors constructed from elements of B in *strictly ascending* order. Every pure alternating tensor constructed from elements of B either has two equal elements (and thus equals 0), or it can be made into exactly one element of B_A by sorting its elements (which keeps the tensor the same if sorting the elements is an even permutation, or flips the sign if sorting the elements is an odd permutation).

[4] See section 5.4 if you need a reminder of what "relation" means.

In either case, B_S and B_A are spanning sets of $\bigodot^n V$ and $\bigwedge^n V$. It's natural to suspect that these sets should be linearly independent as well. Our strategy for proving that they are alternating sets will be similar to our argument on page 250: construct a map that has a nonzero value on exactly one of the pure tensors.

The case for alternating tensors is a bit easier. First, we'll need a couple of preliminary results.

Proposition. *Let $f : V^n \to W$ be a multilinear map, let $\sigma \in S_n$ be a permutation of $\{1, \ldots, n\}$ and let $f_\sigma : V^n \to W$ be the map*

$$f_\sigma(\mathbf{v}_1, \ldots, \mathbf{v}_n) = f(\mathbf{v}_{\sigma(1)}, \ldots, \mathbf{v}_{\sigma(n)}).$$

Then f_σ is also multilinear.

Proof. The restricted map formed by holding all arguments to f_σ constant except the argument in position i is the same as the restricted map formed by holding all arguments to f constant except the argument in position $\sigma^{-1}(i)$. Since the restricted maps from f are all linear, the restricted maps from f_σ must all be linear as well.

□

Proposition. *Suppose $f : V^n \to W$ is any multilinear map over an arbitrary field. Define the* antisymmetrization

$$f_A(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) f(\mathbf{v}_{\sigma(1)}, \ldots, \mathbf{v}_{\sigma(n)}).$$

Then $f_A : V^n \to W$ is an alternating multilinear map.

Proof. f_A is multilinear because it's a sum of maps that (by the previous proposition) are also all multilinear. To see that it's alternating, note that $\mathbf{v}_i = \mathbf{v}_j$, then $f(\mathbf{v}_{\sigma(1)}, \ldots, \mathbf{v}_{\sigma(n)}) = f(\mathbf{v}_{\tau\circ\sigma(1)}, \ldots, \mathbf{v}_{\tau\circ\sigma(n)}$ where τ is the transposition of i and j, so we can divide the right-hand sum in the definition of f_A into a sum over even permutations σ and odd permutations $\tau \circ \sigma$, where each function value occurs twice with opposite signs attached. So the whole sum must equal zero.

□

Proposition. *Let B be a basis of a vector space V over a field $\mathbb{F}$, and let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be distinct elements of B. Then there is an alternating multilinear map $f : V^n \to \mathbb{F}$ such that $f(\mathbf{u}_1, \ldots, \mathbf{u}_n) = 1$ and $f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = 0$ for any other elements $\mathbf{v}_1, \ldots, \mathbf{v}_n \in B$ that are not a rearrangement of $\mathbf{u}_1, \ldots, \mathbf{u}_n$.*

Proof. Let $f' : V^n \to W$ be the multilinear (but not alternating) map where the value of $f'(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ is given as follows: write the inputs as (necessarily unique) linear combinations drawn from B, and then multiply the coefficient of $\mathbf{u}_i$ in the expression for $\mathbf{v}_i$ for all indices $1 \leq i \leq n$. The only nonzero value of this map on arguments drawn from B is $f'(\mathbf{u}_1, \ldots, \mathbf{u}_n) = 1$.

Let f be the antisymmetrization of f' as defined in the last proposition. We know that f is alternating. Furthermore, if $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are elements of B but not a rearrangement of $\mathbf{u}_1, \ldots, \mathbf{u}_n$, then either two of the vectors $\mathbf{v}_i$ are equal, or one of the vectors $\mathbf{v}_i$ does not equal any of $\mathbf{u}_1, \ldots, \mathbf{u}_n$. In either case, $f'(\mathbf{v}_{\sigma(1)}, \ldots, \mathbf{v}_{\sigma(n)}) = 0$ for all permutations σ, so $f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = 0$.

□

Corollary. *Let V be an arbitrary vector space on a field $\mathbb{F}$, let B be a basis of V with an arbitrary total order imposed on its elements, and let $B_A \subset \bigwedge^n V$ be the set of n-component alternating tensors with components drawn from B in strictly ascending order. Then for every element of B_A, there is a map from V to $\mathbb{F}$ whose value on that tensor is 1 and whose value on every other element of B_A is 0.*

Proof. For any alternating tensor $\mathbf{u}_1 \wedge \cdots \wedge \mathbf{u}_n$, this map is the $\bigwedge^n V \to \mathbb{F}$ that corresponds to the alternating map $V^n \to \mathbb{F}$ constructed in the previous proposition.

□

Corollary. *B_A, as defined in the previous corollary, is a basis of $\bigwedge^n V$.*

There's a slight difficulty, though, if we use the same argument for symmetric tensors. One natural approach would be to define, for any multilinear map $f : V^n \to W$, a symmetrization

$$f_S(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \sum_{\sigma \in S_n} f(\mathbf{v}_{\sigma(1)}, \ldots, \mathbf{v}_{\sigma(n)}).$$

This almost works. If f is the multilinear map from V^n to $\mathbb{F}$ that satisfies $f(\mathbf{u}_1, \ldots, \mathbf{u}_n) = 1$ and $f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = 0$ for any other elements $\mathbf{v}_1, \ldots, \mathbf{v}_n \in B$ (including if $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are just $\mathbf{u}_1, \ldots, \mathbf{u}_n$ reordered), then f_S is indeed a symmetric multilinear map, and value is zero on any set of inputs drawn from B that are not some rearrangement of $\mathbf{u}_1, \ldots, \mathbf{u}_n$. The problem, however, is that in certain base fields, $f_S(\mathbf{u}_1, \ldots, \mathbf{u}_n) = 0$ as well.

The problem arises, specifically, with fields with nonzero characteristic: that is, if there's some integer k such that $\underbrace{1 + 1 + \cdots + 1}_{k \text{ times}} = 0$. (We first discussed fields with nonzero characteristic way back in section 1.3.3, giving the example of a field with two elements.) As an example, suppose $\mathbf{u}_1, \mathbf{u}_2$ are two elements of our basis B of V, and we want to construct a symmetric map $f_S : V^3 \to \mathbb{F}$ such that $f_S(\mathbf{u}_1, \mathbf{u}_1, \mathbf{u}_2) = 1$ and f has value zero on all other inputs drawn from B, that don't use $\mathbf{u}_1$ as an input twice and $\mathbf{u}_2$ as an input once. We already have a function $f : V^3 \to \mathbb{F}$ such that $f(\mathbf{u}_1, \mathbf{u}_1, \mathbf{u}_2) = 1$ and $f = 0$ on all other inputs drawn from B, including other rearrangements of $\{\mathbf{u}_1, \mathbf{u}_1, \mathbf{u}_2\}$.

What happens if we use our putative formula for f_S? There are $3! = 6$ permutations on three elements, but there are actually only three ways to rearrange three inputs to f when two of them are identical, so the sum in the formula for f_S involves many duplicate terms. In particular, the two permutations σ such that $\sigma(3) = 1$ (namely the transposition $(1\,3)$ and the cycle $(1\,2\,3)$) give the same inputs to (and thus value of) f, as do the two permutations for which $\sigma(3) = 2$ (namely the transposition $(2\,3)$ and the cycle $(1\,3\,2)$) and the two for which $\sigma(3) = 3$ (namely the identity and the transposition $(1, 2)$). Therefore:

$$\begin{aligned} f_S(\mathbf{u}_1, \mathbf{u}_1, \mathbf{u}_2) &= 2f(\mathbf{u}_2, \mathbf{u}_1, \mathbf{u}_1) + 2f(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_1) + 2f(\mathbf{u}_1, \mathbf{u}_1, \mathbf{u}_2) \\ &= 2f(\mathbf{u}_1, \mathbf{u}_1, \mathbf{u}_2) \end{aligned}$$

where the symbol 2 in a generic field means $1 + 1$. But in a field of characteristic 2 (that is, where $1 + 1 = 0$), then $f_S(\mathbf{u}_1, \mathbf{u}_1, \mathbf{u}_2) = 0$.

In general, if a function has distinct inputs $\mathbf{u}_1, \ldots, \mathbf{u}_k$ occurring $a_1, \ldots, a_k$ times each, then every distinct way of arranging the inputs is given by $a_1! \cdots a_k!$ different permutations (because there are $a_1!$ ways of simply shuffling the $\mathbf{u}_1$ inputs around without touching the others, then $a_2!$ independent ways of shuffling the $\mathbf{u}_2$ inputs around, and

so on), so f_S as we've defined it above will be zero in any field whose characteristic divides $a_1! \cdots a_k!$.

We can solve this difficulty by reducing the number of permutations in the sum that defines f_S, to eliminate the coefficient $a_1! \cdots a_k!$. The following proposition shows how to do this:

Proposition. *Let V be a vector space over an arbitrary field $\mathbb{F}$, let B be a basis of V, and let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be not necessarily distinct elements of B. Then there is a symmetric multilinear function $f : V^n \to \mathbb{F}$ such that $f(\mathbf{u}_1, \ldots, \mathbf{u}_n) = 1$ and f has value 0 on all other inputs from B.*

Proof. Let $f' : V^n \to \mathbb{F}$ be the multilinear function such that $f'(\mathbf{u}_1, \ldots, \mathbf{u}_n) = 1$ and $f'(\mathbf{v}_1, \ldots, \mathbf{v}_n) = 0$ for any other inputs $\mathbf{v}_1, \ldots, \mathbf{v}_n \in B$ (including reorderings of $\mathbf{u}_1, \ldots, \mathbf{u}_n$). Define the equivalence relation $\sigma \sim \tau$ on S_n to be true if $\mathbf{u}_{\sigma(i)} = \mathbf{u}_{\tau(i)}$ for all integers $1 \leq i \leq n$ (it's easy to check that this is in fact an equivalence relation).

Every equivalence class, therefore, contains all the permutations that produce a particular rearrangement of $\mathbf{u}_1, \ldots, \mathbf{u}_n$ when different vectors with equal value are indistinguishable. (For instance, if $\mathbf{u}_1 = \cdots = \mathbf{u}_n$, then every permutation is equivalent to every other; and if $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are all different, then there's one equivalence class containing every permutation.) Let $R \subseteq S_n$ be a set containing one arbitrary representative element from every equivalence class, and define

$$f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \sum_{\sigma \in R} f'(\mathbf{v}_{\sigma(1)}, \ldots, \mathbf{v}_{\sigma(n)}).$$

If $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a rearrangement of $\mathbf{u}_1, \ldots, \mathbf{u}_n$, then the sum on the right contains exactly one term with value 1 (namely, the term for the permutation $\sigma \in R$ such that $\mathbf{v}_{\sigma(i)} = \mathbf{u}_i$ for all indices $1 \leq i \leq n$). Otherwise, f is a sum over values of f' that all evaluate to 0.

Therefore, f is a multilinear function, and it is symmetric on values of B: that is, if $\mathbf{v}_1, \ldots, \mathbf{v}_n \in B$, then $f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = f(\mathbf{v}_{\sigma(1)}, \ldots, \mathbf{v}_{\sigma(n)})$ for all permutations $\sigma \in S_n$, not just those that are in R. So the map $F(\mathbf{v}_1, \ldots, \mathbf{v}_n) \coloneqq f(\mathbf{v}_1, \ldots, \mathbf{v}_n) - f(\mathbf{v}_{\sigma(1)}, \ldots, \mathbf{v}_{\sigma(n)})$ is a multilinear map (remember from page 262 that the map generated from a multilinear map by arbitrarily permuting its arguments is also multilinear, and the difference of two multilinear maps is also multilinear), and its values on the basis B of V are all zero.

Therefore, F is the zero map: that is, $f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = f(\mathbf{v}_{\sigma(1)}, \ldots, \mathbf{v}_{\sigma(n)})$ for all vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ in V, not just in B. So f is symmetric.

□

Corollary. *Let V be a vector space over a field $\mathbb{F}$, let B be a basis of V with an arbitrary total order, and let B_S be the set of pure symmetric tensors $\mathbf{v}_1 \odot \cdots \odot \mathbf{v}_n$ with n components drawn from B in non-strictly ascending order. Then for every element of B_S, there is a linear map from $\bigodot^n V$ to $\mathbb{F}$ with value 1 on that element and 0 on every other.*

Proof. This map is the factoring through $\bigodot^n V$ of the symmetric map $f : V^n \to \mathbb{F}$ constructed in the last proposition.

□

Corollary. *B_S as defined in the last proposition is linearly independent, and thus a basis of $\bigodot^n V$.*