Discover Linear Algebra

A first course in linear algebra

Discover Linear Algebra

A first course in linear algebra

Jeremy Sylvestre
University of Alberta

August 14, 2023

Jeremy Sylvestre is Associate Professor of Mathematics at the University of Alberta's Augustana Campus.

Edition: 2023

Website: Discover Linear Algebra[1]

Website: GitHub source code repository[2]

[1] `sites.ualberta.ca/~jsylvest/books/dla.html`
[2] `github.com/jjrsylvestre/dla/`

Preface

The purpose of this book is to serve as supporting material for a fairly typical introductory course in *Linear Algebra* using a discovery-based pedagogical approach.

Each chapter is organized into sections titled *Discovery guide*, *Terminology and notation*, *Concepts*, *Examples*, and *Theory* (though some chapters have additional *Motivation* and/or *More examples* sections).

The purpose for employing this uniform sectioning scheme is to give the student a uniform flow to encountering each new collection of topics:

Discovery guide initial encounter through discovery- and problem-solving-based activities;

Terminology and notation introduction of the communication tools necessary to begin a more sophisticated conversation about the new topics;

Concepts fuller discussion of the new topics, grounded in reflections on the questions and results of the *Discovery guide* section;

Examples computational examples to assist students with the procedural tasks related to the new topics, as well as additional examples that serve to illustrate certain concepts; and

Theory a more formal and general description of the concepts, with proofs.

Traditional textbooks usually intersperse terminology, concepts, examples, and theory in a linear narrative, and relegate "activities" to the *Exercises* section at the end of the chapter. By organizing the flow of learning in the above-described manner, it is hoped that the process of encountering and re-encountering (and *re*-encountering) the topics in different modes — discovery, reflection and discussion, examples, and theory — and at increasing levels of sophistication will lead to deeper learning.

The organization of topics is fairly typical, under the choice of "late vectors" (though the term **column vector** is used informally in the early chapters). Systems of linear equations are used to motivate matrix theory, up through a basic treatment of determinants and the classical adjoint. Then vectors in $\mathbb{R}^n$ are introduced as the initial model for how a "vector space" should behave, with an emphasis on a geometric understanding of the vector operations. A basic introduction to abstract vector spaces follows. Finally, the topic of matrix forms is broached by a simple treatment of eigenvalues/eigenvectors and diagonalization. Note that even though the concept of **similar matrices** is referenced in this final topic, the topic of "change of basis" has been omitted (see below). However, the full two-semester version[3] of this book does include the topic of change of basis.

[3] `sites.ualberta.ca/~jsylvest/books/DLA`

When using *discovery* as a pedagogical principal, it is not possible to cover as many topics, or to cover each topic with the same breadth, as in a breakneck-paced lecture class. The goal of these notes is not to teach students *a bunch of mathematics* in the particular topic of linear algebra, but instead to teach students *about mathematics* through the discovery of the beautiful and coherent subject of linear algebra. I have tried to distill each topic down to the necessary minimal core of concepts essential to the study of the subject, and have rejected inclusion of peripheral topics and facts or esoteric applications. I do not intend for these notes to be workable for everyone in every kind of linear algebra class. (But since they are released under an open license, they could of course be edited to make them workable for any kind of linear algebra course.)

Jeremy Sylvestre
Camrose, Alberta 2020

Contents

Back Matter

Part I

Systems of Equations and Matrices

CHAPTER 1

Systems of linear equations

1.1 Discovery guide

Discovery 1.1 Sketch the graph of $2x + y = 3$.

(a) What type of graph is it? What is the name of this course again?

(b) *Fill in the blanks:* The connection between the graph and the equation above is that the graph is the collection of ______ that ______ the equation above.

Discovery 1.2

(a) On the same axes as your graph for Discovery 1.1, sketch the graph of $x + y = 1$.

(b) Looking at your graphs, is there any pair of values (x, y) that satisfy *both* equations simultaneously?

Discovery 1.3

(a) On a new set of axes, sketch the graphs of $2x + y = 3$ and $4x + 2y = 4$.

(b) Looking at these two graphs, is there any pair of values (x, y) that satisfy *both* equations simultaneously?

Discovery 1.4 The graph of a linear equation in three variables (e.g., $3x+y-2z = 5$) corresponds to a plane in three-dimensional space.

Suppose you had *three* equations in three variables. Try to imagine the geometric configuration of the corresponding three planes in each of the following situations. You might find it helpful to use three pieces of paper as props.

(a) There is *no* triple of numbers (x, y, z) that satisfies all three plane equations at once.

(b) There are *an infinite number* of triples of numbers (x, y, z) that satisfy all three plane equations at once.

(c) There is *exactly one* triple of numbers (x, y, z) that satisfies all three plane equations at once.

Discovery 1.5 Consider the **system of equations**

$$\left\{\begin{array}{rcrcrcr} x & + & 2y & - & z & = & 5, \\ & & y & + & z & = & -1. \end{array}\right.$$

(a) If $z = 2$, what is y? ... what is x?

(b) If $z = -10$, what is y? . . . what is x?

(c) For these example values of z, why do you think you are being asked to determine the value of y first and then to determine the value of x?

(d) Do you think that, given any arbitrary value for z, you could solve for y and then for x?

(e) The three values of x, y, z that you came up with in Task a together represent *one* solution to the system of equations. The three values of x, y, z that you came up with in Task b together represent *another* solution to the system of equations.

Based on your response to Task d, how many solutions does this system have in total?

(f) If $z = t$, what is y? . . . what is x?

Discovery 1.6 Suppose x and y are "mystery" numbers, but you have a clue to their identities: you know that both $x - 2y = -4$ and $2x + y = 2$ are true.

(a) *Without* determining the values of x and y, answer each of the following with a *number*.

(i) $3(x - 2y) = ?$

(ii) $-2(2x + y) = ?$

(iii) $(x - 2y) + (2x + y) = ?$

(iv) $(2x + y) - 2(x - 2y) = ?$

(b) Algebraically simplify the expression in the last part of Task a, and combine this simplified expression with your numerical answer to that part to solve for y. Then use one of the original equations from the introduction to this activity to solve for x.

Why was that combination of the left-hand sides of the two equations particularly useful for determining the values of x and y?

Discovery 1.7 We can work with a system of equations more efficiently by representing it compactly as an **augmented matrix**. For example,

$$\left\{\begin{array}{rcrcrcr} -2x & + & 2y & - & 5z & = & -1 \\ 3x & & & + & 3z & = & 9 \\ x & - & y & + & 3z & = & 2 \end{array}\right. \longrightarrow \left[\begin{array}{rrr|r} -2 & 2 & -5 & -1 \\ 3 & 0 & 3 & 9 \\ 1 & -1 & 3 & 2 \end{array}\right]$$

Do you understand how this system was turned into a matrix? Now perform the following calculations, but *using the matrix, obtaining a new matrix at each step*.

(a) Change the order of the equations: interchange the first and third equations.

(b) Starting with your new system from Task a, subtract 3 times the first equation from the second equation, and add 2 times the first equation to the third equation.

(c) Starting with your new system from Task b, multiply the second equation by 1/3.

(d) Your final result from Task c, should be a "simplified" matrix. Turn this matrix back into a system of equations and see how much easier it is to solve the system.

1.2 Terminology and notation

linear equation
an equation of the form

number times *variable* plus *number* times *variable* plus …
equals *number*

system of linear equations
a (finite) collection of linear equations

Note 1.2.1 We will often just say **system of equations** to mean a system of linear equations.

solution a set of values that, when substituted in for the variables of a system of equations, satisfy all of the equations in the system simultaneously

solution set
the collection of all possible solutions to a system of equations

consistent system
a system of equations with at least one solution

inconsistent system
a system of equations with no solution

parametric equations
a collection of formulas, based on one or more parameters, for the variables in a linear system that represent all possible solutions to the system (as the parameters vary over all real numbers)

matrix a rectangular array of numbers

augmented matrix
a matrix of the coefficients and constants in a linear system

matrix entry
one of the numbers in a matrix; sometimes also referred to as a **matrix coefficient**

row a horizontal line of entries in a matrix

column a vertical line of entries in a matrix

elementary row operations
the three basic operations that can be applied to an augmented matrix without changing the set of solutions to the corresponding linear system:

(i) swap the positions of two rows,

(ii) multiply a row by a nonzero constant, and

(iii) add a multiple of one row to another.

1.3 Concepts

In this section.

- Subsection 1.3.1 *System solutions*
- Subsection 1.3.2 *Determining solutions*

Goal 1.3.1 *Develop a systematic procedure to determine all combinations (if any) of variable values that solve a system of equations.*

Before we work to realize this goal, let's make sure we understand it.

1.3.1 System solutions

Question 1.3.2 What is a solution, and how do we verify solutions? □

For the system consisting of the two lines in the xy-plane from Discovery 1.1 and Discovery 1.2,

$$\left\{\begin{array}{rcrcr} 2x & + & y & = & 3, \\ x & + & y & = & 1, \end{array}\right.$$

the combination $x = 2$, $y = -1$ is a solution because both equations will be satisfied simultaneously with these values. We verify this by proper "LHS vs RHS" procedure:

$$\begin{aligned} \text{First equation:} \quad & \text{LHS} = 2x + y = 2(2) + (-1) = 3 = \text{RHS}, \\ \text{Second equation:} \quad & \text{LHS} = x + y = 2(2) + (-1) = 2 + (-1) = 1 = \text{RHS}. \end{aligned}$$

Since LHS=RHS in both equations when $x = 2$ and $y = -1$, we have a valid solution to the system. However, the combination $x = 1, y = 1$ is *not* a solution to the system, because at least one of the equations will not be satisfied by these values. Again, we can verify this by proper "LHS vs RHS" procedure:

$$\begin{aligned} \text{First equation:} \quad & \text{LHS} = 2x + y = 2(1) + 1 = 3 = \text{RHS}, \\ \text{Second equation:} \quad & \text{LHS} = x + y = 1 + 1 = 2 \neq \text{RHS}. \end{aligned}$$

While the first equation is satisfied, the second is not, and so this combination of variable values is not a valid solution.

Remark 1.3.3 In the example above and in Discovery guide 1.1 we have seen that systems of linear equations have geometric interpretations: intersecting lines in the xy-plane, or intersecting planes in xyz-space. We can make a similar geometric interpretation for systems with more than 3 variables by imagining "hyperplanes" intersecting in higher-dimensional spaces, but unfortunately our three-dimensional brains cannot actually picture such a thing.

Question 1.3.4 How many solutions can a system have? □

We have seen in Discovery guide 1.1 that there are different possibilities for the number of different solutions a particular system can have.

one unique solution
This is demonstrated by the system formed by the two lines from Discovery 1.1 and Discovery 1.2, as the two lines in these activities only intersected in a single point.

no solutions
This is demonstrated by the two lines in Discovery 1.3, as these two lines were parallel and did not intersect.

an infinite number of solutions

This is demonstrated by the system in Discovery 1.5, as any chosen value of z leads to a new solution by then solving for y and x in turn, and there are infinity of different choices of starting value z.

Question 1.3.5 Are the possibilities considered above the *only* possibilities? Could there be a system that has exactly *seven* different solutions, say? □

We will prove in Chapter 4 (Theorem 4.5.5) that for *every* system there are in fact only these three possibilities as encountered in Discovery guide 1.1.

In Discovery 1.4, you were asked to imagine the geometric configuration of three planes (each represented algebraically by a linear equation in three variables) to realize each of the three possibilities described above. Hopefully you can also imagine how it would be geometrically impossible for three planes to intersect in *exactly* seven points, no more and no fewer.

Question 1.3.6 When a system has an infinite number of solutions, how can we express *all possible solutions* in a compact way? (We certainly cannot *list* all possible solutions.) □

We can use one or more *parameters* to represent the choices that must be made to get to one particular solution, and then use formulas in those parameter(s) to express the patterns of similarity between the different solutions. For example, in Task f of Discovery 1.5, there did not seem to be any restriction on what values the variable z could take and still be part of a solution to the system. So z was set to be an unspecified parameter t, and then y and x could be solved for in terms of this parameter. Choosing different values of t (such as $t = 2$ or $t = -10$, as in the previous parts of the referenced discovery activity) leads to different particular solutions for the system. The infinity of possible solutions to this system is now represented entirely by the infinity of choices available for starting value of the parameter t.

Remark 1.3.7 It may seem silly to trade one variable letter z for another letter t. But these letters represent different kinds of "unknown" quantities. Letter z represents a *variable in an equation whose value we would like to determine*, whereas letter t represents a *parameter whose value we are free to choose*. Remember that mathematical notation is a *tool for communicating ideas*: the letter t is a traditional choice for a parameter in mathematics, and so we switch from letter z to letter t to indicate to the reader (whether that is one's self or someone reading our work) this change in perspective from variable to parameter.

1.3.2 Determining solutions

The first of two core ideas behind how we should go about determining the solutions of a system of equations is contained in Discovery 1.6. The left-hand side of a linear equation looks like a jumble of numbers and letters, but remember that it is just *a* formula *for computing a* single *number*, and that the result of this computation is proposed to always be equal the number on the right-hand side of the equation. So if we algebraically manipulate or combine the left-hand sides of equations in the system, as long as we perform the same manipulation or combination of the corresponding right-hand sides of those equations, then the same variable values that solve the new old system should solve the new system, and vice versa.

We need to be a little bit careful with the kinds of manipulations and combinations we allow ourselves so that our manipulations are *nondestructive*. For example, if we multiplied both left- and right-hand sides of an equation by 0, we would lose all information the original equation contained, since we would be left with just $0 = 0$. In this case, new and old equations would *not* have the same solutions. This is why we restrict ourselves to the **elementary row operations** described in Section 1.2: to ensure our manipulations are always nondestructive.

Rows versus equations. The elementary operations are stated as *row* operations on an augmented matrix, but just replace the word "row" with "equation" in their descriptions and you have the equivalent manipulation of equations in a system.

The second core idea behind solving systems of equations is contained in Discovery 1.7. We should choose sequences of manipulations that will result in a simplified system for which it is easier to determine the solutions. Discovery 1.7 lays out a specific sequence of operations to do this; in the next discovery guide and corresponding chapter we will explore a systematic strategy for performing such simplification.

Finally, Discovery 1.7 contains another important idea: all of the crucial information in a system of equations is contained in its coefficients on variables and the constant on the right-hand side of each equation. We can get rid of the clutter of all the variable letters by turning a system of equations into an **augmented matrix**. We can then perform manipulations of the equations in the system by just performing the corresponding operations on the coefficients in the matrix. You should keep in mind the structure of an augmented matrix: each row represents an equation, and each column (except the last) represents a variable. See the examples below on how row operations correspond to the algebra of equations.

1.4 Examples

In this section.

- Subsection 1.4.1 *Row operations versus equation manipulations*

1.4.1 Row operations versus equation manipulations

Let's examine the operations in Discovery 1.7 in detail, by considering the operations as both *equation manipulations* and *row operations* simultaneously.

In each step, notice how the row↔equation and column↔variable correspondence is preserved.

$$\left[\begin{array}{rrr|r} -2 & 2 & -5 & -1 \\ 3 & 0 & 3 & 9 \\ 1 & -1 & 3 & 2 \end{array}\right] \qquad \left\{\begin{array}{rcrcrcr} -2x & + & 2y & - & 5z & = & -1 \\ 3x & & & + & 3z & = & 9 \\ x & - & y & + & 3z & = & 2 \end{array}\right.$$

Interchange the first and third rows/equations.

$$\left[\begin{array}{rrr|r} 1 & -1 & 3 & 2 \\ 3 & 0 & 3 & 9 \\ -2 & 2 & -5 & -1 \end{array}\right] \qquad \left\{\begin{array}{rcrcrcr} x & - & y & + & 3z & = & 2 \\ 3x & & & + & 3z & = & 9 \\ -2x & + & 2y & - & 5z & = & -1 \end{array}\right.$$

Subtract 3 times the first row/equation from the second row/equation. For the equation version, we do this by performing the same combination of left- and right-hand sides, collecting terms on the left.

$$\begin{array}{ccccccc} (\mathrm{LHS}_2) & - & 3(\mathrm{LHS}_1) & = & (\mathrm{RHS}_2) & - & 3(\mathrm{RHS}_1) \\ (3x+3z) & - & 3(x-y+3z) & = & 9 & - & 3(2) \end{array}$$

This combination leads to new equation

$$0x+3y-6z=3.$$

Notice that when collecting terms, we ended up performing that "subtract 3 times the first from the second" on the coefficients of each variable. So we can achieve the same result in the matrix by performing "subtract 3 times the entry in the first row from the entry in the second row," one column at a time.

$$\left[\begin{array}{rrr|r} 1 & -1 & 3 & 2 \\ 0 & 3 & -6 & 3 \\ -2 & 2 & -5 & -1 \end{array}\right] \qquad \left\{\begin{array}{rcrcrcr} x & - & y & + & 3z & = & 2 \\ & & 3y & - & 6z & = & 3 \\ -2x & + & 2y & - & 5z & = & -1 \end{array}\right.$$

Next, add 2 times the first row/equation to the third row/equation:

$$\begin{array}{ccccccc} (\mathrm{LHS}_3) & + & 2(\mathrm{LHS}_1) & = & (\mathrm{RHS}_3) & + & 2(\mathrm{RHS}_1) \\ (-2x+2y-5z) & + & 2(x-y+3z) & = & -1 & + & 2(2), \end{array}$$

leading to new equation

$$z=3,$$

which we will use to replace the old third row/equation.

$$\left[\begin{array}{rrr|r} 1 & -1 & 3 & 2 \\ 0 & 3 & -6 & 3 \\ 0 & 0 & 1 & 3 \end{array}\right] \qquad \left\{\begin{array}{rcrcrcr} x & - & y & + & 3z & = & 2 \\ & & 3y & - & 6z & = & 3 \\ & & & & z & = & 3 \end{array}\right.$$

Finally, multiply the second row/equation by 1/3.

$$\begin{array}{ccc} (1/3)(\mathrm{LHS}_2) & = & (1/3)(\mathrm{RHS}_2) \\ (1/3)(3y-6z) & = & (1/3)(3) \\ y-2z & = & 1 \end{array}$$

The matrix is modified accordingly.

$$\left[\begin{array}{rrr|r} 1 & -1 & 3 & 2 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & 3 \end{array}\right] \qquad \left\{\begin{array}{rcrcrcr} x & - & y & + & 3z & = & 2 \\ & & y & - & 2z & = & 1 \\ & & & & z & = & 3 \end{array}\right.$$

The final system on the right is much easier to solve: we can see immediately from the third equation that $z = 3$, then can use this in the second equation to determine $y = 7$, and finally can use both of these in the first equation to determine $x = 0$.

A look ahead. In Chapter 2, we will develop a systematic method of simplifying a system in this manner, but working exclusively with augmented matrices. Also, we will take the process a few steps further to make the system as simple as possible. Notice how "back-solving" the system proceeds from bottom-right to top-left. We will use the same process when solving systems using matrices.

Chapter 2

Solving systems using matrices

2.1 Discovery guide

Reminder.

The elementary row operations are

(i) swap rows;

(ii) multiply a row by a non-zero constant; and

(iii) add a multiple of one row to another.

Discovery 2.1 Consider the following system.

$$\left\{\begin{array}{rcrcrcr} 2x & & & - & 2z & = & 4, \\ x & - & y & & & = & 3, \\ 4x & - & 2y & - & 3z & = & 7. \end{array}\right.$$

(a) Convert to an augmented matrix.

(b) Via elementary row operations, obtain a "leading 1" in the first entry of the first row (maybe swap some rows?), then use it to eliminate all other entries in the first column.

(c) Obtain a leading 1 in the second entry of the second row (do not use/alter the first row!), then use it to eliminate all other entries in the second column (yes, you can now alter the first row).

(d) Obtain a leading 1 in the third entry of the third row (do not use/alter first or second rows!), then use it to eliminate all other entries in the third column.

(e) Turn the final augmented matrix back into a system and solve it.

Discovery 2.2 Consider the following system.

$$\left\{\begin{array}{rcrcrcr} 3x & + & 6y & + & 5z & = & -9, \\ 2x & + & 4y & + & 3z & = & -5, \\ 3x & + & 6y & + & 6z & = & -12. \end{array}\right.$$

(a) Convert to an augmented matrix.

(b) Via elementary row operations, obtain a leading 1 in the first entry of the first row (maybe combine first two rows somehow?), then use it to eliminate all other entries in the first column.

(c) Is it possible to obtain a leading 1 in the second entry of the second row?

(d) Obtain a leading 1 in third entry of the second row (do not use/alter the first row!), then use it to eliminate all other entries in the third column.

(e) Assign a parameter to every variable whose column *does not* contain a leading one. Turn the final augmented matrix back into a system and solve it in terms of your parameter(s).

Discovery 2.3 Consider the following system.

$$\left\{\begin{array}{rcrcrcr} x & + & 2y & + & z & = & 2, \\ 2x & + & 5y & + & 2z & = & -3, \\ 2x & + & 4y & + & 2z & = & -1. \end{array}\right.$$

(a) Convert to an augmented matrix.

(b) Use the leading 1 in first entry of the first row to eliminate all other entries in the first column.

(c) Convert the new third row back into an equation. What does this mean about the system?

Discovery 2.4 Consider the following system. Notice that the "equals" column is all zeros. Such a system is called **homogeneous**.

$$\left\{\begin{array}{rcrcrcrcl} 3x_1 & + & 6x_2 & - & 8x_3 & + & 13x_4 & = & 0, \\ x_1 & + & 2x_2 & - & 2x_3 & + & 3x_4 & = & 0, \\ 2x_1 & + & 4x_2 & - & 5x_3 & + & 8x_4 & = & 0. \end{array}\right.$$

Careful. After you've reduced the homogeneous system in this activity, remember that there is still the omitted "equals" column of all zeros.

(a) There is one obvious particular solution to the system. What is it?

(b) Will any row operation ever alter the "equals" column?

(c) Convert the system to a **coefficient matrix** (i.e. omit the "equals" column). Then solve as usual.

Discovery 2.5 In a homogeneous system, what is the relationship between the number of variables, the number of "leading ones" in the most reduced form of the coefficient matrix, and the number of parameters required to solve the system? What pattern of leading ones in a completely reduced coefficient matrix tells you that the corresponding homogeneous system has a single, unique solution?

Discovery 2.6 Consider system

$$\left\{\begin{array}{rcrcrcl} 3x_1 & - & x_2 & + & 4x_3 & = & b_1, \\ x_1 & + & 2x_2 & - & x_3 & = & b_2, \\ 3x_1 & & & + & 3x_3 & = & b_3, \end{array}\right.$$

where the constants of each equation are not specified. For what values of the unknown constants b_1, b_2, b_3 is this system consistent?

To answer this question, row reduce the associated augmented matrix (below) until you are at a point where you can determine conditions on the constants

b_1, b_2, b_3 that ensures the system is consistent.

$$\left[\begin{array}{rrr|c} 3 & -1 & 4 & b_1 \\ 1 & 2 & -1 & b_2 \\ 3 & 0 & 3 & b_3 \end{array}\right].$$

2.2 Terminology and notation

row echelon form
a matrix that has the following properties:

- if a row has nonzero entries, its first nonzero entry is a one (called a **leading one**),
- each leading one occurs in a column that is to the right of the column containing the leading one in the row above it, and
- zero rows appear below all nonzero rows

reduced row echelon form
a row echelon form matrix that also has the following property:

- the leading one of each row has all other entries in the column that contains it equal to zero

Note 2.2.1 The acronyms **REF** and **RREF** are commonly used for **row echelon form** and **reduced row echelon form**, respectively.

row reduction
the process of using elementary row operations to reduce a matrix to REF or RREF

row equivalent matrices
matrices where it is possible to obtain one from the other through a sequence of elementary row operations

rank the number of leading ones in the RREF of the matrix

leading variables
the variables in a linear system whose columns in the RREF of the augmented matrix contain the leading one of some row

free variables
the variables in a linear system that are not leading variables

general solution
a set of parametric equations from which all solutions to a linear system can be obtain by choosing arbitrary values for the parameters

homogeneous system
a linear system in which the “equals” column is all zeros

coefficient matrix
the matrix for a linear system but without the “equals” column

trivial solution
the obvious solution to a homogeneous system obtained by setting all variables to equal zero

nontrivial solution
a solution to a homogeneous system that is not the trivial solution

2.3 Concepts

In this section.

- Subsection 2.3.1 *Reducing matrices*
- Subsection 2.3.2 *Solving systems*

2.3.1 Reducing matrices

In Discovery guide 2.1, you were led through a strategy to simplify an augmented matrix. Below is presented a step-by-step description of the strategy. But first, it is important to stress that your goal is *not* to become an expert row reducer — very few people ever need to know how to row reduce a matrix by hand outside of a linear algebra class. Computers are great at row reducing, and should be used to efficiently solve linear systems in the "real world." Here, we are not interested in learning calculation tricks or short-cuts — we can safely leave those to the experts that program computers to solve linear systems. (Prospective computational experts in the audience of this course can learn such calculation short-cuts in a *numerical methods* course.)

Goal 2.3.1 *Learn, understand, and become reasonably proficient at a* simple*, systematic strategy to reduce a matrix to RREF, so that we can use this knowledge to understand the theory of linear systems and matrices.*

Procedure 2.3.2 Reduce a matrix to RREF.

1. *Obtain a leading one in a column as far to the left as possible, then move the row containing this leading one to the top row. Use this leading one to eliminate (i.e. reduce to zero) all other entries in that column.*

2. *Ignoring the first row, obtain a leading one in a column as far to the left as possible, then move the row containing this new leading one to the second row. Use this new leading one to eliminate all other entries in that column (including in the first row now).*

3. *Ignoring the first and second rows, obtain a leading one in a column as far to the left as possible, then move the row containing this new leading one to the third row. Use this new leading one to eliminate all other entries in that column (including in the first and second rows now).*

4. *Continue in this fashion until all rows either have a leading one or contain all zeros.*

The choice and order of row operations you use to implement this strategy depends on the augmented matrix you start with, and knowing how to proceed is a skill that you will develop through practise and experience.

2.3.2 Solving systems

In the end, we will want to turn our simplified RREF matrix back into a system of equations. When we do this, every **leading one** corresponds to a **leading variable** that has a coefficient of 1, and so is easy to isolate and solve for in terms of the other variables. Another way to think of this is that *a leading variable is* constrained *by the equation in which it appears*, and its value depends on the values of the other variables in that equation. On the other hand, every variable that does not have a leading one in its column of the RREF matrix *cannot* be solved for without going in circles: you cannot solve for variable x in terms of variables y and z, and then turn around and solve for variable y in terms of variables x and z. A variable without a leading one becomes a **free variable**: there are no constraints on its value, and *every* choice of value for that variable leads to one or more solutions (depending on choices of values for other free variables) for the system similarly to Discovery 1.5.

Procedure 2.3.3 Describe the solution set of a linear system. *To determine the solution set of a system of equations from the corresponding RREF matrix, expressed in terms of parametric equations if necessary (if there are free variables), carry out the following steps.*

1. *For each variable column that does not have a leading one, assign a parameter to the corresponding variable. Use different letters for different free variables.*

2. *For each nonzero row, turn the row back into an equation and isolate the leading variable. Substitute in the associated parameter for each free variable that appears in the equation.*

For a homogeneous system, as in Discovery 2.4, there is no need to work with the full augmented matrix, since no elementary row operation will ever change the column of zeros on the right. Instead, we reduce just the coefficient matrix, making sure to remember that we are dealing with a homogeneous system when it is time to convert back to equations and solve the simplified system.

2.4 Examples

In this section.

- Subsection 2.4.1 *Worked examples from the discovery guide*

Here we use our procedures to use matrices to reduce and solve the systems from Discovery guide 2.1. Here are a few things to note about our method.

- We only use the three *elementary* row operations. It sometimes is possible to reduce a bit faster using non-elementary operations such as adding a multiple of a row to a multiple of another row, but remember we are not interested in short-cuts, and using non-elementary operations will get us into trouble in later topics.

- We sometimes perform more than one operation at the same time. This is an acceptable short-cut, as long as we ***never simultaneously modify a row and also use that row to modify another row***.

- We write down the row operation(s) we are using in that reduction step to the right of the matrix, to keep track of what we are doing.

- We don't always have to multiply a row by a fraction to get a leading one — we can sometimes use a difference between entries in a column, and avoid fractions that way.

- There are many different sequences of operations one could use to get from initial augmented matrix to an RREF matrix. The reductions in the examples below are not the only way, nor are they necessarily the *best* way to proceed. As long as we steadily progress toward RREF, that's all that matters.

- We never write an equals sign between matrices when row reducing. We will explore what it means for two matrices to be equal in Discovery guide 4.1, and this here is not it. When you perform a row operation, the result is a *different* matrix than the original, and the two matrices represent *different* systems of equations. However, the two matrices *are* related, and it is to express this relationship that we have the terminology **row equivalent**.

2.4.1 Worked examples from the discovery guide

Example 2.4.1 One unique solution. From Discovery 2.1:

$$\left\{\begin{array}{rcrcrcl} 2x & & & - & 2z & = & 4, \\ x & - & y & & & = & 3, \\ 4x & - & 2y & - & 3z & = & 7. \end{array}\right.$$

We form the augmented matrix for the system, and reduce.

$$\left[\begin{array}{rrr|r} 2 & 0 & -2 & 4 \\ 1 & -1 & 0 & 3 \\ 4 & -2 & -3 & 7 \end{array}\right] \begin{array}{l} R_1 \leftrightarrow R_2 \\ \\ \\ \end{array}$$

$$\longrightarrow \left[\begin{array}{rrr|r} 1 & -1 & 0 & 3 \\ 2 & 0 & -2 & 4 \\ 4 & -2 & -3 & 7 \end{array}\right] \begin{array}{l} \\ R_2 - 2R_1 \\ R_3 - 4R_1 \end{array}$$

$$\longrightarrow \left[\begin{array}{rrr|r} 1 & -1 & 0 & 3 \\ 0 & 2 & -2 & -2 \\ 0 & 2 & -3 & -5 \end{array}\right] \begin{array}{l} \\ \frac{1}{2}R_2 \\ \\ \end{array}$$

$$\longrightarrow \left[\begin{array}{rrr|r} 1 & -1 & 0 & 3 \\ 0 & 1 & -1 & -1 \\ 0 & 2 & -3 & -5 \end{array}\right] \begin{array}{l} R_1 + R_2 \\ \\ R_3 - 2R_2 \end{array}$$

$$\longrightarrow \left[\begin{array}{rrr|r} 1 & 0 & -1 & 2 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & -1 & -3 \end{array}\right] \begin{array}{l} \\ \\ -R_3 \end{array}$$

$$\longrightarrow \left[\begin{array}{rrr|r} 1 & 0 & -1 & 2 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & 3 \end{array}\right] \begin{array}{l} R_1 + R_3 \\ R_2 + R_3 \\ \\ \end{array}$$

$$\longrightarrow \left[\begin{array}{rrr|r} 1 & 0 & 0 & 5 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 3 \end{array}\right]$$

Every variable column has a leading one, so there are no free variables and no parameters are required. We can solve for each variable as a specific number, so the system has one unique solution: $x = 5$, $y = 2$, and $z = 3$. □

Example 2.4.2 Infinite number of solutions. From Discovery 2.2:

$$\left\{\begin{array}{rcrcrcr} 3x & + & 6y & + & 5z & = & -9, \\ 2x & + & 4y & + & 3z & = & -5, \\ 3x & + & 6y & + & 6z & = & -12. \end{array}\right.$$

We form the augmented matrix for the system, and reduce.

$$\begin{aligned}
&\left[\begin{array}{ccc|c} 3 & 6 & 5 & -9 \\ 2 & 4 & 3 & -5 \\ 3 & 6 & 6 & -12 \end{array}\right]\begin{array}{l} R_1-R_2 \\ \\ \\ \end{array}\\
\longrightarrow&\left[\begin{array}{ccc|c} 1 & 2 & 2 & -4 \\ 2 & 4 & 3 & -5 \\ 3 & 6 & 6 & -12 \end{array}\right]\begin{array}{l} \\ R_2-2R_1 \\ R_3-3R_1 \end{array}\\
\longrightarrow&\left[\begin{array}{ccc|c} 1 & 2 & 2 & -4 \\ 0 & 0 & -1 & 3 \\ 0 & 0 & 0 & 0 \end{array}\right]\begin{array}{l} \\ -R_2 \\ \\ \end{array}\\
\longrightarrow&\left[\begin{array}{ccc|c} 1 & 2 & 2 & -4 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 0 & 0 \end{array}\right]\begin{array}{l} R_1-2R_2 \\ \\ \\ \end{array}\\
\longrightarrow&\left[\begin{array}{ccc|c} 1 & 2 & 0 & 2 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 0 & 0 \end{array}\right]
\end{aligned}$$

The second column does not contain a leading one, so variable y is free and we assign to it a parameter: $y = t$. We can then use the simplified system

$$\left\{\begin{array}{rcrcrcr} x & + & 2y & & & = & 2, \\ & & & & z & = & -3, \\ & & & & 0 & = & 0. \end{array}\right.$$

to solve for $x = 2 - 2t$ and $z = -3$. In parametric form, the **general solution** of the system can be expressed as

$$x = 2 - 2t, \qquad y = t, \qquad z = -3,$$

and every **particular solution** to the system can be obtained by choosing a value for t. For example, the particular solution associated to $t = 3$ is

$$x = -4, \qquad y = 3, \qquad z = -3,$$

and the particular solution associated to $t = -\sqrt{2}$ is

$$x = 2 + 2\sqrt{2}, \qquad y = -\sqrt{2}, \qquad z = -3.$$

□

Example 2.4.3 No solution. From Discovery 2.3:

$$\left\{\begin{array}{rcrcrcr} x & + & 2y & + & z & = & 2, \\ 2x & + & 5y & + & 2z & = & -3, \\ 2x & + & 4y & + & 2z & = & -1. \end{array}\right.$$

We form the augmented matrix for the system, and reduce.

$$\begin{aligned}
&\left[\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 2 & 5 & 2 & -3 \\ 2 & 4 & 2 & -1 \end{array}\right]\begin{array}{l} \\ R_2-2R_1 \\ R_3-2R_1 \end{array}\\
\longrightarrow&\left[\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 1 & 0 & -7 \\ 0 & 0 & 0 & -5 \end{array}\right]\begin{array}{l} R_1-2R_2 \\ \\ -\frac{1}{5}R_3 \end{array}
\end{aligned}$$

$$\longrightarrow \left[\begin{array}{ccc|c} 1 & 0 & 1 & 16 \\ 0 & 1 & 0 & -7 \\ 0 & 0 & 0 & 1 \end{array}\right] \begin{array}{l} R_1 - 16R_3 \\ R_2 + 7R_3 \\ \end{array}$$

$$\longrightarrow \left[\begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right]$$

Here we have a leading one in the “equals” column. If we turn that third row back into an equation, we have

$$0x + 0y + 0z = 1,$$

but there are no possible values of x, y, z that satisfy this equation. Therefore, the system is **inconsistent**. (Of course, we could have seen that this would happen right from the second matrix, and could have stopped there. But we went all the way to RREF to have another example demonstrating the row reduction strategy. In practice, we should stop reducing as soon as we can see that the system will be inconsistent.) □

Example 2.4.4 A homogeneous system. From Discovery 2.4:

$$\left\{\begin{array}{rcrcrcrcl} 3x_1 & + & 6x_2 & - & 8x_3 & + & 13x_4 & = & 0, \\ x_1 & + & 2x_2 & - & 2x_3 & + & 3x_4 & = & 0, \\ 2x_1 & + & 4x_2 & - & 5x_3 & + & 8x_4 & = & 0. \end{array}\right.$$

For a **homogeneous** system, we only reduce the **coefficient matrix**, since elementary row operations will never change an “equals” columns that contains all zeros.

$$\left[\begin{array}{cccc} 3 & 6 & -8 & 13 \\ 1 & 2 & -2 & 3 \\ 2 & 4 & -5 & 8 \end{array}\right] \begin{array}{l} R_1 \leftrightarrow R_2 \\ \\ \end{array}$$

$$\longrightarrow \left[\begin{array}{cccc} 1 & 2 & -2 & 3 \\ 3 & 6 & -8 & 13 \\ 2 & 4 & -5 & 8 \end{array}\right] \begin{array}{l} \\ R_2 - 3R_1 \\ R_3 - 2R_1 \end{array}$$

$$\longrightarrow \left[\begin{array}{cccc} 1 & 2 & -2 & 3 \\ 0 & 0 & -2 & 4 \\ 0 & 0 & -1 & 2 \end{array}\right] \begin{array}{l} \\ -\frac{1}{2}R_2 \\ \end{array}$$

$$\longrightarrow \left[\begin{array}{cccc} 1 & 2 & -2 & 3 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & -1 & 2 \end{array}\right] \begin{array}{l} R_1 + 2R_3 \\ \\ R_3 + R_2 \end{array}$$

$$\longrightarrow \left[\begin{array}{cccc} 1 & 2 & 0 & -1 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

To solve, remember that this is just the *coefficient* matrix for the simplified system, so all columns correspond to a variable, and the “equals” column is still all zeros but does not appear. We have two free variables, corresponding to the lack of leading one in the second and fourth columns. So set parameters $x_2 = s$ and $x_4 = t$. The first two rows turn into equations

$$\begin{array}{rcrcrcrcl} x_1 & + & 2x_2 & & & - & x_4 & = & 0, \\ & & & & x_3 & - & 2x_4 & = & 0, \end{array}$$

from which we obtain the general solution in parametric form

$$\begin{aligned} x_1 &= -2x_2 + x_4 & x_2 &= s, & x_3 &= 2x_4 & x_4 &= t. \\ &= -2s + t, & & & &= 2t, & & \end{aligned}$$

□

Example 2.4.5 Correspondence between the solutions to homogeneous and nonhomogeneous systems with the same coefficient matrix. Consider the homogeneous system

$$\left\{\begin{array}{rcrcrcl} 3x & + & 6y & + & 5z & = & 0, \\ 2x & + & 4y & + & 3z & = & 0, \\ 3x & + & 6y & + & 6z & = & 0. \end{array}\right.$$

As in the previous example, to solve we work with just the coefficient matrix

$$\begin{bmatrix} 3 & 6 & 5 \\ 2 & 4 & 3 \\ 3 & 6 & 6 \end{bmatrix}.$$

But notice that *this is the same coefficient matrix as for the system in Discovery 2.2*, and the same row reduction sequence we used to solve that system in Example 2.4.2 would reduce this coefficient matrix to

$$\begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

And from here we also take the same steps as in Example 2.4.2 to solve this system. Assign parameter t to free variable y, and use the simplified homogeneous system

$$\left\{\begin{array}{rcrcrcl} x & + & 2y & & & = & 0, \\ & & & & z & = & 0, \\ & & & & 0 & = & 0. \end{array}\right.$$

to solve for $x = -2t$ and $z = 0$. Let's compare the parametric forms of the solutions to the original nonhomogeneous system from Discovery 2.2 and the corresponding homogeneous system solved here.

Nonhomogeneous	*Homogeneous*
$x = 2 + (-2)t$	$x = (-2)t$
$y = 0 + t$	$y = t$
$z = -3 + 0t$	$z = 0t$

We have added some zeros and ts to emphasize the similarity between the solutions. To interpret this similarity, remember that every value of t provides a particular solution to the systems. When $t = 0$, the corresponding solutions are $\{x = 2, y = 0, z = -3\}$ for the nonhomogeneous system and the trivial solution for the homogeneous system. For every other value of t, it seems that the corresponding solution for the nonhomogeneous system is equal to that "initial" particular solution $\{x = 2, y = 0, z = -3\}$ plus the corresponding homogeneous solution values. In Chapter 4 we will see that this pattern emerges in every nonhomogeneous system (see Lemma 4.5.4). □

2.5 Theory

In this section.

2.5.1 Reduced matrices

While there are many different sequences of row operations we could use to obtain a row equivalent RREF matrix, we have the following.

Theorem 2.5.1 Uniqueness of RREF. *For each matrix, there is one unique RREF matrix to which it is row equivalent.*

Remark 2.5.2 The same is not true about REF. When we are row reducing, there is usually a point where we reach REF but are not yet at RREF. From this point on as we further progress toward RREF, every matrix we produce will be both in REF and row equivalent to the original matrix. So a matrix can be row equivalent to many REF matrices.

Remark 2.5.3 We have defined the rank of a matrix to be the number of leading ones in the RREF of the matrix. If we did not have uniqueness of RREFs, there would be ambiguity in this definition from the possibility that different RREFs for a given starting matrix could have different numbers of leading ones. With the above theorem, we now know that there is no such possibility; the mathematical jargon for this certainty is to say that the definition of rank is **well-defined**.

Though rank is defined in terms of the RREF of a matrix, from our experience row reducing we can see that row operations cannot increase or decrease the number of leading ones that we will ultimately end up with.

Proposition 2.5.4 Rank from REF. *The rank of a matrix is equal to the number of leading ones in any REF matrix to which it is row equivalent.*

2.5.2 Solving systems using matrices

Using row operations to simplify and solve systems of equations works precisely because of the following.

Theorem 2.5.5 *Row equivalent matrices represent systems of equations that have the same solution set.*

Proof. We will delay proving this theorem until after we have developed more matrix theory. (See Theorem 6.5.10 in Subsection 6.5.4.) ■

The reason for this is that elementary row operations do not change the information inherently contained in the equations represented by the rows of a matrix. They modify and combine how this information is expressed, but no new information can be introduced through the elementary row operations, and also no information is ever lost.

When determining solution sets, our experience in Discovery guide 2.1 leads us to the following.

Proposition 2.5.6 Patterns of consistent/inconsistent systems.

1. *A system is inconsistent precisely when the RREF for its augmented matrix has a leading one in the "equals" column.*

2. *A consistent system has one unique solution precisely when the RREF for*

its augmented matrix has a leading one in every *variable column.*

3. *A consistent system has infinite solutions when it requires parameters to solve; that is, when the RREF for its augmented matrix has at least one variable column that does* not *contain a leading one.*

Warning 2.5.7 Ending up with a row of zeros in the RREF for a system's augmented matrix ***does not*** indicate that parameters will be needed. It is possible for the RREF matrix for a consistent system to *both* satisfy Statement 2 of Proposition 2.5.6 *and* to have a row of zeros.

Check your understanding. Can you write down an example of such an RREF matrix as desribed in the Warning? In order to be able to do this, what must be true about the "size" of the matrix?

We can restate Statement 2 and Statement 3 of Proposition 2.5.6 using the notion of **rank**: *a consistent system has a unique solution precisely when the rank of its augmented matrix is equal to the number of variables, and has infinite solutions precisely when the rank of its augmented matrix is strictly less than the number of variables*. For the infinite solutions case, we can be precise about the number of parameters required.

Proposition 2.5.8 Number of required parameters. *For a consistent system, the number of parameters required to express the general solution in parametric form satisfies*

$$(\textit{number of parameters}) = (\textit{number of variables}) - (\textit{rank of augmented matrix}).$$

We should also record what we learned about homogeneous systems in Discovery 2.4.

Theorem 2.5.9 Consistency of homogeneous systems. *A homogeneous system always has the trivial solution* $x_1 = 0, x_2 = 0, \ldots, x_n = 0$ *in its solution set. Thus, every homogeneous system is consistent.*

In light of this, for a homogeneous system, we can ignore Statement 1 of Proposition 2.5.6. Also, from our experience solving homogeneous systems so far, in Statement 2 and Statement 3 of Proposition 2.5.6 we can replace the words "augmented matrix" with "coefficient matrix".

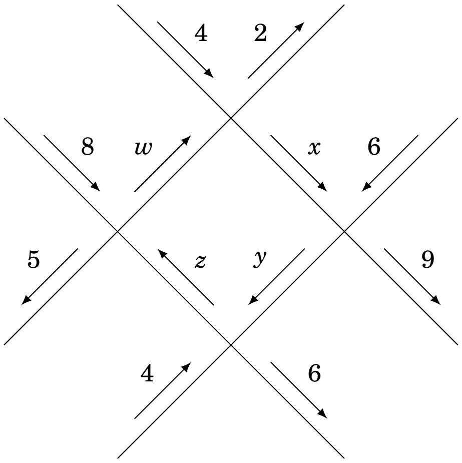
CHAPTER 3

Using systems of equations

3.1 Discovery guide

In this set of discovery activities, we look at some places where linear systems naturally arise.

Discovery 3.1 Use the Law of Conservation (in this case, *flow in equals flow out* at each point of intersection) in the flow network below to set up a system of equations to determine the internal flow rates. (Do not solve your system.)



Discovery 3.2 Set up a system of equations to balance the chemical equation:

$$a\mathrm{NH}_3 + b\mathrm{O}_2 \to c\mathrm{NO} + d\mathrm{H}_2\mathrm{O}.$$

Do not solve your system.

No shortcuts. You might have learned some procedure for balancing a chemical equation in high school. We are not interested in that procedure. We would like to see how attempting to balance a chemical equation has a linear system at its root.

Discovery 3.3 Any two (distinct) points in the Cartesian plane determine a unique line. Set up a system of equations that would let you solve for the slope and y-intercept of the line $y = mx + b$ that passes through the points $(-3,4)$ and $(2,-1)$ (but do not solve the system). Write down the augmented matrix for your system.

Discovery 3.4 Any three (distinct, non-collinear) points in the Cartesian plane determine a unique parabola. Set up a system of equations that would let you solve for the coefficients a, b, c of the parabola $y = ax^2 + bx + c$ that passes through the points $(-1, -4)$, $(1, 0)$, and $(2, 5)$ (but do not solve the system). Write down the augmented matrix for your system.

Discovery 3.5 Based on the previous two activities and their answers, how many points are necessary to determine a unique degree n polynomial $y = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$? If you knew such points

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots,$$

and you used these points to create a linear system to determine the coefficients of the polynomial, what would be the pattern in the rows of the resulting augmented matrix for the system?

3.2 Examples

In this section.

3.2.1 A simple example

Some of the first recorded uses of systems of equations in human history (without all the modern algebraic symbolism, of course) were applications to agriculture and commerce.

Problem 3.2.1 Nutrient profiles in horse feed. You have determined from recommendations in reputable reference sources that your large working horse requires 1150 g of protein, 36 g of calcium, and 25 g of phosphorous daily. You have had samples of your hay, grain, and pasture analyzed and have determine their nutritional components as percentages by mass.

	Hay	Grain	Pasture
Protein	8.2	13.9	4.1
Calcium	0.46	0.06	0.15
Phosphorous	0.21	0.45	0.07

How much of each feed type should your horse consume daily?

Solution. Let H, G, P represent the amount in kilograms of the three types of feed that the horse will be fed. Then each nutritional requirement leads to an equation.

$$\begin{array}{lrcrcrcl} \text{Protein:} & 0.082H & + & 0.139G & + & 0.041P & = & 1.150 \\ \text{Calcium:} & 0.0046H & + & 0.0006G & + & 0.0015P & = & 0.036 \\ \text{Phosphorous:} & 0.0021H & + & 0.0045G & + & 0.0007P & = & 0.025 \end{array}$$

Multiplying each equation by 10^4 to clear all decimals, we obtain an augmented matrix for the system,

$$\left[\begin{array}{ccc|c} 820 & 1390 & 410 & 11500 \\ 46 & 6 & 15 & 360 \\ 21 & 45 & 7 & 250 \end{array}\right]$$

If one were to solve this system, it would be revealed that the horse needs to eat close to 18 kg of pasture and be fed about 1.7 kg of hay and close to 2 kg of grain daily. □

3.2.2 Flow in networks

In a traffic network, fluid network, communications network, etc., matter cannot be created or destroyed. So at each node we can always apply some sort of law of conservation: the number of units entering the node (whether cars, litres of fluid, data packets, etc.) must be equal to the number of units exiting the node.

Let's apply this to the network in Discovery 3.1, starting at the top node and working clockwise to form *flow-in-equals-flow-out* equations.

$$\begin{aligned} 4+w &= 2+x \\ x+6 &= 9+y \\ y+4 &= z+6 \\ 8+z &= w+5 \end{aligned}$$

In order to facilitate conversion to an augmented matrix, we usually write a system of equations with all the variable terms on the left and collect all the constant values on the right.

$$\left\{\begin{array}{rcrcrcrcr} w & - & x & & & & & = & -2 \\ & & x & - & y & & & = & 3 \\ & & & & y & - & z & = & 2 \\ -w & & & & & & z & = & -3 \end{array}\right. \implies \left[\begin{array}{rrrr|r} 1 & -1 & 0 & 0 & -2 \\ 0 & 1 & -1 & 0 & 3 \\ 0 & 0 & 1 & -1 & 2 \\ -1 & 0 & 0 & 1 & -3 \end{array}\right]$$

Looking at the network diagram in Discovery 3.1, notice how all the external legs are known, and the internal legs are unknown. In trying to measure the behaviour of a system, it might be necessary to try to be as as unintrusive as possible, so you might be confined to measuring behaviour at points leading in or out of the overall system. Unfortunately, if you try to solve the system above, you will find that at least one more measurement of one of the internal legs is necessary in order to come to a definite solution.

3.2.3 Balancing chemical equations

Similarly to network analysis, there is a law of conservation at play in a chemical reaction since atoms are not created, destroyed, or changed to other kinds of atoms. So all of the atoms that make up the reactant particles must also be present in the product particles.

Let's apply this to the chemical reaction in Discovery 3.2, analyzing each atom in turn to balance the number of that atom in the reactant particles with the number of that atom in the product particles.

$$\begin{array}{ll} \text{Nitrogen:} & a = c \\ \text{Hydrogen:} & 3a = c + 2d \\ \text{Oxygen:} & 2b = c + d \end{array}$$

Again, we move all the variables to one side, obtaining in this case a homogeneous system, and then convert to a matrix.

$$\left\{\begin{array}{rcrcrcrcr} a & & & - & c & & & = & 0 \\ 3a & & & - & c & - & 2d & = & 0 \\ & & 2b & - & c & - & d & = & 0 \end{array}\right. \implies \left[\begin{array}{rrrr} 1 & 0 & -1 & 0 \\ 3 & 0 & -1 & -2 \\ 0 & 2 & -1 & -1 \end{array}\right]$$

This system must be consistent because it is homogeneous, but we have four variables and only three equations, so the solution will require a parameter. This makes sense physically, because we could always increase the number of reactant particles and just produce a larger number of product particles, but the parametric equations in the system solution will constrain the numbers of particles to be in balance *relative to each other*. But usually we prefer to describe the reaction as simply as possible by choosing the parameter value to be the smallest positive integer that clears all fractions that may have arisen in the solving process.

3.2.4 Polynomial interpolation

It is a fundamental principle in plane geometry that given two distinct points there is one unique line that passes through those points. And this principle continues to higher degree polynomials.

Note. A line is a degree-one polynomial.

For three points in the plane with disinct x-values there exists one unique parabola that passes through those points (where we consider a line as a degenerate form of parabola in the case that the three points are collinear). For four points with distinct x-values, there exists a unique cubic polynomial whose graph passes through all four points. And so on.

It may seem that this is not a *linear* problem, since polynomial functions involve powers of the variable x. But *x is not the variable here* — the unknown coefficients that define the particular polynomial function are what we are trying to solve for.

To illustrate how linear algebra can solve this problem, let's work through the associated discovery activities from Discovery guide 3.1.

Example 3.2.2 Linear interpolation. In Discovery 3.3, we would like to determine the line $y = mx + b$ that passes through the points $(-3,4)$ and $(2,-1)$. A point in the plane is on a particular line precisely when its coordinates satisfy the equation that defines the line. Requiring this gives us two equations, one for each point:

$$\begin{aligned} 4 &= m\cdot(-3)+b, \\ -1 &= m\cdot 2 + b. \end{aligned}$$

We already have the variables to one side, so we will just flip the equations around. However, we have chosen to put the variables in the order b, m to emphasize a pattern that will become evident as we do more examples.

$$\left\{\begin{array}{rcrcr} b & - & 3m & = & 4 \\ b & + & 2m & = & -1 \end{array}\right. \qquad \Longrightarrow \qquad \left[\begin{array}{rr|r} 1 & -3 & 4 \\ 1 & 2 & -1 \end{array}\right]$$

Solving this system would lead to one unique solution, as expected. □

Example 3.2.3 Quadratic interpolation. In Discovery 3.4, we would like to determine the parabola $y = ax^2 + bx + c$ that passes through the points $(-1,-4)$, $(1,0)$, and $(2,5)$. Again, each point leads to an equation by requiring that the point's coordinates satisfy the parabola's defining equation.

$$\begin{aligned} -4 &= a(-1)^2 + b(-1) + c \\ 0 &= a(1)^2 + b(1) + c \\ 5 &= a(2)^2 + b(2) + c \end{aligned}$$

Again, we will reverse the order of the variables to highlight the patterns.

$$\left\{\begin{array}{rcrcrcr} c & + & (-1)b & + & (-1)^2a & = & -4 \\ c & + & 1b & + & 1^2a & = & 0 \\ c & + & 2b & + & 2^2a & = & 5 \end{array}\right. \qquad \Longrightarrow \qquad \left[\begin{array}{rrr|r} 1 & -1 & (-1)^2 & -4 \\ 1 & 1 & 1^2 & 0 \\ 1 & 2 & 2^2 & 5 \end{array}\right]$$

And again, solving this system would lead to one unique solution, as expected. □

General interpolation. Now let's set up the solution to the general polynomial interpolation problem, as in Discovery 3.5. We have undetermined, degree-n polynomial equation $y = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ that we would to make pass through the points $(x_1, y_1), (x_2, y_2), \ldots, (x_{n+1}, y_{n+1})$.

Note. We always need one more point than the degree of the polynomial, because that is the number of coefficients in the polynomial.

Hopefully the pattern is obvious now, allowing us to proceed directly to the corresponding augmented matrix.

$$\left[\begin{array}{ccccc|c} 1 & x_1 & x_1^2 & \cdots & x_1^n & y_1 \\ 1 & x_2 & x_2^2 & \cdots & x_2^n & y_2 \\ & & \vdots & & & \vdots \\ 1 & x_{n+1} & x_{n+1}^2 & \cdots & x_{n+1}^n & y_{n+1} \end{array}\right]$$

3.3 Terminology and notation

Vandermonde matrix

an $m \times n$ matrix where the entries in each row form a sequence $1, x_i, x_i^2, x_i^3, \ldots, x_i^n$ for some number x_i, so that the full matrix has form

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_m^n \end{bmatrix}$$

3.4 Theory

In this section.

- Subsection 3.4.1 *Polynomial interpolation*

3.4.1 Polynomial interpolation

We have seen that Vandermonde matrices naturally arise as coefficient matrices in attempting to solve polynomial interpolation problems. Our geometric fact that these problems always have solutions can be formulated as an algebraic property of these matrices.

Theorem 3.4.1 Consistency of Vandermonde matrices. *A linear system whose coefficient matrix is a Vandermonde matrix is always consistent as long as the number of equations is no more than the number of variables and the second column contains no repeat entries. In the case that the number of equations is equal to the number of variables, there is one unique solution to the system.*

Corollary 3.4.2 Consistency of the polynomial interpolation problem. *Given $n+1$ points in the plane with different x-values, there is one unique polynomial of degree n or less that passes through all of the points.*

CHAPTER 4

Matrices and matrix operations

4.1 Discovery guide

Discovery 4.1 Consider matrices

$$A = \begin{bmatrix} 1 & 2 & 3 \\ -1 & 3 & 2 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 1 \\ -1 & 4 \\ 1 & 0 \end{bmatrix}, \qquad C = \begin{bmatrix} -6 & 1 \\ 1 & 2 \end{bmatrix}.$$

For each matrix, how would you describe its **size** (or **dimensions**)?

Discovery 4.2 Consider matrices

$$A = \begin{bmatrix} 1 & 1 & 3 \\ -1 & 3 & 2 \end{bmatrix}, \qquad B = \begin{bmatrix} x^2 & 2x+3 & 3 \\ -1 & 3 & 2 \end{bmatrix},$$
$$C = \begin{bmatrix} 1 & 1 & 3 \\ x^2 & 3 & 2 \end{bmatrix}, \qquad D = \begin{bmatrix} x^2 & 2x+3 \\ -1 & 3 \end{bmatrix}.$$

(a) For what value(s) of x is B equal to A? C equal to A? D equal to A?

(b) Discuss what it means for two matrices to be equal.

Discovery 4.3 Consider matrices

$$A = \begin{bmatrix} 1 & 2 & 3 \\ -1 & 3 & 2 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 2 & 1 \\ -1 & 0 & 4 \end{bmatrix}, \qquad C = \begin{bmatrix} -6 & 1 \\ 1 & 2 \end{bmatrix}.$$

(a) What should $A+B$ mean? What should $A-B$ mean?

(b) What should $3A$ mean?

(c) Now let's consider the sum $A+C$.

- **(i)** Compute $A+C$. Call this result matrix D. What are the dimensions of D?
- **(ii)** Now compute $D-A$. Do this numerically, not algebraically; that is, forget where your result matrix D came from and actually compute $D-A$ using the same procedure that you used to subtract matrices in Task a. What are the dimensions of this result?
- **(iii)** Now let's remember that $D = A+C$. Algebraically, what result would you expect from computing $(A+C)-A$? Does your numerical computation in the previous step agree with your algebraic expectation?

(Keep in mind your answer to what it means for two matrices to be equal from Task 4.2.b.)

(iv) Given how things worked out, how do you feel about performing $A+C$ in the first place?

Discovery 4.4 The number zero is important in algebra, it lets us do things like the following.

$$\begin{aligned} a+5 &= 7 \\ a+5-5 &= 7-5 \\ a+0 &= 2 \\ a &= 2. \end{aligned}$$

The critical step for us right now is the last simplification of the left-hand side:

$$a+0=a.$$

(a) What matrix do you think will act like zero in matrix addition? Is the answer different for different dimensions?

(b) What will be the result if you multiply this special "zero" matrix by a number (similarly to Task 4.3.b)?

Discovery 4.5

(a) Use your idea from Task b of Discovery 4.2 to turn the following *single* matrix equation into a system of *two* equations in the unknowns c and d. (Don't bother to actually solve for the values of c and d.)

$$\begin{bmatrix} c+2d \\ 3d \end{bmatrix} = \begin{bmatrix} 5 \\ -3 \end{bmatrix}$$

Careful: What sizes are the two matrices above?

(b) Now do the reverse of Task a: write the following system of equations as a *single* matrix equation using a *column* matrix on each side of the equation:

$$\left\{ \begin{array}{rcrcrcr} x_1 & - & 3x_2 & - & x_3 & = & -4, \\ -2x_1 & + & 7x_2 & + & 2x_3 & = & 9. \end{array} \right.$$

Again, be careful about the sizes of your matrices! If you have an equals sign between two matrices, they must adhere to your principle from Task b of Discovery 4.2.

(c) The simplest system of equations is one equation in one unknown, i.e.

$$ax=b.$$

But we don't usually just think of this as *left-hand side* and *right-hand side*, we think of it in the pattern

$$\text{coefficient} \times \text{unknown} = \text{constant}.$$

Can we represent the system from Task b in a similar pattern using a matrix equation

$$A\mathbf{x}=\mathbf{b}?$$

(i) What should the **coefficient matrix** A be?

(ii) What should the **(column) matrix of unknowns x** be?

(iii) What should the **(column) matrix of constants b** be?

(d) On the left-hand side of the matrix equation $A\mathbf{x} = \mathbf{b}$, the operation *matrix-times-matrix* should compute to a *single* matrix. What size of matrix should this multiplication result be?

Hint. The result of computing $A\mathbf{x}$ must make sense in the matrix equality $A\mathbf{x} = \mathbf{b}$, per the pattern of **matrix equality** you described in Task 4.2.b.

(e) Finally, we want $A\mathbf{x} = \mathbf{b}$ to represent in one matrix equation the full system of two number equations from Task b. We already came up with a matrix equation to represent that system in Task b. Looking at your matrices A and $\mathbf{x}$ from Task c, and comparing with the left-hand side of your matrix equation from Task b, what procedure should be used to carry out the operation *matrix A times column* $\mathbf{x}$?

(f) The values $x_1 = 2$, $x_2 = 1$, $x_3 = 3$, represent a solution to the system in Task b. Verify this by carrying out the multiplication $A\mathbf{x}$, using your calculation procedure from Task e, and with the unknowns x_1, x_2, x_3 in the column matrix $\mathbf{x}$ replaced by these solution values. Then compare your calculation result with $\mathbf{b}$.

Discovery 4.6 Consider

$$A = \begin{bmatrix} 1 & -3 & -1 \\ -2 & 7 & 2 \end{bmatrix}, \qquad X = \begin{bmatrix} 2 & 0 & 2 \\ 1 & 3 & 0 \\ 3 & -1 & -2 \end{bmatrix}.$$

Compute the product AX by considering X as a collection of three columns

$$X = \begin{bmatrix} | & | & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \\ | & | & | \end{bmatrix}$$

and using the procedure for "matrix times column" that you developed in Discovery 4.5.

Discovery 4.7 We all know that 3 times 5 and 5 times 3 have the same result. Algebraically, we write that $ab = ba$ is true for all numbers a, b. What about matrices?

(a) Try it with matrices

$$A = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}, \qquad B = \begin{bmatrix} 3 & 2 \\ 1 & -1 \end{bmatrix}.$$

(b) Look back at matrices A and X from Discovery 4.6, where you computed the matrix product AX. Does multiplying XA in the opposite order even make sense?

Discovery 4.8 Considering the previous three activities about matrix multiplication, what patterns have you observed about the required sizes of the two matrices involved for things to work out?

In particular, if A has m rows and n columns, and B has k rows and ℓ columns, what relationship must there be between these numbers for the *matrix-times-columns* calculation method to make sense when computing AB? And in that case, what size will the resulting product matrix AB be?

Discovery 4.9 In the following, assume A, B are square matrices.

(a) What do you think A^2 means? A^3?

(b) Explain why the formula $(AB)^2 = A^2 B^2$ is *wrong*. What is the correct formula?

Hint. What does $(AB)^2$ *mean*? Then consider Discovery 4.7.

(c) Explain why the formula $(A+B)^2 = A^2 + 2AB + B^2$ is *wrong*. What is the correct formula?

Hint. FOIL.

4.2 Terminology and notation

$(i,j)^{\text{th}}$ **entry of a matrix**
the entry in the i^{th} row and j^{th} column of a matrix

size (or dimensions) of a matrix
the number of rows and columns in a matrix, usually written $m \times n$ to mean m rows and n columns

equal matrices
matrices with the same size, and the same numbers in corresponding entries

matrix addition
the new matrix obtained from two old matrices of identical sizes by adding corresponding entries

scalar multiple
the new matrix obtained from an old matrix obtained by multiplying every entry by the same number k; the common **scale factor** k is called a **scalar**

Remark 4.2.1 We will encounter the geometric origin of the word **scalar** in Chapter 11.

zero matrix
a matrix where every entry is zero, written $\mathbf{0}$

column vector
a matrix consisting of a single column

row vector
a matrix consisting of a single row

vector of unknowns
a column vector containing all of the variables in a system

vector of constants
a column vector containing all of the constants from the right-hand sides of the equations in a system

square matrix
a matrix with the same number of columns as rows

main diagonal
the diagonal of entries in a square matrix from top left to bottom right

transpose the new matrix obtained from an old matrix by turning rows into columns and columns into rows; we usually write A^{T} to mean the transpose of the matrix A

4.3 Concepts

In this section.

- Subsection 4.3.4 *Basic matrix operations*
- Subsection 4.3.5 *The zero matrix*
- Subsection 4.3.6 *Linear systems as matrix equations*
- Subsection 4.3.7 *Matrix multiplication*
- Subsection 4.3.9 *Transpose*

4.3.1 Matrix entries

Matrices are big, unwieldy things, so we often use a letter as a placeholder for a matrix, just as we might use a letter to represent a number in algebra. We usually use uppercase letters for matrices, as in Discovery guide 4.1. (Though sometimes we use a boldface lowercase letter to represent a column or row vector, as in Discovery 4.5.) When we want to refer to a specific entry in a matrix, we identify it by two indices: its row number and its column number, in that order. For example, the $(2,1)^{\text{th}}$ entry of matrix A of Discovery 4.2 is -1. When we have a matrix represented by an uppercase letter and want to also use letters to represent its entries, we usually use the lowercase version of the same letter, with the row and column indices in subscript. For example, for the matrix A of Discovery 4.2, the $(2,1)^{\text{th}}$ entry is $a_{21} = -1$. Sometimes we might write $[A]_{ij}$ to refer to the $(i,j)^{\text{th}}$ entry of matrix A, particularly when instead of a single letter inside the square brackets, we have a formula of letters.

4.3.2 Matrix dimensions

Matrices have an obvious notion of size, but we need two numbers to describe it: the number of rows and the number of columns. Again, by convention we always list number of rows first. For example, matrix A of Discovery 4.2 is size 2×3, meaning it has 2 rows and 3 columns. For a square matrix, the two numbers describing the size of A are equal, so we might just say that a square matrix A has size n to mean it is $n \times n$.

4.3.3 Matrix equality

In Discovery 4.2, you explored what it means for two matrices to be equal. In algebra involving numbers, we write $a = b$ when variables a and b represent the same number. That is, a and b are equal when they represent the same piece of information. Similarly, two “variable” matrices are equal when they represent the same information. In particular, two matrices are equal when they have the same numbers in corresponding entries. But size is also important here: in Discovery 4.2, matrix D can never be equal to matrix A no matter what value we choose for variable x, because A will always contain more information than D in its extra third column. So even before we compare entries, we require equal matrices to have the same size.

4.3.4 Basic matrix operations

In Discovery 4.3, you probably decided that addition and subtraction of matrices should be carried out in the obvious ways: we should just add or subtract corresponding entries. See Example 4.4.1 and Example 4.4.2.

For matrices that have different sizes, it may be tempting to “fill out” the smaller matrix with zeros so that it can be added to the larger. *But this would* *add more information to the smaller matrix that it's not supposed to have*, creating a *different* matrix prior to the addition. So we should resist this temptation; we will only ever add or subtract matrices that have the same size, and addition/subtraction of matrices of different sizes will remain **undefined**.

When we multiply a number a by 2 to get $2a$, we are doubling the value of a. In other words, we are *scaling* a by a scale factor (or **scalar**) of 2. Similarly, we can use a scalar to "scale" a matrix by multiplying every entry in the matrix by that number. If A is a matrix and k is a scalar (i.e. a number), then kA is the **scalar multiple** of A by k. See Example 4.4.3.

4.3.5 The zero matrix

The number zero plays a special role with respect to addition of numbers: it is the only number that has no effect when it is added to another number. For addition of matrices of a particular size, there is only one kind of matrix that has the same effect: a matrix filled with all zeros. We call such a matrix the **zero matrix**, and write **0** to represent it.

Remark 4.3.1 There are many zero matrices, one of every possible size of matrix. However, we still often say *the* zero matrix, because we are usually referring to the zero matrix of a particular size.

The zero matrix will allow us to do the matrix version of the algebra in the preamble to Discovery 4.4, since subtracting a matrix from itself will obviously result in the zero matrix. For more properties of the zero matrix, see Proposition 4.5.1 in Subsection 4.5.1.

4.3.6 Linear systems as matrix equations

Consider the system in Task b of Discovery 4.5:

$$\left\{\begin{array}{rcrcrcr} x_1 & - & 3x_2 & - & x_3 & = & -4, \\ -2x_1 & + & 7x_2 & + & 2x_3 & = & 9. \end{array}\right. \tag{$*$}$$

We would like to replace these two equations by a single matrix equation, which is easy enough to do:

$$\begin{bmatrix} x_1 - 3x_2 - x_3 \\ -2x_1 + 7x_2 + 2x_3 \end{bmatrix} = \begin{bmatrix} -4 \\ 9 \end{bmatrix}. \tag{$**$}$$

Note that both of these column matrices are 2×1 matrices — even though the entries of the left-hand matrix seem to contain a lot of numbers, *each row has only a single entry* because these formulas are calculation recipes that compute a *single number* out of several numbers, some known and some unknown.

To make such a matrix equation more resemble the basic linear equation pattern of

$$\text{coefficient} \times \text{unknown} = \text{constant},$$

we collect all the system coefficients into a **coefficient matrix**, all the variables into the **(column) vector of unknowns**, and all the right-hand constants into the **(column) vector of constants**:

$$A = \begin{bmatrix} 1 & -3 & -1 \\ -2 & 7 & 2 \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} -4 \\ 9 \end{bmatrix},$$

respectively.

Remark 4.3.2 It may seem more natural to write the vector of unknowns as a row vector instead of a column vector, but it is preferable mathematically to have all of the vectors involved be (roughly) the same *kind* of vector (even though they are often not *exactly* the same kind of vector, since they might not have the same size).

We would like to express the system in ($*$) as one matrix equation $A\mathbf{x} = \mathbf{b}$, and to do this we need to decide how A times $\mathbf{x}$ should work. But we already know how to represent the system as a single matrix equation (see ($**$)), so we should have

$$A\mathbf{x} = \begin{bmatrix} x_1 - 3x_2 - x_3 \\ -2x_1 + 7x_2 + 2x_3 \end{bmatrix},$$

or

$$\begin{bmatrix} 1 & -3 & -1 \\ -2 & 7 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 - 3x_2 - x_3 \\ -2x_1 + 7x_2 + 2x_3 \end{bmatrix}.$$

We can now see how a matrix times a column should proceed: multiply the entries of the first row of the matrix against the corresponding entries in the column, add these products, and put the result in the first entry of the result column matrix. Then multiply the second row of the matrix against the column in the same fashion and put the result in the second entry of the result column matrix. And so on, if the matrix has more than two rows. See Subsection 4.3.7 below for a more detailed description on this process.

With the matrix product $A\mathbf{x}$ defined in this way, the single matrix equation $A\mathbf{x} = \mathbf{b}$ now contains all the same information as the multiple linear equations of the original system.

4.3.7 Matrix multiplication

We can extend this *row-times-column* calculation procedure to define multiplication of two matrices (instead of just a matrix and a column vector) by thinking of the second matrix as a collection of columns,

$$B = \begin{bmatrix} | & | & & | \\ \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_\ell \\ | & | & & | \end{bmatrix} \quad \Longrightarrow \quad AB = \begin{bmatrix} | & | & & | \\ A\mathbf{b}_1 & A\mathbf{b}_2 & \cdots & A\mathbf{b}_\ell \\ | & | & & | \end{bmatrix}. \qquad (***)$$

This *matrix-times-columns* way of defining matrix multiplication will be very useful later. But right now, let's drill down to individual entries of the result AB.

Let's first consider the case of a $1 \times n$ row vector $\mathbf{a}$ times an $n \times 1$ column vector $\mathbf{b}$. In this case,

$$\mathbf{ab} = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} a_1b_1 + a_2b_2 + \cdots + a_nb_n \end{bmatrix}. \qquad (\dagger)$$

Notice that the result is a 1×1 matrix containing just a single entry.

Now let's consider a matrix A times a column $\mathbf{b}$, where we consider A as being made of row vectors. Then,

$$A\mathbf{b} = \begin{bmatrix} — & \mathbf{a}_1 & — \\ — & \mathbf{a}_2 & — \\ & \vdots & \\ — & \mathbf{a}_m & — \end{bmatrix} \mathbf{b} = \begin{bmatrix} \mathbf{a}_1\mathbf{b} \\ \mathbf{a}_2\mathbf{b} \\ \vdots \\ \mathbf{a}_m\mathbf{b} \end{bmatrix},$$

where each entry $\mathbf{a}_i\mathbf{b}$ in the result on the right is calculated by the *row-times-column* pattern from (†). However, we do not actually have a 1×1 matrix in each entry, but instead place the *number* that would be the sole entry in $\mathbf{a}_i\mathbf{b}$.

Finally, we can extend this to the case of matrix A times matrix B, by

$$AB = \begin{bmatrix} — & \mathbf{a}_1 & — \\ — & \mathbf{a}_2 & — \\ & \vdots & \\ — & \mathbf{a}_m & — \end{bmatrix} \begin{bmatrix} | & | & & | \\ \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_\ell \\ | & | & & | \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1\mathbf{b}_1 & \mathbf{a}_1\mathbf{b}_2 & \cdots & \mathbf{a}_1\mathbf{b}_\ell \\ \mathbf{a}_2\mathbf{b}_1 & \mathbf{a}_2\mathbf{b}_2 & \cdots & \mathbf{a}_2\mathbf{b}_\ell \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_m\mathbf{b}_1 & \mathbf{a}_m\mathbf{b}_2 & \cdots & \mathbf{a}_m\mathbf{b}_\ell \end{bmatrix}.$$

Pattern. The $(i,j)^{\text{th}}$ entry of matrix product AB is the result of a *row-times-column* calculation, as in (†), using the i^{th} row of A and the j^{th} column of B.

In order for each *row-times-column* calculation to work, we need the number of entries in a row of A to match up with the number of entries in a column of B. (Just as in the definition of matrix addition, we do not “fill out” a matrix with extra entries if these numbers do not match.) But the number of entries in a row of A is the number of columns of A, and the number of entries in a column of B is the number of rows of B.

Pattern. If A is $m \times n$ and B is $k \times \ell$, we can only compute AB if n and k are the same; otherwise, we say that the product AB is **undefined**. In the case that n and k are the same, the product AB has size $m \times \ell$.

An easy way to remember this is that if we want to multiply

$$m \times n \quad \text{times} \quad k \times \ell,$$

it will only work if the “inside” dimensions n and k match, and result will be the “outside” dimensions $m \times \ell$.

In Discovery 4.7, you found that one of the familiar rules of algebra is *not* true for matrix algebra: *matrices* cannot *be multiplied in any order*, because different orders of multiplication might yield different results. In fact, for non-square matrices, often one of the two orders of multiplication is not even *defined*.

Warning 4.3.3 When manipulating algebraic expressions where the letters represent matrices, be careful *not* to inadvertently use the algebra rule $BA = AB$, because it is *not* true for matrices.

4.3.8 Matrix powers

As you probably decided in Discovery 4.9, we define powers of matrices in the usual way: A^2 means “A times A,” A^3 means “A times A times A,” and so on.

Warning 4.3.4

- To compute A^2, you need to carry out the computation AA using the “row times column” definition of matrix multiplication. Just squaring every entry of A will *not* give you the correct result! And similarly for A^3, A^4, etc. — you need to carry out all the iterated multiplications. See Subsection 4.4.2 for example calculations.

- As in the second pattern discussed in Subsection 4.3.8, we can only compute the product $A^2 = AA$ if the number of columns of A is equal to the number of rows of A. That is, ***matrix powers are only defined for square matrices***.

The fact that reversing the order of matrix multiplication can produce a different result adds some extra wrinkles to the algebra of matrix powers. In

Discovery 4.9.b and Discovery 4.9.c, we need to be careful about order of multiplication. By definition, $(AB)^2$ means $(AB)(AB)$, but we cannot simplify this to $A^2B^2 = (AA)(BB)$, because order of multiplication matters, and we so we cannot in general change the order of multiplication of the inner B and A. Similarly, when using FOIL to expand $(A+B)^2 = (A+B)(A+B)$ (which *is* valid matrix algebra, see Subsection 4.5.1), for the O part of FOIL we get AB and for the I part we get BA, but these cannot be combined into $2AB$ in general because *order matters for matrix multiplication*.

4.3.9 Transpose

There is one more matrix operation that we did not explore in Discovery guide 4.1: the **transpose** of a matrix. To compute the transpose of a particular matrix A, take the entries of the first row of A and write them as the entries of the first *column* in a new matrix. Then take the entries of the second row of A and write them as the entries of the second column in the new matrix. And so on. The resulting new matrix is called the **transpose of** A, and we write A^{T} to mean this new matrix obtained from the old matrix A. See Subsection 4.4.5 for examples of computing transposes.

It is not possible at this stage to explain why we might want to use such an operation. If we are thinking of matrices as coefficient or augmented matrices of linear systems, why would we want all the coefficients in a particular equation in a system to become the coefficients attached to a particular variable in a new system? However, the transpose is such a simple operation that it is useful to include its properties in our development at this early stage.

Here are some things to notice about the operation of transpose as you look at the examples in Subsection 4.4.5. First, since we are taking rows of A and making them columns in A^{T}, the number of columns of A^{T} must be the number of rows of A. Also, the number of entries in a row of A becomes the number of entries in a column of A^{T}, so the same must be true about the number of rows of A^{T} versus the number of columns of A. That is, if A is size $m \times n$, then A^{T} is size $n \times m$. Second, instead of turning rows of A into columns of A^{T}, notice that we could take the *columns* of A and use them as *rows* in a new matrix, and the result would be the same as A^{T}. This symmetry means that if we compute the transpose of A^{T}, we will be back at A.

4.4 Examples

In this section.

4.4.1 Basic matrix operations

Here are some basic examples of matrix addition, subtraction, and scalar multiplication. For subtraction, watch out for double negatives!

Example 4.4.1 Matrix addition.

$$\begin{bmatrix} 1 & -2 \\ 3 & 4 \\ -5 & 6 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 1 & -2 \\ 11 & 0 \end{bmatrix} = \begin{bmatrix} 1+0 & -2+1 \\ 3+1 & 4+(-2) \\ -5+11 & 6+0 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 4 & 2 \\ 6 & 6 \end{bmatrix}$$

□

Example 4.4.2 Matrix subtraction.

$$\begin{bmatrix} 1 & -2 & 3 \\ 0 & -4 & -5 \end{bmatrix} - \begin{bmatrix} 0 & 1 & 1 \\ -2 & 11 & -1 \end{bmatrix} = \begin{bmatrix} 1-0 & -2-1 & 3-1 \\ 0-(-2) & -4-11 & -5-(-1) \end{bmatrix}$$
$$= \begin{bmatrix} 1 & -3 & 2 \\ 2 & -15 & -4 \end{bmatrix}$$

□

Example 4.4.3 Scalar multiplication of a matrix.

$$(-5)\begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix} = \begin{bmatrix} -5 & 10 \\ 15 & -20 \end{bmatrix}$$

□

4.4.2 Matrix multiplication

Example 4.4.4 A detailed multiplication example. Let's compute the matrix product AB, for

$$A = \begin{bmatrix} 3 & -2 \\ 1 & 0 \\ -4 & 5 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix}.$$

Notice that the sizes of A (3×2) and B (2×2) are compatible for multiplication in the order AB, and that the result will be size 3×2. First let's multiply A onto the columns of B.

$$\begin{bmatrix} 3 & -2 \\ 1 & 0 \\ -4 & 5 \end{bmatrix}\begin{bmatrix} 1 \\ -3 \end{bmatrix} = \begin{bmatrix} 3\cdot 1+(-2)\cdot(-3) \\ 1\cdot 1+0\cdot(-3) \\ -4\cdot 1+5\cdot(-3) \end{bmatrix} = \begin{bmatrix} 9 \\ 1 \\ -19 \end{bmatrix}$$
$$\begin{bmatrix} 3 & -2 \\ 1 & 0 \\ -4 & 5 \end{bmatrix}\begin{bmatrix} -2 \\ 4 \end{bmatrix} = \begin{bmatrix} 3\cdot(-2)+(-2)\cdot 4 \\ 1\cdot(-2)+0\cdot 4 \\ -4\cdot(-2)+5\cdot 4 \end{bmatrix} = \begin{bmatrix} -14 \\ -2 \\ 28 \end{bmatrix}$$

Combining these two computations, we get

$$AB = \begin{bmatrix} 3 & -2 \\ 1 & 0 \\ -4 & 5 \end{bmatrix}\begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix} = \begin{bmatrix} 9 & -14 \\ 1 & -2 \\ -19 & 28 \end{bmatrix}.$$

With some practise at matrix multiplication, you should be able to compute a product AB directly without doing separate computations for each column of the second matrix.

In this matrix multiplication example, notice that it does not make sense to even consider the possibility that $BA = AB$ because the sizes of B and A are not compatible for multiplication in the order BA, and so BA is undefined! □

Check your understanding. Is it *never* true that $BA = AB$? It should be obvious that it will be true if A is a square zero matrix and B is a square matrix of the same size. Can you come up with an example of 2×2 matrices A and B where neither is the zero matrix, and $BA = AB$ is true?

Example 4.4.5 Matrix powers. Since powers of matrices only work for *square matrices*, the power A^2 is undefined for the 3×2 matrix A in the previous matrix multiplication example. But we can compute B^2 for the 2×2 matrix B from that example.

$$\begin{aligned} B^2 = BB &= \begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix} \\ &= \begin{bmatrix} 1\cdot 1 + (-2)(-3) & 1(-2) + (-2)\cdot 4 \\ -3\cdot 1 + 4(-3) & (-3)(-2) + 4\cdot 4 \end{bmatrix} = \begin{bmatrix} 7 & -10 \\ -15 & 22 \end{bmatrix} \end{aligned}$$

To compute B^3, we can compute either of

$$\begin{aligned} B^3 = BBB = (BB)B = B^2B & \\ &= \begin{bmatrix} 7 & -10 \\ -15 & 22 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix} \\ &= \begin{bmatrix} 7\cdot 1 + (-10)(-3) & 7(-2) + (-10)\cdot 4 \\ -15\cdot 1 + 22(-3) & -15(-2) + 22\cdot 4 \end{bmatrix} \\ &= \begin{bmatrix} 37 & -54 \\ -81 & 118 \end{bmatrix}, \end{aligned}$$

or

$$\begin{aligned} B^3 = BBB = B(BB) = BB^2 & \\ &= \begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix} \begin{bmatrix} 7 & -10 \\ -15 & 22 \end{bmatrix} \\ &= \begin{bmatrix} 1\cdot 7 + (-2)(-15) & 1(-10) + (-2)\cdot 22 \\ (-3)\cdot 7 + 4(-15) & -3(-10) + 4\cdot 22 \end{bmatrix} \\ &= \begin{bmatrix} 37 & -54 \\ -81 & 118 \end{bmatrix}, \end{aligned}$$

and the result is the same. □

4.4.3 Combining operations

Example 4.4.6 Computing matrix formulas involving a combination of operations. Let's compute both $A(B + kC)$ and $AB + k(AC)$, for

$$A = \begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 2 \\ 1 & -1 \end{bmatrix}, \qquad C = \begin{bmatrix} 5 & 5 \\ -2 & -2 \end{bmatrix}, \qquad k = 3.$$

Keep in mind that *operations inside brackets should be performed first*, and as usual multiplication (both matrix and scalar) should be performed before addition (unless there are brackets to tell us otherwise).

$$\begin{aligned} A(B + kC) &= \begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix} \left(\begin{bmatrix} 0 & 2 \\ 1 & -1 \end{bmatrix} + 3 \begin{bmatrix} 5 & 5 \\ -2 & -2 \end{bmatrix} \right) \\ &= \begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix} \left(\begin{bmatrix} 0 & 2 \\ 1 & -1 \end{bmatrix} + \begin{bmatrix} 15 & 15 \\ -6 & -6 \end{bmatrix} \right) \end{aligned}$$

$$
\begin{aligned}
&= \begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix} \begin{bmatrix} 15 & 17 \\ -5 & -7 \end{bmatrix} \\
&= \begin{bmatrix} 25 & 31 \\ -65 & -79 \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
AB + k(AC) &= \begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix} \begin{bmatrix} 0 & 2 \\ 1 & -1 \end{bmatrix} + 3\left(\begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix} \begin{bmatrix} 5 & 5 \\ -2 & -2 \end{bmatrix} \right) \\
&= \begin{bmatrix} -2 & 4 \\ 4 & -10 \end{bmatrix} + 3 \begin{bmatrix} 9 & 9 \\ -23 & -23 \end{bmatrix} \\
&= \begin{bmatrix} -2 & 4 \\ 4 & -10 \end{bmatrix} + \begin{bmatrix} 27 & 27 \\ -69 & -69 \end{bmatrix} \\
&= \begin{bmatrix} 25 & 31 \\ -65 & -79 \end{bmatrix}
\end{aligned}
$$

□

Hopefully you're not surprised that we got the same final result for both the formulas $A(B+kC)$ and $AB+k(AC)$. From our familiar rules of algebra, we expect to be able to multiply A inside the brackets in the first expression, and then rearrange the order of multiplication by A and by k. However, we need to be careful — our "familiar" rules of algebra come from operations with *numbers*, and matrix algebra involves operations with *matrices*: addition, subtraction, and *two different* kinds of multiplication, scalar and matrix. We should not *blindly* expect all of our "familiar" rules of algebra to apply to matrix operations. We've already seen that the matrix version of the familiar rule $ba = ab$ is *not* true for matrix multiplication! In Subsection 4.5.1, we list the rules of algebra that *are* valid for matrix operations (which is *most* of our familiar rules from the algebra of numbers), and for some of the rules, in that same subsection we verify that they are indeed valid for matrices.

4.4.4 Linear systems as matrix equations

4.4.4.1 A first example

Example 4.4.7 A system as a matrix equation. Let's again consider the system from Task b of Discovery 4.5. To solve, we row reduce the associated augmented matrix to RREF as usual.

$$
\left[\begin{array}{rrr|r} 1 & -3 & -1 & -4 \\ -2 & 7 & 2 & 9 \end{array}\right] \qquad \xrightarrow[\text{reduce}]{\text{row}} \qquad \left[\begin{array}{rrr|r} 1 & 0 & -1 & -1 \\ 0 & 1 & 0 & 1 \end{array}\right]
$$

Variable x_3 is free, so assign a parameter $x_3 = t$. Then we can solve to obtain the general solution is parametric form,

$$
x_1 = -1 + t, \qquad x_2 = 1, \qquad x_3 = t.
$$

Let's check a couple of particular solutions against the matrix *equation* $A\mathbf{x} = \mathbf{b}$ that represents the system. Recall that for this system, $\mathbf{x}$ is the 3×1 column vector that contains the variables x_1, x_2, x_3. The particular solutions associated to parameter values $t = 0$ and $t = 3$ are

$$
t = 0: \qquad x_1 = -1, \qquad x_2 = 1, \qquad x_3 = 0;
$$

and

$$
t = 3: \qquad x_1 = 2, \qquad x_2 = 1, \qquad x_3 = 3.
$$

Let's collect the $t = 0$ solution values into the vector $\mathbf{x}$ and check $A\mathbf{x}$ versus $\mathbf{b}$:

$$\text{LHS} = A\mathbf{x} = \begin{bmatrix} 1 & -3 & -1 \\ -2 & 7 & 2 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -1-3+0 \\ 2+7+0 \end{bmatrix} = \begin{bmatrix} -4 \\ 9 \end{bmatrix} = \mathbf{b} = \text{RHS}.$$

So the solution to the linear system we got by row reducing did indeed give us a vector solution $\mathbf{x}$ to the matrix equation $A\mathbf{x} = \mathbf{b}$. Let's similarly check the $t = 3$ solution, as in Task f of Discovery 4.5:

$$\text{LHS} = A\mathbf{x} = \begin{bmatrix} 1 & -3 & -1 \\ -2 & 7 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 2-3-3 \\ -4+7+6 \end{bmatrix} = \begin{bmatrix} -4 \\ 9 \end{bmatrix} = \mathbf{b} = \text{RHS}.$$

Again, our system solution gives us a solution to the matrix equation. □

Check your understanding. Carry out the same verification as in Example 4.4.7 for the *general* solution to the system, with the parameter t left variable.

4.4.4.2 Expressing system solutions in vector form

We may use matrices and matrix algebra to express the solutions to solutions as column vectors. In particular, we can expand solutions involving parameters into a **linear combination** of column vectors. Expressing solutions this way allows us to see the effect of each parameter on the system.

Let's re-examine the systems in the examples from Section 2.4 as matrix equations, and express their solutions in vector form.

Example 4.4.8 Solutions in vector form: one unique solution. The system from Discovery 2.1 can be expressed in the form $A\mathbf{x} = \mathbf{b}$ for

$$A = \begin{bmatrix} 2 & 0 & -2 \\ 1 & -1 & 0 \\ 4 & -2 & -3 \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 4 \\ 3 \\ 7 \end{bmatrix}.$$

We solved this system in Example 2.4.1 and determined that it had one unique solution, $x = 5$, $y = 2$, and $z = 3$. In vector form, we write this solution as

$$\mathbf{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \\ 3 \end{bmatrix}.$$

□

Example 4.4.9 Solutions in vector form: an infinite number of solutions. The system from Discovery 2.2 can be expressed in the form $A\mathbf{x} = \mathbf{b}$ for

$$A = \begin{bmatrix} 3 & 6 & 5 \\ 2 & 4 & 3 \\ 3 & 6 & 6 \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} -9 \\ -5 \\ -12 \end{bmatrix}.$$

We solved this system in Example 2.4.2, and determined that it had an infinite number of solutions. We expressed the general solution to the system using parametric equations

$$x = 2 - 2t, \qquad y = t, \qquad z = -3,$$

In vector form, we expand this solution as

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

Notice how the solution is the sum of a **constant part**

$$\begin{bmatrix} 2 \\ 0 \\ -3 \end{bmatrix}$$

and a **variable part**

$$t \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}.$$

Further notice how the constant part is a particular solution to the system — it is the “initial” particular solution associated to the parameter value $t = 0$. □

Example 4.4.10 Solutions in vector form: a homogenous system. The system from Discovery 2.4 is homogeneous, so it can be expressed in the form $A\mathbf{x} = \mathbf{0}$ for

$$A = \begin{bmatrix} 3 & 6 & -8 & 13 \\ 1 & 2 & -2 & 3 \\ 2 & 4 & -5 & 8 \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix},$$

where $\mathbf{0}$ is the 3×1 zero column vector. We solved this system in Example 2.4.4, and determined that it had an infinite number of solutions. We expressed the general solution to the system using parametric equations

$$x_1 = -2s + t, \qquad x_2 = s, \qquad x_3 = 2t, \qquad x_4 = t.$$

In vector form, we expand this solution as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -2s+t \\ s \\ 2t \\ t \end{bmatrix} = \begin{bmatrix} -2s+t \\ s+0t \\ 0s+2t \\ 0s+t \end{bmatrix} = \begin{bmatrix} -2s \\ s \\ 0s \\ 0s \end{bmatrix} + \begin{bmatrix} t \\ 0t \\ 2t \\ t \end{bmatrix} = s \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} 1 \\ 0 \\ 2 \\ 1 \end{bmatrix}.$$

This time, the solution is a sum of *two* variables parts,

$$s \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad \text{and} \qquad t \begin{bmatrix} 1 \\ 0 \\ 2 \\ 1 \end{bmatrix},$$

since there are two parameters. And there is *no* constant part to the general solution, because if we set both parameters to zero we obtain the trivial solution $\mathbf{x} = \mathbf{0}$. A homogeneous system will always work out this way. (So it would be more accurate to say that the general solution to the system from Discovery 2.4 has *trivial* constant part, instead of saying it has *no* constant part.) □

Example 4.4.11 Solutions in vector form: patterns for homogeneous and nonhomogeneous systems with the same coefficient matrix. In

Example 2.4.5, we solved a homogenous system $A\mathbf{x} = \mathbf{0}$ with

$$A = \begin{bmatrix} 3 & 6 & 5 \\ 2 & 4 & 3 \\ 3 & 6 & 6 \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix},$$

and found an infinite number of solutions, with general solution expressed parametrically as

$$x = -2t, \qquad y = t, \qquad z = 0.$$

In vector form, we express this as

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

This homogeneous system has the same coefficient matrix as in Example 4.4.9 above, so it is not surprising that their general solutions are related. In particular, notice that both systems *have the same variable part*, but that the nonhomogeneous system from Example 4.4.9 has a non-trivial constant part. □

Compare. the pattern in Example 4.4.11 with the pattern in Example 2.4.5.

4.4.5 Transpose

Example 4.4.12 Computing transposes. Let's compute some transposes.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \qquad B = \begin{bmatrix} -1 & 2 & 3 \\ 5 & 0 & 4 \\ 6 & 7 & 1 \end{bmatrix} \qquad C = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \end{bmatrix}$$

$$A^{\mathrm{T}} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \qquad B^{\mathrm{T}} = \begin{bmatrix} -1 & 5 & 6 \\ 2 & 0 & 7 \\ 3 & 4 & 1 \end{bmatrix} \qquad C^{\mathrm{T}} = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \end{bmatrix}$$

The matrix A is size 2×3, so when we turn rows into columns to compute A^{T}, we end up with a 3×2 result. Matrices B and C are square, so each of their transposes end up being the same size as the original matrix. But also, the numbers for the entries in B and C were chosen to emphasize some patterns in the transposes of square matrices. In interchanging rows and columns in B, notice how entries to the upper right of the main diagonal move to the "mirror image" position in the lower left of the main diagonal, and vice versa. So for square matrices, we might think of the transpose as "reflecting" entries in the main diagonal, while entries right on the main diagonal end up staying in place. Finally, we might consider this same "reflecting-in-the-diagonal" view of the transpose for C, except C has the *same* entries in corresponding "mirror image" entries on either side of the diagonal, and so we end up with $C^{\mathrm{T}} = C$. □

4.5 Theory

In this section.

4.5.1 Rules of matrix algebra

When we want to work algebraically with letters that represent matrices, most of the familiar rules from the algebra of numbers still hold. We collect many of these rules of matrix algebra in the list below. We will not prove that *all* of these rules are valid, but we will verify some of the rules to demonstrate the general pattern of their proofs. For some of the proofs we will be more rigorous than others, but in all of the proofs we want to verify that the matrix on the left-hand side is equal to the one on the right-hand side.

Proposition 4.5.1 Matrix algebra. *The following are valid rules of matrix algebra. In each statement, assume that A, B, C are arbitrary matrices and $\mathbf{0}$ is a zero matrix, all of appropriate sizes so that the matrix operations can be carried out. In particular, in any rule involving a matrix power, the matrices involved are assumed to be square. Also assume that k and m are scalars, and that p and q are positive integers.*

1. *Basic rules of addition and multiplication.*
   (a) $B + A = A + B$
   (b) $A + (B + C) = (A + B) + C$
   (c) $A(B + C) = AB + AC$
   (d) $(A + B)C = AC + BC$
   (e) $A(BC) = (AB)C$
2. *Rules involving scalar multiplication.*
   (a) $k(A + B) = kA + kB$
   (b) $(k + m)A = kA + mA$
   (c) $(kA)B = k(AB)$
   (d) $A(kB) = k(AB)$
   (e) $k(mA) = (km)A$
   (f) $A - B = A + (-1)B$
3. *Rules involving a zero matrix.*
   (a) $A + \mathbf{0} = A$
   (b) $A - A = \mathbf{0}$
   (c) $A\mathbf{0} = \mathbf{0}$
   (d) $\mathbf{0}B = \mathbf{0}$
   (e) $k\mathbf{0} = \mathbf{0}$
4. *Rules involving matrix powers.*
   (a) $A^p A^q = A^{p+q}$
   (b) $(A^p)^q = A^{pq}$
   (c) $(kA)^p = k^p A^p$
   (d) $\mathbf{0}^p = \mathbf{0}$
5. *Rules involving the transpose.*
   (a) $(A^{\mathrm{T}})^{\mathrm{T}} = A$
   (b) $(A + B)^{\mathrm{T}} = A^{\mathrm{T}} + B^{\mathrm{T}}$
   (c) $(kA)^{\mathrm{T}} = kA^{\mathrm{T}}$
   (d) $(AB)^{\mathrm{T}} = B^{\mathrm{T}} A^{\mathrm{T}}$
   (e) $(A^p)^{\mathrm{T}} = (A^{\mathrm{T}})^p$
   (f) $\mathbf{0}^{\mathrm{T}} = \mathbf{0}$

Proof of Rule 1.b. First, it's important to remember what equality of matrices means, so that we know what we should be verifying: *two matrices are equal when they have the same size and the same entries*. And to be sure, while the formulas on the left- and right-hand sides of the rule under consideration each *involve* three matrices, the formulas themselves each represent a *single* matrix.

Let's also make sure we understand the difference between the left- and right-hand sides. On the left, the brackets tell us that we should add B and C first, and then add A to that result. The brackets on the right tell us that we should add A and B first, and then add C to that result. Next, let's compare sizes. To be able to add A,B,C, they must be all the same size, and then the result of adding them (in any combination) will also be that common size. So the left- and right-hand results will be the same size of matrix. Finally, let's make sure each entry in the left-hand result is the same as the corresponding entry in the right-hand result. Since we don't actually know what the entries are or even how many entries there are, we cannot verify this entry by entry. So we work in general instead: consider what the $(i,j)^{\text{th}}$ entry of each side must be, where i,j is a pair of row and column indices, in terms of the entries of A,B,C. For this, you might want to review the conventions on referring to matrix entries described in Subsection 4.3.1. On the left, we have

$$[B+C]_{ij} = b_{ij} + c_{ij},$$

and so

$$\big[A+(B+C)\big]_{ij} = [A]_{ij} + [B+C]_{ij} = a_{ij} + (b_{ij} + c_{ij}).$$

A similar process on the right gives us

$$\big[(A+B)+C\big]_{ij} = [A+B]_{ij} + [C]_{ij} = (a_{ij} + b_{ij}) + c_{ij}.$$

Since we know from high-school algebra that addition of ordinary numbers satisfies the associativity rule

$$a + (b+c) = (a+b) + c,$$

we can see that the $(i,j)^{\text{th}}$ entries of the matrices represented by the formulas on the left- and right-hand sides of this rule will always match. ■

Proof of Rule 2.c. First, since scalar multiplication does not change the size of a matrix, if A and B are compatible sizes for multiplication, then so are kA and B, and the sizes of $(kA)B$ and $k(AB)$ will be the same. Next, consider the $(i,j)^{\text{th}}$ entry of each side. Write $\mathbf{a}_i$ for the i^{th} row of A and $\mathbf{b}_j$ for the j^{th} column of B. Using the *row-times-column* pattern of matrix multiplication, and noticing that the i^{th} row of kA is just $k\mathbf{a}_i$, we have

$$[\text{LHS}]_{ij} = \big[(kA)B\big]_{ij} = (k\mathbf{a}_i)\mathbf{b}_j, \qquad [\text{RHS}]_{ij} = \big[k(AB)\big]_{ij} = k(\mathbf{a}_i\mathbf{b}_j).$$

So these two entries will be equal if the rule

$$(k\mathbf{a})\mathbf{b} = k(\mathbf{a}\mathbf{b})$$

is always true for $1 \times n$ row vector $\mathbf{a}$ and $n \times 1$ column vector $\mathbf{b}$. In this new rule, both sides are size 1×1, and indeed we have

$$\begin{aligned}
\text{New LHS} &= (k\mathbf{a})\mathbf{b} \\
&= \left(k\begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}\right)\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \\
&= \begin{bmatrix} ka_1 & ka_2 & \cdots & ka_n \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}
\end{aligned}$$

$$= \left[(ka_1)b_1 + (ka_2)b_2 + \cdots + (ka_n)b_n\right],$$

and

$$\begin{aligned}
\text{New RHS} &= k(\mathbf{ab}) \\
&= k\left(\begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}\right) \\
&= k\left[a_1b_1 + a_2b_2 + \cdots + a_nb_n\right] \\
&= \left[k(a_1b_1) + k(a_2b_2) + \cdots + k(a_nb_n)\right].
\end{aligned}$$

We can now clearly see that the two sides will be equal using the associativity rule $(ka)b = k(ab)$ for numbers from high-school algebra. ■

Proof of Rule 4.a. We can prove this rule in the same manner as the corresponding rule for powers of numbers from high-school algebra, without worrying about individual entries of the matrices on either side of the equality. On the left we are separately multiplying together p factors of A and q factors of A, and then multiplying those two results together. Rule 1.e says that we can multiply all of these factors of A together in any combinations and get the same result. Since there are $p+q$ factors of A all together, the result will be the same as A^{p+q}. ■

Remark 4.5.2

- Algebra rules are not handed down from on high, they represent patterns where two different sequences of computations always produce the same result. For example, we can see that

  $$2(3+5) = 2\cdot 3 + 2\cdot 5,$$

  not from algebra but from computation:

  $$\text{LHS} = 2(3+5) = 2(8) = 16, \qquad \text{RHS} = 2\cdot 3 + 2\cdot 5 = 6 + 10 = 16.$$

  This example of different computations yielding the same result did not depend on the numbers $2, 3, 5$ but on the pattern of the sequences of computations, and we capture this pattern algebraically in terms of letters as the **distributive rule** $a(b+c) = ab + ac$. The algebra rules above capture similar universal patterns of different sequences of matrix operations that always produce the same result.

- In the rules, the letters A, B, C are *placeholders* for any arbitrary matrices. When we use these rules, we might need to apply them where a *whole formula* of letters that computes to a single matrix takes the place of one of A, B, C. For example, see the first step of the FOIL example below.

- In the preamble to the proposition, we stated that *most* of the familiar rules from the algebra of numbers still hold for matrices. But there is one important rule that *does not* hold! Remember that ***order of matrix multiplication matters***: AB and BA are *not* equal in general.

- As you read the rules, think about the point of the rule. For example, consider Rule 1.b. Matrix addition is defined as an operation between *two* matrices. If we write something like $A+B+C$, it is ambiguous what is meant. Does it mean that $A+B$ should be performed first, and then C added to that result? Or should $B+C$ be performed first, and then A added to that result? Mathematical notation is about *communication* of mathematical ideas, patterns, and computations. *Ambiguity in communication is bad.* To resolve the ambiguity in writing $A+B+C$, we would require brackets to communicate which order of successive additions is meant. But the point of Rule 1.b is that *it doesn't matter* — either meaning will yield the same end result. Rule 1.e establishes a similar pattern for matrix multiplication.

- Also, as you read the rules, try to think of the pattern each one is expressing *in words*. For example, for Rule 3.a, reading out "A plus zero equals A" is a lot less clear than interpreting the rule as "adding the zero matrix to any matrix has no effect."

- Rule 1.c and Rule 1.d are not redundant because *order of matrix multiplication matters*. In particular, it's important to be careful when using these rules to factor a common multiple out of a sum. For example, $AX+BX$ *cannot* be factored as $X(A+B)$, because then X is multiplying on the left when originally it was multiplying both A and B on the right. The correct factorization is $AX+BX=(A+B)X$. Even worse, $AX+XB$ *cannot be factored at all*.

- Because of Rule 2.f, all of the rules that involve addition are also valid for subtraction (with the obvious exception of commutivity Rule 1).

- There are two things to note about the rules involving the transpose. First, in Rule 5.f, the zero matrices on either side of the equality are not necessarily of the same size (unless they are both square). Second, notice how a transpose of a product reverses the order of multiplication in Rule 5.d. This happens because in the product AB we are multiplying rows of A against columns of B. If we were to compute the product of transposes $A^{\mathrm{T}}B^{\mathrm{T}}$, we would be multiplying rows of A^{T} (i.e. *columns* of A) against columns of B^{T} (i.e. *rows* of B). Obviously these two computations won't compare, and we need to reverse the order to $B^{\mathrm{T}}A^{\mathrm{T}}$ so that rows of B^{T} (i.e. *columns* of B) multiply against columns of A^{T} (i.e. *rows* of A), similarly to AB.

Example 4.5.3 Using the rules. Here is an example of using some of the basic rules to justify a slightly more involved rule like FOIL. Assume A,B,Y,Z are square matrices of the same size.

$$\begin{aligned}(A+B)(Y+Z) &= A(Y+Z)+B(Y+Z) && \text{(i)}\\ &= (AY+AZ)+(BY+BZ) && \text{(ii)}\\ &= AY+AZ+BY+BZ && \text{(iii)}\end{aligned}$$

Here are the justifications for the numbered steps, using the algebra rules in Proposition 4.5.1.

(i) right-distributive Rule 1.d, with $C=Y+Z$;

(ii) left-distributive Rule 1.c, used twice;

(iii) brackets can be omitted by associativity Rule 1.b.

□

4.5.2 Linear systems as matrix equations

In the examples expressing system solutions in terms of column vectors in Subsection 4.4.4, we noticed a pattern: the general solution of a consistent system can always be expressed as a *constant part* plus a *variable part*, where the constant part is a particular solution to the system (corresponding to setting all parameters to 0) and the variable part involves the parameters. This is true even for

- a system with one unique solution (as in Example 4.4.8), in which case we consider the variable part to be zero; and
- a homogeneous system (as in Example 4.4.10), in which case we consider the constant part to be zero (i.e. the trivial solution).

We further saw that the pattern goes a bit deeper when we explored the pattern between the solutions to Example 4.4.9 and Example 4.4.11. These two systems had the *same coefficient matrix*, but one was nonhomogeneous and the other was homogeneous. We saw that the two solutions *have exactly the same variable part*. This pattern will always emerge for a consistent system.

Lemma 4.5.4 Homogeneous/nonhomogeneous solution set patterns. *If $\mathbf{x}_1$ is a particular solution to system $A\mathbf{x} = \mathbf{b}$, then every other solution to this system can be expressed as the sum of $\mathbf{x}_1$ and some solution to the corresponding homogeneous system $A\mathbf{x} = \mathbf{0}$.*

Proof. We have solution $\mathbf{x}_1$ to system $A\mathbf{x} = \mathbf{b}$. By definition, this means that the matrix equation defining the system is valid when we substitute $\mathbf{x} = \mathbf{x}_1$. That is, we know that $A\mathbf{x}_1 = \mathbf{b}$. Suppose we have another solution $\mathbf{x}_2$. Again, this means that $A\mathbf{x}_2 = \mathbf{b}$ is also true. We would like to show that $\mathbf{x}_2$ is equal to the sum of $\mathbf{x}_1$ and some solution to the homogeneous system $A\mathbf{x} = \mathbf{0}$. Set $\mathbf{x}_0 = \mathbf{x}_2 - \mathbf{x}_1$. We claim that $\mathbf{x} = \mathbf{x}_0$ is a solution to $A\mathbf{x} = \mathbf{0}$. Let's verify:

$$\text{LHS} = A\mathbf{x}_0 = A(\mathbf{x}_2 - \mathbf{x}_1) = A\mathbf{x}_2 - A\mathbf{x}_1 = \mathbf{b} - \mathbf{b} = \mathbf{0} = \text{RHS}.$$

So $\mathbf{x}_0$ is a solution to the homogeneous system. Furthermore,

$$\mathbf{x}_1 + \mathbf{x}_0 = \mathbf{x}_1 + (\mathbf{x}_2 - \mathbf{x}_1) = (\mathbf{x}_1 - \mathbf{x}_1) + \mathbf{x}_2 = \mathbf{0} + \mathbf{x}_2 = \mathbf{x}_2.$$

Thus, $\mathbf{x}_2$ is equal to the sum of $\mathbf{x}_1$ and a solution to the homogeneous system (i.e. $\mathbf{x}_0$), as desired. ■

We can also use the matrix algebra viewpoint of linear systems to definitively answer Question 1.3.4.

Theorem 4.5.5 None, one, or infinite solutions. *There are exactly three possibilities for the number of solutions to a linear system: no solution, one unique solution, or an infinite number of solutions.*

Proof. We have seen in examples that it is possible for a system to have no solution, and that it is also possible for a system to have one unique solution. We will argue that an infinite number of solutions is the only remaining possibility. If we are not in one of the first two cases, then our system must be consistent and must have more than one solution. That is, there must be at least two different solutions. Pick two different solutions, label them $\mathbf{x}_1$ and $\mathbf{x}_2$, and set $\mathbf{x}_0 = \mathbf{x}_2 - \mathbf{x}_1$. The same algebra as in the proof of Lemma 4.5.4 verifies that $\mathbf{x}_0$ is a solution to the homogeneous system $A\mathbf{x} = \mathbf{0}$. Let t be a parameter. We claim that for every possible value of the parameter t, $\mathbf{x}_1 + t\mathbf{x}_0$ is a solution to $A\mathbf{x} = \mathbf{b}$. Let's verify:

$$\text{LHS} = A(\mathbf{x}_1 + t\mathbf{x}_0) = A\mathbf{x}_1 + tA\mathbf{x}_0 = \mathbf{b} + t\mathbf{0} = \mathbf{b} + \mathbf{0} = \mathbf{b} = \text{RHS}.$$

If $\mathbf{x}_0$ were secretly the zero vector, then $\mathbf{x}_1 + t\mathbf{x}_0$ would always equal $\mathbf{x}_1$ no matter the value of t. But since $\mathbf{x}_1$ and $\mathbf{x}_2$ are *different* solutions to $A\mathbf{x} = \mathbf{b}$, we have $\mathbf{x}_0 = \mathbf{x}_2 - \mathbf{x}_1 \neq \mathbf{0}$, and so different values of t produce different column vectors $\mathbf{x}_1 + t\mathbf{x}_0$. Each of these column vectors is a solution to $A\mathbf{x} = \mathbf{b}$, as verified above, and so since there are infinity of possible values for t, there are infinite different possibilities for $\mathbf{x}_1 + t\mathbf{x}_0$, and so infinite possible solutions to $A\mathbf{x} = \mathbf{b}$.

Note. The expression $\mathbf{x}_1 + t\mathbf{x}_0$ in the proof may not represent *all possible* solutions to the system, since the system may require more than one parameter to solve. But the need for *at least* one parameter in solving a system guarantees that there will be an infinite number of solutions.

■

Chapter 5

Matrix inverses

5.1 Discovery guide

Discovery 5.1 The number one is important in algebra, it lets us do things like

$$\begin{aligned} 5a &= 15 \\ \frac{1}{5} \cdot 5 \cdot a &= \frac{1}{5} \cdot 15 \\ 1a &= 3 \\ a &= 3. \end{aligned}$$

The critical step for us right now is the last simplification of the left-hand side:

$$1a = a.$$

(a) What matrix do you think will act similarly in matrix algebra for 2×2 matrices to how the number 1 acts in number algebra? To answer this, try to fill in the first matrix below so that the matrix equality is always true, no matter the values of a, b, c, d.

$$\begin{bmatrix} \square & \square \\ \square & \square \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

(b) Write I for your 2×2 matrix from Task a (for the ***I*** in **Identity matrix**).

- **(i)** Does $IA = A$ work for every 2×2 matrix A? For every 2×3 matrix A? For every $2 \times \ell$ matrix A, no matter the number ℓ of columns?
- **(ii)** Does $BI = B$ also work for every 2×2 matrix B? For every $\ell \times 2$ matrix B?

(c) *Extend:* What is the 3×3 version of I? The 4×4 version? The $n \times n$ version?

Discovery 5.2 In the preamble to Discovery 5.1, there were two ingredients necessary to make the algebra work:

- there is a special number 1 so that $1a = a$ for all numbers a; and
- for a nonzero number like 5, there is a **multiplicative inverse** 1/5 so that $(1/5) \cdot 5 = 1$.

Multiplicative inverses are very useful in algebra, so we would also like to have them in matrix algebra.

(a) Consider

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

Can you determine

$$B = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

so that $BA = I$? If so, check that $AB = I$ also.

Solve, don't guess. Don't just guess at matrix B, use the definition of **matrix equality** applied to BA and I to set up equations and try to solve for unknown entries a, b, c, d.

(b) Consider

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

Can you determine

$$B = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

so that $BA = I$?

Discovery 5.2 demonstrates that some square matrices have multiplicative inverses (i.e. are **invertible**) and some do not (called **singular** in this case). If square matrix A is invertible, write A^{-1} for its inverse. (But *never* write $1/A$!) This inverse is *defined* by its relationship to A and I: A^{-1} is the square matrix of the same size as A so that both $AA^{-1} = I$ and $A^{-1}A = I$ are true.

Check your understanding. Do you understand why we must assert that *both* $AA^{-1} = I$ and $A^{-1}A = I$ are true in defining the inverse A^{-1}? Maybe look back at Discovery 4.7.

Discovery 5.3 In the following, assume A, B, C are square *invertible* matrices, all of the same dimension, and assume that k is a nonzero scalar. Do *not* just look up the answers in the rest of this chapter, try to come up with them yourselves.

For this activity, it might be helpful to think of the pattern of the inverse in the following way: ***given a square matrix M, the inverse of M is the square matrix of the same size that can fill both of the boxes below to create true matrix equalities.*.**

$$M\boxed{} = I \qquad\qquad \boxed{}M = I \qquad\qquad (*)$$

(a) What do you think is the inverse of A^{-1}? In other words, if you use $M = A^{-1}$ in $(*)$, what single choice of matrix can be used to fill in both boxes?

(b) Determine a formula for the inverse of kA in terms of k and A^{-1}. In other words, if you use $M = kA$ in $(*)$, what formula involving k and A^{-1} can be used to fill in both boxes?

(c) Explain why the formula for the inverse of the product AB is *not* $A^{-1}B^{-1}$. Then determine a correct formula in terms of A^{-1} and B^{-1}. (Again, to determine the correct formula for $(AB)^{-1}$, use $M = AB$ in $(*)$, and then try to figure out what single formula you can enter into both boxes so that both left-hand sides reduce to I.)

(d) *Extend:* Determine a formula for the inverse of the product ABC in terms of the inverses A^{-1}, B^{-1}, and C^{-1}.

(e) What do you think A^{-2} means? There are two possibilities because the notation implies the application of two different processes: squaring and inverting. Do they both work out to be the same? Try with A given below. (For convenience, its inverse is also given.)

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

Discovery 5.4

(a) In algebra, when $AB = AC$ we would usually conclude that $B = C$. Try this out for the matrices below.

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad B = \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix} \qquad C = \begin{bmatrix} -1 & -1 \\ 2 & 3 \end{bmatrix}$$

What is it about matrix A that is making the usual algebra of “cancellation” fail?

Hint. Think about the “hidden” algebra behind the cancellation $ab = ac \implies b = c$ for numbers.

(b) In what circumstance is the algebra $AB = AC \implies B = C$ valid? What explicit algebra steps go into this deduction?

(c) Is the algebra $AB = CA \implies B = C$ ever valid?

Discovery 5.5 If we have a linear system $A\mathbf{x} = \mathbf{b}$ with a square and invertible coefficient matrix A, we can use matrix algebra to solve the system instead of row reducing, ***by using*** A^{-1} ***to isolate*** $\mathbf{x}$**.**

Here is an invertible 3×3 matrix A and its inverse:

$$A = \begin{bmatrix} 0 & 1 & -2 \\ 1 & 2 & 0 \\ -2 & -4 & 1 \end{bmatrix}, \qquad A^{-1} = \begin{bmatrix} -2 & -7 & -4 \\ 1 & 4 & 2 \\ 0 & 2 & 1 \end{bmatrix}.$$

Use matrix algebra (not row reducing!) to solve the system $A\mathbf{x} = \mathbf{b}$ for

$$\mathbf{b} = \begin{bmatrix} -1 \\ 1 \\ 3 \end{bmatrix}.$$

Now use the same method to solve the system $A\mathbf{x} = \mathbf{b}$ for

$$\mathbf{b} = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}.$$

Discovery 5.6 In general, for system $A\mathbf{x} = \mathbf{b}$ with a coefficient matrix A that is square and invertible, how many solutions does the system have? Justify your answer.

Hint. How many solutions did each of the systems in Discovery 5.5 have? Why?

5.2 Terminology and notation

identity matrix
a square matrix with ones down the main diagonal and zeros everywhere else, usually represented by the letter I

Remark 5.2.1 There are many identity matrices, one of every possible size of square matrix. But we still often say *the* identity matrix, because we are usually referring to the identity matrix of a particular size. If we need to be clear about what size of identity matrix, we will write I_n to mean the $n \times n$ identity matrix.

inverse (of a square matrix A)
a square matrix B of the same size as A so that both $BA = I$ and $AB = I$ are true, where I is the identity matrix of the same size as A and B

invertible matrix
a square matrix for which an inverse exists

singular matrix
a square matrix for which no inverse exists

5.3 Concepts

In this section.

5.3.1 The identity matrix

The number one plays a special role with respect to multiplication of numbers: it is the only number that has no effect when it is multiplied against another number. In multiplication of matrices, there is only one kind of matrix that has the same effect: a square matrix with all ones down the main diagonal and zeros in every other entry. We call this matrix the **identity matrix**, and write I to represent it.

The identity matrix is to multiplication what the zero matrix is to addition, and it will allow us to (sometimes) do the matrix version of the algebra in the preamble to Discovery 5.1. Except there is one wrinkle that we will explore in this chapter and next: while we can always “cancel” a matrix to the zero matrix by subtracting, unfortunately we will *not* always be able to “cancel” a matrix to the identity by “dividing.”

5.3.2 Inverse matrices

If a is a nonzero number, we can use the inverse $a^{-1} = 1/a$ to multiply a to 1. In algebra, we often use this fact to "cancel" a number from an algebraic expression. In matrix algebra, we can attempt to do the same thing for square matrices. For a square matrix A, we would like to find a square matrix of the same size that multiplies A to the identity I (where the identity is the matrix version of the number 1). If we can determine such an inverse for A, we write A^{-1} for it. Note that we need this inverse to multiply A to I from *both* sides, because order of multiplication matters. That is, we need to be sure that *both* $A^{-1}A = I$ and $AA^{-1} = I$.

The only *number* that doesn't have an inverse is 0. However, we saw in Discovery 5.2 that some *nonzero matrices* do not have inverses (i.e. are **singular**). While the singular example in Discovery 5.2.b only had one nonzero entry, it is possible to come up with examples of singular matrices that have *no* entries that are zero — see Subsection 5.4.1 for one example.

A look ahead.

- We will see in Section 5.5 that a square matrix can have only *one* inverse matrix (Theorem 5.5.2), so writing A^{-1} to mean *the* inverse of an invertible matrix A is unambiguous. We will also see in Section 6.5 that it is enough to check only *one* of $BA = I$ and $AB = I$ in order to know that $B = A^{-1}$ (Proposition 6.5.4 and Proposition 6.5.6).
- We will also see that a square matrix is singular when it has some relationship to another matrix that has too many zero entries to be invertible.
  - In Chapter 6, we learn that a square matrix is singular precisely when its RREF has too many zeros (Theorem 6.5.2).
  - In a future linear algebra course, you may learn that a matrix is singular precisely when it is **similar** to one that has too many zeros (a fact closely related to Theorem 21.6.3).

5.3.3 Matrix division

In Chapter 4, we defined the operations of addition, subtraction, and multiplication of matrices (as well as scalar multiplication), *but we did not define* division *of matrices*. Matrix inverses are similar to division in that they can be used to cancel a square matrix to I, the matrix version of 1. But we need to be careful with this analogy, because *order of matrix multiplication matters*. With numbers, when we write division as a fraction a/b, it doesn't matter if we mean ab^{-1} or $b^{-1}a$ because both orders of multiplication yield the same result. For matrices, it is ambiguous to write a fraction A/B because it is not clear whether AB^{-1} or $B^{-1}A$ should be meant, and it matters because AB^{-1} and $B^{-1}A$ might not yield the same result.

Warning 5.3.1 There is no such operation as division of matrices. Never write fractions of matrices in matrix algebra — use matrix inverses instead.

5.3.4 Cancellation

In the algebra of numbers, if we have an equation $ab = ac$, we would usually cancel the a from both sides to conclude that $b = c$. In Discovery 5.4, we see that this doesn't always work for matrices. When we cancel a from both sides of $ab = ac$, what we are really doing algebraically is to *divide* both sides by a.

However, we cannot do this if a is 0, and similarly we cannot cancel the A from $AB = AC$ if A is not invertible. If it *is* invertible, however, then we may cancel A by applying A^{-1} to both sides:

$$\begin{aligned} AB &= AC \\ A^{-1}AB &= A^{-1}AC \\ IB &= IC \\ B &= C. \end{aligned}$$

When we apply algebraic manipulations to an equation, we need to make sure we perform the *exact same* operation on both sides of the equals sign. If we are introducing a new matrix into an equation by multiplication, we need to make sure we multiply the new matrix from the same side on both sides of the equation, because *order of matrix multiplication matters*! So, when we were faced with $AB = CA$ in Task c of Discovery 5.4, it would be incorrect to cancel A here *even if A is invertible*, because to cancel it on the left-hand side of the equation we need to multiply by A^{-1} on the left, and to cancel on the right-hand side we need to multiply A^{-1} on the right. These are *different* operations, and doing one on one side of the equation and the other on the other side would violate the equals sign. If we try to do both, we just go in circles:

$$\begin{aligned} AB &= CA \\ A^{-1}AB &= A^{-1}CA && \text{(i)} \\ IB &= A^{-1}CA && \text{(ii)} \\ B &= A^{-1}CA \\ BA^{-1} &= A^{-1}CAA^{-1} && \text{(iii)} \\ BA^{-1} &= A^{-1}CI \\ BA^{-1} &= A^{-1}C. \end{aligned}$$

In the above steps:

(i) the same operation (i.e. multiplication by A^{-1} *on the left*) must be applied to both sides;

(ii) the A^{-1} and A on the right do not cancel to I; and

(iii) the same operation (i.e. multiplication by A^{-1} *on the right*) must be applied to both sides.

In more detail, when we have $A^{-1}CA$ on the right above, unfortunately we cannot cancel the A^{-1} and A to I because we don't have $A^{-1}A = I$, we have a C between A and its inverse, and *order of matrix multiplication matters*! And notice that after all this algebra, in the last line we are no further ahead than when we began.

5.3.5 Solving systems using inverses

As we explored in Discovery 5.5, when the coefficient matrix of a linear system is square and invertible, we can solve the system by matrix algebra instead of row reducing. In this case, we can use the inverse to cancel the coefficient matrix from the left-hand side of the system equation $A\mathbf{x} = \mathbf{b}$:

$$A\mathbf{x} = \mathbf{b}$$

$$\begin{aligned} A^{-1}A\mathbf{x} &= A^{-1}\mathbf{b} \\ I\mathbf{x} &= A^{-1}\mathbf{b} \\ \mathbf{x} &= A^{-1}\mathbf{b}. \end{aligned}$$

We will see in the next chapter that inverting a matrix is the same amount of work as row reducing it, so solving a system this way is not a shortcut method. But it can be faster if you want to solve several systems with the same coefficient matrix A but different vectors of constants $\mathbf{b}$, as we had in Discovery 5.5, so that you only do the row reducing work (in computing A^{-1}) once.

5.4 Examples

In this section.

- Subsection 5.4.1 *Inverses of* 2×2 *matrices*
- Subsection 5.4.2 *Solving systems using inverses*
- Subsection 5.4.3 *Solving other matrix equations using inverses*

5.4.1 Inverses of 2×2 matrices

There is a general formula for the inverse of a 2×2 formula:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \implies \qquad A^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

The formula $ad - bc$ in the denominator of the scalar multiple in this inverse formula is called the **determinant** of A. Clearly the formula does not work when the determinant of A is 0, since we cannot divide by zero. In fact, in Chapter 6 it will be possible for us to prove that A is *not* invertible when $ad - bc = 0$. There are similar formulas for inverses of larger matrices, but they are too complicated to write down explicitly. We will study the general theory of determinants and related inversion formulas in Chapters 8–10.

Example 5.4.1 Using the 2×2 inversion formula. Matrix A below is invertible, and its inverse is given. Watch for double negatives when computing $ad - bc$!

$$A = \begin{bmatrix} -5 & 1 \\ -3 & 2 \end{bmatrix} \qquad \implies \qquad \begin{aligned} A^{-1} &= \frac{1}{(-5)(2)-(1)(-3)}\begin{bmatrix} 2 & -1 \\ 3 & -5 \end{bmatrix} \\ &= -\frac{1}{7}\begin{bmatrix} 2 & -1 \\ 3 & -5 \end{bmatrix} \\ &= \begin{bmatrix} -2/7 & 1/7 \\ -3/7 & 5/7 \end{bmatrix}. \end{aligned}$$

Let's check that we have the correct inverse. To keep the computations simple, we'll leave the $-1/7$ as a scalar multiple when expressing A^{-1}.

$$\begin{aligned} A^{-1}A &= \left(-\frac{1}{7}\begin{bmatrix} 2 & -1 \\ 3 & -5 \end{bmatrix}\right)\begin{bmatrix} -5 & 1 \\ -3 & 2 \end{bmatrix} & AA^{-1} &= \begin{bmatrix} -5 & 1 \\ -3 & 2 \end{bmatrix}\left(-\frac{1}{7}\begin{bmatrix} 2 & -1 \\ 3 & -5 \end{bmatrix}\right) \\ &= -\frac{1}{7}\begin{bmatrix} 2 & -1 \\ 3 & -5 \end{bmatrix}\begin{bmatrix} -5 & 1 \\ -3 & 2 \end{bmatrix} & &= -\frac{1}{7}\begin{bmatrix} -5 & 1 \\ -3 & 2 \end{bmatrix}\begin{bmatrix} 2 & -1 \\ 3 & -5 \end{bmatrix} \end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{7}\begin{bmatrix} -10+3 & 2-2 \\ -15+15 & 3-10 \end{bmatrix} &\qquad &= -\frac{1}{7}\begin{bmatrix} -10+3 & 5-5 \\ -6+6 & 3-10 \end{bmatrix} \\
&= -\frac{1}{7}\begin{bmatrix} -7 & 0 \\ 0 & -7 \end{bmatrix} &\qquad &= -\frac{1}{7}\begin{bmatrix} -7 & 0 \\ 0 & -7 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} &\qquad &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}
\end{aligned}
$$

So, we have both $A^{-1}A = I$ and $AA^{-1} = I$, as required. □

Example 5.4.2 Sometimes the 2×2 inversion formula does not apply. Consider matrix

$$B = \begin{bmatrix} 3 & 6 \\ 1 & 2 \end{bmatrix}.$$

For this matrix, we have

$$ad - bc = 3 \cdot 2 - 6 \cdot 1 = 6 - 6 = 0.$$

So even though none of the entries of B are 0, it is not invertible. □

5.4.2 Solving systems using inverses

Just as we can solve the numerical equation $ax = b$ as $x = a^{-1}b$, we can solve a system of equations that is represented as a matrix equation $A\mathbf{x} = \mathbf{b}$ using A^{-1}.

Example 5.4.3 Consider the system

$$\left\{ \begin{array}{rcrcr} -5x & + & y & = & 3, \\ -3x & + & 2y & = & -2. \end{array} \right.$$

The coefficient matrix for this system is

$$A = \begin{bmatrix} -5 & 1 \\ -3 & 2 \end{bmatrix},$$

which is conveniently the matrix for which we have already computed the inverse using the 2×2 inversion formula in Subsection 5.4.1. So we can solve the system as

$$
\begin{aligned}
A\mathbf{x} = \mathbf{b} \qquad \implies \qquad \mathbf{x} &= A^{-1}\mathbf{b} \\
&= \left(-\frac{1}{7}\begin{bmatrix} 2 & -1 \\ 3 & -5 \end{bmatrix} \right) \begin{bmatrix} 3 \\ -2 \end{bmatrix} \\
&= -\frac{1}{7}\begin{bmatrix} 2 & -1 \\ 3 & -5 \end{bmatrix} \begin{bmatrix} 3 \\ -2 \end{bmatrix} \\
&= -\frac{1}{7}\begin{bmatrix} 8 \\ 19 \end{bmatrix} \\
&= \begin{bmatrix} -8/7 \\ -19/7 \end{bmatrix},
\end{aligned}
$$

so that the system has one unique solution $x = -8/7$, $y = -19/7$. □

5.4.3 Solving other matrix equations using inverses

We can similarly use matrix algebra and inverses to solve matrix equations in general.

Example 5.4.4 Consider the matrix equation

$$3\begin{bmatrix} 1 & 1 \\ -1 & 2 \end{bmatrix} + X\begin{bmatrix} 0 & -3 \\ 2 & 1 \end{bmatrix} = I.$$

Suppose we would like to solve this equation for the unknown 2×2 matrix X, where I is the 2×2 identity matrix.

One approach to this problem would be to express X in terms of unknown entries,

$$X = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

and then set up four equations in the four unknowns a, b, c, d. This would lead to a system of equations that we could row reduce and solve. But it's easier just to use ordinary (matrix) algebra. Set

$$W = \begin{bmatrix} 1 & 1 \\ -1 & 2 \end{bmatrix}, \qquad Z = \begin{bmatrix} 0 & -3 \\ 2 & 1 \end{bmatrix},$$

substitute these definitions into the given equation, and isolate X algebraically:

$$\begin{aligned} 3W + XZ &= I \\ XZ &= I - 3W \\ XZZ^{-1} &= (I - 3W)Z^{-1} \\ X &= (I - 3W)Z^{-1}. \end{aligned}$$

Of course, this method wouldn't work if Z was not invertible, but it is, and we can calculate

$$\begin{aligned} I - 3W &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 3\begin{bmatrix} 1 & 1 \\ -1 & 2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 3 & 3 \\ -3 & 6 \end{bmatrix} \\ &= \begin{bmatrix} -2 & -3 \\ 3 & -5 \end{bmatrix}, \end{aligned} \qquad \begin{aligned} Z^{-1} &= \frac{1}{0 \cdot 1 - (-3) \cdot 2}\begin{bmatrix} 1 & 3 \\ -2 & 0 \end{bmatrix} \\ &= \frac{1}{6}\begin{bmatrix} 1 & 3 \\ -2 & 0 \end{bmatrix}. \end{aligned}$$

From this we obtain

$$X = (I - 3W)Z^{-1} = \begin{bmatrix} -2 & -3 \\ 3 & -5 \end{bmatrix}\left(\frac{1}{6}\begin{bmatrix} 1 & 3 \\ -2 & 0 \end{bmatrix}\right) = \frac{1}{6}\begin{bmatrix} 4 & -6 \\ 13 & 9 \end{bmatrix} = \begin{bmatrix} 2/3 & -1 \\ 13/6 & 3/2 \end{bmatrix}.$$

□

5.5 Theory

In this section.

- Subsection 5.5.1 *Properties of the identity matrix*
- Subsection 5.5.2 *Properties of the inverse*

5.5.1 Properties of the identity matrix

Here are some important facts about the identity matrix and inverses of matrices. You could consider this proposition as a continuation of Proposition 4.5.1.

Proposition 5.5.1 Algebra rules involving the identity matrix. *Let I represent the $n \times n$ identity matrix.*

1. *For every $m \times n$ matrix A and every $n \times k$ matrix B, we have $AI = A$ and $IB = B$.*
2. *For every positive integer p, we have $I^p = I$.*
3. *An identity matrix is its own inverse; i.e. $I^{-1} = I$.*
4. *An identity matrix is equal to its own transpose; i.e. $I^{\mathrm{T}} = I$.*

Proof. We will leave the proof of these properties up to you, the reader. ■

5.5.2 Properties of the inverse

And now some first properties of the inverse. We will explore inverses more in the next chapter.

Theorem 5.5.2 Uniqueness of inverses. *A square matrix is either singular or has one* unique *inverse.*

Proof. A square matrix either has an inverse (i.e. is invertible) or it doesn't (i.e. is singular). We would like to know that in the invertible case, there can be only *one* inverse. So suppose that A is a square matrix, and that B is an inverse for A. Then, by definition we have both $BA = I$ and $AB = I$ (see Section 5.2). What if we had another inverse for A? Suppose C was also an inverse for A, so that both $CA = I$ and $AC = I$ were true. Here, all of A,B,C,I are square of the same size. But then,

$$\begin{aligned} C &= CI && \text{(i)} \\ &= C(AB) && \text{(ii)} \\ &= (CA)B && \text{(iii)} \\ &= IB && \text{(iv)} \\ &= B && \text{(v),} \end{aligned}$$

with justifications

(i) Rule 1 of Proposition 5.5.1;

(ii) B is an inverse for A;

(iii) Rule 1.e of Proposition 4.5.1;

(iv) C is an inverse for A; and

(v) Rule 1 of Proposition 5.5.1.

So C and B must actually be the *same* inverse for A. Since we can apply the same reasoning to any inverse for A, there can only be one inverse for A. ■

Proposition 5.5.3 Singularity of zero matrices. *A square zero matrix is always singular.*

Proof. It should be obvious from Rule 3.c and Rule 3.d of Proposition 4.5.1 that it is impossible for $A = \mathbf{0}$ to work in the definition of **inverse** from Section 5.2. ■

Let's record the formula for 2×2 inverses that we encountered in Subsection 5.4.1.

Proposition 5.5.4 2×2 **inversion formula.** *Consider general* 2×2 *matrix* $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$. *If* $ad - bc \neq 0$, *then* A *is invertible with inverse*

$$A^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Proof idea. You can check by direct computation that these two matrices multiply to the identity matrix, in either order. ■

A look ahead. We will further explore the formula $ad - bc$ and its connection to invertibility of matrices in subsequent chapters.

Here are the properties of inverses we explored in Discovery 5.3. We have changed some of the letters to avoid confusion with the A and B in the definition of **inverse** in Section 5.2.

Proposition 5.5.5 Algebra rules involving inverses.

1. *If* M *is an invertible square matrix, then its inverse* M^{-1} *is also invertible with inverse* $(M^{-1})^{-1} = M$.
2. *If* M *is an invertible square matrix, then for every nonzero scalar* k *the scalar multiple* kM *is also invertible with inverse* $(kM)^{-1} = k^{-1}M^{-1}$.
3. *If* M *and* N *are both invertible square matrices of the same size, then their product* MN *is also invertible with inverse* $(MN)^{-1} = N^{-1}M^{-1}$.
4. *If* $M_1, M_2, \ldots, M_{\ell-1}, M_\ell$ *are all invertible square matrices of the same size, then their product*
$$M_1 M_2 \cdots M_{\ell-1} M_\ell$$
is also invertible with inverse
$$(M_1 M_2 \cdots M_{\ell-1} M_\ell)^{-1} = M_\ell^{-1} M_{\ell-1}^{-1} \cdots M_2^{-1} M_1^{-1}.$$
5. *If* M *is an invertible square matrix, then for every positive integer* ℓ *the power* M^ℓ *is also invertible with inverse* $\left(M^\ell\right)^{-1} = \left(M^{-1}\right)^\ell$.

Proof of Statement 1. We have a square matrix $A = M^{-1}$ and would like to determine an inverse B for it, so that both $BA = I$ and $AB = I$ are true. But we already know this is true for $B = M$, since then

$$BA = MM^{-1} = I, \qquad \text{and} \qquad AB = M^{-1}M = I.$$

■

Proof of Statement 2. We have a square matrix $A = kM$, with $k \neq 0$, and would like to determine an inverse B for it. Let's try $B = k^{-1}M^{-1}$:

$$\begin{aligned} BA &= (k^{-1}M^{-1})(kM) \\ &= (k^{-1}k)(M^{-1}M) \\ &= 1I \\ &= I, \end{aligned} \qquad \begin{aligned} AB &= (kM)(k^{-1}M^{-1}) \\ &= (kk^{-1})(MM^{-1}) \\ &= 1I \\ &= I, \end{aligned}$$

where in the first steps we have applied Rule 2.c and Rule 2.d of Proposition 4.5.1.

Since both $BA = I$ and $AB = I$ are true, then $B = k^{-1}M^{-1}$ is the inverse of $A = kM$. ■

Proof of Statement 3. We have a square matrix $A = MN$ and would like to determine an inverse B for it. Let's try $B = N^{-1}M^{-1}$:

$$\begin{aligned} BA &= (N^{-1}M^{-1})(MN) \\ &= N^{-1}(M^{-1}M)N \\ &= N^{-1}IN \\ &= I, \end{aligned} \qquad \begin{aligned} AB &= (MN)(N^{-1}M^{-1}) \\ &= M(NN^{-1})M^{-1} \\ &= MIM^{-1} \\ &= I, \end{aligned}$$

where in the first steps we have applied Rule 1.e of Proposition 4.5.1.

Since both $BA = I$ and $AB = I$ are true, then $B = N^{-1}M^{-1}$ is the inverse of $A = MN$.

Order matters. In this proof, we were able to interchange the order of scalar multiplication and matrix multiplication because of the rules for scalar multiplication in Proposition 4.5.1. However, it would have been *incorrect* to try to make similar manipulations in the proof of Statement 3, because *order of matrix multiplication matters*!

■

Proof of Statement 4. We leave this proof to you, the reader. ■

Proof of Statement 5. This is the special case of Statement 4 where each of $M_1, M_2, \ldots, M_{\ell-1}, M_\ell$ is equal to M. ■

Remark 5.5.6 In light of Statement 5 of the proposition, for an invertible matrix M and a positive integer k we can write M^{-k} to mean *either* the inverse $(M^k)^{-1}$ or the power $(M^{-1})^k$, since they are the same. This answers the question in Discovery 5.3.e.

We can turn some of the statements of Proposition 5.5.5 around to create new facts about singular (i.e. non-invertible) matrices.

Proposition 5.5.7 Singular products have singular factors.

1. *If the product MN is singular, where M and N are square matrices of the same size, then at least one of M, N must be singular.*

2. *If the product*

$$M_1 M_2 \cdots M_{\ell-1} M_\ell$$

is singular, where $M_1, M_2, \ldots, M_{\ell-1}, M_\ell$ are square matrices of all the same size, then at least one of these matrices must be singular.

3. *If some power M^ℓ is singular, where M is a square matrix and ℓ is a positive integer, then M must be singular.*

Proof of Statement 1. If both M and N were invertible, then Statement 3 of Proposition 5.5.5 says that the product MN would be invertible. But we are assuming that the product MN is singular, so it is not possible for *both* M and N to be invertible. ■

Outline of proof for Statement 2. The proof of this statement is similar to the one above for Statement 1, relying on Statement 4 of Proposition 5.5.5 instead. We leave the details to you, the reader. ■

Outline of proof for Statement 3. This proof again is similar to that above for Statement 1, relying on Statement 5 of Proposition 5.5.5 instead. Alternatively, one could view this as the special case of Statement 2 of the current proposition, where each factor M_i is taken to be equal to M. ■

We did not explore this in our discovery guide, but we can add properties of the inverse with respect to the transpose.

Proposition 5.5.8 Inverse of a transpose. *If A is invertible, then so is A^{T}, with*

$$\left(A^{\mathrm{T}}\right)^{-1} = \left(A^{-1}\right)^{\mathrm{T}}.$$

Proof. Suppose A is an invertible square matrix, and write B for $(A^{-1})^{\mathrm{T}}$. If we can show that both $BA^{\mathrm{T}} = I$ and $A^{\mathrm{T}}B = I$, then by definition we will have shown that A^{T} is invertible, and by Theorem 5.5.2 we will have shown that the inverse of A^{T} is $B = (A^{-1})^{\mathrm{T}}$. Let's check the first required equality:

$$\begin{aligned} \text{LHS} &= BA^{\mathrm{T}} \\ &= (A^{-1})^{\mathrm{T}}A^{\mathrm{T}} && \text{(i)} \\ &= (AA^{-1})^{\mathrm{T}} && \text{(ii)} \\ &= I^{\mathrm{T}} && \text{(iii)} \\ &= I && \text{(iv)} \\ &= \text{RHS}, \end{aligned}$$

with justifications

(i) definition of B;

(ii) Rule 5.d from Proposition 4.5.1;

(iii) definition of inverse;

(iv) Rule 4 from Proposition 5.5.1.

The verification of $AB = I$ is similar, and we leave it up to you, the reader. ■

Using Statement 5 of Proposition 5.5.5 along with Proposition 5.5.8, we can expand the scope of our algebra rules for matrix powers.

Proposition 5.5.9 Algebra involving matrix powers with negative exponents. *With the convention that A^0 should be equal to I for any invertible square matrix A, the matrix algebra rules involving matrix powers in Proposition 4.5.1 (including the property of the transpose relative to powers in rule Rule 5.e) and in Proposition 5.5.1 remain valid for* all *integers p and q, positive or negative (or zero).*

Finally, we will record the observation of Discovery 5.6.

Proposition 5.5.10 Consistency of invertible coefficient matrix. *If the coefficient matrix for a linear system is square and invertible, then the system has* *one unique solution.*

Proof. Consider system $A\mathbf{x} = \mathbf{b}$ where the coefficient matrix A is square and invertible. Then we can apply A^{-1} to both sides of this matrix equation just as in Subsection 5.3.5 and in Example 5.4.3, to isolate $\mathbf{x} = A^{-1}\mathbf{b}$. Thus, $\mathbf{x} = A^{-1}\mathbf{b}$ is the only possible solution to the system. ■

A look ahead. It follows from a fact in the next chapter (Theorem 6.5.2) that the logic of Proposition 5.5.10 goes the other way as well: if a system with a square coefficient matrix has one unique solution, then that coefficient matrix must be invertible.

CHAPTER 6

Elementary matrices

6.1 Discovery guide

Discovery 6.1 Consider the matrices

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad E = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & 0 & 2 & -1 \\ 1 & 2 & 3 & 4 \\ 0 & -1 & 0 & 3 \end{bmatrix}.$$

(a) Remind yourself using the *row-times-column* pattern of matrix multiplication why $IA = A$ is true.

(b) Notice how E is only one entry different from I. How does this change the process of computing EA compared to computing IA?

Think of multiplication by E as "transforming" A into the result matrix EA. How could you describe the transformation in this particular example?

Hint. In the "transformation" $A \to EA$, which rows of A stay the same, and which rows change? For the rows that change, how exactly do they change?

(c) Do you think the same thing will happen when computing E times some other matrix?

(d) We know that $EI = E$. But then consider EI in terms of the first two parts of this discovery activity. So in terms of row operations, what is the relationship between E and I?

Discovery 6.2 Create a 3×3 matrix E' so that for every $3 \times n$ matrix A, the result of $E'A$ is the same as performing the row operation "multiply row 3 by -4" on A.

Hint. What was the pattern you identified in Discovery 6.1.d?

Discovery 6.3 Create a 3×3 matrix E'' so that for every $3 \times n$ matrix A, the result of $E''A$ is the same as performing the row operation "swap rows 1 and 2" on A.

Hint. What was the pattern you identified in Discovery 6.1.d?

Matrices E, E', E'' from the discovery activities so far are called **elementary matrices**. As the preceding activities demonstrate, every elementary row operation has a corresponding elementary matrix.

Discovery 6.4 Suppose we were to take a $3 \times \ell$ matrix A and compute

$$E''E'EA = E''\big(E'(EA)\big),$$

where E, E', E'' are as in Activities 6.1–6.3. How can we interpret this matrix multiplication result in terms of row operations? (Careful of the order of operations!)

Discovery 6.5 Consider $B = \begin{bmatrix} 1 & 0 & -3 \\ 0 & 0 & 2 \\ 0 & 1 & 0 \end{bmatrix}$.

(a) Determine elementary matrices E_1, E_2, E_3 so that $E_3E_2E_1B$ is the identity matrix.

(b) The matrix B happens to be invertible. Manipulate the formula $E_3E_2E_1B = I$ algebraically to obtain a formula for B^{-1} involving your elementary matrices.

(c) Tack an identity matrix I onto the right end of your formula for B^{-1} from Task b. (Recall that multiplying by I has no effect.)

Using this new, modified formula for B^{-1} as inspiration, come up with a procedure to use *only row operations* (and not elementary matrices) to compute the inverse of a square matrix.

Hint. Where did your elementary matrices E_1, E_2, E_3 come from? And what are they now “doing” to the identity matrix, and in what order?

Discovery 6.6 Consider the general 2×2 matrix $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$.

(a) Assume that $a \neq 0$. Use the method you developed in Discovery 6.5 to determine the inverse of A.

(b) Where there any other assumptions about the entries of A (besides $a \neq 0$) that you needed to make for this to work? Why?

Hint. Division by zero is undefined.

(c) Repeat for the other case: assume $a = 0$.

Discovery 6.7 Complete the following tasks for each of the three types of elementary row operations, one at a time:

(i) swap two rows;

(ii) multiply a row by a nonzero constant;

(iii) add a multiple of one row to another.

(a) Suppose someone has performed the row operation you are currently considering on a matrix:

$$A \xrightarrow[\text{op}]{\text{row}} A'.$$

If you know only the operation and the result A', how can you recover the original matrix A?

$$A' \xrightarrow{?} A$$

(b) Suppose we consider Task a with $A = I$:

$$I \xrightarrow[\text{op}]{\text{row}} EI \xrightarrow{\text{(a)}} E'EI,$$

where

- (a) is the same “reverse” row operation you came up with in Task a
- E is the elementary matrix corresponding to the original row operation you are currently considering
- and E' is the elementary matrix corresponding to the (a) row operation.

According to Task a, what should the final result $E'EI$ be? What does this say in general about the inverse of an elementary matrix of the type you are currently considering?

6.2 Terminology and notation

elementary matrix

a matrix obtained from the identity matrix by a single elementary row operation

6.3 Concepts

In this section.

Even though the title of this chapter is *Elementary matrices*, it is really another about *matrix inverses*.

Goal 6.3.1 *Obtain criteria that can be used to determine whether a square matrix is invertible, and develop a method to compute inverses of invertible square matrices.*

Suppose A and B are square matrices of the same size, and A is invertible. Start with B, multiply on the left by A to get AB, and then multiply that result on the left by A^{-1} to get $A^{-1}AB = IB = B$, which is right back where we started. The point being that an inverse matrix A^{-1} *undoes* or *reverses* multiplication by A. So if we want to understand inverses, we need to understand how to reverse matrix multiplication.

Now, our motivation for defining matrix multiplication in the way that we did (i.e. rows times columns) was so that we could use matrix multiplication to represent a system of equations by a single matrix equation $A\mathbf{x} = \mathbf{b}$, with both the vector of unknowns $\mathbf{x}$ and the vector of constants $\mathbf{b}$ as column vectors. (See Discovery 4.5, and more generally Chapter 4.) Furthermore, for a system $A\mathbf{x} = \mathbf{b}$ with a square invertible coefficient matrix A, we can solve the system *either* by row reducing *or* by reversing the multiplication by A and algebraically isolating $\mathbf{x} = A^{-1}\mathbf{b}$. So there must be a connection between row operations, matrix multiplication, and matrix inverses. And elementary matrices are precisely that connection.

6.3.1 Elementary matrices

In Discovery guide 6.1–6.3, we discovered that we can create special square matrices so that multiplying another matrix by that special matrix (on the left) has the same effect as performing an elementary row operation, and we called these special matrices **elementary matrices**. So if E is an elementary matrix and A is another matrix of a compatible size (but not necessarily square), then the result of computing the matrix product EA is the same as performing some elementary row operation on A.

Applying this same reasoning with A replaced by I, we see that $EI = E$ must be the same result as applying that elementary row operation on the identity.

This gives us an easy way to produce an elementary matrix for a particular elementary row operation.

Procedure 6.3.2 To create the elementary matrix associated to a specific row operation. *Perform the desired elementary row operation on the identity matrix of the appropriate size.*

See Subsection 6.4.1 for some examples.

If each elementary row operation can be achieved by multiplication by an elementary matrix, then a *sequence* of row operations can be achieved can be achieved by iterated multiplication by elementary matrices, as in Discovery 6.4. For example, suppose we were to perform the following sequence of operations on some $3 \times \ell$ matrix A:

$$A \xrightarrow{R_2+2R_1} A' \xrightarrow{-4R_3} A'' \xrightarrow{R_1 \leftrightarrow R_2} A'''.$$

The first operation is the same as that corresponding to the elementary matrix E from Discovery 6.1, so the first result A' is equal to EA. Similarly, the second operation is the same as that corresponding to the elementary matrix E' from Discovery 6.2, but this second operation is *being applied to the first result A'*. So the second result A'' is equal to

$$E'A' = E'(EA).$$

Finally, the third operation is the same as that corresponding to the elementary matrix E'' from Discovery 6.3, and this third operation is *being applied to the second result A''*. So the third result A''' is equal to

$$E''A'' = E''\bigl(E'(EA)\bigr).$$

So our sequence of row operations is

$$A \xrightarrow{R_2+2R_1} EA \xrightarrow{-4R_3} E'EA \xrightarrow{R_1 \leftrightarrow R_2} E''E'EA,$$

where each new elementary matrix corresponds to the operation of the preceding arrow. ***Notice the order of the elementary matrices in the final product*** — the elementary matrices appear in right-to-left order compared to the order that the operations have been performed. Make sure you understand why this is so.

In Discovery 6.5, we examined this kind of correspondence between row operations and elementary matrices in a row *reduction* process. It is possible to row reduce the matrix B in that activity to the identity matrix in three operations, represented by elementary matrices E_1, E_2, E_3:

$$B \xrightarrow[\text{operation}]{\text{first}} E_1B \xrightarrow[\text{operation}]{\text{second}} E_2E_1B \xrightarrow[\text{operation}]{\text{third}} E_3E_2E_1B.$$

See Subsection 6.4.2 for another example of determining elementary matrices corresponding to the steps in a row reduction process.

6.3.2 Inverses by elementary matrices

As discussed above, in Discovery 6.5 we reduced a matrix B to the identity matrix in three operations. In terms of elementary matrices, this means that $E_3E_2E_1B = I$, where E_1, E_2, E_3 are the elementary matrices corresponding to the three operations in the reduction sequence.

Assuming that matrix B is invertible, we could use B^{-1} to manipulate this equality:

$$I = E_3E_2E_1B$$

$$\begin{aligned} \implies \qquad IB^{-1} &= (E_3E_2E_1B)B^{-1} \\ \implies \qquad B^{-1} &= E_3E_2E_1(BB^{-1}) \\ &= E_3E_2E_1I \\ &= E_3E_2E_1. \end{aligned}$$

So if a matrix is invertible, we can compute its inverse by row reducing it to the identity matrix and then multiplying together the elementary matrices that correspond to the steps in that row reduction, in the proper order. But there is a more direct way, as we will see in Subsection 6.3.5 below.

Remark 6.3.3 There are many different sequences of row operations that could reduce a matrix to its RREF, and so when a matrix is invertible there are many different ways we could compute its inverse via a product of elementary matrices. These different ways can even involve different numbers of elementary matrices.

6.3.3 Inverses of elementary matrices

As we explored in Discovery 6.7, every elementary row operation has a *reverse* operation.

Operation	swap two rows $R_i \leftrightarrow R_j$
Reverse operation	swap the rows again $R_i \leftrightarrow R_j$
Reverse of the reverse	swap the rows again $R_i \leftrightarrow R_j$

Figure 6.3.4 Reversing row swaps.

Operation	multiply a row by a nonzero constant $R_i \to kR_i$
Reverse operation	divide that row by the constant $R_i \to \frac{1}{k}R_i$
Reverse of the reverse	divide that row by the reciprocated constant $R_i \to \frac{1}{1/k}R_i = kR_i$

Figure 6.3.5 Reversing row scales.

Operation	add a multiple of one row to another $R_i \to R_i + kR_j$
Reverse operation	subtract that multiple of the one row from the other $R_i \to R_i + (-k)R_j$
Reverse of the reverse	subtract that negative multiple of the one row from the other $R_i \to R_i + \bigl(-(-k)\bigr)R_j = R_i + kR_j$

Figure 6.3.6 Reversing row combinations.

In each case, performing an operation on a matrix and then performing the reverse operation on that result will return you to the original matrix. Also notice that in each case the reverse operation of a reverse operation is the original operation. So, if E is the elementary matrix corresponding to some operation, and E' is the elementary matrix corresponding to the reverse operation, then also E corresponds to the reverse of the operation of E'.

If we perform these operations on the identity matrix, we get

$$I \xrightarrow{\text{operation}} EI \xrightarrow[\text{operation}]{\text{reverse}} E'EI,$$

$$I \xrightarrow[\text{operation}]{\text{reverse}} E'I \xrightarrow{\text{operation}} EE'I.$$

But in both situations we should be back at the identity matrix, because the second operation reverses the first. Thus, $E'E = I$ and $EE' = I$, which by definition says that E' is the inverse of E. Hence, *every elementary matrix is invertible*, and *the inverse of an elementary matrix is the elementary matrix corresponding to the reverse operation*.

6.3.4 Decomposition of invertible matrices

Let's go back to the matrix B from Discovery 6.5, for which we obtained matrix equality $E_3E_2E_1B = I$ for some particular elementary matrices E_1, E_2, E_3. We have just learned in the preceding subsection (Subsection 6.3.3) that elementary matrices are invertible, so we can use the algebra of matrix inverses to isolate B as

$$B = E_1^{-1}E_2^{-1}E_3^{-1}.$$

Check your understanding. Do you understand why the inverses of the elementary matrices appear in the reverse order on the right-hand side? Carry out the steps in the matrix algebra

$$\begin{aligned} E_3E_2E_1B &= I \\ \to \quad B &= E_1^{-1}E_2^{-1}E_3^{-1} \end{aligned}$$

yourself if you are unsure.

Now, from the preceding subsection we know that each of $E_1^{-1}, E_2^{-1}, E_3^{-1}$ is also an elementary matrix. So if we describe the pattern of the formula $B = E_1^{-1}E_2^{-1}E_3^{-1}$ in words, we might choose to ignore the inverses and say that *B can be expressed as a product of elementary matrices*. Since a product of elementary matrices represents performing the corresponding elementary row operations in sequence (on the identity matrix, if you like), we might say that a square matrix is invertible precisely when it represents some *sequence* of elementary row operations, and so inverting it is the same as trying to *reverse* that sequence of operations.

6.3.5 Inverses by row reduction

Still working with the matrix B from Discovery 6.5, consider the formula

$$B^{-1} = E_3E_2E_1I,$$

from the computation in Subsection 6.3.2. We could simplify away the identity matrix (as we did above), but as is often the case in mathematics, *simplifying hides patterns*. Remember where the elementary matrices E_1, E_2, E_3 came from — our starting point in the computation above was the formula $E_3E_2E_1B = I$, which we obtained from the fact that these elementary matrices represented the steps taken to reduce B to the identity. So when we compare the two formulas

$$E_3E_2E_1B = I, \qquad\qquad E_3E_2E_1I = B^{-1},$$

we realize that *the same sequence of operations that reduces B to I can be used to "unreduce" I to B^{-1}.*

Now, it is inefficient to first row reduce a matrix to I, and then unreduce I to B^{-1} afterward, because we will be doing the same operations, in the same order, in both parts of the process. It would be faster to do both at once, one operation at a time.

Procedure 6.3.7 Computing an inverse. *To compute the inverse of a square matrix A, augment that matrix with the identity matrix and row reduce until the identity matrix is obtained on the left where there initially was A. The matrix on the right where there was initially I will now be A^{-1}.*

$$\left[\begin{array}{c|c} A & I \end{array}\right] \quad \xrightarrow[\textit{reduce}]{\textit{row}} \quad \left[\begin{array}{c|c} I & A^{-1} \end{array}\right]$$

If it is not possible to obtain the identity on the left (i.e. if the RREF of A is not I), then A is not invertible.

The last statement of the procedure will be justified by Theorem 6.5.2 in Subsection 6.5.2. See Subsection 6.4.3 for an example of carrying out this procedure.

Pattern. Note that Procedure 6.3.7 keeps track of the elementary matrices involved in row reducing a matrix A for us, and automatically applies them to the identity (effectively multiplying them together) to produce the inverse on the right:

$$\begin{aligned}
\left[\begin{array}{c|c} A & I \end{array}\right] &\xrightarrow[\text{operation}]{\text{first}} \left[\begin{array}{c|c} E_1A & E_1I \end{array}\right] \\
&\xrightarrow[\text{operation}]{\text{second}} \left[\begin{array}{c|c} E_2E_1A & E_2E_1I \end{array}\right] \\
&\quad\vdots \\
&\xrightarrow[\text{operation}]{\text{last}} \left[\begin{array}{c|c} E_\ell\cdots E_2E_1A & E_\ell\cdots E_2E_1I \end{array}\right] = \left[\begin{array}{c|c} I & A^{-1} \end{array}\right]. \qquad (*)
\end{aligned}$$

6.4 Examples

In this section.

- Subsection 6.4.1 *Elementary matrices and their inverses*
- Subsection 6.4.2 *Decomposing an invertible matrix and its inverse into elementary matrices*
- Subsection 6.4.3 *Inversion by row reduction*

6.4.1 Elementary matrices and their inverses

Let's see examples of forming the elementary matrix that corresponds to an elementary row operation, and then determining its inverse, for each of the three kinds of elementary operations. We use Procedure 6.3.2 to form these elementary matrices.

Let's do some 4 × 4 examples.

Example 6.4.1 Swapping rows. Consider the operation of swapping the second and fourth rows of a $4 \times n$ matrix A. We can achieve the same result with a matrix product EA where E is a 4×4 elementary matrix. To obtain E, we perform the desired operation on the identity matrix:

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \xrightarrow[\text{rows}]{\text{swap}} \quad E = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

To obtain the inverse E^{-1}, we perform the reverse operation. But that's just swapping the same two rows back again:

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \xrightarrow[\text{rows}]{\text{swap}} \quad E^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

So, in this case, the inverse elemenatary matrix is the same as the original. □

Example 6.4.2 Multiplying a row by a constant. Now consider the operation of swapping the second row of a $4 \times n$ matrix A by 5. We can achieve the same result with a matrix product EA where E is a 4×4 elementary matrix. To obtain E, we perform the desired operation on the identity matrix:

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \xrightarrow[\text{second row}]{\text{multiply}} \quad E = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

To obtain the inverse E^{-1}, we perform the reverse operation, which in this case is dividing the second row by 5 (which is the same as multiplying the second row by 1/5):

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \xrightarrow[\text{second row}]{\text{divide}} \quad E^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/5 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

□

Example 6.4.3 Combining rows. Finally, consider the operation of adding double the first row to the third row of a $4 \times n$ matrix A. We can achieve the same result with a matrix product EA where E is a 4×4 elementary matrix. To obtain E, we perform the desired operation on the identity matrix:

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \xrightarrow[\text{rows}]{\text{combine}} \quad E = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Once again, to obtain the inverse E^{-1}, we perform the reverse operation, which in this case is *subtracting* double the first row from the third:

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \xrightarrow[\text{rows}]{\text{combine}} E^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Comparing E and E^{-1} in this case, notice how the 2 becomes negated, which is actually the *additive* inverse of the number two (since $2+(-2)=0$ and $(-2)+2=0$). This connection between inverting matrix multiplication and inverting numerical addition is important in more advanced abstract algebra. □

Notice. In all three examples, we *always start at the identity matrix to create an elementary matrix*, even when computing the inverse of an elementary matrix.

6.4.2 Decomposing an invertible matrix and its inverse into elementary matrices

Again, let's do a 4×4 example. As we row reduce, we'll keep track of the corresponding elementary matrices. But that also means we need to make sure we are performing *elementary* row operations, and only performing one at a time — no shortcuts!

Example 6.4.4 Consider 4×4 matrix

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 5 & 3 & 0 \\ 0 & 1 & 0 & 0 \\ -2 & 0 & 0 & 1 \end{bmatrix}.$$

Row reduce.

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 5 & 3 & 0 \\ 0 & 1 & 0 & 0 \\ -2 & 0 & 0 & 1 \end{bmatrix} R_4 + 2R_1 \qquad \left(E_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix} \right)$$

$$\longrightarrow \quad E_1A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 5 & 3 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} R_2 \leftrightarrow R_3 \qquad \left(E_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right)$$

$$\longrightarrow \quad E_2E_1A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 5 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} R_3 - 5R_2 \qquad \left(E_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -5 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right)$$

$$\longrightarrow \quad E_3E_2E_1A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tfrac{1}{3}R_3 \qquad \left(E_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right)$$

$$\longrightarrow \quad E_4E_3E_2E_1A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Notice in this process that each elementary matrix is newly obtained by applying a row operation to the identity matrix, not *by applying a row operation to the previous elementary matrix in the sequence.*

We now have $E_4E_3E_2E_1A = I$, which suggests that

$$\begin{aligned} A^{-1} &= E_4E_3E_2E_1 \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -5 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & \frac{1}{3} & -\frac{5}{3} & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix}. \end{aligned}$$

To check that this is really is the correct inverse for A, you can check that this matrix multiplied against A in the other order also results in the identity matrix (i.e. that $A(E_4E_3E_2E_1) = I$ as well).

Also, with some matrix algebra, from $E_4E_3E_2E_1A = I$ we can isolate

$$\begin{aligned} A &= E_1^{-1}E_2^{-1}E_3^{-1}E_4^{-1} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -2 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 5 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \end{aligned}$$

Recall that each of these inverse elementary matrices can each be obtained from the identity matrix using the corresponding reverse operation. You may check that the result of multiplying these inverses together is A. □

6.4.3 Inversion by row reduction

Let's illustrate Procedure 6.3.7 using the matrix A from Subsection 6.4.2 above. Since A is 4×4, we augment A with the 4×4 identity matrix and then row reduce, being careful to apply our row operations *through the entire augmented rows*.

Example 6.4.5 We would like to compute the inverse of

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 5 & 3 & 0 \\ 0 & 1 & 0 & 0 \\ -2 & 0 & 0 & 1 \end{bmatrix}.$$

Augment with I and reduce.

$$\left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 5 & 3 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ -2 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{array}\right] R_4 + 2R_1$$

$$\longrightarrow \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 5 & 3 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 2 & 0 & 0 & 1 \end{array}\right] R_2 \leftrightarrow R_3$$

$$\longrightarrow \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 5 & 3 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 0 & 0 & 1 \end{array}\right] R_3 - 5R_2$$

$$\longrightarrow \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 3 & 0 & 0 & 1 & -5 & 0 \\ 0 & 0 & 0 & 1 & 2 & 0 & 0 & 1 \end{array}\right] \tfrac{1}{3}R_3$$

$$\longrightarrow \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & \frac{1}{3} & -\frac{5}{3} & 0 \\ 0 & 0 & 0 & 1 & 2 & 0 & 0 & 1 \end{array}\right]$$

The matrix on the right is now our desired inverse,

$$A^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & \frac{1}{3} & -\frac{5}{3} & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix},$$

which agrees with our calculation of A^{-1} using elementary matrices in Example 6.4.4. □

6.5 Theory

In this section.

As mentioned, elementary matrices are precisely the connection we need between systems of equations and row operations on one hand and matrix multiplication and inverses on the other.

6.5.1 Inverses of elementary matrices

Let's first record an important property of elementary matrices we encountered in Section 6.3.

Lemma 6.5.1 Elementary is invertible. *Every elementary matrix is invertible, and its inverse is also an elementary matrix.*

Proof. We have already essentially proved this statement in Subsection 6.3.3. ■

6.5.2 Inverses versus row operations

Now let's connect inverses to row reduction via elementary matrices.

Theorem 6.5.2 Characterizations of invertibility. *For a square matrix A, the following are equivalent.*

1. *Matrix A is invertible.*
2. *Every linear system that has A as a coefficient matrix has one unique solution.*
3. *The homogeneous system $A\mathbf{x} = \mathbf{0}$ has only the trivial solution.*
4. *There is some linear system that has A as a coefficient matrix and has one unique solution.*
5. *The rank of A is equal to the size of A.*
6. *The RREF of A is the identity.*
7. *Matrix A can be expressed as a product of some number of elementary matrices.*

Proof. We will show that each statement of the theorem implies the next.

Whenever Statement 1 is true, then so is Statement 2. We have already seen in Proposition 5.5.10 that when A is invertible, then also every system with A as coefficient matrix will have one unique solution.

Whenever Statement 2 is true, then so is Statement 3. If Statement 2 is true about A, then every system $A\mathbf{x} = \mathbf{b}$ with A as coefficient matrix has one unique solution. In particular, the homogeneous system $A\mathbf{x} = \mathbf{0}$ (i.e. in the case that $\mathbf{b} = \mathbf{0}$) has one unique solution. But we know that a homogeneous system always has the trivial solution $\mathbf{x} = \mathbf{0}$, so that must be the one unique solution.

Whenever Statement 3 is true, then so is Statement 4. We need to verify that there is *at least one* example of a system $A\mathbf{x} = \mathbf{b}$ that has one unique solution. But we are already assuming that the homogeneous system $A\mathbf{x} = \mathbf{0}$ has one unique solution, so the required example is provided by taking $\mathbf{b} = \mathbf{0}$.

Whenever Statement 4 is true, then so is Statement 5. Suppose that Statement 4 is true, so that there is at least one example of a system $A\mathbf{x} = \mathbf{b}$ that has one unique solution. Imagine trying to solve this system by row reducing the associated augmented matrix:

$$\left[\begin{array}{c|c} A & \mathbf{b} \end{array}\right] \xrightarrow[\text{reduce}]{\text{row}} \left[\begin{array}{c|c} \text{RREF}(A) & \mathbf{b}' \end{array}\right],$$

where $\mathbf{b}'$ is whatever appears in the "equals" column after all of our row operations. When we have arrived at the RREF of A in the coefficient matrix part on the left, and are ready to solve the simplified solution, there should not be be any free variables. Because free variables would lead to parameters, and hence infinite solutions, whereas we are assuming that this particular system has only one unique solution. So *every* column in the RREF of A must have a leading one. By definition, the rank of A is equal to the number of leading ones in its RREF, and so for this A the rank is equal to the number of columns. But A is square, so the number of columns is the same as the size of A.

Whenever Statement 5 is true, then so is Statement 6. If A is square, so is its RREF, and both matrices have the same size. And if the rank of A is equal to its size, then every column in the RREF of A must have a leading one, and these leading ones must march down the diagonal of A. In a RREF matrix, a column that contains a leading one must have every other entry equal to zero. Thus, the RREF of A must be the identity matrix.

Whenever Statement 6 is true, then so is Statement 7. Suppose that Statement 6 is true, so that A can be reduced to the identity. That is, A can be reduced to I by some sequence of elementary row operations. Each of these operations has a corresponding elementary matrix, so there is some collection of elementary matrices $E_1, E_2, \ldots, E_{\ell-1}, E_\ell$ so that

$$E_\ell E_{\ell-1} \cdots E_2 E_1 A = I. \tag{$*$}$$

Recall. The elementary matrices need to be multiplied in reverse order because we apply the first row operation to A by multiplying $E_1 A$, and then the second operation is applied to *that result* by multiplying $E_2(E_1 A)$. And so on.

Now, by Lemma 6.5.1, each of $E_1, E_2, \ldots, E_{\ell-1}, E_\ell$ is invertible. Therefore, so is the product

$$E_\ell E_{\ell-1} \cdots E_2 E_1,$$

with inverse

$$E_1^{-1} E_2^{-1} \cdots E_{\ell-1}^{-1} E_\ell^{-1}$$

(Rule 4 of Proposition 5.5.5).

Using this inverse, we can isolate A in ($*$) above:

$$\begin{aligned} E_\ell \cdots E_2 E_1 A &= I \\ (E_\ell \cdots E_2 E_1)^{-1} (E_\ell \cdots E_2 E_1) A &= (E_\ell \cdots E_2 E_1)^{-1} I \\ I A &= (E_\ell \cdots E_2 E_1)^{-1} \\ A &= E_1^{-1} E_2^{-1} \cdots E_\ell^{-1}. \end{aligned}$$

So, we have A expressed as a product of the *inverses* of a collection of elementary matrices. But by Lemma 6.5.1, each of these inverses is actually an elementary matrix as well, and so we really have A expressed as a product of a collection of elementary matrices, as desired.

Whenever Statement 7 is true, then so is Statement 1. If A is equal to a product of elementary matrices, then since each of those elementary matrices is invertible (Lemma 6.5.1), their product (and hence A) is also invertible (Rule 4 of Proposition 5.5.5).

Conclusion. We now have a circle of logical deductions. Starting with the knowledge that any one of the seven statements is true for a particular matrix A, we can deduce from the logic above that the next statement is true for A, and then from that, that the next statement is true for A, and so on. When we get to the last statement, the logic above then requires that the first statement will also be true for A, and we can continue from there on to the second statement, and so on, until we are sure that *all* statements are true for A. Therefore, the seven statements are equivalent. ■

Remark 6.5.3

- In this theorem, the claim that these seven statements are equivalent for a particular matrix A means that if we know that any *one* of the statements is true for A, then it must be that *all seven* statements are true for A. For example, if we had a square matrix that we were able to row reduce to the identity, then the theorem tells us that that matrix must be invertible, that every linear system with that matrix as coefficient matrix has one unique solution, and so on. On the other hand, if we know that any one of the statements is *false* for a particular matrix A, then it must be that *all seven* statements are *false* for A. As soon as one statement is known to be false for a particular square matrix, it becomes impossible for any of the other statements to be true for that matrix, since knowing that this other statement is true implies that *all seven* statements are true for it, including the original statement that we already knew was false. And a statement cannot be both true and false for a particular matrix A.

- It may seem unneccessary or even redundant to have *all three* of Statements 2–4 included in the list in Theorem 6.5.2, but these statements are definitely not the same. The equivalence of Statement 1 and Statement 2 tells us that when a matrix is invertible, then every system corresponding to that coefficient matrix has one unique solution, and vice versa. But the reverse connection would be difficult to use in practice: would you want to check that *every* system with a particular square coefficient matrix has one unique solution in order to conclude that the matrix is invertible? There are infinity of possible systems to check! The equivalence of Statement 4 and Statement 1 makes the reverse logic easier in practice: if you have just *one* example of a linear system with a square coefficient matrix that has one unique solution, then you can conclude that the matrix is invertible. Even better, the equivalence of Statement 3 and Statement 1 tells you that you can just check the corresponding *homogeneous* system as your one example of a system with that particular coefficient matrix that has only one unique solution. Furthermore, the equivalence of Statement 2 and Statement 4 tells you that once you know *one* example of a system with that particular coefficient matrix that has only one unique solution, then you can conclude without checking that *every* system with that coefficient matrix has only one unique solution.

- In the proof of Theorem 6.5.2, the most important link is the one between Statement 6 and Statement 7, as this equivalence provides the link between row reducing and elementary matrices. In practice, the link between Statement 7 and Statement 1 is also important, as it helps us to compute the inverse of a matrix. But in further developing matrix theory, the most important link is the one between Statement 1 and Statement 3, as it will allow us to obtain further general properties of inverses. In particular, these statements will figure into the proofs of the propositions in the next subsection.

6.5.3 More properties of inverses

Using our new connections between inverses and row operations, we can expand our knowledge about inverses in general.

Proposition 6.5.4 Left inverse is inverse. *Suppose A and B are square matrices of the same size such that $BA = I$. Then A is invertible with $A^{-1} = B$.*

Proof. We are assuming that we have square matrices A and B so that $BA = I$. We would first like to check that A is invertible. By Theorem 6.5.2, we can instead check that the homogeneous system $A\mathbf{x} = \mathbf{0}$ has only the trivial solution. So suppose that $\mathbf{x}_0$ is a solution to this system, so that $A\mathbf{x}_0 = \mathbf{0}$. But then we can carry out two different simplifications of $BA\mathbf{x}_0$, one using the assumption $BA = I$ and one using the assumption $A\mathbf{x}_0 = \mathbf{0}$:

$$\begin{aligned} BA\mathbf{x}_0 &= (BA)\mathbf{x}_0 & \qquad BA\mathbf{x}_0 &= B(A\mathbf{x}_0) \\ &= I\mathbf{x}_0 & &= B\mathbf{0} \\ &= \mathbf{x}_0, & &= \mathbf{0}. \end{aligned}$$

Since both simplifications are correct, we have $\mathbf{x}_0 = \mathbf{0}$. So what we have discovered is that because there exists a matrix B so that $BA = I$, then whenever we think we have a solution $\mathbf{x}_0$ to the system $A\mathbf{x} = \mathbf{0}$, that solution turns out to be the trivial solution. Thus, $A\mathbf{x} = \mathbf{0}$ must have *only* the trivial solution, and hence A is invertible (Theorem 6.5.2).

Now that we know that A is invertible, we can use its inverse to manipulate the equality $BA = I$:

$$\begin{aligned} BA &= I \\ (BA)A^{-1} &= IA^{-1} \\ B(AA^{-1}) &= A^{-1} \\ BI &= A^{-1} \\ B &= A^{-1}. \end{aligned}$$

So, we have that A is invertible and $A^{-1} = B$, as desired. ■

Remark 6.5.5 Recall that by definition, to verify that a matrix B is the inverse of a matrix A, we would need to check that *both* $BA = I$ *and* $AB = I$ are true. We needed both orders of multiplication in the definition of **inverse matrix** because order of matrix multiplication matters, and we couldn't be sure that both BA and AB would produce the same result. Via the theory of elementary matrices, we now have the above proposition that allows us to check an inverse by *only* checking one order of multiplication: $BA = I$.

There is nothing special about $BA = I$ versus $AB = I$. The previous and following propositions combine to tell us we only need to verify *only one* of $BA = I$ or $AB = I$ to check that B is the inverse of A.

Proposition 6.5.6 Right inverse is inverse. *Suppose A and B are square matrices of the same size such that $AB = I$. Then A is invertible with $A^{-1} = B$.*

Proof. Here, we are assuming that we have square matrices A and B so that $AB = I$, and we again would like to know that A is invertible and that $B = A^{-1}$. However, instead of appealing back to Theorem 6.5.2, we can use Proposition 6.5.4 *with the roles of A and B reversed*: since $AB = I$, Proposition 6.5.4 says that B must be invertible and that $A = B^{-1}$. But inverses are themselves invertible (Rule 1 of Proposition 5.5.5), so A is invertible with

$$A^{-1} = (B^{-1})^{-1} = B,$$

as desired. ■

In Proposition 5.5.5, we learned that products and powers of invertible matrices are always invertible. It turns out that a product of matrices can *only* be invertible if the matrices making up the product are all invertible, and a power of a matrix can *only* be invertible if the base matrix is invertible.

Proposition 6.5.7 Invertible products have invertible factors.

1. *If the product MN is invertible, where M and N are square matrices of the same size, then both M and N must be invertible.*
2. *If the product*
$$M_1M_2\cdots M_{\ell-1}M_\ell$$
is invertible, where $M_1, M_2, \ldots, M_{\ell-1}, M_\ell$ are square matrices all of the same size, then each of $M_1, M_2, \ldots, M_{\ell-1}, M_\ell$ must be invertible.
3. *If power M^ℓ is invertible, where M is a square matrix and ℓ is positive integer, then M must be invertible.*

Proof of Statement 1. Suppose that MN is invertible. Then it has an inverse; let's call it X instead of $(MN)^{-1}$. By definition, this means that

$$X(MN) = I.$$

Using Rule 1.e of Item 1, we may rewrite

$$(XM)N = I.$$

Applying Proposition 6.5.4 with $B = XM$ and $A = N$, we may conclude that N is invertible with inverse

$$N^{-1} = XM.$$

Similarly, since X is the inverse of MN, we may write

$$(MN)X = I$$

and rewrite

$$M(NX) = I.$$

Applying Proposition 6.5.6 this time, we may conclude that M is invertible with inverse

$$M^{-1} = NX.$$

■

Proof of Statement 2. We leave the proof of this statement to you, the reader. ■

Proof of Statement 3. This is the special case of Statement 2 where each of $M_1, M_2, \ldots, M_{\ell-1}, M_\ell$ is equal to M. ■

As in Proposition 5.5.7, we can turn the statements of Proposition 6.5.7 around to create new facts about singular (i.e. non-invertible) matrices. Note that the statements below are *new* statements about singular matrices, related but *not* equivalent to the statements in Proposition 5.5.7.

Proposition 6.5.8 Product of singular is singular.

1. *If one or both of M or N are singular, where M and N are square matrices of the same size, then the product MN will also be singular.*

2. *If one or more of the matrices $M_1, M_2, \ldots, M_{\ell-1}, M_\ell$ are singular, where $M_1, M_2, \ldots, M_{\ell-1}, M_\ell$ are square matrices all of the same size, then the product*

$$M_1 M_2 \cdots M_{\ell-1} M_\ell$$

will also be singular.

3. *If M is a singular square matrix, then every power M^ℓ ($\ell \geq 1$) will also be singular.*

Proof of Statement 1. If the product MN were invertible, then Statement 1 of Proposition 6.5.7 says that each of M and N would have to be invertible. But we are assuming that at least one of them is not, so it is not possible for the product MN to be invertible. ■

Proof of Statement 2. The proof of this statement is similar to the one above for Statement 1, relying on Statement 2 of Proposition 6.5.7 instead. We leave the details to you, the reader. ■

Proof of Statement 3. This proof again is similar to that above for Statement 1, relying on Statement 3 of Proposition 6.5.7 instead. Alternatively, one could view this as the special case of Statement 2 of the current proposition, where each factor M_i is taken to be equal to M. ■

Finally, we can use the link between Statement 1 and Statement 6 of Theorem 6.5.2 to make Proposition 5.5.4 more precise.

Proposition 6.5.9 2×2 **invertibility.** *The general* 2×2 *matrix* $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ *is invertible if* $ad - bc \neq 0$*, and is singular if* $ad - bc = 0$*.*

A look ahead. *We will encounter a version of Proposition 6.5.9 that is valid for every size of square matrix in Chapter 10 (see Theorem 10.5.3).*

Proof outline. We explored this in Discovery 6.6.

Start with the matrix $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ and row reduce to see whether it is possible to get to the identity. But in the operations we choose, we need to be careful not to divide by zero, because the variable entries could be any values, including some zero. So it will be necessary to break into cases, such as $a = 0$ versus $a \neq 0$, and the row reduction steps chosen will differ in the different cases. Ultimately, it will be possible to get the identity as the RREF of A precisely when $ad - bc \neq 0$, and it will be impossible when $ad - bc = 0$. From here, we may appeal to the equivalence of Statement 1 and Statement 6 of Theorem 6.5.2. ■

6.5.4 Solution sets of row equivalent matrices

Elementary matrices also give us the tool we need to prove that row equivalent matrices represent systems with the same solution set. We first recorded the following as Theorem 2.5.5 in Subsection 2.5.2, but did not prove it there. We repeat the theorem here, and include a proof.

Theorem 6.5.10 *Row equivalent matrices represent systems of equations that have the same solution set.*

Proof. Consider systems $A_1\mathbf{x} = \mathbf{b}_1$ and $A_2\mathbf{x} = \mathbf{b}_2$, where augmented matrices

$$A_1' = \left[\begin{array}{c|c} A_1 & \mathbf{b}_1 \end{array}\right], \qquad A_2' = \left[\begin{array}{c|c} A_2 & \mathbf{b}_2 \end{array}\right]$$

are row equivalent. Then there exists a sequence of elementary row operations that can be applied to A_1' to produce A_2'. If we set E to be the *product* of all the elementary matrices corresponding to the operations in this sequence, then we have $A_2' = EA_1'$. Because of the way matrix multiplication acts on columns, we then have

$$\left[\begin{array}{c|c} A_2 & \mathbf{b}_2 \end{array}\right] = E \left[\begin{array}{c|c} A_1 & \mathbf{b}_1 \end{array}\right] = \left[\begin{array}{c|c} EA_1 & E\mathbf{b}_1 \end{array}\right],$$

and so we also have

$$A_2 = EA_1, \qquad \mathbf{b}_2 = E\mathbf{b}_1.$$

Furthermore, we know that every elementary matrix is invertible (Lemma 6.5.1), and that products of invertible matrices are invertible (Statement 4 of Proposition 5.5.5), so we conclude that E is invertible. Therefore, we also have

$$A_1 = E^{-1}A_2, \qquad \mathbf{b}_1 = E^{-1}\mathbf{b}_2.$$

We are now in a position to verify that a solution to one system is also a solution to the other system.

A solution to system $A_1\mathbf{x} = \mathbf{b}_1$ is also a solution to $A_2\mathbf{x} = \mathbf{b}_2$. Suppose $\mathbf{x} = \mathbf{x}_1$ solves system $A_1\mathbf{x} = \mathbf{b}_1$, so that $A_1\mathbf{x}_1 = \mathbf{b}_1$ is true. Then,

$$A_2\mathbf{x}_1 = (EA_1)\mathbf{x}_1 = E(A_1\mathbf{x}_1) = E\mathbf{b}_1 = \mathbf{b}_2.$$

Thus, $\mathbf{x} = \mathbf{x}_1$ is also a solution to system $A_2\mathbf{x} = \mathbf{b}_2$.

A solution to system $A_2\mathbf{x} = \mathbf{b}_2$ is also a solution to $A_1\mathbf{x} = \mathbf{b}_1$. Suppose $\mathbf{x} = \mathbf{x}_2$ solves system $A_2\mathbf{x} = \mathbf{b}_2$, so that $A_2\mathbf{x}_2 = \mathbf{b}_2$ is true. Then,

$$A_1\mathbf{x}_2 = (E^{-1}A_2)\mathbf{x}_2 = E^{-1}(A_2\mathbf{x}_2) = E^{-1}\mathbf{b}_2 = \mathbf{b}_1.$$

Thus, $\mathbf{x} = \mathbf{x}_2$ is also a solution to system $A_1\mathbf{x} = \mathbf{b}_1$.

Conclusion. Since we have now shown that every solution of one system is a solution to the other system, both systems must have exactly the same solution set. ■

CHAPTER 7

Special forms of square matrices

7.1 Discovery guide

Recall that the **main diagonal** of a square matrix refers to the entries on the diagonal from top left to bottom right. Here are some special types of *square* matrices for consideration.

scalar matrix
a scalar multiple of the identity matrix

diagonal matrix
all entries *not* on the main diagonal are zero

upper triangular matrix
all entries *below* the main diagonal are zero

lower triangular matrix
all entries *above* the main diagonal are zero

symmetric matrix
a matrix that is equal to its own transpose

Discovery 7.1 Carry out the following tasks for *each* of the special types of matrices defined above. Think in general, and consider *every* possible size of matrix, not just 2×2 and 3×3! You don't need to *prove* each answer, but you should be able to articulate an informal justification for each answer that doesn't rely on examples (unless it's a *counter*example).

Tip. When considering the questions in this activity for *symmetric* matrices, rather than trying to figure things out with examples, it is much easier to work algebraically with a letter A representing an arbitrary symmetric matrix, and use the definition of symmetric: $A^{\mathrm{T}} = A$.

(a) Write down both a 2×2 and a 3×3 example of the type. Is it clear why this type of matrix has been given its particular name?

(b) Does the (square) zero matrix have this type? Does an identity matrix? Does every 1×1?

(c) If A is a matrix of this type, is every scalar multiple of A also of this type? Is A^{T} of this type?

(d) If A and B are matrices of this type and of the same size, is their sum of this type? Their product? A power (with a positive exponent)?

(e) *[Omit this task for symmetric matrices.]*

Recall that a matrix is invertible if and only if its RREF is the identity matrix. Based on this, can you come up with a simple condition by which you can determine whether a matrix of this type is invertible or not?

(f) If A is an invertible matrix of this type, is its inverse also of this type?

Hint (for symmetric matrices). For the case of symmetric matrices, it will be too complicated to work by examples. Instead, consider the formula $(A^{-1})^{\mathrm{T}} = (A^{\mathrm{T}})^{-1}$ from Proposition 5.5.8 and the definition of **symmetric matrix** above.

(g) Come up with a condition or set of conditions on the entries a_{ij} of a square matrix A by which you can determine whether or not A is of this type.

Hint. Here is an example of the type of condition we're looking for, using the identity matrix: a square matrix A is equal to the identity matrix if $a_{ii} = 1$ for all indices i, and $a_{ij} = 0$ for all pairs of indices i, j with $i \neq j$.

Discovery 7.2 Consider matrices

$$D = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & 1 & 1 \\ -1 & -1 & -1 \\ 1 & 1 & 1 \end{bmatrix}.$$

(a) Compute DA. *Describe the pattern:* multiplying a matrix on the *left* by a diagonal matrix is the same as __________.

(b) Compute AD. *Describe the pattern:* multiplying a matrix on the *right* by a diagonal matrix is the same as __________.

Discovery 7.3 Consider the upper triangular matrix

$$U = \begin{bmatrix} 2 & 1 & 1 \\ 0 & 3 & 1 \\ 0 & 0 & 5 \end{bmatrix}.$$

(a) Can you decompose U into a sum $U = D + P$ of a diagonal matrix D and a “purely” upper triangular matrix P?

(b) Can you decompose U into a product $U = DR$ of a diagonal matrix D and an upper triangular matrix R in REF?

Discovery 7.4 Consider the upper triangular matrix

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

(a) Compute N^2, N^3, and N^4. Do you notice a pattern?

(b) Without computing, what is N^5? N^{99}?

(c) Make a conjecture (i.e. a guess based on previous examples) about what will happen if you compute powers of a 5×5 matrix of a similar form, with all entries equal to 0 except for a line of 1s down the first “superdiagonal.”

Discovery 7.5 This activity will guide you through proving that the sum of two diagonal matrices is again diagonal.

Suppose that A and B are diagonal matrices of the same size. (But do *not* assume that they have a particular size like 2×2 or 3×3 or etc.)

(a) Describe what our assumption that A is diagonal means about the entries of A in terms of your answer to Discovery 7.1.g. Then do the same for B.

(b) Decide exactly what you need to check in order to be sure that the sum $A+B$ is diagonal, in terms of your answer to Discovery 7.1.g. Then carry out that check, using your answer to Task a.

Discovery 7.6 This activity will guide you through proving that the sum of two symmetric matrices is again symmetric. Unlike the proof in Discovery 7.5, we will not need to consider individual entries, since the definition of **symmetric matrix** does not refer to individual entries like the definition of **diagonal matrix** does.

Suppose that A and B are symmetric matrices of the same size. (But do *not* assume that they have a particular size like 2×2 or 3×3 or etc.)

(a) Express what it means for A to be symmetric in mathematical notation, using the symbols A, ${}^{\mathrm{T}}$, and $=$. Then do the same for B.

(b) Express what it would mean for the sum $A+B$ to be symmetric in mathematical notation, similarly to Task a.

(c) Your expressions from Task a are *things we are assuming to be true*. Your expression from Task b is *the condition that needs to be verified*. Carry out this verification, *making sure to use proper LHS vs RHS procedure*. In this verification, you will need to use your assumed knowledge from Task a as well as an algebra rule from Proposition 4.5.1.

7.2 Terminology and notation

scalar matrix
a square matrix that is equal to a scalar multiple of the identity matrix

diagonal matrix
a square matrix where all entries that are not on the main diagonal are equal to zero

upper triangular matrix
a square matrix where all entries that are below the main diagonal are equal to zero

lower triangular matrix
a square matrix where all entries that are above the main diagonal are equal to zero

symmetric matrix
a square matrix that is equal to its own transpose

7.3 Concepts

In this section.

- Subsection 7.3.1 *Algebra with scalar matrices*
- Subsection 7.3.2 *Inverses of special forms*
- Subsection 7.3.3 *Decompositions using special forms*

After writing down examples of these special forms of square matrices in Discovery 7.1, it should be obvious what these kinds of matrices “look” like. But we need to appreciate the difference between our conceptions and the technical definitions of these forms. For example, when we think of an example of an upper triangular matrix, we are likely to focus on the entries on and above the main diagonal, because those are what form the “upper triangular” shape, and all the other entries below the main diagonal are zero. But the technical definition of **upper triangular matrix** provided in Section 7.2 focuses on those zero entries below the main diagonal, and *does not mention the entries on or above the main diagonal at all*.

Unlike a conception, a technical definition aims to capture the minimum information necessary to identify an instance of the concept. For the purposes of identifying an upper triangular matrix, the entries on or above the main diagonal are irrelevant and only the zeros below the main diagonal matter, because if any of those entries were nonzero the matrix in question would most certainly *not* be upper triangular. But this minimalism in making technical definitions can sometimes have surprising side effects, as we discovered in Discovery 7.1. For example, a diagonal matrix is, by definition, also *both* upper and lower triangular, because its entries below *and* above the main diagonal are all zero. As an extreme example, a square zero matrix is simultaneously *all three* of diagonal, upper triangular, and lower triangular.

Question 7.3.1 Why are these special forms important? □

At this stage, we can state a few reasons why we might be interested in identifying these matrix forms with special names.

- For the diagonal and triangular forms, the fact that many of their entries are zero makes computing with them especially easy, whether with respect to matrix operations or with respect to solving systems.
- With regards to solving systems, any square matrix in REF (or RREF) must be upper triangular. And lower triangular is just the transposed version of upper triangular, so it seems reasonable to identify it along with the upper triangular form.
- Symmetric matrices play a special role in the geometry of the plane, of space, and of higher-dimensional "hyperspaces," as you may discover in a second course in linear algebra.
- Finally, for each of these forms (including symmetric), you discovered in Discovery 7.1 that *adding or scalar multiplying matrices of the form resulted in another matrix of the same form*. This was also true for products, powers and inverses, except that a product of two symmetric matrices may not be symmetric. The fact that matrix operations on these forms produce results of the same form is an important property in more advanced abstract algebra.

7.3.1 Algebra with scalar matrices

In Discovery 7.1, you might have noticed how certain rules of matrix algebra apply to scalar matrices:

$kI + mI = (k+m)I$

(Rule 2.b of Proposition 4.5.1);

$kI - mI = (k-m)I$

(Rule 2.b and Rule 2.f of Proposition 4.5.1 combined);

$(kI)(mI) = (km)I$

(Rule 2.c and Rule 2.d of Proposition 4.5.1 combined with Rule 2 of Proposition 5.5.1);

$(kI)^p = k^p I$

(Rule 4.c of Proposition 4.5.1 combined with Rule 2 of Proposition 5.5.1);

$(kI)^{-1} = k^{-1} I$

(Rule 2 of Proposition 5.5.5 combined with Rule 3 of Proposition 5.5.1).

So for scalar matrices there seems to be a pattern: *the matrix operation can be achieved by just performing the corresponding scalar operation*. This essentially gives us a way to "inject" the algebra of numbers into the algebra of square matrices of any given size, an extremely important notion in more advanced abstract algebra that you may encounter a taste of in a second linear algebra course.

7.3.2 Inverses of special forms

In Discovery 7.1.e, we examined the invertibility of these various forms of matrices. In Theorem 6.5.2 we learned that a matrix is invertible only if it can be reduced to the identity. Now, scalar matrices, diagonal matrices, and upper triangular matrices are already pretty close to being reduced, but we can see that if any of the diagonal entries of these forms of matrix is zero, then there will be no hope of getting a leading one in that column, and so we won't be able to reduce to the identity. Thus, *a scalar, diagonal, or upper triangular matrix is only invertible if its diagonal entries are all nonzero*. And, since the transpose of a lower triangular is upper triangular, and since taking a transpose does not affect invertibility (Proposition 5.5.8), then the same is true about lower triangular matrices. Analyzing the invertibility of symmetric matrices is a little more complicated, but in Discovery 7.1.f, we discovered that for each of these special forms (including symmetric matrices), the inverse of a matrix of that form is also of that form.

7.3.3 Decompositions using special forms

In Discovery 7.3 we discovered that an upper triangular matrix can be decomposed into a sum or a product of a diagonal matrix with a special kind of upper triangular matrix. Using the matrix from that discovery activity as an example, we have

$$\begin{bmatrix} 2 & 1 & 1 \\ 0 & 3 & 1 \\ 0 & 0 & 5 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \tag{$*$}$$

$$\begin{bmatrix} 2 & 1 & 1 \\ 0 & 3 & 1 \\ 0 & 0 & 5 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & \frac{1}{3} \\ 0 & 0 & 1 \end{bmatrix}. \tag{$**$}$$

The special upper triangular matrix in the product decomposition in (∗∗) is called a **unipotent matrix** because its powers will always have that line of ones down the main diagonal. The special upper triangular matrix in the sum decomposition in (∗) is called a **nilpotent matrix** because its powers will always have that line of zeros down the main diagonal, and in fact, just like the nilpotent matrices you analyzed in Discovery 7.4, if you raise this matrix to an exponent equal to its size you will get the zero matrix!

A look ahead. Nilpotent matrices play an important role in more advanced theory of matrix forms, which you might encounter in a second linear algebra course.

7.4 Examples

In this section.

7.4.1 Computation patterns

Here we will concentrate mostly on computational patterns involving diagonal matrices. (Computations involving upper triangular or lower triangular matrices are somewhat similar — see further below.)

Example 7.4.1 Matrix operations involving diagonal matrices. Let's look at each of a sum, product, power, and inverse involving diagonal matrices, in the 3 × 3 case.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} + \begin{bmatrix} 4 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 6 \end{bmatrix} = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 9 \end{bmatrix}$$

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

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{bmatrix} = \begin{bmatrix} 1^2 & 0 & 0 \\ 0 & 2^2 & 0 \\ 0 & 0 & 3^2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \end{bmatrix}$$

□

We can easily identify some patterns in the above example.

- We add diagonal matrices by adding corresponding diagonal entries.
- We multiply diagonal matrices by multiplying corresponding diagonal entries.
- We exponentiate a diagonal matrix by exponentiating each of the diagonal entries by the same exponent.
- We invert a diagonal matrix by inverting (i.e. taking the reciprocal of) each of the diagonal entries.

We have some of the same patterns for upper and lower triangular matrices, at least for the diagonal entries. We'll demonstrate with some upper triangular example computations.

Example 7.4.2 Basic matrix operations involving upper triangular matrices.

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{bmatrix} + \begin{bmatrix} 4 & 1 & 1 \\ 0 & -2 & 1 \\ 0 & 0 & 6 \end{bmatrix} = \begin{bmatrix} 5 & 2 & 2 \\ 0 & 0 & 2 \\ 0 & 0 & 9 \end{bmatrix}$$

$$\begin{aligned} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 4 & 1 & 1 \\ 0 & -2 & 1 \\ 0 & 0 & 6 \end{bmatrix} &= \begin{bmatrix} 4+0+0 & 1-2+0 & 1+1+0 \\ 0+0+0 & 0-4+0 & 0+2+6 \\ 0+0+0 & 0+0+0 & 0+0+18 \end{bmatrix} \\ &= \begin{bmatrix} 4 & -1 & 2 \\ 0 & -4 & 8 \\ 0 & 0 & 18 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{bmatrix}^2 &= \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 1+0+0 & 1+2+0 & 1+1+3 \\ 0+0+0 & 0+4+0 & 0+2+3 \\ 0+0+0 & 0+0+0 & 0+0+9 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 3 & 5 \\ 0 & 4 & 5 \\ 0 & 0 & 9 \end{bmatrix} \\ &= \begin{bmatrix} 1^2 & 3 & 5 \\ 0 & 2^2 & 5 \\ 0 & 0 & 3^2 \end{bmatrix} \end{aligned}$$

□

Computing the inverse of an upper triangular matrix is not as simple as for a diagonal matrix — some row reduction will be required, using Procedure 6.3.7.

Example 7.4.3 Inverse of an upper triangular matrix. Augment with the identity and reduce.

$$\left[\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 2 & 2 & 0 & 1 & 0 \\ 0 & 0 & 3 & 0 & 0 & 1 \end{array}\right] \begin{array}{l} \\ \frac{1}{2}R_2 \\ \frac{1}{3}R_3 \end{array} \longrightarrow \left[\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 & 0 & \frac{1}{3} \end{array}\right] \begin{array}{l} R_1 - R_2 \\ \\ \\ \end{array}$$

$$\longrightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 1 & 1 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 & 0 & \frac{1}{3} \end{array}\right] \begin{array}{l} \\ R_2 - R_3 \\ \\ \end{array} \longrightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 1 & 0 & 0 & \frac{1}{2} & -\frac{1}{3} \\ 0 & 0 & 1 & 0 & 0 & \frac{1}{3} \end{array}\right]$$

With this reduction, we have calculated that

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -\frac{1}{2} & 0 \\ 0 & \frac{1}{2} & -\frac{1}{3} \\ 0 & 0 & \frac{1}{3} \end{bmatrix}.$$

□

Again, in these two examples we see the same patterns on the main diagonal as for diagonal matrices. Products, powers, and inverses of lower triangular matrices would be similar.

Remark 7.4.4 More patterns with diagonal matrices. In the example calculations of Discovery 7.2, we also found the following patterns.

- Multiplying a matrix A on the *left* by a diagonal matrix D multiplies each *row* of A by the corresponding diagonal entry of D.
- Multiplying a matrix A on the *right* by a diagonal matrix D multiplies each *column* of A by the corresponding diagonal entry of D.

A look ahead. The second of the patterns described in Remark 7.4.4 will be important in Chapter 22.

7.5 Theory

In this section.

- Subsection 7.5.1 *Algebra of special forms*
- Subsection 7.5.2 *Invertibility of special forms*

Here we record properties of these special forms of matrices relative to the various matrix operations.

7.5.1 Algebra of special forms

First, we summarize some of the algebra of working with these forms. We have already explored proving parts of the proposition below in Discovery 7.5 and Discovery 7.6, so below we provide similar proofs for a couple more parts.

Proposition 7.5.1

1. *The result of adding two diagonal matrices, scalar multiplying a diagonal*

matrix, multiplying two diagonal matrices, taking an inverse of a diagonal matrix, or taking a power (positive or negative) of a diagonal matrix is always a diagonal matrix.

2. *Statement 1 remains true*
   - *when all occurrences of the word "diagonal" are replaced by "scalar," or*
   - *when all occurrences of the word "diagonal" are replaced by "upper triangular," or*
   - *when all occurrences of "diagonal" are replaced by "lower triangular."*
3. *Statement 1 remains true when all occurrences of the word "diagonal" are replaced by "symmetric,"* except *that the product of two symmetric matrices may not be symmetric.*

Partial proof of Statement 2. We will prove that the result of scalar multiplying an upper triangular matrix is again upper triangular. As we discovered in Discovery 7.1.g, an upper triangular matrix U is characterized by having all entries u_{ij} equal to 0 for $i > j$ (i.e. entries below the main diagonal). The scalar multiple kU has entries $[kU]_{ij} = ku_{ij}$, so if $u_{ij} = 0$ for $i > j$, then also $ku_{ij} = 0$ for $i > j$, and the matrix kU is also upper triangular. ■

Partial proof of Statement 3. We will prove that the inverse of an invertible, symmetric matrix is again symmetric. So suppose that A is both invertible and symmetric. By definition of symmetry, this means that A is equal to its own transpose. We would like to verify that A^{-1} is also symmetric; that is, that A^{-1} is equal to its own transpose. Let's do that, using proper LHS vs RHS procedure for the proposed equality $(A^{-1})^{\mathrm{T}} = A^{-1}$:

$$\begin{aligned} \text{LHS} &= (A^{-1})^{\mathrm{T}} & \\ &= (A^{\mathrm{T}})^{-1} & \text{(i)} \\ &= (A)^{-1} & \text{(ii)} \\ &= \text{RHS}, & \end{aligned}$$

with justifications

(i) Proposition 5.5.8; and

(ii) $A^{\mathrm{T}} = A$ by symmetric assumption.

■

7.5.2 Invertibility of special forms

Finally, we record our observations about the invertibility of some of these special forms. The following fact was already discussed in Subsection 7.3.2, so we will not formally prove it.

Proposition 7.5.2 *An upper or lower triangular matrix is invertible precisely when the entries on its main diagonal are all nonzero.*

Special case. Since scalar and diagonal matrices are just particular forms of triangular matrix, Proposition 7.5.2 applies to scalar and diagonal matrices as well.

CHAPTER 8

Determinants

8.1 Discovery guide

Discovery 8.1 Consider the generic 2×2 matrix A and the "mixed up" version A_{mix}:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \qquad\qquad A_{\mathrm{mix}} = \begin{bmatrix} d & -c \\ -b & a \end{bmatrix}.$$

(a) Compute $AA_{\mathrm{mix}}^{\mathrm{T}}$. Then fill in the blank.

$$AA_{\mathrm{mix}}^{\mathrm{T}} = (\qquad\qquad)I \tag{$*$}$$

(b) Modify equation $(*)$ algebraically to fill in the blank.

$$A(\qquad\qquad) = I \tag{$**$}$$

(c) Recall that if the product of two square matrices is equal to I, then those matrices must be inverses of each other (Proposition 6.5.4 and Proposition 6.5.6). With this knowledge, compare equation $(**)$ with Proposition 5.5.4.

(d) What needs to be true about a, b, c, d for the algebra in Task b to be valid? Why?

The goal of this and the next two discovery guides (along with the corresponding chapters) is to develop something similar to the results of the first discovery activity above for larger square matrices. First, we will start by extending the 2×2 formula $ad - bc$. This formula *determines* whether a 2×2 matrix is invertible or not, so we call it the *determinant* of the matrix.

We will actually start back at 1×1 matrices, and build up from there.

Discovery 8.2 Consider the generic 1×1 matrix $A = \begin{bmatrix} a \end{bmatrix}$.

(a) The inverse of $A = \begin{bmatrix} a \end{bmatrix}$ is $A^{-1} = \begin{bmatrix} \qquad\qquad \end{bmatrix}$, but this only works if $\qquad\qquad$.

(b) So *before* attempting to compute A^{-1}, we can *determine* whether this attempt will be successful by looking at the matrix $A = \begin{bmatrix} a \end{bmatrix}$ and considering the *single number* $\qquad$.

(Make sure your response is always a number!)

To build up to larger matrices, we need to take it step-by-step.

Discovery 8.3 For an $n \times n$ matrix with $n > 1$, the $(i,j)^{\text{th}}$ **minor** (denoted M_{ij}) is the determinant of the smaller submatrix obtained by removing the row and column that contain the $(i,j)^{\text{th}}$ entry.

Since you know how to compute 1×1 determinants, you can now compute all four minors $(M_{11}, M_{12}, M_{21}, M_{22})$ of the matrix

$$\begin{bmatrix} -1 & 3 \\ -4 & 2 \end{bmatrix}.$$

Discovery 8.4 The $(i,j)^{\text{th}}$ **cofactor** of a matrix (denoted C_{ij}) is defined to be the $(i,j)^{\text{th}}$ minor, *except* that we multiply it by -1 when $i+j$ is odd. That is, $C_{ij} = (-1)^{i+j} M_{ij}$. Compute all four cofactors $(C_{11}, C_{12}, C_{21}, C_{22})$ for the matrix from Discovery 8.3. (You've already computed the minors, now you just need to make some of them negative.)

Discovery 8.5 We now initially define the **determinant** of a matrix to be a combination of entries and cofactors along the first row. To compute the determinant, multiply each entry in the first row by its own cofactor, and then add all these together. For a 2×2 matrix, the formula is

$$\det A = a_{11} C_{11} + a_{12} C_{12}.$$

Use this formula to compute the determinant of the matrix from Discovery 8.3.

Discovery 8.6 Use $\det A = a_{11} C_{11} + a_{12} C_{12}$ to compute the determinant of the generic 2×2 matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Surprised?

Discovery 8.7 Compute the determinant of the 3×3 matrix

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

Use the same sort of "**cofactor expansion** along the first row" as before; that is, "entry times cofactor plus entry times cofactor plus entry times cofactor . . . " along the first row, except now your cofactor calculations will involve 2×2 determinants.

Tip. In light of Discovery 8.6, just use the $ad - bc$ formula to calculate determinants of 2×2 submatrices.

Discovery 8.8 For this activity, use the same matrix as Discovery 8.7.

(a) Try computing a cofactor expansion along a different row.

(b) Now try along a column.

What did you find in these calculations? Make a conjecture about cofactor expansions along different rows or columns in a matrix in general.

Discovery 8.9 Recall the cofactor formula: $C_{ij} = (-1)^{i+j} M_{ij}$. The $(-1)^{i+j}$ part will be 1 when $i+j$ is even and -1 when $i+j$ is odd. In a 2×2 matrix this makes a pattern: $\left[\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}\right]$.

Make similar matrices of +/− for the patterns of cofactor signs in 3×3 and 4×4 matrices.

Discovery 8.10

(a) Using your finding from Discovery 8.8 as appropriate, come up with simple formulas for the determinant of diagonal matrices, upper triangular matrices, and lower triangular matrices.

Hint. In these special matrices, there are some rows/columns that are easier to use in a cofactor expansion than others.

(b) What is $\det \mathbf{0}$? ... $\det I$? Are the answers the same for every size of zero/ identity matrix?

8.2 Terminology and notation

$(i,j)^{\text{th}}$ minor of a square matrix A
the determinant of the smaller square matrix obtained from A by removing the i^{th} row and the j^{th} column
— written M_{ij}

$(i,j)^{\text{th}}$ cofactor of a square matrix A
equal to either the corresponding minor of A or its negative, depending on whether $i+j$ is even or odd
— written C_{ij}

cofactor expansion along the i^{th} row of square matrix A
the formula $a_{i1}C_{i1}+a_{i2}C_{i2}+\cdots+a_{in}C_{in}$, where C_{ij} denotes the $(i,j)^{\text{th}}$ cofactor of A

cofactor expansion along the j^{th} column of square matrix A
the formula $a_{1j}C_{1j}+a_{2j}C_{2j}+\cdots+a_{nj}C_{nj}$, where again C_{ij} denotes the $(i,j)^{\text{th}}$ cofactor of A

determinant
the common value of all cofactor expansions of a particular square matrix
— written $\det A$

$\det A$ notation to represent the value of the determinant of a square matrix A

Alternative determinant notation. When computing cofactor expansions, we are often performing determinant calculations inside determinant calculations, and it becomes awkward to have det symbols littered throughout our intermediate calculations. So we will also write $|A|$ to mean the determinant of a matrix, especially for actual matrices. For example,

$$A=\begin{bmatrix}1&2&3\\4&5&6\\7&8&9\end{bmatrix}\qquad\implies\qquad\det A=\begin{vmatrix}1&2&3\\4&5&6\\7&8&9\end{vmatrix}.$$

8.3 Concepts

In this section.

- Subsection 8.3.1 *Definition of the determinant*
- Subsection 8.3.2 *Determinants of* 1×1 *matrices*
- Subsection 8.3.3 *Determinants of* 2×2 *matrices*
- Subsection 8.3.4 *Determinants of larger matrices*
- Subsection 8.3.5 *Determinants of special forms*

In Discovery 8.1, we discovered that for every 2×2 matrix A there is a related matrix A' so that the product AA' is a scalar multiple of the identity matrix. Call this scalar δ for now. If $\delta\neq 0$, we can do some algebra to get

$$AA'=\delta I\qquad\implies\qquad A(\delta^{-1}A')=I,$$

which means that A must be invertible with $A^{-1} = \delta^{-1}A'$ (Proposition 6.5.6).

Goal 8.3.1 *For a square matrix A of* any *size, determine a scalar δ and a matrix A' so that $AA' = \delta I$.*

Now, we could achieve this goal by *always* choosing $\delta = 0$ and $A' = 0$, but that won't help us replicate for larger matrices the patterns we discovered in the 2×2 case. We will find that there is a very particular procedure to achieve this goal that works for every square matrix *and* recovers the 2×2 case above, so we will tackle the goal in two parts:

1. determine the scalar δ for each square matrix A, and then

2. determine how to construct the matrix A' that goes along with it.

The process of producing the scalar δ is then a *function* on square matrices. For a particular square matrix A, we will call the output δ of this function the **determinant of** A, and usually write $\det A$ instead of δ.

Idea 8.3.2 *If $AA' = (\det A)I$, then in the case that $\det A \neq 0$, from*

$$A\bigl((\det A)^{-1}A'\bigr) = I$$

and Proposition 6.5.6 we know both that A is invertible and its inverse must be $(\det A)^{-1}A'$, as in the 2×2 case discussed above. Also, we will learn in Chapter 10 that when $\det A = 0$, then A must *be singular. So the value of the determinant of a matrix will* determine *whether or not it is invertible.*

For now, we will concentrate on the first step and learn how to compute determinants, as it turns out that the "companion" matrix A' will be constructed out of determinants of submatrices of A. We will discuss this special matrix and complete our goal in Chapter 10.

8.3.1 Definition of the determinant

It may seem from Section 8.2 that the definition of determinant is *circular* — we define the determinant in terms of entries and cofactors (via cofactor expansions), where cofactors are defined in terms of minors, which are defined in terms of ... determinants? But the key word in the definition of **minor** is *smaller* — determinants are defined *recursively* in terms of smaller matrices. In Discovery guide 8.1, after first exploring the determinant of a 2×2 matrix as motivation, we started afresh with a precise definition of the 1×1 determinant, and then defined the 2×2 determinant in terms of 1×1 determinants. Then the 3×3 determinant is defined in terms of 2×2 determinants, and so on. As we will see in examples in Section 8.4, computing a determinant from this recursive definition will involve unpacking it in terms of determinants of one smaller size, then unpacking those in terms of determinants of one size smaller again, and so on. Technically, this process should continue until we are down to a bunch of 1×1 determinants, but since there is a simple formula for a 2×2 determinant, in direct computations we will stop there.

Warning 8.3.3 Computing determinants by cofactor expansions is *extremely* inefficient, whether by hand or by computer. For example, for a 10×10 matrix, the recursive process of a cofactor expansion could eventually require you to compute more than 1.8 million 2×2 determinants. In the next chapter we will discover that we can also compute determinants by ... you guessed it, row reduction! (And there are other, more efficient methods for determinants by computer — we will leave those to a *numerical methods* course.) But again, the goal of this course is *not* to turn *you* into a super-efficient computer. We want to understand and be somewhat proficient at computing determinants by cofactor expansions so that we can think about and understand them in the abstract while we develop the *theory* of determinants.

8.3.2 Determinants of 1×1 matrices

Consider the general 1×1 matrix $A = \begin{bmatrix} a \end{bmatrix}$. We should expect the invertibility of A to be completely *determined* by the value of the single entry a, since that is all the information that A contains. And that is precisely the case, as A is invertible when $a \neq 0$, with $A^{-1} = \begin{bmatrix} a^{-1} \end{bmatrix}$, and A is singular when $a = 0$, because then A would be the zero matrix. Since entry a *determines* the invertibility of A, we set $\det \begin{bmatrix} a \end{bmatrix} = a$.

8.3.3 Determinants of 2×2 matrices

In Discovery 8.6, we calculated the determinant of the general 2×2 matrix to be

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc,$$

using a cofactor expansion along the first row. (We leave it up to you, the reader, to check that a cofactor expansion along a column or along the second row yields the same result.) And we already verified by row reducing that a 2×2 matrix is invertible precisely when $ad - bc \neq 0$ (Proposition 6.5.9).

8.3.4 Determinants of larger matrices

In Discovery 8.2–8.7, we used an **inductive** process to build up from computing 1×1 determinants to 3×3 determinants. The inductive process continues for larger matrices to provide a formula for the determinant of an $n \times n$ matrix for every n via a cofactor expansion along the first row:

$$\det A = a_{11}C_{11} + a_{12}C_{12} + a_{13}C_{13} + \cdots + a_{1n}C_{1n}. \tag{$*$}$$

And we saw in Discovery 8.8 that can replace the cofactor expansion in $(*)$ with a cofactor expansion along any row or column of our choosing and get the same result.

Inductive versus recursive. Induction and **recursion** are two sides of the same coin. Both are step-by-step processes. In an inductive process, we *build up* step-by-step, using the results of the previous step to create the process for the next step. Theoretically, we imagine this process could continue forever, effectively establishing *all infinity* of the possible steps/cases. In a recursive process, we *work backwards* from a particular step/case, repeatedly decomposing the current case into a process/calculation of the type of the previous case. In the case of calculations or algorithms, an inductive process usually leads to a recursive algorithm. If you undertake further studies in mathematics and/or computing science you will encounter induction and recursion frequently.

While we have a convenient general formula for 2×2 matrices in terms of the four entry variables, we certainly wouldn't want to attempt to write out a general formula for the determinant of a 5×5 matrix in *twenty-five* entry variables. Instead, for matrices larger than 2×2, computing a determinant for a specific matrix from a cofactor expansion is a **recursive** process, since cofactors are just minor determinants with some sign changes. A cofactor expansion for an $n \times n$ matrix requires n cofactor calculations. Each of those cofactor calculations is a determinant calculation of some $(n-1)\times(n-1)$ "submatrices". Each of those determinants, if calculated by cofactor expansion, will require $n-1$ determinant calculations of various $(n-2)\times(n-2)$ "submatrices". And so on. As you can see, the number of calculations involved grows out of hand quite quickly, even for single-digit values of n.

We will work through some 3×3 and 4×4 cofactor expansions in Section 8.4, but we will develop a more efficient determinant calculation procedure based on row operations in Chapter 9. For now, let's record the cofactor sign patterns from Discovery 8.9. Remember that a cofactor is equal to either the corresponding minor determinant or its negative, depending on whether the sum $i+j$ of row and column indices is even or odd. This extra "sign" portion of the cofactor formula in terms of minor determinants will alternate from entry to entry, since as we move along a row or along a column, only one of i or j will change, and so $i+j$ will flip from even to odd or vice versa. So the cofactor signs follow the patterns,

$$3\times 3\colon \begin{bmatrix} + & - & + \\ - & + & - \\ + & - & + \end{bmatrix}, \qquad 4\times 4\colon \begin{bmatrix} + & - & + & - \\ - & + & - & + \\ + & - & + & - \\ - & + & - & + \end{bmatrix}, \qquad 5\times 5\colon \begin{bmatrix} + & - & + & - & + \\ - & + & - & + & - \\ + & - & + & - & + \\ - & + & - & + & - \\ + & - & + & - & + \end{bmatrix}, \tag{8.3.1}$$

and so on.

8.3.5 Determinants of special forms

In Discovery 8.10, we examined the determinant of diagonal and triangular matrices. Let's consider the case of a diagonal matrix:

$$D = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & d_n \end{bmatrix}.$$

A cofactor expansion along the first column will look like

$$d_1 C_{11} + 0\cdot C_{21} + 0\cdot C_{31} + \cdots + 0\cdot C_{n1}.$$

Because of all of those zero entries, the only cofactor we actually need to compute is C_{11}, and the cofactor expansion collapses to just the entry d_1 times its cofactor. But the cofactor sign of the $(1,1)$ entry is positive, so we really just get d_1 times its minor determinant:

$$\det D = d_1 \begin{vmatrix} d_2 & 0 & \cdots & 0 \\ 0 & d_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & d_n \end{vmatrix}.$$

This minor determinant is again a diagonal matrix, so we can again expand along the first column to get a similar result. And the pattern will continue until we finally get down to a 1×1 minor

$$\det D = d_1 d_2 \begin{vmatrix} d_3 & 0 & \cdots & 0 \\ 0 & d_4 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & d_n \end{vmatrix} = d_1 d_2 d_3 \begin{vmatrix} d_4 & 0 & \cdots & 0 \\ 0 & d_5 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & d_n \end{vmatrix}$$

$$= \cdots = d_1 d_2 \cdots d_{n-2} \begin{vmatrix} d_{n_1} & 0 \\ 0 & d_n \end{vmatrix} = d_1 d_2 \cdots d_{n-1} \left|[d_n]\right|$$
$$= d_1 d_2 \cdots d_{n-1} d_n.$$

So *the determinant of a diagonal matrix is equal to the product of its diagonal entries*.

What if we apply this pattern to an $n \times n$ scalar matrix kI? Since such a matrix has the entry k repeated down the diagonal n times, the determinant will be n factors of k multiplied together, so that $\det(kI) = k^n$. Applying this formula to the zero matrix ($k = 0$) and the identity matrix ($k = 1$), we have

$$\det \mathbf{0} = 0, \qquad\qquad \det I = 1.$$

When computing the determinant of an upper triangular matrix, a similar pattern of computation as in the diagonal case would arise, because choosing to always expand along the first column would result in diagonal entry times an upper triangular minor determinant. And the same pattern would repeat for lower triangular matrices, but for those it is best to expand along the first row.

8.4 Examples

In this section.

- Subsection 8.4.1 *Determinants of* 2×2 *matrices*
- Subsection 8.4.2 *Minors and cofactors of* 3×3 *matrices*
- Subsection 8.4.3 *Determinants of* 3×3 *matrices*
- Subsection 8.4.4 *Minors and cofactors of* 4×4 *matrices*
- Subsection 8.4.5 *Determinants of* 4×4 *matrices*

8.4.1 Determinants of 2×2 matrices

An easy way to remember the 2×2 determinant formula is with a crisscross pattern, as illustrated below for general 2×2 matrix $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$.

$$\det A = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

Example 8.4.1 Determinant of a 2×2 matrix. For $A = \left[\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right]$, we have

$$\det A = \begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix} = 1 \cdot 4 - 2 \cdot 3 = 4 - 6 = -2.$$

□

Watch out for double negatives! The next example illustrates the occurrence of a double negative in a determinant calculation.

Example 8.4.2 Another 2×2 determinant. For $A = \left[\begin{smallmatrix} 1 & 2 \\ -3 & 4 \end{smallmatrix}\right]$, we have

$$\det A = \begin{vmatrix} 1 & 2 \\ -3 & 4 \end{vmatrix} = 1 \cdot 4 - 2 \cdot (-3) = 4 + 6 = 10.$$

□

8.4.2 Minors and cofactors of 3×3 matrices

8.4.2.1 Minors

A minor determinant is just a *one-size-smaller* determinant. To obtain that smaller matrix, we remove one row and one column. Usually we specify which to remove by focusing on a single entry and removing the row and column that contain the entry.

Example 8.4.3 Minor determinants in a 3×3 matrix. Let's compute a couple of minor determinants in the matrix from Discovery 8.7:

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

The notation M_{11} means the minor associated to the $(1,1)$ entry, so we should remove both the first row and the first column, leaving behind a 2×2 matrix.

$$M_{11} = \begin{vmatrix} \enclose{horizontalstrike,verticalstrike}{3} & \enclose{horizontalstrike}{1} & \enclose{horizontalstrike}{0} \\ \enclose{verticalstrike}{-2} & -2 & 1 \\ \enclose{verticalstrike}{0} & 1 & -1 \end{vmatrix} = \begin{vmatrix} -2 & 1 \\ 1 & -1 \end{vmatrix}$$

We can now compute this minor determinant using the $ad - bc$ pattern for 2×2 determinants.

$$M_{11} = \begin{vmatrix} \searrow\!-2 & 1\!\swarrow \\ \swarrow\!1 & -1\!\searrow \end{vmatrix} = (-2) \cdot (-1) - 1 \cdot 1 = 2 - 1 = 1.$$

Now let's try the M_{23} minor determinant. This time we should remove the second row and the third column.

$$M_{23} = \begin{vmatrix} 3 & 1 & \enclose{verticalstrike}{0} \\ \enclose{horizontalstrike}{-2} & \enclose{horizontalstrike}{-2} & \enclose{horizontalstrike,verticalstrike}{1} \\ 0 & 1 & \enclose{verticalstrike}{-1} \end{vmatrix} = \begin{vmatrix} 3 & 1 \\ 0 & 1 \end{vmatrix}$$

Again, from here we compute this minor determinant using the $ad - bc$ pattern.

$$M_{23} = \begin{vmatrix} \searrow\!3 & 1\!\swarrow \\ \swarrow\!0 & 1\!\searrow \end{vmatrix} = 3 \cdot 1 - 1 \cdot 0 = 3 - 0 = 3.$$

□

8.4.2.2 Cofactors

A cofactor just takes a minor determinant and (sometimes) flips its sign: when the corresponding entry is at an "even" position then the cofactor is equal to the minor determinant value, and when the corresponding entry is at an "odd" position then the sign is flipped.

Example 8.4.4 Cofactors in a 3×3 matrix. Let's continue Example 8.4.3 above. The minor determinant M_{11} corresponds to the $(1,1)$ entry in the matrix, which is at an "even" position since $1 + 1 = 2$ is even. So the corresponding cofactor value is equal to the minor determinant value:

$$C_{11} = M_{11} = 1.$$

But the minor determinant M_{23} corresponds to the $(2,3)$ entry in the matrix, which is at an "odd" position since $2+3=5$ is odd. So the corresponding cofactor value is equal to the negative of the minor determinant value:

$$C_{23} = -M_{23} = -3.$$

□

8.4.3 Determinants of 3×3 matrices

For a 3×3 matrix, we choose a *single* row or column and perform a cofactor expansion. It's usually best to choose the row or column with the most zeros, since for a zero entry the "entry times cofactor" part of the expansion for that entry will be zero no matter the value of the cofactor, and we don't actually have to calculate that cofactor. Also, we will use our cofactor sign patterns from Subsection 8.3.4 (see Pattern (8.3.1)), instead of calculating $(-1)^{i+j}$ explicitly.

Example 8.4.5 Determinant of a 3×3 matrix along a row. Let's compute the determinant of the matrix from Discovery 8.7:

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

Any of the first row or column or the third row or column would be good choices as they all contain a zero entry. Let's choose the third row, since it also contains some 1s, which will simplify things a bit. Notice how we have also annotated that row with the cofactor sign pattern.

$$\det A = \left|\begin{array}{ccc} 3 & 1 & 0 \\ -2 & -2 & 1 \\ \hline \multicolumn{1}{|c}{0^+} & 1^- & \multicolumn{1}{c|}{-1^+} \\ \hline \end{array}\right|$$

Now expand along that third row.

$$\det A = 0\cdot \begin{vmatrix} \cancel{3} & 1 & 0 \\ \cancel{-2} & -2 & 1 \\ \cancel{0} & \cancel{1} & \cancel{-1} \end{vmatrix} - 1\cdot \begin{vmatrix} 3 & \cancel{1} & 0 \\ -2 & \cancel{-2} & 1 \\ \cancel{0} & \cancel{1} & \cancel{-1} \end{vmatrix} + (-1)\cdot \begin{vmatrix} 3 & 1 & \cancel{0} \\ -2 & -2 & \cancel{1} \\ \cancel{0} & \cancel{1} & \cancel{-1} \end{vmatrix}$$

The minus sign between the first two terms in the expansion is the proper cofactor sign for the middle entry of the third row. Also, recall that a cofactor for an entry involves the minor for that entry — the determinant of the smaller matrix obtained by removing the row and column in which that entry sits. We have indicated each removal of a row or column by a strike-through. Since A is 3×3, all of its minors are 2×2 determinants that we can compute with our crisscross pattern. However, since the $(3,1)$ entry is 0, there is no need to compute the $(3,1)$ minor.

$$\det A = 0 - 1\cdot \begin{vmatrix} 3 & 0 \\ -2 & 1 \end{vmatrix} + (-1)\cdot \begin{vmatrix} 3 & 1 \\ -2 & -2 \end{vmatrix}$$

Using our crisscross pattern for 2×2 determinants, we can now compute

$$\begin{aligned} \det A &= 0 - 1\cdot\bigl[3\cdot 1 - 0\cdot(-2)\bigr] + (-1)\cdot\bigl[3\cdot(-2) - 1\cdot(-2)\bigr] \\ &= -3 + (-1)(-4) \\ &= 1. \end{aligned}$$

□

Just to check, let's compute the determinant in the above example again using a cofactor expansion along the second column.

Example 8.4.6 Determinant of a 3×3 matrix along a column. Let's again compute the determinant of the matrix from Discovery 8.7, but this time along the middle column.

$$\det A = \left|\begin{array}{r|c|r} 3 & 1^- & 0 \\ -2 & -2^+ & 1 \\ 0 & 1^- & -1 \end{array}\right|$$

Expand along the chosen column.

$$\det A = -1 \cdot \begin{vmatrix} \cancel{3} & \cancel{1} & \cancel{0} \\ -2 & -\cancel{2} & 1 \\ 0 & \cancel{1} & -1 \end{vmatrix} + (-2) \cdot \begin{vmatrix} 3 & \cancel{1} & 0 \\ \cancel{-2} & \cancel{-2} & \cancel{1} \\ 0 & \cancel{1} & -1 \end{vmatrix} - 1 \cdot \begin{vmatrix} 3 & \cancel{1} & 0 \\ -2 & -\cancel{2} & 1 \\ \cancel{0} & \cancel{1} & \cancel{-1} \end{vmatrix}$$

In the expansion, the negative sign in front of the first term and the minus sign between the second and third terms are from the cofactor sign pattern for the second column.

Now reduce to a combination of 2×2 determinants.

$$\det A = -1 \cdot \begin{vmatrix} -2 & 1 \\ 0 & -1 \end{vmatrix} + (-2) \cdot \begin{vmatrix} 3 & 0 \\ 0 & -1 \end{vmatrix} - 1 \cdot \begin{vmatrix} 3 & 0 \\ -2 & 1 \end{vmatrix}$$

Apply the 2×2 crisscross pattern.

$$\begin{aligned} \det A &= (-1)(2-0) + (-2)(-3-0) - 1 \cdot (3-0) \\ &= -2 + 6 - 3 \\ &= 1. \end{aligned}$$

□

In the end, we got the same result as our first calculation, which is not a coincidence — see Theorem 8.5.1.

8.4.4 Minors and cofactors of 4×4 matrices

Applying the *one-size-smaller* pattern again, a minor determinant in a 4×4 matrix is the determinant of a 3×3 matrix obtained by removing one row and one column. And again cofactor values are equal to minor determinant values, except that we flip the signs for values associated to "odd" positions with the 4×4 matrix.

Example 8.4.7 Consider the matrix

$$\begin{bmatrix} -1 & 3 & 1 & 0 \\ -5 & 6 & 7 & 8 \\ 2 & -2 & -2 & 1 \\ 2 & 0 & 1 & -1 \end{bmatrix}.$$

To compute the M_{21} minor determinant, we remove the second row and the first column.

$$M_{21} = \begin{vmatrix} -\cancel{1} & 3 & 1 & 0 \\ \cancel{-5} & \cancel{6} & \cancel{7} & \cancel{8} \\ \cancel{2} & -2 & -2 & 1 \\ \cancel{2} & 0 & 1 & -1 \end{vmatrix} = \begin{vmatrix} 3 & 1 & 0 \\ -2 & -2 & 1 \\ 0 & 1 & -1 \end{vmatrix}$$

You might recognize this 3×3 matrix as the same as the one from the examples in Subsection 8.4.3, so we already know its determinant. Also, the $(2,1)$ entry in the original 4×4 matrix is in an "odd" position since $2+1=3$ is odd, so must flip the sign to obtain the C_{21} cofactor value from the M_{21} minor determinant value:

$$M_{21}=1, \qquad\qquad C_{21}=-M_{21}=-1.$$

□

8.4.5 Determinants of 4×4 matrices

Finally, here is a 4×4 example. We'll do one with a few zeros, so that it doesn't get too out of hand.

Example 8.4.8 Determinant of a 4×4 matrix. Consider

$$A=\begin{bmatrix} 1 & -1 & 2 & 1\\ 2 & 0 & 1 & 1\\ 0 & 1 & 0 & -3\\ 1 & -2 & -1 & 0 \end{bmatrix}.$$

Let's choose the third row, as that has two zero entries.

$$\det A=\left|\begin{array}{cccc} 1 & -1 & 2 & 1\\ 2 & 0 & 1 & 1\\ \hline \multicolumn{1}{|c}{0^+} & 1^- & 0^+ & \multicolumn{1}{c|}{-3^-}\\ \hline 1 & -2 & -1 & 0 \end{array}\right|$$

The cofactor expansion along the chosen row will involve only two 3×3 minor determinant calculations — minor determinants M_{31} and M_{33} will not be needed, since their corresponding entries are 0.

$$\det A=0\cdot M_{31}-1\cdot\begin{vmatrix} 1 & \cancel{-1} & 2 & 1\\ 2 & \cancel{0} & 1 & 1\\ \cancel{0} & \cancel{1} & \cancel{0} & \cancel{-3}\\ 1 & \cancel{-2} & -1 & 0 \end{vmatrix}+0\cdot M_{33}-(-3)\cdot\begin{vmatrix} 1 & -1 & 2 & \cancel{1}\\ 2 & 0 & 1 & \cancel{1}\\ \cancel{0} & \cancel{1} & \cancel{0} & \cancel{-3}\\ 1 & -2 & -1 & \cancel{0} \end{vmatrix}$$

Next we choose a row or column in each of the remaining minor determinants.

$$\det A=-\left|\begin{array}{cc|c|} 1 & 2 & 1^+\\ 2 & 1 & 1^-\\ 1 & -1 & 0^+ \\ \cline{3-3} \end{array}\right|+3\cdot\left|\begin{array}{ccc} 1 & -1 & 2\\ \hline \multicolumn{1}{|c}{2^-} & 0^+ & \multicolumn{1}{c|}{1^-}\\ \hline 1 & -2 & -1 \end{array}\right|$$

Notice how the cofactor signs in the chosen row/column follow the 3×3 pattern, *not* the 4×4 pattern from the original matrix.

Now expand each of these 3×3 minor determinants.

$$\det A=-\left(1\cdot\begin{vmatrix} \cancel{1} & \cancel{2} & \cancel{1}\\ 2 & 1 & \cancel{1}\\ 1 & -1 & \cancel{0} \end{vmatrix}-1\cdot\begin{vmatrix} 1 & 2 & \cancel{1}\\ \cancel{2} & \cancel{1} & \cancel{1}\\ 1 & -1 & \cancel{0} \end{vmatrix}+0\cdot M_{33}\right)$$

$$+3\cdot\left(-2\cdot\begin{vmatrix} \cancel{1} & -1 & 2\\ \cancel{2} & \cancel{0} & \cancel{1}\\ \cancel{1} & -2 & -1 \end{vmatrix}+0\cdot M_{22}-1\cdot\begin{vmatrix} 1 & -1 & \cancel{2}\\ \cancel{2} & \cancel{0} & \cancel{1}\\ 1 & -2 & \cancel{-1} \end{vmatrix}\right)$$

Now reduce to a combination of 2×2 determinants.

$$\det A = -\left(\begin{vmatrix} 2 & 1 \\ 1 & -1 \end{vmatrix} - \begin{vmatrix} 1 & 2 \\ 1 & -1 \end{vmatrix} \right) + 3 \cdot \left(-2 \cdot \begin{vmatrix} -1 & 2 \\ -2 & 1 \end{vmatrix} - \begin{vmatrix} 1 & -1 \\ 1 & -2 \end{vmatrix} \right)$$

Finally, we can apply the 2×2 criss-cross pattern as illustrated above.

$$\begin{aligned} \det A &= -\big((-2-1)-(-1-2)\big) + 3\Big(-2\big(1-(-4)\big) - \big(-2-(-1)\big)\Big) \\ &= -(-3+3) + 3(-10+1) \\ &= -27. \end{aligned}$$

□

8.5 Theory

In this section.

- Subsection 8.5.1 *Basic properties of determinants*

8.5.1 Basic properties of determinants

The following justifies our definition of the determinant as the *common* value of all cofactor expansions of a matrix.

Theorem 8.5.1 Uniformity of cofactor expansions. *Every cofactor expansion of a given square matrix, whether along a row or a column, evaluates to the same value.*

Proof. The proof of this theorem is beyond the scope of this course; instead see [2, 3]. ■

Finally, let's record the determinants of special forms of matrices we discussed in Subsection 8.3.5. However, we omit the proofs since we have already considered in detail the patterns behind the proofs in that earlier discussion.

Proposition 8.5.2 Determinants of special forms.

1. *For a matrix that is diagonal or triangular, the determinant is equal to the product of the diagonal entries.*
2. *For a scalar matrix,* $\det(kI) = k^n$.
3. $\det \mathbf{0} = 0$ *for a square zero matrix.*
4. $\det I = 1$.

CHAPTER 9

Determinants versus row operations

9.1 Discovery guide

Discovery 9.1 What is $\det A$ if A is a square matrix with a row of zeros? Explain by referring to a cofactor expansion.

Discovery 9.2 Consider the matrix

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

(a) Compute the determinant by cofactor expansion *along the first row*.

(b) Now swap the first and second rows, and compute the determinant of the new matrix by cofactor expansion *along the second row* (which will now have the entries of first row of the original matrix). Why do you think you got the answer you did?

Hint. Do you remember the cofactor sign patterns? If not, see Pattern (8.3.1).

(c) Do you think the same thing will happen if you swap the second and third rows of the original matrix?

(d) What about if you swap the first and third rows of the original matrix?

(e) What if you swap the 1st and 2nd rows of the original matrix, then swap the 2nd and 3rd rows of that matrix, and then swap the 1st and 2nd rows of that matrix? Do you want to change your answer to Task d?

(f) *Complete the rule:* If B is obtained from A by swapping two rows, then $\det B$ is related to $\det A$ by .

(g) *Complete the rule:* If E is an elementary matrix of the "swap two rows" type, then $\det E =$.

Hint. How do you create an elementary matrix?

Discovery 9.3 Suppose A is a square matrix with two identical rows. What happens to the matrix when you swap those two identical rows? According to Discovery 9.2, what is supposed to happen to the determinant when you swap rows? What can you conclude about $\det A$?

Discovery 9.4 Consider the matrix from Discovery 9.2.

(a) Multiply the first row by 7, and compute the determinant of the new matrix. Do you think the same will happen if you multiplied some other row of the matrix by 7? Explain by referring to cofactor expansions.

(b) *Complete the rule:* If B is obtained from A by multiplying one row by k, then $\det B$ is related to $\det A$ by .

(c) *Complete the rule:* If E is an elementary matrix of the "multiply a row by k" type, then $\det E =$.

Hint. How do you create an elementary matrix?

(d) Suppose you multiply the *whole* matrix by 7. What happens to the determinant in that case?

Hint. How many rows are you multiplying by 7?

(e) *Complete the rule:* For scalar k and $n \times n$ matrix A, $\det(kA) =$.

Hint. If you multiply a *whole* matrix by a scalar, you are in effect multiplying *every row* by that scalar.

Discovery 9.5 Suppose A is a square matrix where one row is equal to a multiple of another. Combine your answer to Discovery 9.3 with a rule from Discovery 9.4 to determine $\det A$.

Discovery 9.6 Consider the generic 3×3 matrix

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

Its determinant is $a_{11}C_{11} + a_{12}C_{12} + a_{13}C_{13}$.

Suppose we add k times the second row to the first:

$$\begin{bmatrix} a_{11}+ka_{21} & a_{12}+ka_{22} & a_{13}+ka_{23} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

(a) Has this row operation changed the cofactors of entries in the first row?

(b) Write out the cofactor expansion along the first row for the new matrix. Then use some algebra to express this cofactor expansion as:

$$(\text{some formula}) + k(\text{some other formula}).$$

The first "some formula" should look familiar. Can you craft a 3×3 matrix so that "some other formula" can be similarly interpreted?

(c) What is the value of the "some other formula" part from Task b?

Hint. Discovery 9.3

(d) *Complete the rule:* If B is obtained from A by adding a multiple of one row to another, then $\det B$ is related to $\det A$ by .

(e) *Complete the rule:* If E is an elementary matrix of the "add a multiple of one row to another" type, then $\det E =$.

Hint. How do you create an elementary matrix?

9.2 Concepts

In this section.

We would like to connect determinants to invertibility, and as always row operations are the way to do so.

9.2.1 Swapping rows: effect on determinant

Swapping adjacent rows. In Discovery 9.2, we first explored what happens to a determinant if we swap *adjacent* rows in a matrix, and we discovered the following. Suppose we take square matrix A and swap row i with row $i+1$, which are adjacent, obtaining new matrix A'. Compared to a cofactor expansion of $\det A$ along row i, a cofactor expansion of $\det A'$ along row $i+1$ has all the same entries and minor determinants, because the $(i+1)^{\text{th}}$ row in A' now contains the entries from the i^{th} row in A, and vice versa. However, the cofactor signs along the $(i+1)^{\text{th}}$ row are all the opposite of those along the i^{th} row. Therefore, all the terms in the cofactor expansions of $\det A$ and $\det A'$ are negatives of each other, and so $\det A' = -\det A$. We concluded that ***swapping adjacent rows changes the sign of the determinant.***

Swapping (possibly) non-adjacent rows. Now, it might seem that we might sometimes get $\det A'$ to be *equal* to $\det A$ if we swapped non-adjacent rows. In particular, if we swapped two rows that were separated by a single other row (as in Discovery 9.2.d), the two rows would have the same pattern of cofactor signs, and our thinking above might would lead to $\det A' = \det A$ in this case. However, it turns out that *any swap of rows can be achieved by an* odd *number of consecutive adjacent row swaps*, and an odd number of sign changes will have a net result of changing the sign. So **any *swap of a pair of rows changes the sign of the determinant***.

Matrices with two identical rows. In Discovery 9.3, we paused to consider a consequence of this effect of swapping rows on the determinant. Suppose a square matrix has two identical rows. If we swap those two particular rows, then from our discussion above we expect the determinant of the new matrix obtained from this operation to be the negative of the determinant of the original matrix. But if those rows are identical, then swapping them has no effect and the determinants of the new and old matrices should be equal. Since the only number that remains unchanged when its sign is changed is zero, we conclude that ***a square matrix with two (or more) identical rows has determinant* 0.**

Corresponding elementary matrices. Recall that elementary matrices are obtained from the identity by a single row operation. So if we take the identity matrix (which has determinant 1) and swap two rows to obtain the elementary matrix that corresponds to that operation, then that elementary matrix must have determinant -1.

9.2.2 Multiplying rows: effect on determinant

Multiplying a single row by a scalar. In Discovery 9.4, we first explored what happens to a determinant if we multiply a *single* row in a matrix, and we discovered the following. Suppose we take square matrix A and multiply row i by the constant k, obtaining new matrix A'. Compared to a cofactor expansion of $\det A$ along row i, a cofactor expansion of $\det A'$ along row i has all the same minor determinants, because the entries in all the other rows are still the same as in A. However, when we add up all the "entry times cofactor" terms in a cofactor expansion of $\det A'$ along row i, there is the new common factor of k from the scaled entries of that row. If we factor that common k out, we are left with exactly the cofactor expansion of $\det A$ along row i. Hence, ***multiplying a* single *row in a matrix scales the determinant by the same factor***.

Multiplying a *whole* matrix by a scalar. In Discovery 9.4.d and Discovery 9.4.e, we also considered what happens if we multiply a *whole* matrix by a constant. But scalar multiplying a matrix is the same as multiplying *every* row by that scalar. If multiplying a *single* row by k changes the determinant by a factor of k, then multiplying *every* row by k must change the determinant by n factors of k, where n is the size of the matrix (and hence the number of rows). That is, for a square $n \times n$ matrix A and a scalar k, we have $\det(kA) = k^n \det A$.

Warning 9.2.1 It is very common for students to forget this lesson and *incorrectly* remember the formula as $\det(kA)$ being equal to $k \det A$, just because that "looks" correct. Don't be one of those students!

Matrices with proportional rows. Let's pause again to consider a consequence of this effect of multiplying a row by a constant on the determinant. Suppose A is a matrix where one row is equal to a multiple (by k, say) of another row, as in Discovery 9.5. We can multiply that row by $1/k$ to obtain matrix A' with determinant $\det A' = (1/k) \det A$. But now A' has two identical rows, so $\det A' = 0$, which forces $\det A = 0$. So we can extend our fact about matrices with some identical rows to matrices with some *proportional* rows: ***a matrix with two (or more) proportional rows has determinant*** 0.

Corresponding elementary matrices. Again, let's consider elementary matrices corresponding to this type of operation. If we take the identity matrix (which has determinant 1) and multiply a row by a nonzero constant k to obtain the elementary matrix that corresponds to that operation, then that elementary matrix must have determinant $k \cdot 1 = k$.

9.2.3 Combining rows: effect on determinant

Now we move to the operation of adding a multiple of one row to another, explored in Discovery 9.6. This is the most complicated of the three operations, so we will just consider the 3×3 case, as in the referenced discovery activity. Consider the general 3×3 matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

As in the discovery activity, suppose we add k times the second row to the first, to get

$$A' = \begin{bmatrix} a_{11}+ka_{21} & a_{12}+ka_{22} & a_{13}+ka_{23} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

The cofactors, C_{11}, C_{12}, C_{13}, along the first row of A' are exactly the same as the cofactors along the first row in A, since calculating those cofactors only involve entries in the second and third rows, which have not changed. If we do a cofactor expansion of $\det A'$ along the first row, we get

$$\begin{aligned} \det A' &= (a_{11}+ka_{21})C_{11} + (a_{12}+ka_{22})C_{12} + (a_{13}+ka_{23})C_{13} \\ &= a_{11}C_{11} + ka_{21}C_{11} + a_{12}C_{12} + ka_{22}C_{12} + a_{13}C_{13} + ka_{23}C_{13} \\ &= (a_{11}C_{11} + a_{12}C_{12} + a_{13}C_{13}) + k(a_{21}C_{11} + a_{22}C_{12} + a_{23}C_{13}) \\ &= (\det A) + k(a_{21}C_{11} + a_{22}C_{12} + a_{23}C_{13}), \end{aligned}$$

In the second term of the last line, we have sort of a "mixed" cofactor expansion for A, where the entries are from the second row but the cofactors are from the first row. This mixed expansion is definitely not equal to $\det A$ or $\det A'$, but could it be the determinant of some new matrix A''? To have the same first-row cofactors as A, this new A'' matrix would have to have the same second and third rows as A, since those entries are what are used to calculate the first-row cofactors. If we also repeat the second row entries from A in the first row of A'', so that

$$A'' = \begin{bmatrix} a_{21} & a_{22} & a_{23} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

then a cofactor expansion of $\det A''$ along the first row gives us exactly the "mixed" cofactor expansion in the second term of our last expression for $\det A'$ above. However, A'' has two identical rows, so its determinant is 0. We can now continue our calculation from above:

$$\begin{aligned} \det A' &= \det A + k(a_{21}C_{11} + a_{22}C_{12} + a_{23}C_{13}) \\ &= \det A + k(\det A'') \\ &= \det A + k \cdot 0 \\ &= \det A. \end{aligned}$$

This result is fairly surprising: while the two simpler row operations affected the determinant, ***the elementary row operation combining rows has no effect at all on the determinant***.

A look ahead. Recall Goal 8.3.1 from last chapter: for a given matrix B we are trying to determine a scalar δ and a special matrix B' so that $BB' = \delta I$. (The scalar δ will end up being $\det B$). We will build the special matrix B' in the next chapter, but we will need to remember the discovery we have made here that ***a "mixed" cofactor expansion always evaluates to*** 0.

Corresponding elementary matrices. Just as with the other two row operations, we can apply what we've learned to elementary matrices. If we take the identity matrix and add a multiple of one row to another to obtain the elementary matrix that corresponds to that operation, then that elementary matrix must have the same determinant as the identity, which is 1.

9.2.4 Column operations and the transpose

You could imagine that an alien civilization might also develop the theory of linear algebra, but perhaps with some cosmetic differences. Perhaps they prefer to write their equations vertically, and so when they convert equations to augmented matrices, a *column* represents an equation and a *row* contains the coefficients in each equation for a particular variable. In essence, their matrix theory is the *transpose* of ours. They would then proceed to "column reduce" matrices in order to solve the underlying system, instead of row reducing. Since determinants can be computed by cofactor expansions along either rows or columns and yield the same result, and since the cofactor sign patterns we determined in Pattern (8.3.1) are symmetric in the main diagonal (i.e. the pattern is unchanged by a transpose), this alien development of linear algebra would discover all the same things about the relationships between *column* operations and the determinant as we have about *row* operations and the determinant. We have recorded all these facts about column operations in Subsection 9.4.1, alongside the corresponding facts about row operations. And there is one more fact about the transpose, which is the bridge between our matrix theory and the alien matrix theory: ***a transpose has no effect on the determinant***. You can easily see why this would be true, since a cofactor expansion along a *column* in A^{T} would work out the same as a cofactor expansion along a *row* in A.

9.2.5 Determinants by row reduction

The relationships between row operations and the determinant that we have explored in Discovery guide 9.1 and described above provide us with another method of computing determinants. An REF for a square matrix must always be upper triangular, since the leading ones must be either on or to the right of the main diagonal. So when row reducing there is always a point where we reach an upper triangular matrix. And from Statement 1 of Proposition 8.5.2 we know that determinants of upper triangular matrices are particularly easy to compute. So starting with any square matrix, we can row reduce to upper triangular, keeping track of how the determinant has changed at each step, and then work backwards from the determinant of the upper triangular matrix to determine the determinant of the original matrix. We'll save doing an example for Subsection 9.3.1.

Warning 9.2.2 When using this method, it is really important to stick to *elementary* row operations. In learning to row reduce, you may have discovered that you can perform operations of the kind $R_i \to kR_i + mR_j$ and still get the correct set of solutions to the corresponding system. However, this kind of operation is *not* elementary — it is actually a combination of *two* elementary operations performed at once, and *will* change the determinant. It's best just to avoid operations of this kind for determinant calculations.

9.3 Examples

In this section.

9.3.1 Determinants by row reduction

As discussed in Warning 8.3.3, determinants by cofactor expansions are extremely inefficient for matrices larger than 3×3. Here we provide an example of using the row reduction method to compute a determinant.

Example 9.3.1 Using row reduction to compute a determinant. Let's recompute the determinant of

$$A = \begin{bmatrix} 1 & -1 & 2 & 1 \\ 2 & 0 & 1 & 1 \\ 0 & 1 & 0 & -3 \\ 1 & -2 & -1 & 0 \end{bmatrix},$$

the same matrix from Example 8.4.8.

First, let's row reduce. For the purposes of describing our thinking in using the matrix reduction calculation to determine the determinant of A, we'll label our matrices as we go.

$$A = \begin{bmatrix} 1 & -1 & 2 & 1 \\ 2 & 0 & 1 & 1 \\ 0 & 1 & 0 & -3 \\ 1 & -2 & -1 & 0 \end{bmatrix} \begin{matrix} \\ R_2 - 2R_1 \\ \\ R_4 - R_1 \end{matrix} \longrightarrow A_1 = \begin{bmatrix} 1 & -1 & 2 & 1 \\ 0 & 2 & -3 & -1 \\ 0 & 1 & 0 & -3 \\ 0 & -1 & -3 & -1 \end{bmatrix} \; R_2 \leftrightarrow R_3$$

$$\longrightarrow A_2 = \begin{bmatrix} 1 & -1 & 2 & 1 \\ 0 & 1 & 0 & -3 \\ 0 & 2 & -3 & -1 \\ 0 & -1 & -3 & -1 \end{bmatrix} \begin{matrix} \\ \\ R_3 - 2R_2 \\ R_4 + R_2 \end{matrix} \longrightarrow A_3 = \begin{bmatrix} 1 & -1 & 2 & 1 \\ 0 & 1 & 0 & -3 \\ 0 & 0 & -3 & 5 \\ 0 & 0 & -3 & -4 \end{bmatrix} \begin{matrix} \\ \\ -\frac{1}{3}R_3 \\ \\ \end{matrix}$$

$$\longrightarrow A_4 = \begin{bmatrix} 1 & -1 & 2 & 1 \\ 0 & 1 & 0 & -3 \\ 0 & 0 & 1 & -\frac{5}{3} \\ 0 & 0 & -3 & -4 \end{bmatrix} \begin{matrix} \\ \\ \\ R_4 + 3R_3 \end{matrix} \longrightarrow A_5 = \begin{bmatrix} 1 & -1 & 2 & 1 \\ 0 & 1 & 0 & -3 \\ 0 & 0 & 1 & -\frac{5}{3} \\ 0 & 0 & 0 & -9 \end{bmatrix}$$

We would need one more operation to get to REF, but we are already at upper triangular so we don't need to bother. And notice that we didn't bother clearing entries *above* leading ones, since our goal was to get to an upper triangular matrix, which only requires entries *below* leading ones to be cleared.

Now we'll work backwards to determine $\det A$.

A_5 This last matrix is upper triangular, so its determinant is equal to the product of the diagonal entries: $\det A_5 = 1 \cdot 1 \cdot 1 \cdot (-9) = -9$.

A_4 Matrix A_5 was produced from A_4 by an operation that does not change the determinant, so $\det A_4$ must be -9 as well.

A_3 Matrix A_4 was produced from A_3 by multiplying a row, so $\det A_4 = -\frac{1}{3} \det A_3$. Solving for $\det A_3$, we get $\det A_3 = -3 \cdot (-9) = 27$.

A_2 Matrix A_3 was produced from A_2 by an operation that does not change the determinant, so $\det A_2$ must be 27 as well.

A_1 Matrix A_2 was produced from A_1 by swapping rows, so these two determinants have opposite signs. Thus, $\det A_1 = -27$.

A Matrix A_1 was produced from A by a pair of operations, neither of which changes the determinant, so finally we have $\det A = -27$.

This analysis agrees with the calculation of $\det A$ by cofactor expansion in Example 8.4.8. □

9.3.2 Matrices of determinant zero

Example 9.3.2 Recognizing $\det A = 0$. Here are a few examples of recognizing matrices that have determinant 0.

$$\text{(i)}\ \begin{bmatrix} 1 & -1 & 2 & 1 \\ 0 & 1 & 0 & -3 \\ 0 & 1 & 0 & -3 \\ 1 & 2 & 3 & 4 \end{bmatrix} \qquad \text{(iii)}\ \begin{bmatrix} 1 & -1 & 0 & 1 \\ 4 & 1 & 0 & -3 \\ -1 & 1 & 0 & -3 \\ 1 & 2 & 0 & 4 \end{bmatrix}$$

$$\text{(ii)}\ \begin{bmatrix} 1 & -1 & 2 & 1 \\ 0 & 1 & 5 & 5 \\ 7 & 1 & 0 & -3 \\ -2 & 2 & -4 & -2 \end{bmatrix} \qquad \text{(iv)}\ \begin{bmatrix} 1 & -1 & 2 & -1 \\ 0 & 1 & 5 & 1 \\ 7 & 1 & 0 & 1 \\ -2 & 2 & -4 & 2 \end{bmatrix}$$

The first matrix has two identical rows, the second matrix has two proportional rows ($R_4 = -2R_1$), the third matrix has a column of zeros, and the fourth matrix has two identical columns. So the determinant of each of these matrices is 0. □

9.4 Theory

In this section.

Here we will recap all of the facts we discussed in Section 9.2, as well as add in a fact from Discovery 9.1. We have already adequately discussed the ideas behind most of these facts, so for most of them we will not include a proof.

9.4.1 Effect of row operations on the determinant

We begin by recording a fact that helped us in our exploration of the effect of swapping rows on the determinant.

Lemma 9.4.1 *Any row swap can be achieved by a sequence of an* odd *number of adjacent row swaps.*

Proof idea. Suppose you want to swap rows R and R' in a matrix using only adjacent row swaps, where R appears higher in the matrix than R', and they are separated by m other rows. First move R down, one adjacent row swap at a time, until it is in the position just above R'. Then swap R and R', which are now adjacent. Finally, move R' up, one adjacent row swap at a time, until it is in the original position of R. Count the number of adjacent swaps that have been made as an expression in m, and notice that it is odd. ■

Here are all the things we learned in Discovery guide 9.1.

Proposition 9.4.2 Determinants versus row operations. *The following are true for every square matrix.*

1. *If there is a row of zeros, then the determinant is* 0.
2. *If two rows are swapped, then*

$$\det(\textit{new matrix}) = -\det(\textit{old matrix}).$$

3. *If there are two identical rows, then the determinant is* 0.

4. *If a row is multiplied by constant k, then*

$$\det(\textit{new matrix}) = k \det(\textit{old matrix}).$$

5. *If a whole matrix A is scalar multiplied by a constant k, then* $\det(kA) = k^n \det A$*, where n is the size of the matrix.*

6. *If there are two proportional rows, then the determinant is* 0.

7. *If a multiple of one row is added to another row, then*

$$\det(\textit{new matrix}) = \det(\textit{old matrix}).$$

And here are our connections between rows and columns with respect to the determinant.

Lemma 9.4.3 Determinant of a transpose. *For every square matrix A,* $\det(A^{\mathrm{T}}) = \det A$.

Proposition 9.4.4 Determinants versus column operations. *The statements of Proposition 9.4.2 remain true when every instance of the word "row" is replaced by the word "column."*

9.4.2 Determinants of elementary matrices

Finally, we'll record our discoveries about the determinants of elementary matrices.

Proposition 9.4.5

1. *An elementary matrix corresponding to swapping rows has determinant* -1.

2. *An elementary matrix corresponding to multiplying a row by a constant k has determinant k.*

3. *An elementary matrix corresponding to adding a multiple of one row to another has determinant* 1.

CHAPTER 10

Determinants, the adjoint, and inverses

10.1 Discovery guide

Reminder.

The effects of the elementary row operations on the determinant are:

swapping rows
det(new) = − det(old);

multiplying a row by constant k
det(new) = k det(old);

adding a multiple of one row to another
det(new) = det(old).

Discovery 10.1 Consider the general 3×3 matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

Each entry a_{ij} has a corresponding cofactor C_{ij}, creating a **matrix of cofactors**

$$C_A = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix}.$$

The ***transpose*** of this matrix is called the (classical) **adjoint** of A.

(a) Write out the $(1,1)$ entry of the product AC_A^{T} as a formula in the entries of A and C_A. Does the result look familiar?

(b) What do you think the other diagonal entries of AC_A^{T} are?

(c) Write out the $(1,2)$ entry of the product AC_A^{T} as a formula in the entries of A and C_A. Does the result look familiar? What did we discover about "mixed" cofactor expansions in Discovery 9.6 and Subsection 9.2.3?

(d) What do you think the other non-diagonal entries of AC_A^{T} are?

(e) Refer back to Discovery 8.1. Have we finally achieved Goal 8.3.1?

Discovery 10.2

(a) Suppose $\det A = 0$. If you apply some elementary row operation to A, what is the determinant of the new matrix? (Consider each of the three kinds of operations.)

(b) If $\det A = 0$ and you perform a whole sequence of row operations to A, what is the determinant of the last matrix in the sequence?

(c) Recall that if A is invertible, then it can be row reduced to I (Theorem 6.5.2). If $\det A = 0$, could A be invertible?

Hint. Use your answer to Task b.

(d) Conversely, if A is invertible, could $\det A = 0$?

Hint. No need to think about row reducing — combine your answer to Task c with some logical thinking.

Discovery 10.3

(a) Suppose $\det A \neq 0$. Is there any elementary row operation you can apply to A so that the new matrix has determinant 0? (Consider each of the three kinds of operations.)

(b) If $\det A \neq 0$ and you perform a whole sequence of row operations to A, could the last matrix in the sequence have determinant 0?

(c) Recall that if a matrix is singular (that is, not invertible), then it is *not* possible to row reduce it to I (Theorem 6.5.2), and so its RREF must have a row of zeros. If $\det A \neq 0$, could A be singular?

Hint. Use your answer to Task b.

(d) Conversely, if A is singular, is $\det A \neq 0$ possible?

Hint. No need to think about row reducing — combine your answer to Task c with some logical thinking.

Discovery 10.4 Recall that for matrix A and *elementary* matrix E, the result of EA is the same as the result of performing an elementary row operation on A (namely, the operation corresponding to E). Verify the formula

$$\det(EA) = (\det E)(\det A) \tag{$*$}$$

for *each* of the three types of elementary matrices E (assuming A to be a square matrix of the same size as E).

Helpful notes.

- To verify a formula, consider LHS and RHS *separately*, and argue that they equal the same value. Do *not* work with the proposed equality directly, since you don't *know* it's an equality yet.
- Do *not* just use examples; think abstractly instead.
- For each type of E, on the LHS consider the product of *matrices* EA and how its determinant compares to $\det A$ using the rules for how row operations affect determinant (explored in Discovery guide 9.1, and recalled for you at the top of this activity section). For this, think of $\det A = \det(\text{old})$ and $\det(EA) = \det(\text{new})$. Then, on the RHS, consider the value of $\det E$ and the corresponding product of *numbers* $(\det E)(\det A)$.

Discovery 10.5 In this activity, we will verify the general formula

$$\det(MN) = (\det M)(\det N) \tag{$**$}$$

in the case that M is invertible (assuming M and N to be square matrices of the same size).

(a) Recall that every invertible matrix can be expressed as a product of elementary matrices (Theorem 6.5.2). For now, suppose that M (which we have assumed invertible) can be expressed as a product of *three* elementary matrices, say $M = E_1E_2E_3$. Use formula ($*$) to verify that

$$\det(E_1E_2E_3N) = (\det E_1)(\det E_2)(\det E_3)(\det N).$$

Hint. Start with the LHS and apply formula ($*$) one step at a time. In applying formula ($*$), what are you using for E and for A at each step?

(b) Now use formula ($*$) to verify that

$$(\det E_1)(\det E_2)(\det E_3)(\det N) = \big(\det(E_1E_2E_3)\big)(\det N).$$

(c) Make sure you understand why parts (a) and (b) together verify formula ($**$) for this M.

(d) Do you think the calculations in this activity would work out similarly no matter how many E_i's are required to express M as a product of elementary matrices?

Discovery 10.6 If matrix A is invertible, by definition this means that $AA^{-1} = I$ (as well as $A^{-1}A = I$).

(a) Determine the value of $\det(AA^{-1})$ from the equality $AA^{-1} = I$.

(b) Starting with your answer to Task a, use formula ($**$) from Discovery 10.5 to obtain a formula for $\det(A^{-1})$ in terms of $\det A$.

Recall. A fraction 1/*A* does not make sense for matrices. However, $\det A$ is just a number, so you can do all the normal algebra you would like with it!

Discovery 10.7 In this discovery activity, we extend formula ($**$) to also be valid in case that M is singular (assuming M and N to be square matrices of the same size).

Recall that if M is singular (i.e. not invertible), then every product MN is singular (Statement 1 of Proposition 6.5.8).

Combine this with your answer to Discovery 10.3.d to verify formula ($**$) in the case that M is singular.

Reminder. To verify a formula, consider LHS and RHS *separately*, and argue that they equal the same value. Do *not* work with the equality directly, since you don't *know* it's an equality yet. Do *not* just use examples; think abstractly instead.

10.2 Terminology and notation

The following definitions apply to a square matrix A.

matrix of cofactors
the matrix obtained by replacing all the entries of A with the corresponding cofactors of A, denoted C_A

(classical) adjoint matrix
the transpose of the matrix of cofactors of A, denoted $\operatorname{adj} A$

10.3 Concepts

In this section.

Recall that in Section 8.3 we set a goal for ourselves: given a square matrix A, determine a matrix A' so that AA' is a scalar multiple of the identity (Goal 8.3.1). The adjoint finally fulfills this goal.

10.3.1 The classical adjoint

Before we dive in, a note about the adjective "classical." In a second course in linear algebra, you will probably learn that square matrices have a different kind of "adjoint" matrix that is completely unrelated to the adjoint we will discuss here. (The word "adjoint" gets used a lot in mathematics for many different concepts.) So we are attaching the adjective "classical" to the adjoint matrix we define here to distinguish it from that other one.

Terminology. Actually, a better adjective might be "algebraic" for this version of **adjoint matrix**, as that other kind of adjoint matrix could reasonably be called the "geometric" adjoint matrix.

Let's remind ourselves how determinants are defined, by cofactor expansions. For matrix $A = \left[a_{ij}\right]$, the cofactor expansion of $\det A$ along row i is

$$\det A = a_{i1}C_{i1} + a_{i2}C_{i2} \cdots + a_{in}C_{in},$$

where the C_{ij} are the associated cofactors. This pattern of a sum of products sure looks like matrix multiplication, where we are multiplying the i^{th} row of A against a column of some matrix. Since each position in A has a corresponding cofactor, we can create a **matrix of cofactors** $C_A = \left[C_{ij}\right]$. Except the pattern of indices for the C_{ij} in the cofactor expansion above progresses along a *row* of this cofactor matrix, whereas when we multiply matrices we multiply rows against *columns*. However, we know a way to turn rows into columns — the transpose. We call the transpose of the matrix of cofactors the **(classical) adjoint of** A, and write $\operatorname{adj} A$ to mean C_A^{T}.

In Discovery 10.1, we explored what happens when we multiply out A times $\operatorname{adj} A$. We only worked with the 3×3 case, but the same patterns would emerge for any size matrix. Remember that in a product like $A(\operatorname{adj} A)$ we get the $(i,j)^{\text{th}}$ entry by multiplying the i^{th} row of the first matrix against the j^{th} column of the second matrix. Since the second matrix is a transpose, its j^{th} column will be the j^{th} *row* of the matrix of cofactors C_A. Thus, for each diagonal entry (that is, where $j = i$), we will be multiplying a row of A against the corresponding row of cofactors, and we'll get the value of $\det A$ repeated down the diagonal of $A(\operatorname{adj} A)$. On the other hand, for an off-diagonal entry (that is, where $j \neq i$), we'll get a row of A multiplied against the cofactors associated to a *different* row. In our analysis of the operation of combining rows in Subsection 9.2.3, we determined that *a "mixed" cofactor expansion always evaluates to* 0. So all off-diagonal entries of $A(\operatorname{adj} A)$ are 0, and this product matrix is diagonal. Moreover, since the same value $\det A$ is repeated down the diagonal, this product matrix is in fact scalar:

$$A(\operatorname{adj} A) = (\det A)I.$$

As mentioned at the start of this section, this fulfills Goal 8.3.1, with $\delta = \det A$ and $A' = \operatorname{adj} A$. In particular, this gives us a formula for the inverse of any matrix that has nonzero determinant:

$$A^{-1} = \frac{1}{\det A} \operatorname{adj} A.$$

Remark 10.3.1 Just as cofactor expansions are an inefficient means to compute determinants, calculating an inverse using the adjoint formula above is very inefficient, since computing an adjoint for an $n \times n$ matrix involves computing n^2 determinants of $(n-1)\times(n-1)$ matrices. You are much better off computing an inverse by row reducing, as in Subsection 6.3.5 and Subsection 6.4.3. However, the above formula is useful for further developing the theory of solving systems by inverses, as we will soon see.

10.3.2 Determinants determine invertibility

Part of our motivation for developing determinants was to make sense of the $ad - bc$ formula that determines whether a 2×2 matrix is invertible, and obtain a similar formula for larger square matrices. In completing Goal 8.3.1 by obtaining the formula $A(\operatorname{adj} A) = (\det A)I$, we learn that whenever $\det A \neq 0$ then $A\left[(\det A)^{-1}(\operatorname{adj} A)\right] = I$, and so A is invertible (Proposition 6.5.6).

To repeat, we now know that ***if*** $\det A \neq 0$***, then*** A ***must be invertible***. Logically, that raises three related questions.

Question 10.3.2

- If A is invertible, must $\det A$ be nonzero?
- If $\det A = 0$, must A be singular?
- If A is singular, must $\det A$ be zero?

□

In the study of logic, the statement version of these three questions are called the **converse**, **inverse**, and **contrapositive**, respectively, of the original **conditional statement** that states:

> If $\det A \neq 0$, then A is invertible.

And the study of logic tells us that *the answers to these three questions are* not *necessarily all affirmative just because the original statement is true*. So in Discovery 10.2 and Discovery 10.3 we considered these questions, as well as the original statement, by considering the effects of row reducing on the determinant. Here is what we discovered, in the order we considered them in those two discovery activities, relying on our knowledge that a square matrix is invertible if and only if its RREF is the identity matrix (Theorem 6.5.2)

If $\det A = 0$*.* Since no elementary row operation can change a zero determinant to a nonzero one, the RREF of A must also have determinant 0. But then the RREF of A cannot be I, since $\det I = 1$. So A is not invertible.

If A is invertible. Then $\det A$ cannot be zero, since then A wouldn't be invertible, as we just argued in the previous point.

If $\det A$ *is nonzero.* Since no elementary row operation can change a nonzero determinant to a zero determinant (multiplying a row by 0 is not an elementary operation), the RREF for A must also have nonzero determinant. But then that RREF cannot have a row of zeros, because then its determinant would be 0. Since it is square, that RREF matrix must have all of its leading ones, making it the identity matrix, and so A is invertible.

If A is singular. Then $\det A$ must be zero, since if it were nonzero then A would be invertible, as we just argued in the previous point.

10.3.3 Determinants versus matrix multiplication: case of elementary matrices

In Discovery 10.4, we considered $\det(EA)$ for E an elementary matrix and A a square matrix. Since there are three different kinds of elementary matrices, we had three different cases to consider. In each case, we were able to combine the appropriate part of Proposition 9.4.2 on the one hand with the appropriate part of Proposition 9.4.5 on the other, in order to verify

$$\det(EA) = (\det E)(\det A) \tag{$*$}$$

is true in all cases of the type of elementary matrix E. (For the details of these three cases, see the proof for Lemma 10.5.5, which appears in Subsection 10.5.3.)

Expressed in words, the equality above represents the pattern that ***a determinant of a product is the product of the determinants***, at least in the case where the first matrix in the product is elementary (for now).

10.3.4 Determinants versus matrix multiplication: invertible case

In Discovery 10.5, we progressed to considering determinants of a product of matrices where the first matrix in the product is invertible. In particular, this means that the first matrix can be expressed somehow as a product of elementary matrices (Theorem 6.5.2), and so we can unravel the determinant of this product one elementary matrix at a time, using the result of the previous subsection at each step.

As in Discovery 10.5, consider matrix N and invertible matrix M, where M can be expressed as a product of *three* elementary matrices, $M = E_1E_2E_3$. The we can repeatedly use our rule $(*)$ from the elementary matrix case in Subsection 10.3.3 above to obtain

$$\begin{aligned}
\det(MN) &= \det(E_1E_2E_3N) \\
&= (\det E_1)\big(\det(E_2E_3N)\big) && \text{(i)} \\
&= (\det E_1)(\det E_2)\big(\det(E_3N)\big) && \text{(ii)} \\
&= (\det E_1)(\det E_2)(\det E_3)(\det N) && \text{(iii)} \\
&= (\det E_1)\big(\det(E_2E_3)\big)(\det N) && \text{(iv)} \\
&= \big(\det(E_1E_2E_3)\big)(\det N) && \text{(v)} \\
&= (\det M)(\det N),
\end{aligned}$$

with justifications

(i) apply rule $(*)$ with $E = E_1$ and $A = E_2E_3N$;

(ii) apply rule $(*)$ with $E = E_2$ and $A = E_3N$;

(iii) apply rule $(*)$ with $E = E_3$ and $A = N$;

(iv) apply rule $(*)$ with $E = E_2$ and $A = E_3$; and

(v) apply rule $(*)$ with $E = E_1$ and $A = E_2E_3$.

Of course, this sort of calculation could be repeated no matter how many elementary matrices went into a product expression for M. So we can make our final statement of the last subsection a little stronger: ***a determinant of a product is the product of the determinants***, at least in the case where the first matrix in the product is invertible (for now).

10.3.5 Determinants versus matrix multiplication: singular case

Finally, in Discovery 10.7 we considered the determinant of a product of matrices where the first matrix in the product is singular. It is fairly straightforward to verify that again, in this case, ***a determinant of a product is the product of the determinants*** whenever the first matrix in the product is singular. (See the proof of the singular case for Statement 1 of Proposition 10.5.6, which will appear in Subsection 10.5.3.)

10.3.6 Determinants versus matrix multiplication: all cases

The considerations in Subsection 10.3.4 and Subsection 10.3.5 together verify the universal pattern

$$\det(MN) = (\det M)(\det N)$$

for square matrices M and N of the same size, no matter whether M is invertible or singular, and so the pattern that ***a determinant of a product is the product of the determinants*** is true in *all* cases. In more sophisticated mathematical language, we say that the determinant function is **multiplicative**.

10.3.7 Determinant of an inverse

In Discovery 10.6, we used the fact that the determinant is multiplicative to investigate the relationship between the determinants of an invertible matrix and its inverse. By definition of inverse, we have $AA^{-1} = I$. Since the product AA^{-1} is the same matrix as the identity, it must have the same determinant, so

$$\det(AA^{-1}) = 1$$

(Statement 4 of Proposition 8.5.2).

As well, we know that $\det A \neq 0$, since A is invertible. So,

$$\begin{aligned} \det(AA^{-1}) &= 1 \\ (\det A)\big(\det(A^{-1})\big) &= 1 \\ \det(A^{-1}) &= \frac{1}{\det A}. \end{aligned}$$

Thus, ***the determinant of an inverse is the inverse of the determinant***.

Careful. Remember that we never write fractions or reciprocals of matrices. However, $\det A$ is not a matrix — it is a *number* that we are assuming is nonzero in this case, so we are justified in writing and using its reciprocal in these calculations.

10.3.8 Cramer's rule

While the adjoint inversion formula is not a good choice for computing inverses, it does have applications. Here is one application to solving systems. Remember that if $A\mathbf{x} = \mathbf{b}$ is a linear system with a square, invertible coefficient matrix A, then there is one unique solution $\mathbf{x} = A^{-1}\mathbf{b}$. Using the adjoint inversion formula, we get

$$\mathbf{x} = A^{-1}\mathbf{b} = \frac{1}{\det A}(\operatorname{adj} A)\mathbf{b}. \tag{$**$}$$

As usual, let's consider this solution formula in the case that A is 3 × 3, in which case both $\mathbf{x}$ and $\mathbf{b}$ are 3 × 1:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

The product $(\operatorname{adj} A)\mathbf{b}$ will be a column matrix, whose top entry is obtained by multiplying the first row of $\operatorname{adj} A$ against the column $\mathbf{b}$. But $\operatorname{adj} A$ is the transpose of the matrix of cofactors for A, so the first row of $\operatorname{adj} A$ contains the cofactors from the first *column* of A, and we have

$$\big[(\operatorname{adj} A)\mathbf{b}\big]_{11} = C_{11}b_1 + C_{21}b_2 + C_{31}b_3. \tag{$***$}$$

This looks like a cofactor expansion of some determinant! The cofactors are from the first column of A, so their values only depend on the *second* and *third* columns of A. But the entries are from $\mathbf{b}$, so if we replace the first column in A with $\mathbf{b}$ to get a new matrix

$$A_1 = \begin{bmatrix} | & | & | \\ \mathbf{b} & \mathbf{a}_2 & \mathbf{a}_3 \\ | & | & | \end{bmatrix},$$

then a cofactor expansion of $\det A_1$ along the first column gives us exactly the expression in (∗∗∗) for the first entry in the product $(\operatorname{adj} A)\mathbf{b}$. Using this in (∗∗), and considering only the top entry on both sides, we get

$$x_1 = \frac{1}{\det A}\big[(\operatorname{adj} A)\mathbf{b}\big]_{11} = \frac{\det A_1}{\det A}.$$

Similar calculations would tell us that $x_2 = (\det A_2)/(\det A)$, where A_2 is the matrix obtained by replacing the second column of A by $\mathbf{b}$, and similarly for x_3. And the same pattern emerges for larger systems.

We will work out an example of using Cramer's rule in Subsection 10.4.3.

10.4 Examples

In this section.

- Subsection 10.4.1 *The* 2×2 *case*
- Subsection 10.4.2 *Computing an inverse using the adjoint*
- Subsection 10.4.3 *Cramer's rule*

10.4.1 The 2×2 case

Let's compute the adjoint of the general 2×2 matrix $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$. First, the minors.

$$M_{11} = \begin{vmatrix} \cancel{a} & \cancel{b} \\ \cancel{c} & d \end{vmatrix} = d \qquad M_{12} = \begin{vmatrix} \cancel{a} & \cancel{b} \\ c & \cancel{d} \end{vmatrix} = c$$

$$M_{21} = \begin{vmatrix} \cancel{a} & b \\ \cancel{c} & \cancel{d} \end{vmatrix} = b \qquad M_{22} = \begin{vmatrix} a & \cancel{b} \\ \cancel{c} & \cancel{d} \end{vmatrix} = a$$

In the matrix of cofactors for a 2×2 matrix, the off-diagonal cofactors become negative, and then the adjoint is the transpose of that.

$$C_A = \begin{bmatrix} d & -c \\ -b & a \end{bmatrix} \qquad\qquad \operatorname{adj} A = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Compare. Look back at A_{mix} and its transpose from Discovery 8.1 and compare with this general 2×2 adjoint.

The inverse of A is then the reciprocal of the determinant times the adjoint, so that

$$A^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix},$$

as promised in Proposition 5.5.4.

10.4.2 Computing an inverse using the adjoint

As mentioned, using the adjoint to compute the inverse of a matrix is not very efficient for matrices larger than 2×2. In most cases, you are better off using the row reduction method. However, there *are* situations where you might want to use the adjoint instead, as in the example below.

Example 10.4.1 Using the adjoint to compute an inverse. It can be tedious to row reduce a matrix with variable entries. Consider

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

To row reduce X, our first step would be to obtain a leading one in the first column. We might choose to perform $R_1 \to \frac{1}{x}R_1$, except that this operation would be invalid in the case that $x = 0$. Or we might choose to perform $R_2 \to \frac{1}{x-1}R_2$, except that this operation would be invalid in the case that $x = 1$. So to row reduce X we would need to consider three different cases, $x = 0$, $x = 1$, and $x \neq 0, 1$, performing different row reduction sequences in each of these cases. And when we get to the point of trying to obtain a leading one in the second column, we might discover there are even more cases to consider.

So instead we will attempt to compute the inverse of X using the adjoint. First, the minors.

$$M_{11} = \begin{vmatrix} 0 & x \\ x & 1 \end{vmatrix} = -x^2 \qquad M_{12} = \begin{vmatrix} x-1 & x \\ 0 & 1 \end{vmatrix} = x-1 \qquad M_{13} = \begin{vmatrix} x-1 & 0 \\ 0 & x \end{vmatrix} = x^2 - x$$

$$M_{21} = \begin{vmatrix} 1 & -1 \\ x & 1 \end{vmatrix} = 1+x \qquad M_{22} = \begin{vmatrix} x & -1 \\ 0 & 1 \end{vmatrix} = x \qquad M_{23} = \begin{vmatrix} x & 1 \\ 0 & x \end{vmatrix} = x^2$$

$$M_{31} = \begin{vmatrix} 1 & -1 \\ 0 & x \end{vmatrix} = x \qquad M_{32} = \begin{vmatrix} x & -1 \\ x-1 & x \end{vmatrix} = x^2 + x - 1 \qquad M_{33} = \begin{vmatrix} x & 1 \\ x-1 & 0 \end{vmatrix} = 1 - x$$

We obtain the matrix of cofactors by making certain minor determinants negative, according to the 3×3 pattern of cofactor signs, and then the adjoint is the transpose.

$$C_X = \begin{bmatrix} -x^2 & 1-x & x^2-x \\ -1-x & x & -x^2 \\ x & -x^2-x+1 & 1-x \end{bmatrix}$$

$$\operatorname{adj} X = \begin{bmatrix} -x^2 & -1-x & x \\ 1-x & x & -x^2-x+1 \\ x^2-x & -x^2 & 1-x \end{bmatrix}$$

To compute the inverse of X, we still need its determinant. But we already have all the cofactors, so a cofactor expansion will be easy. Let's do a cofactor expansion of $\det X$ along the third row. (Remember that the cofactors already have the appropriate signs, so we are just summing "entry times cofactor" terms.)

$$\det X = 0x + x(-x^2 - x + 1) + 1(1 - x) = 1 - x^3 - x^2$$

Notice. This determinant is not zero for either $x = 0$ or $x = 1$, so we really *would* have had to consider all those different cases if we chose to compute X^{-1} by row reducing.

□

Finally, we obtain a formula for the inverse of X that is valid for every value of x for which the determinant is nonzero,

$$X^{-1} = \frac{1}{1 - x^3 - x^2} \begin{bmatrix} -x^2 & -1-x & x \\ 1-x & x & -x^2-x+1 \\ x^2-x & -x^2 & 1-x \end{bmatrix}.$$

10.4.3 Cramer's rule

Example 10.4.2 Using Cramer's rule to compute individual variable values in a system of equations. Consider the system

$$\left\{\begin{array}{rcrcrcrcl} x_1 & - & x_2 & + & 2x_3 & + & x_4 & = & 1, \\ 2x_1 & & & + & x_3 & + & x_4 & = & 1, \\ & & x_2 & & & - & 3x_4 & = & 0, \\ x_1 & - & 2x_2 & - & x_3 & & & = & 1, \end{array}\right.$$

with coefficient matrix and vector of constants,

$$A = \begin{bmatrix} 1 & -1 & 2 & 1 \\ 2 & 0 & 1 & 1 \\ 0 & 1 & 0 & -3 \\ 1 & -2 & -1 & 0 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}.$$

Conveniently, we have already computed $\det A = -27$ in Example 8.4.8 (and again in Example 9.3.1). Since $\det A \neq 0$, we know that A is invertible and so the system has one unique solution. Suppose we want to know the value of x_2 in the solution. We can form the matrix A_2, where the second column of A is replaced by $\mathbf{b}$,

$$A_2 = \begin{bmatrix} 1 & 1 & 2 & 1 \\ 2 & 1 & 1 & 1 \\ 0 & 0 & 0 & -3 \\ 1 & 1 & -1 & 0 \end{bmatrix},$$

and compute $\det A_2$ by a cofactor expansion along the third row (expanding the corresponding 3×3 minor determinant along the first row),

$$\begin{aligned} \det A_2 &= -(-3) \begin{vmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 1 & 1 & -1 \end{vmatrix} \\ &= 3\left(1 \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} - 1 \begin{vmatrix} 2 & 1 \\ 1 & -1 \end{vmatrix} + 2 \begin{vmatrix} 2 & 1 \\ 1 & 1 \end{vmatrix}\right) \\ &= 3\big((-1-1) - (-2-1) + 2(2-1)\big) \\ &= 9. \end{aligned}$$

Thus, the value of x_2 in the one unique solution to the system is

$$x_2 = \frac{\det A_2}{\det A} = \frac{9}{-27} = -\frac{1}{3}.$$

If we also want to know the value of x_4, we form the matrix A_4, where the fourth column of A is replaced by $\mathbf{b}$,

$$A_4 = \begin{bmatrix} 1 & -1 & 2 & 1 \\ 2 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & -2 & -1 & 1 \end{bmatrix},$$

and compute $\det A_4$ (again by a cofactor expansion along the third row, followed by an expansion along the first row of the corresponding 3×3 minor determinant),

$$\det A_4 = -1 \begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 1 \\ 1 & -1 & 1 \end{bmatrix}$$

$$
\begin{aligned}
&= -1\left(1\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} - 2\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} + 1\begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix}\right)\\
&= -\big((1+1) - 2(2-1) + (-2-1)\big)\\
&= 3.
\end{aligned}
$$

Thus, the value of x_4 in the one unique solution to the system is

$$x_4 = \frac{\det A_4}{\det A} = \frac{3}{-27} = -\frac{1}{9}.$$

□

10.5 Theory

In this section.

We have discussed the reasoning behind many of the below facts in Section 10.3, so we will omit some of the formal proofs.

10.5.1 Adjoints and inverses

First, we record the adjoint inversion formula we have discovered.

Theorem 10.5.1 Inversion by adjoint. *If* $\det A \neq 0$ *then* A *is invertible, with* $A^{-1} = \frac{1}{\det A}\operatorname{adj} A$.

Remark 10.5.2 Based on our computations for the 2×2 case in Subsection 10.4.1, if A is 2×2 then the statement of the theorem above is exactly the same as Proposition 5.5.4.

10.5.2 Determinants determine invertibility

As we saw in Subsection 10.3.2, there is a stronger connection between the determinant and invertibility, which we now state here more formally by adding a new statement to Theorem 6.5.2.

Theorem 10.5.3 Characterizations of invertibility. *For a square matrix* A*, the following are equivalent.*

1. *Matrix* A *is invertible.*
2. *Every linear system that has* A *as a coefficient matrix has one unique solution.*
3. *The homogeneous system* $A\mathbf{x} = \mathbf{0}$ *has only the trivial solution.*
4. *There is some linear system that has* A *as a coefficient matrix and has one unique solution.*
5. *The rank of* A *is equal to the size of* A.

6. *The RREF of A is the identity.*

7. *Matrix A can be expressed as a product of some number of elementary matrices.*

8. *The determinant of A is nonzero.*

In particular, a square matrix is invertible if and only if its determinant is nonzero.

Remark 10.5.4 In the last sentence of the theorem, the connecting phrase "if and only if" between the two conditions is just a different way to say that the two conditions are equivalent. And recall that conditions are **equivalent** when they have to be either all true or all false at the same time. Rephrasing in terms of the "all false" scenario, we could also say that ***a square matrix is singular if and only if its determinant is zero***.

10.5.3 Determinant formulas

Here we collect the determinant formulas from Subsections 10.3.3–10.3.7. First we look at a special case, previously considered in Discovery 10.4 and Subsection 10.3.3, of the multiplicative formula for determinants.

Lemma 10.5.5 Determinant is multiplicative: elementary case. *If E is an elementary matrix and A is a square matrix of the same size, then*

$$\det(EA) = (\det E)(\det A). \tag{$*$}$$

Proof. There are three cases to consider here, based on the type of elementary matrix we are dealing with.

Case E represents swapping rows. The product EA represents the result of swapping two rows in A, so

$$\det(EA) = -\det A$$

(Part 2 of Proposition 9.4.2).

But also $\det E = -1$ (Part 1 of Proposition 9.4.5), so

$$(\det E)(\det A) = (-1)(\det A) = -\det A$$

as well.

This establishes $(*)$ in this case.

Case E represents multiplying a row by k. The product EA represents the result of multiplying that row of A by k, so

$$\det(EA) = k\det A$$

(Part 4 of Proposition 9.4.2).

But also $\det E = k$ (Part 2 of Proposition 9.4.5), so

$$(\det E)(\det A) = k\det A$$

as well.

This establishes $(*)$ in this case.

Case E represents adding a multiple of one row to another. The product EA represents the result of adding a multiple of a row to another in A, so $\det(EA)$ is equal to $\det A$. But also $\det E = 1$ (Part 3 of Proposition 9.4.5), so

$$\det(EA) = \det A = (1)(\det A) = (\det E)(\det A),$$

establishing $(*)$ in this case. ■

With the above lemma established, we can consider the general multiplicative formula for determinants.

Proposition 10.5.6 Determinant is multiplicative: general case. *A determinant of a product of square matrices is the product of the determinants of those matrices. In particular, the following hold.*

1. *If M and N are square matrices of the same size, then*

$$\det(MN) = (\det M)(\det N).$$

2. *If $M_1, M_2, \ldots, M_{\ell-1}, M_\ell$ are square matrices of the same size, then*

$$\det(M_1 M_2 \cdots M_{\ell-1} M_\ell) = (\det M_1)(\det M_2)\cdots(\det M_{\ell-1})(\det M_\ell).$$

Proof outline for Statement 1. There are two cases to consider.

Case M is invertible. In this case, M can be expressed as a product of elementary matrices (Theorem 10.5.3), and so Lemma 10.5.5 can be repeatedly applied to obtain the desired equality.

In Discovery 10.5 and Subsection 10.3.4, we worked under the assumption that M could be expressed as a product of *three* elementary matrices, but the calculations and logic used there would work no matter how many elementary matrices were required in a product expression for M.

Comment. A more formal proof would require using the principal of mathematical induction on the number of elementary matrices required in a product expansion for M, but that is beyond the scope of this book.

Case M is singular. In this case, $\det M = 0$ by our newly added statement in the list of Theorem 10.5.3, so we have

$$\text{RHS} = (\det M)(\det N) = 0 \cdot \det N = 0$$

as well. But we also know that if M is singular, then the product MN must also be singular (Statement 1 of Proposition 6.5.8). So again we can apply the equivalence of Statement 1 and Statement 8 of Theorem 10.5.3 to obtain

$$\text{LHS} = \det(MN) = 0.$$

Since both LHS and RHS are equal to 0, they are equal to each other. ■

Proof outline for Statement 2. This result can be obtained by repeated applications of the formula in Statement 1, one M_i at a time.

Comment. Again, a more formal proof would require mathematical induction on the number of matrices in the product.

■

Remark 10.5.7 We can now understand the formula $\det(kA) = k^n \det A$ as a special case of Proposition 10.5.6. Using $M = kI$ and $N = A$, we have

$$\det(kA) = \det\bigl((kI)A\bigr) = \bigl(\det(kI)\bigr)(\det A) = k^n \det A.$$

(See Statement 2 of Proposition 8.5.2.)

Lemma 10.5.5 and the proof of Proposition 10.5.6 connect to Proposition 9.4.2 (which includes the formula $\det(kA) = k^n \det A$ as one of its statements) by the fact that an $n \times n$ scalar matrix kI is the product of n elementary matrices, one for each of the n operations *multiply row R_j by k*.

Proposition 10.5.8 Determinant of an inverse. *The determinant of an inverse is the inverse of the determinant. That is, if N is an invertible matrix then* $\det(A^{-1}) = (\det A)^{-1}$.

10.5.4 Cramer's rule

Finally, we formally record Cramer's rule (discussed in Subsection 10.3.8).

Theorem 10.5.9 Cramer's rule. *If system $A\mathbf{x} = \mathbf{b}$ has invertible square coefficient matrix A, then the value of variable x_j in the one unique solution to the system is*

$$x_j = \frac{\det A_j}{\det A},$$

where A_j is the matrix obtained by replacing the j^{th} column of A by $\mathbf{b}$.

Part II

Vector Spaces

CHAPTER 11

Introduction to vectors

11.1 Discovery guide

Discovery 11.1

(a) Plot points $P(1,2)$ and $Q(3,-1)$ in the xy-plane. Draw an arrow from P to Q. This arrow is called the **directed line segment** $\overrightarrow{PQ}$.

(b) Fill in the **components** for this directed line segment:

$$\overrightarrow{PQ} = (\Delta x, \Delta y) = (\qquad\qquad , \qquad\qquad).$$

(c) On the same axes you've been working on, plot the point R that has the same coordinates as the components of $\overrightarrow{PQ}$, and draw the directed line segment $\overrightarrow{OR}$ where O is the origin. What do you notice about this arrow?

(d) This "common" arrow for $\overrightarrow{PQ}$ and $\overrightarrow{OR}$ (and all arrows in the xy-plane just like it) is called a **vector**. Let's label this vector **v**. Draw another "copy" of **v** so that its **initial point** is $S(-2,1)$. What will be the **terminal point** for this copy of **v**?

Discovery 11.2

(a) Draw the vector $\mathbf{u} = (2,3)$ with its initial point at $P(1,1)$. Label this vector on your diagram, and label its terminal point as Q. Now draw the vector $\mathbf{v} = (3,-1)$ with its initial point at Q. Label this vector on your diagram, and label this second terminal point as R. Draw in vector **w** corresponding to $\overrightarrow{PR}$.

(b) Compute the components of **w** using a $(\Delta x, \Delta y)$ calculation between its initial and terminal points in your diagram, similarly to Task 11.1.b.

Looking at the components of **u** and **v**, what do you notice about the components of **w**? Based on this, we should call **w** the $\quad$ of **u** and **v**, and write $\mathbf{w} = \mathbf{u} \qquad \mathbf{v}$.

(c) Now work in the reverse order: on the same diagram you've been working on, draw **v** starting at P, then draw **u** starting at that terminal point. Where did the second terminal point end up? Turn this into a rule for vector algebra: $\qquad\qquad$.

Hint. Your rule should be about the different ways to combine **u** and **v** that you've explored so far in this activity.

(d) What shape do the four vectors on the outside make? And what is $\overrightarrow{PR}$ relative to that shape (geometrically)?

Discovery 11.3

(a) How should you draw the vector $\mathbf{0} = (0,0)$?

(b) What happens if you draw $\mathbf{0}$ tail-to-head or head-to-tail with another vector (as in Discovery 11.2)? Turn this into a rule for vector algebra: .

Discovery 11.4 We would reasonably expect $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$ in vector algebra.

(a) Draw a geometric representation of this rule on a set of axes for $\mathbf{v} = (2,1)$ (use the origin as the initial point for $\mathbf{v}$).

(b) What should the components of $-\mathbf{v}$ be?

(c) On the same set of axes as before, draw $-\mathbf{v}$ with its initial point at the origin.

Discovery 11.5 Draw an arbitrary vector in the plane, and label it $\mathbf{u}$. Then draw another arbitrary vector with its initial point at the terminal point of $\mathbf{u}$ (but maybe have this new vector head off in a new direction). Label this second vector $\mathbf{v}$. Now draw in the sum vector $\mathbf{w} = \mathbf{u} + \mathbf{v}$, similarly to Discovery 11.2.

(a) Which of your three vectors represents $\mathbf{w} - \mathbf{u}$?

(b) Draw in another vector for $\mathbf{u} - \mathbf{w}$.

(c) What is the point of this activity?

Discovery 11.6

(a) Draw a representative diagram for the vector sum $\mathbf{v} + \mathbf{v}$ using $\mathbf{v} = (2,1)$ (start with the first initial point at the origin). What are the components of this sum vector?

From both the geometry of what you've drawn, and the result for the components of the sum vector $\mathbf{v} + \mathbf{v}$, do you think it is reasonable to write $2\mathbf{v}$ to mean $\mathbf{v} + \mathbf{v}$?

(b) Now draw each of the following, and determine their components: $3\mathbf{v}$, $-2\mathbf{v}$, $\frac{1}{2}\mathbf{v}$, $-\frac{5}{4}\mathbf{v}$.

Discovery 11.7 Draw an arbitrary representative diagram for $\mathbf{w} = \mathbf{u} + \mathbf{v}$ (similarly to how you started Discovery 11.5). On the same set of axes, draw a diagram for $2\mathbf{u} + 2\mathbf{v}$, and compare with $2\mathbf{w}$. Express what you've discovered as a rule of vector algebra, with 2 replaced by variable k: .

Discovery 11.8 On a set of xy-axes, draw the **standard basis vectors**: $\mathbf{e}_1 = (1,0)$ and $\mathbf{e}_2 = (0,1)$, along with the vector $\mathbf{v} = (5,2)$. Then draw a geometric representation of $\mathbf{v}$ as a **linear combination** $\mathbf{v} = 5\mathbf{e}_1 + 2\mathbf{e}_2$.

Discovery 11.9 All the vectors we've encountered so far are **two-dimensional vectors**. Let's bump everything up a dimension.

(a) Using $\mathbf{u} = (1,1,0)$ and $\mathbf{v} = (1,-1,2)$, draw the following on a set of xyz-axes: $\mathbf{u}$, $\mathbf{v}$, $\mathbf{u} + \mathbf{v}$, $-\mathbf{u}$, $\mathbf{v} - \mathbf{u}$, $2\mathbf{v}$.

(b) Now compute the components of each of the vectors from the previous part of this activity.

We can't draw pictures of n-dimensional vectors if $n > 3$, but we can do all the same algebra.

Discovery 11.10 Complete the following vector algebra formulas.

- $(u_1, u_2, \ldots, u_n) + (v_1, v_2, \ldots, v_n) =$
- $(u_1, u_2, \ldots, u_n) - (v_1, v_2, \ldots, v_n) =$
- $-(v_1, v_2, \ldots, v_n) =$
- $k(v_1, v_2, \ldots, v_n) =$
- $\mathbf{0} =$

We can use vectors to represent other kinds of "displacements" besides position displacements. Vectors can be used to represent change between states of any collection of related variables.

Discovery 11.11 An investor sinks \$10,000 into stock in each of companies A, B, C, and D. After a year, the various items in her portfolio have the following values: company A, \$10,475; company B, \$11,240; company C, \$9,756; company D, \$10,054.

Represent the "displacement" in the collection of values of the investor's holdings, from initial state

$$(A, B, C, D) = (10000, 10000, 10000, 10000)$$

to terminal state

$$(A, B, C, D) = (10475, 11240, 9756, 10054)$$

as a four-dimensional vector.

Discovery 11.12 If we write two-dimensional vectors in the form $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$, instead of the form $\mathbf{u} = (u_1, u_2)$, then we can use matrix algebra to do computations with vectors.

(a) Does each rule of vector algebra that we've discovered today have a counterpart rule in matrix algebra?

(b) Will the same be true for the algebra of higher-dimensional vectors? (That is, if we consider using $n \times 1$ column matrices to represent n-dimensional vectors?)

11.2 Terminology and notation

directed line segment
a line segment between two points with an assigned direction from one of the points to the other

Remark 11.2.1 We usually visualize a directed line segment as an arrow.

initial point (of a directed line segment)
the first point in a directed line segment (at the tail of the arrow)

terminal point (of a directed line segment)
the second point in a directed line segment (at the head of the arrow)

components (of a directed line segment)
the list of the *changes* in coordinates between initial point and terminal point

vector the ordered collection of components of a directed line segment

Remark 11.2.2

- We won't make too much of a fuss about the technical definition of a vector, especially since we will vastly increase the number of things we allow ourselves to call **vector** in Chapter 15.
- Notationally, we will typeset variables representing vectors in boldface, just as we did previously for column vectors in the context of matrices and systems of equations.

two-dimensional vector
a vector with two components $\mathbf{v} = (v_1, v_2)$, corresponding to a directed line segment in the plane

three-dimensional vector
a vector with three components $\mathbf{v} = (v_1, v_2, v_3)$, corresponding to a directed line segment in space

n-dimensional vector
a vector with n components $\mathbf{v} = (v_1, v_2, \ldots, v_n)$

two-dimensional space ($\mathbb{R}^2$)
the collection of all two-dimensional vectors

three-dimensional space ($\mathbb{R}^3$)
the collection of all three-dimensional vectors

n-dimensional space ($\mathbb{R}^n$)
the collection of all n-dimensional vectors

zero vector
the vector $\mathbf{0} = (0, 0, \ldots, 0)$

Remark 11.2.3 We refer to $\mathbb{R}^2$ as **two-dimensional space** because, just like a map, the plane has two sets of directions — north/south and east/west. We refer to $\mathbb{R}^3$ as **three-dimensional space** because we still have the north/south and east/west sets of directions in the xy-plane, but we add a third set of directions of up/down along the z-axis. In analogy with this, we refer to $\mathbb{R}^4$ as **four-dimensional space**, $\mathbb{R}^5$ as **five-dimensional space**, and so on.

A look ahead. In Chapter 19, we will make the concept of **dimension** of a **vector space** more precise.

vector addition (of vectors u and v of the same dimension)
given a directed line segment corresponding to $\mathbf{u}$, create a directed line segment corresponding to $\mathbf{v}$ with initial point at the terminal point for the segment for $\mathbf{u}$, and then the sum vector $\mathbf{u}+\mathbf{v}$ corresponds to the directed line segment from the initial point for $\mathbf{u}$ to the terminal point for $\mathbf{v}$

negative (of a vector v)
given a directed line segment for $\mathbf{v}$, the negative vector $-\mathbf{v}$ corresponds to the same segment but in the opposite direction

vector subtraction
the result of adding a vector $\mathbf{u}$ to the negative of another $\mathbf{v}$: $\mathbf{u}-\mathbf{v} = \mathbf{u}+(-\mathbf{v})$

scalar multiple (of a vector v by scalar k)
given a directed line segment for $\mathbf{v}$, the scalar multiple $k\mathbf{v}$ corresponds to the directed line segment that has the same initial point and changes position in the same direction, but whose length has been scaled so that the terminal point is $|k|$ times as far from the initial point as the terminal point for $\mathbf{u}$; if k is negative then the terminal point is also moved to the "other side" of the initial point

parallel vectors
nonzero vectors that are scalar multiples of one another

linear combination of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$
a sum of scalar multiples of the vectors: $k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + \ldots + k_m\mathbf{v}_m$

standard basis vectors (in $\mathbb{R}^n$)
the vectors

$$\begin{aligned} \mathbf{e}_1 &= (1,0,0,\ldots,0), \\ \mathbf{e}_2 &= (0,1,0,\ldots,0), \\ &\ldots, \\ \mathbf{e}_n &= (0,0,\ldots,0,1) \end{aligned}$$

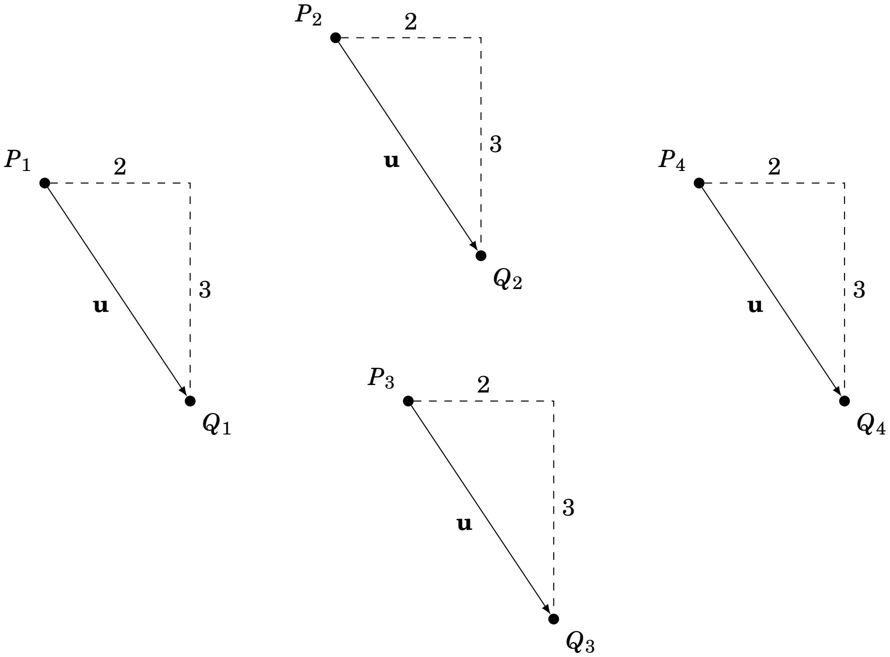
Remark 11.2.4 In physics, it is common to use $\mathbf{i}$ and $\mathbf{j}$ to mean $\mathbf{e}_1$ and $\mathbf{e}_2$ in the plane, and to use $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ to mean $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ in space. However, this alphabetic naming scheme would have to wrap back around to $\mathbf{a}$ in 19 dimensions, and in 27 dimensions there wouldn't be enough letters in the alphabet. So we will (mostly) stick with the $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ naming scheme.

11.3 Concepts

In this section.

11.3.1 Vectors

A directed line segment (or arrow) could be thought of dynamically as describing a change in position, from the initial point to the terminal point. A two-dimensional vector in the plane or a three-dimensional vector in space captures just the *change* part of "change in position," leaving the *position* part (that is, the initial and terminal points) unspecified. For example, in the plane, the instructions "move two units right and three units down" describe a way to change positions, but don't actually specify *from where* or *to where* the change in position is occurring. So a vector corresponds to an infinite number of directed line segments, where each of these directed line segments has a different initial point but all of them require the same "change" to change positions from initial point to terminal point. Continuing our example, *every* change in position between some initial and terminal points in the plane that requires moving two units right and three units down can be represented by the same vector.



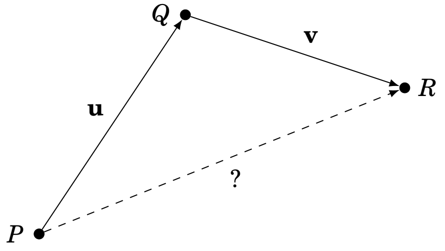
We describe a two-dimensional vector in the plane with a pair of numerical **components**: the change in x and the change in y. If $P(x_1,y_1)$ and $Q(x_2,y_2)$ are points in the plane, then the vector associated to the directed line segment $\overrightarrow{PQ}$ has components $\mathbf{v} = (\Delta x, \Delta y) = (x_2 - x_1, y_2 - y_1)$. (A three-dimensional vector in space requires a third component as well: the change in z.)

Notice what happens when we use the origin $O(0,0)$ as the initial point and an arbitrary point $R(x,y)$ as the terminal point in a directed line segment: the vector associated to $\overrightarrow{OR}$ is then $(x-0, y-0) = (x,y)$. So when the initial point is the origin, the components of the vector are exactly the coordinates of the terminal point. In Discovery 11.1 we saw that this works in reverse as well. That is, if we have a vector $\mathbf{v} = (v_1, v_2)$, then we could consider the numbers v_1, v_2 as coordinates of a point $R(x,y)$ with $x = v_1$ and $y = v_2$, and then the vector associated to $\overrightarrow{OR}$ is just $\mathbf{v}$ again. In this way, every vector corresponds to *one unique* directed line segment with initial point at the origin, and so that is sort of the "natural position" of the vector as a directed line segment. But we will find that it is often convenient to consider other directed line segments that correspond to a particular vector.

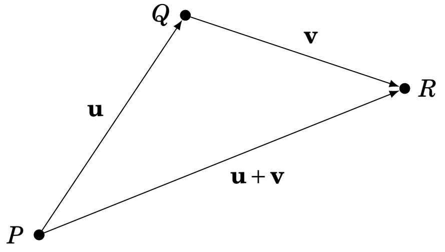
We live in a three-dimensional world (or, at least, it appears that way to us), and our little human brains cannot visualize points or arrows in four- or higher-dimensional spaces. However, we can still describe such imaginary objects using our experience from two- and three-dimensional points and vectors. For example, if we had two points P and Q in an imaginary four-dimensional space, they would each require four coordinates, so we would describe them as $P(w_1,x_1,y_1,z_1)$ and $Q(w_2,x_2,y_2,z_2)$. Then the vector corresponding to the directed line segment $\overrightarrow{PQ}$ would have four components and we would compute it as $\mathbf{v} = (\Delta w, \Delta x, \Delta y, \Delta z) = (w_2 - w_1, x_2 - x_1, y_2 - y_1, z_2 - z_1)$.

11.3.2 Vector addition

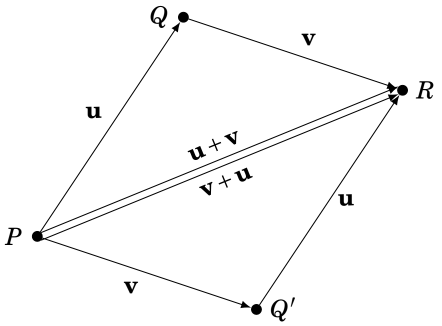
A vector describes a change in position. If we chain two changes in position together, by making the initial point of the second vector the same as the terminal point of the first vector, then we could consider the overall change in position.



If these are points and vectors in the plane, then clearly the change in position from P to R will be described by the *total net* change in x and the *total net* change in y, as we discovered in Discovery 11.2. So, in the diagram above, we obtain the components for the vector corresponding to $\overrightarrow{PR}$ by adding corresponding components of $\mathbf{u}$ and $\mathbf{v}$. That is, if $\mathbf{u} = (u_1, u_2)$ and $\mathbf{v} = (v_1, v_2)$, then the components of the dashed arrow labelled with a question mark are $(u_1 + v_1, u_2 + v_2)$. For obvious reasons, we call this the **sum** of $\mathbf{u}$ and $\mathbf{v}$.



In Discovery 11.2 we also considered the result of interchanging the order of a pair of vectors that have been chained together.



In the diagram above, the vector for $\overrightarrow{PQ'}$ is the same as that for $\overrightarrow{QR}$, because they represent the same change in position, just with a different initial point. Accordingly, we have labelled both vectors with $\mathbf{v}$. And the same applies to $\mathbf{u}$ with respect to $\overrightarrow{PQ}$ and $\overrightarrow{Q'R}$.

The diagram illustrates that if we start at P and chain together the change-of-position instructions contained in vectors $\mathbf{u}$ and $\mathbf{v}$, the order that we do so does not matter — the overall change in position will be from P to R. Thus, *the order of vector addition doesn't matter*. Algebraically, we could have predicted that this would be the case because it doesn't matter what order you add components: the identities $u_1 + v_1 = v_1 + u_1$ and $u_2 + v_2 = v_2 + v_2$ are both valid. But it's useful conceptually to have the above geometric picture of vector addition because, whether you believe this about yourself or not, *humans are spatial thinkers*. And the geometric version of the vector identity $\mathbf{v} + \mathbf{u} = \mathbf{u} + \mathbf{v}$ makes a pretty picture of a parallelogram, so we call it the **parallelogram rule**.

For three-dimensional vectors, we can imagine diagrams like the ones we have drawn above floating in space, and the parallelogram rule would hold there as well. In higher dimensions, we cannot draw pictures, but we could imagine that they are similar. At any rate, the algebra of vector addition is the same in any dimension: for $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ and $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ in $\mathbb{R}^n$, we have

$$\mathbf{u} + \mathbf{v} = (u_1 + v_1, u_2 + v_2, \ldots, u_n + v_n).$$

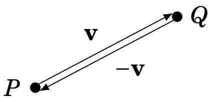
11.3.3 The zero vector

There is one special change in position that is unlike any other — the one where the initial and terminal points are the same, so that there is actually *no* change in position. In two dimensions, this means there is no change in either x or y, so the components are $(0,0)$. Similarly, in any number of dimensions we have the **zero vector** $\mathbf{0} = (0, 0, \ldots, 0)$.

As we explored in Discovery 11.3, if we chain together a vector $\mathbf{v}$, representing some change in position, with the zero vector, which represents no change, then the net result is just the change of $\mathbf{v}$. That is, $\mathbf{v} + \mathbf{0} = \mathbf{v}$, and also $\mathbf{0} + \mathbf{v} = \mathbf{v}$.

11.3.4 Vector negatives and vector subtraction

If we move from P to Q, and from there move from Q back to P, the net result is no change in position, which is represented by the zero vector. This means if we add the vector corresponding to $\overrightarrow{PQ}$ to the one corresponding to $\overrightarrow{QP}$, the result is $\mathbf{0}$. So if we label the vector for $\overrightarrow{PQ}$ as $\mathbf{v}$, it seems reasonable to label the vector for $\overrightarrow{QP}$ as $-\mathbf{v}$, the **negative** of $\mathbf{v}$, so that we have $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$.



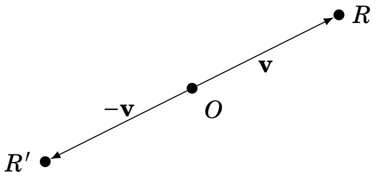
If we are to have $\mathbf{v}+(-\mathbf{v})=\mathbf{0}$, and the components of $\mathbf{0}$ are all 0, then since we add vectors by adding corresponding components, the components of $-\mathbf{v}$ must be the negatives of the components of $\mathbf{v}$. For example, if $\mathbf{v}=(v_1,v_2)$ in the plane, then $-\mathbf{v}=(-v_1,-v_2)$. In any dimension, we have

$$\mathbf{v}=(v_1,v_2,\dots,v_n) \qquad \Longrightarrow \qquad -\mathbf{v}=(-v_1,-v_2,\dots,-v_n).$$

This relationship between the components of $\mathbf{v}$ and $-\mathbf{v}$ will lead to an identity between the negative and a certain scalar multiple of a vector in Subsection 11.3.5 below.

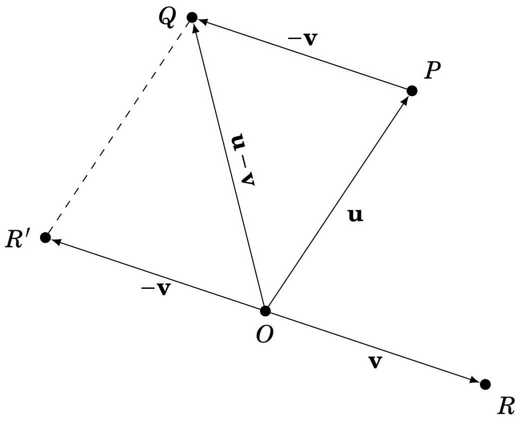
A look ahead. In Section 11.2, we have defined a negative vector as having the opposite direction to the original. However, when we introduce *abstract* vectors in Chapter 15, we won't have geometric notions like "opposite direction," so we will need to rely on the algebraic condition $\mathbf{v}+(-\mathbf{v})=\mathbf{0}$ to know what a "negative vector" should be.

Remembering that the "natural" position for a vector is with its tail at the origin, it's useful to visualize negatives in the following manner.

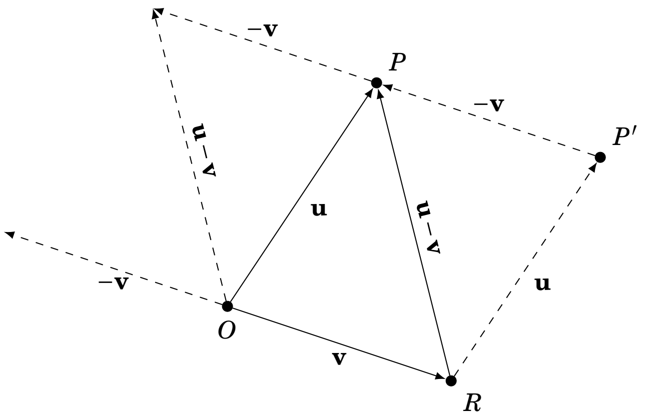


That is, the negative of a vector will change positions by the same distance but in the opposite direction.

To subtract vectors, we add to a vector the negative of another.

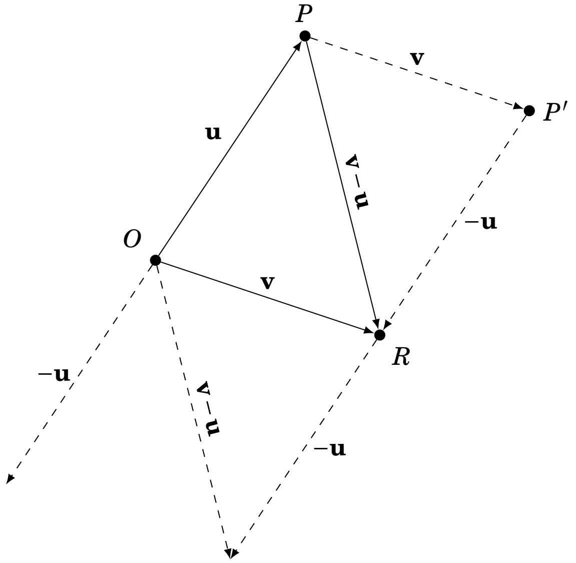


Here, the diagonal vector labelled $\mathbf{u}-\mathbf{v}$ is obtained by adding $\mathbf{u}$ and $-\mathbf{v}$. As we explored in Discovery 11.5, we get an interesting pattern if we draw in another copy of the vector labelled $\mathbf{u}-\mathbf{v}$ with its initial point at R.



Triangle $\triangle RP'P$ creates the vector addition pattern $\mathbf{u}+(-\mathbf{v})=\mathbf{u}-\mathbf{v}$. But notice that $\triangle ORP$ creates a vector addition pattern starting at O and ending up at P, by $\mathbf{v}+(\mathbf{u}-\mathbf{v})=\mathbf{u}$. So we can think of a difference of two vectors as *a vector that runs between the heads of the two vectors in the difference when they share the same initial point*. Algebraically, we can think of the $\mathbf{v}$ and $-\mathbf{v}$ cancelling in the expression $\mathbf{v}+(\mathbf{u}-\mathbf{v})$, leaving just $\mathbf{u}$.

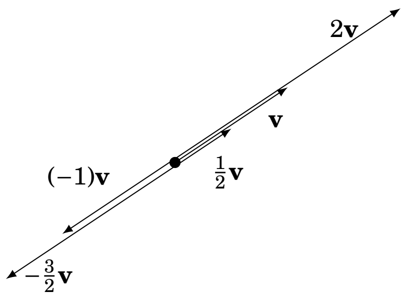
Of course, there are two vectors that run between the heads of $\mathbf{u}$ and $\mathbf{v}$, namely $\mathbf{u}-\mathbf{v}$ and its negative.



Now $\triangle PP'R$ creates a vector addition pattern starting at P and ending up at R, so that $\mathbf{v}+(-\mathbf{u})=\mathbf{v}-\mathbf{u}$. But also $\triangle OPR$ creates a vector addition pattern starting at O and ending up at R, so that $\mathbf{u}+(\mathbf{v}-\mathbf{u})=\mathbf{v}$. And finally, the fact that $\mathbf{u}-\mathbf{v}$ and $\mathbf{v}-\mathbf{u}$ both run between P and R, but in opposite directions, verifies geometrically that $-(\mathbf{u}-\mathbf{v})=\mathbf{v}-\mathbf{u}$, as we would expect algebraically.

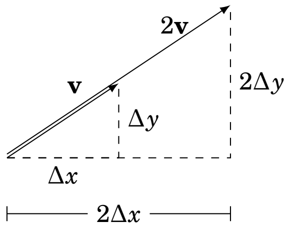
11.3.5 Scalar multiplication

Geometrically, when we scalar multiply a vector we "stretch" or *scale* its length by the scale factor. (If this scale factor is negative, then we also flip the vector around in the opposite direction.) Here are some examples.



Notice how each of these vectors either points in the same direction as or in the opposite direction to **v**. In particular, they are all **parallel** to one another. This happens precisely when the vectors are scalar multiples of one another.

Thinking of the vectors in the diagram above as vectors in the plane, if we scale **v** by a factor of 2, then our knowledge of similar triangles tells us that the change in both x and y must be double.



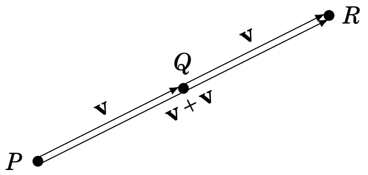
So if $\mathbf{v} = (v_1, v_2)$, then $2\mathbf{v} = (2v_1, 2v_2)$. This relationship between original vector **v** and scaled vector $k\mathbf{v}$ holds in general, in any dimension, and even for negative k:

$$\mathbf{v} = (v_1, v_2, \dots, v_n) \qquad \implies \qquad k\mathbf{v} = (kv_1, kv_2, \dots, kv_n).$$

In the case that $k = -1$, we obtain the identity $(-1)\mathbf{v} = -\mathbf{v}$, as promised earlier.

Remark 11.3.1 It may seem redundant to write $(-1)\mathbf{v} = -\mathbf{v}$, don't both sides mean the same thing? In terms of the effect on components of **v**, yes they are the same. However, when we explore *abstract* vectors in Chapter 15, we won't have components or the geometric notion of "opposite direction" as means of seeing this equality, and so there will initially be a subtle difference between the idea of a vector having an *additive* negative (so that $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$) and the operation of scalar multiplying a vector by the particular scalar -1.

We can connect scalar multiplication to addition, as we explored in Discovery 11.6. If we add a vector to itself, then the sum vector will be twice as long as the original.



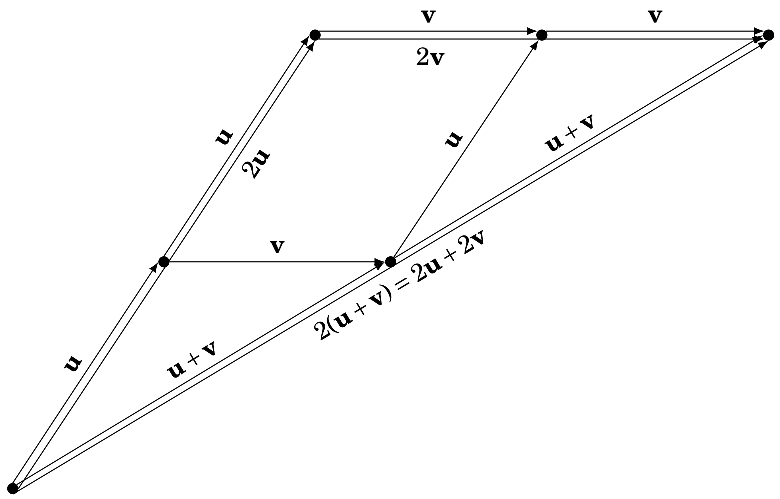
So we have $\mathbf{v} + \mathbf{v} = 2\mathbf{v}$.

11.3.6 Vector algebra

We have already discovered a few rules of vector algebra, such as

$$\mathbf{v} + \mathbf{u} = \mathbf{u} + \mathbf{v}, \qquad -(\mathbf{v} - \mathbf{u}) = \mathbf{u} - \mathbf{v}, \qquad \mathbf{v} + \mathbf{v} = 2\mathbf{v}, \qquad (-1)\mathbf{v} = -\mathbf{v}.$$

In Discovery 11.7, we explored a version of the distributive rule $k(\mathbf{u}+\mathbf{v}) = k\mathbf{u}+k\mathbf{v}$ in the case $k = 2$.



We will provide more rules of vector algebra as Proposition 11.5.1 in Subsection 11.5.1. In Discovery 11.12, we decided that the algebra of vectors is the same as the algebra of column matrices (which we have already been referring to as **column vectors**), so we should be able to anticipate a number of the vector algebra rules that will appear in that proposition.

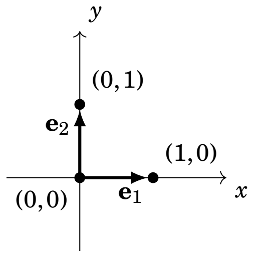
A look ahead. The fact that the algebra of column matrices matches exactly with the algebra (and geometry) of vectors is an important pattern, and recognizing this pattern is the first step to employing the most powerful tool of mathematics: *abstraction*. In Chapter 15 and beyond, we will extract these common algebraic patterns into an abstract concept of a **vector space**, and then use logic to deduce important properties of *all* collections of mathematical objects that follow the same algebraic patterns.

Warning 11.3.2 ***There is no multiplication operation for vectors!***

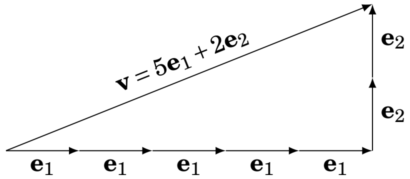
Algebraically, vectors in $\mathbb{R}^n$ are the same as column matrices, and you cannot multiply two column matrices together because their sizes do not match up (except in $\mathbb{R}^1$, but let's ignore that for now). This also means that ***you cannot square a vector, you cannot square-root a vector, you cannot invert a vector, and you cannot divide by a vector***. Do not try to use any of these operations in vector algebra! In Chapters 12–13, we will encounter some operations tied to the geometry of vectors that we will call "vector products" and for which we will use multiplication-like notation, but they will be for very specific geometric purposes and do not really correspond to our idea of multiplication in the algebra of numbers.

11.3.7 The standard basis vectors

In Discovery 11.8, we encountered two very special vectors in the plane, $\mathbf{e}_1 = (1,0)$ and $\mathbf{e}_2 = (0,1)$. These two vectors could be considered the *fundamental* changes in position in the plane — $\mathbf{e}_1$ represents a change by one unit right, and $\mathbf{e}_2$ represents a change by one unit up.



Any change in position can be built out of these two fundamental changes in position. Using the example in Discovery 11.8, the vector $\mathbf{v} = (5,2)$ represents a change in position by 5 units right and 2 units up. We can achieve the "5 units right" part with $5\mathbf{e}_1 = (5,0)$ and the "2 units up part" with $2\mathbf{e}_2 = (0,2)$. To get the total change in position represented by $\mathbf{v}$, we can combine these two building blocks in the **linear combination** $\mathbf{v} = 5\mathbf{e}_1 + 2\mathbf{e}_2$.



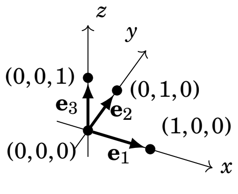
As you can imagine, every vector in the plane can be decomposed into a linear combination of $\mathbf{e}_1$ and $\mathbf{e}_2$ in this manner: for $\mathbf{v} = (v_1, v_2)$, we have $\mathbf{v} = v_1\mathbf{e}_1 + v_2\mathbf{e}_2$. For this reason, the two vectors $\mathbf{e}_1, \mathbf{e}_2$ together are called the **standard basis vectors** in $\mathbb{R}^2$, as they form a basis from which every other vector can be constructed. To use an analogy with chemistry, these two vectors are the basic *atoms* of $\mathbb{R}^2$, and every other vector in $\mathbb{R}^2$ is a *molecule* built out of a specific combination of these atoms. Since there are only two fundamental directions in $\mathbb{R}^2$ (right/left and up/down), it is not surprising that we need only two basis vectors to represent all possible directions. This is the reason we call vectors in $\mathbb{R}^2$ **two-dimensional vectors**.

Note. Left is not considered a different direction from right, it is just the opposite (or negative) direction: as $\mathbf{e}_1$ points right, $-\mathbf{e}_1$ points left. And similarly, up and down are not different directions, just opposite. So there are only two fundamental directions in the plane, not four.

In $\mathbb{R}^3$, there are *three* fundamental directions, two horizontal and one vertical. We might use navigational terminology for the two horizontal directions and describe them as north/south and east/west, and then we can refer to the vertical direction as up/down. So we need *three* standard basis vectors in $\mathbb{R}^3$,

$$\mathbf{e}_1 = (1,0,0), \qquad \mathbf{e}_2 = (0,1,0), \qquad \mathbf{e}_3 = (0,0,1),$$

which we can visualize as below.



Note. In this diagram, you should view the z-axis as coming straight up out of the xy-plane.

As before, any vector in $\mathbb{R}^3$ can be decomposed as a linear combination of these three fundamental vectors. For example, the vector $(1,-1,2)$ decomposes as

$$(1,-1,2) = 1\mathbf{e}_1 + (-1)\mathbf{e}_2 + 2\mathbf{e}_3.$$

And we can repeat all this in $\mathbb{R}^n$ for any value of n, where there are n standard basis vectors,

$$\mathbf{e}_1 = (1,0,0,\ldots,0), \qquad \mathbf{e}_2 = (0,1,0,\ldots,0), \qquad \mathbf{e}_n = (0,0,\ldots,0,1).$$

In fact, given a vector $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ in $\mathbb{R}^n$ (whether $n = 2$ or $n = 3$ or higher), when we try to decompose

$$\mathbf{v} = \boxed{}\,\mathbf{e}_1 + \boxed{}\,\mathbf{e}_2 + \cdots + \boxed{}\,\mathbf{e}_n,$$

we find that there is only *one unique* combination of scalar values that can fill in the blanks and create an equality between $\mathbf{v}$ on the left and the linear combination on the right:

$$\mathbf{v} = v_1\mathbf{e}_1 + v_2\mathbf{e}_2 \cdots + v_n\mathbf{e}_n.$$

A look ahead. The fact that every vector decomposes uniquely as a linear combination of basis vectors is an important feature of the standard basis of $\mathbb{R}^n$ that we will see repeated when we explore the concept of **basis** in abstract vector spaces.

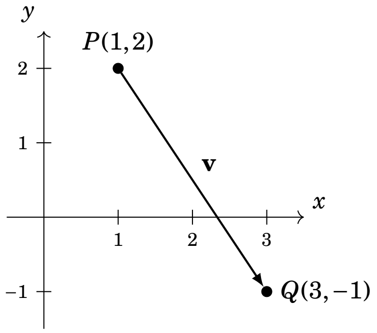
11.4 Examples

In this section.

- Subsection 11.4.1 *Vectors in* $\mathbb{R}^n$
- Subsection 11.4.2 *Vector operations*

11.4.1 Vectors in $\mathbb{R}^n$

Following Discovery 11.1, consider the vector $\mathbf{v}$ in $\mathbb{R}^2$ (that is, in the plane) that represents changing position from $P(1,2)$ to $Q(3,-1)$.

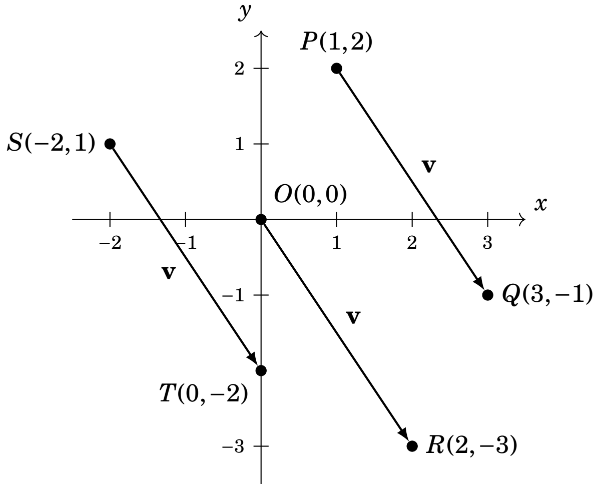


We can compute the components of $\mathbf{v}$ by computing the change in x and the change in y in moving from P to Q:

$$\mathbf{v} = (\Delta x, \Delta y) = (3-1, -1-2) = (2,-3).$$

We can see by looking at their coordinates that moving from point P to Q requires moving 2 units right to get from $x = 1$ to $x = 3$ and moving 3 units down to get from $y = 2$ to $y = -1$, and our calculation of $\mathbf{v}$ above agrees.

The same vector with some other initial point will also have a terminal point that is 2 units to the right and 3 units down from the initial point. In particular, if we take the initial point to be the origin, then the terminal point will have coordinates $(2,-3)$, same as the components of $\mathbf{v}$.



Notice that these different representations of the vector $\mathbf{v}$ are parallel and have the same length.

Vectors can be similarly computed from pairs of points by subtracting coordinates in any dimension. For example, we compute the vector that represents changing position in space from $P(1,2,-3)$ to $Q(3,-1,0)$ by

$$\begin{aligned}\mathbf{v} &= (\Delta x, \Delta y, \Delta z)\\ &= \big(3-1,-1-2,0-(-3)\big)\\ &= (2,-3,3).\end{aligned}$$

Another example in four-dimensional space, with

$$P(1,2,-3,-4), \qquad\qquad Q(1,-1,1,-1),$$

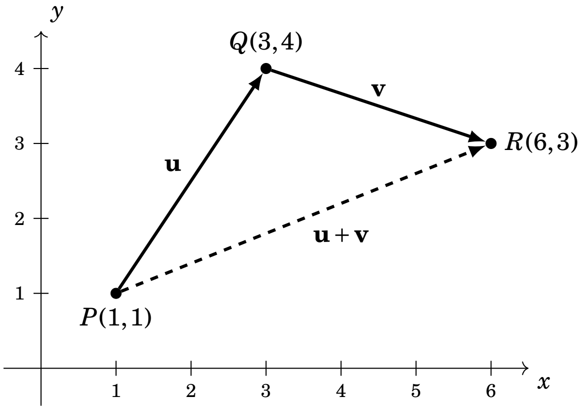
yields

$$\begin{aligned}\mathbf{v} = \overrightarrow{PQ} &= (\Delta x_1, \Delta x_2, \Delta x_3, \Delta x_4)\\ &= \big(1-1,-1-2,1-(-3),-1-(-4)\big)\\ &= (0,-3,4,3).\end{aligned}$$

11.4.2 Vector operations

Here we'll work through some of the computations of Discovery guide 11.1, and provide the accompanying diagrams.

Example 11.4.1 Vector addition in $\mathbb{R}^2$. In Discovery 11.2, we were tasked with geometrically adding vectors $\mathbf{u} = (2,3)$ and $\mathbf{v} = (3,-1)$ in the plane, starting at initial point $P(1,1)$.



To add vectors geometrically, we put them head-to-tail. The vector $\mathbf{u} = (2,3)$ instructs us to move 2 units right and 3 units up, so starting at $P(1,1)$ we end up at $Q(3,4)$. Then the vector $\mathbf{v} = (3,-1)$ instructs us to move 3 units right and 1 unit down, so starting at Q we end up at $R(6,3)$. The sum vector $\mathbf{u}+\mathbf{v}$ represents the *overall* change from P to R, which is 5 units right and 2 units up, so that $\mathbf{u}+\mathbf{v} = (5,2)$. We can also add the vectors algebraically by

$$\begin{aligned}\mathbf{u}+\mathbf{v} &= (2,3)+(3,-1)\\ &= \bigl(2+3, 3+(-1)\bigr)\\ &= (5,2).\end{aligned}$$

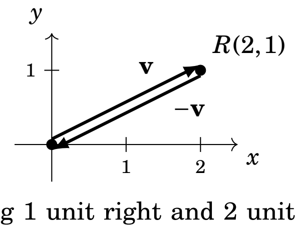
Adding the vectors algebraically is obviously faster and easier than drawing a diagram, but it's good to have a mental picture of the geometric version of addition — it will help conceptually later on. □

Example 11.4.2 Vector addition in higher dimensions. Our geometric picture and algebraic computation of addition are similar for three-dimensional vectors in space. In $\mathbb{R}^n$ with $n > 3$, we can't draw a picture but we could imagine vector addition would take same the familiar triangle shape, and the algebraic computations are similar again. For example,

$$\begin{aligned}(1,2,3,4,5)+(6,-2,4,0,1) &= \bigl(1+6, 2+(-2), 3+4, 4+0, 5+1\bigr)\\ &= (7,0,7,4,6)\end{aligned}$$

in $\mathbb{R}^5$. □

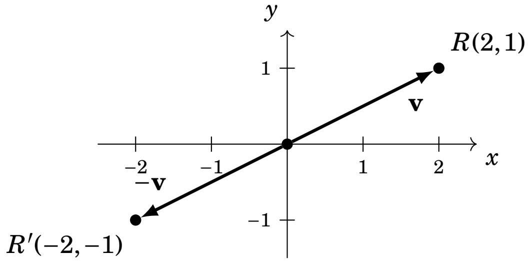
Example 11.4.3 Negative vectors. In Discovery 11.4, we explored the concept of a negative vector as the vector that will return us to our initial point, after changing positions along vector $\mathbf{v} = (2,1)$ in the plane, starting at the origin. Recall that if a vector has its initial point at the origin, then the terminal point has coordinates equal to the components of the vector.



If $\mathbf{v}$ represents moving 1 unit right and 2 units up, then to return to our original position we must move 1 unit left and 2 units down, so that $-\mathbf{v} = (-2,-1)$. Of course, the components of $-\mathbf{v}$ do not depend on what initial point we choose — we would need to make the same reverse change of position no matter where $\mathbf{v}$

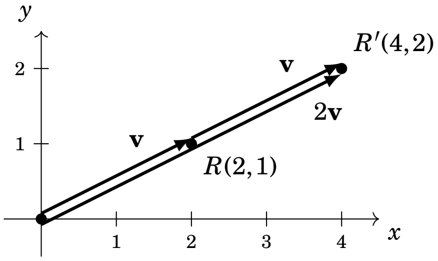
started.

As in Subsection 11.3.4, it is helpful to have a mental picture of a negative vector where its initial point is the same as for the original vector. In this orientation, the vector and its negative are parallel but oppositely directed.



□

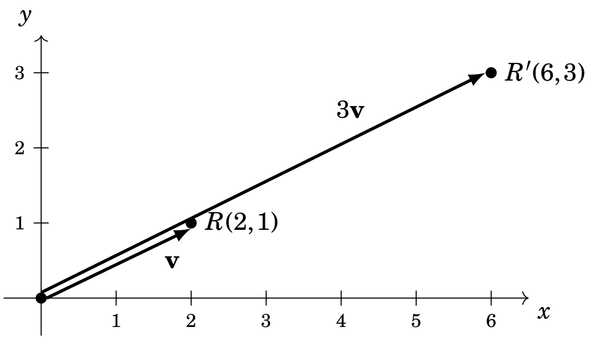
Example 11.4.4 Scalar multiplication. In Discovery 11.6, we explored scalar multiplication geometrically in the plane, using $\mathbf{v} = (2,1)$, initially by relating scalar multiplication to addition.

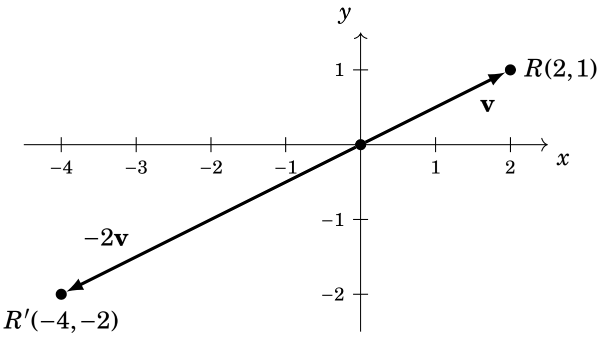


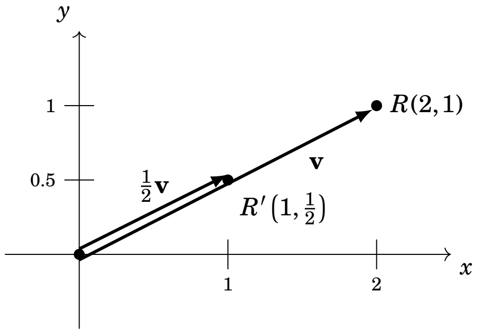
The above diagram illustrates that $\mathbf{v}+\mathbf{v} = 2\mathbf{v}$, which we can also confirm algebraically:

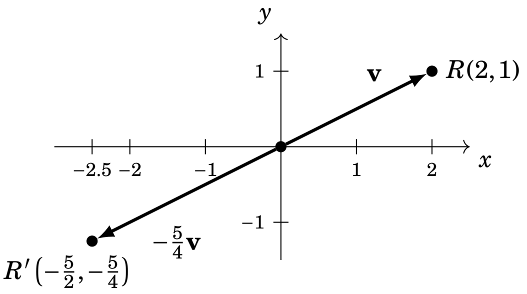
$$\begin{aligned} \mathbf{v}+\mathbf{v} &= (2+2,1+1) \\ &= (2,4) \\ &= 2(2,1) \\ &= 2\mathbf{v}. \end{aligned}$$

Geometrically, the scalar multiples $3\mathbf{v}$, $-2\mathbf{v}$, $\frac{1}{2}\mathbf{v}$, and $-\frac{5}{4}\mathbf{v}$ are all parallel to $\mathbf{v}$ but with lengths stretched or compressed by the scale factor. Additionally, a negative scalar multiple flips the vector around in the opposite direction.









Since the initial point is the origin, each vector above has components equal to the coordinates of its terminal point. In particular, we have

$$3\mathbf{v} = 3(2,1) = (6,3), \qquad -2\mathbf{v} = -2(2,1) = (-4,-2),$$
$$\frac{1}{2}\mathbf{v} = \frac{1}{2}(2,1) = \left(1,\frac{1}{2}\right), \qquad -\frac{5}{4}\mathbf{v} = -\frac{5}{4}(2,1) = \left(-\frac{5}{2},-\frac{5}{4}\right).$$

In higher dimensions, scalar multiplication works in exactly the same way algebraically — we just multiply each component of the vector by the scale factor. For example, for $\mathbf{v} = (1,-2,3,-4,5)$ in $\mathbb{R}^5$, we have

$$-17\mathbf{v} = (-17,34,-51,68,-85).$$

□

11.5 Theory

In this section.

- Subsection 11.5.1 *Vector algebra*

11.5.1 Vector algebra

Here we list the basic rules of algebra for vectors in $\mathbb{R}^n$. There is no need to prove these rules like we did for the rules of matrix algebra in Subsection 4.5.1, because we know from Discovery 11.12 that vectors in $\mathbb{R}^n$ can be converted into column matrices and then the vector operations of addition, negation, and scalar multiplication all work as with matrices. And so, since the following rules are all valid when the vectors are replaced by column matrices, they are all valid for vectors in $\mathbb{R}^n$.

Proposition 11.5.1 Rules of vector algebra in $\mathbb{R}^n$. *The following are valid rules of vector algebra. In each statement, assume that* $\mathbf{u}, \mathbf{v}, \mathbf{w}$ *are arbitrary vectors and* $\mathbf{0}$ *is a zero vector, all of the same dimension. Also assume that* k *and* m *are scalars.*

1. *Rules of vector addition.*

   (a) $\mathbf{v} + \mathbf{u} = \mathbf{u} + \mathbf{v}$

   (b) $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$

2. *Rules involving scalar multiplication.*

   (a) $k(\mathbf{u} + \mathbf{v}) = k\mathbf{u} + k\mathbf{v}$

   (b) $(k + m)\mathbf{v} = k\mathbf{v} + m\mathbf{v}$

   (c) $k(m\mathbf{v}) = (km)\mathbf{v}$

   (d) $1\mathbf{v} = \mathbf{v}$

   (e) $(-1)\mathbf{v} = -\mathbf{v}$

   (f) $\mathbf{u} - \mathbf{v} = \mathbf{u} + (-1)\mathbf{v}$

3. *Rules involving a zero vector.*

   (a) $\mathbf{v} + \mathbf{0} = \mathbf{v}$

   (b) $\mathbf{v} - \mathbf{v} = \mathbf{0}$

   (c) $k\mathbf{0} = \mathbf{0}$

CHAPTER 12

Geometry of vectors

12.1 Discovery guide

Discovery 12.1

(a) Draw the vector $\mathbf{u} = (3,2)$ in the xy-plane, then draw a representation of the decomposition $\mathbf{u} = 3\mathbf{e}_1 + 2\mathbf{e}_2$, where $\mathbf{e}_1$ and $\mathbf{e}_2$ are the standard basis vectors in $\mathbb{R}^2$.

Then call on the help of some dead Greek dude to help you compute the length of $\mathbf{u}$.

(b) Does the same method work to determine the length of $\mathbf{w} = (3,-2)$? (And what is the point of checking this case?)

(c) In general, the formula for the length of a two-dimensional vector $\mathbf{v} = (v_1, v_2)$ is .

(d) The same sort of formula works for in three or more dimensions. Fill in the general formulas below.

- The length of $\mathbf{v} = (v_1, v_2, v_3)$ is .
- The "length" of $\mathbf{v} = (v_1, v_2, v_3, v_4)$ is .
- The "length" of $\mathbf{v} = (v_1, v_2, \dots, v_n)$ is .

The quantity for which we developed formulas in Discovery 12.1 is called the **norm** of $\mathbf{v}$, and is denoted $\|\mathbf{v}\|$. (We don't use the word "length" for $n > 3$ — how do you measure length in four dimensions?)

Discovery 12.2

(a) Rewrite your last, general formula from Discovery 12.1.d:

$$\text{for } \mathbf{v} = (v_1, v_2, \dots, v_n), \quad \|\mathbf{v}\| = \qquad .$$

Now square this formula:

$$\text{for } \mathbf{v} = (v_1, v_2, \dots, v_n), \quad \|\mathbf{v}\|^2 = \qquad .$$

(b) Describe the pattern of your formula for $\|\mathbf{v}\|^2$ *in words* without using any letter variables:

the square of the norm of a vector is equal to

.

Discovery 12.3 In this activity, make sure you can answer the questions for *all* dimensions, and make sure you can justify your answer using the formula for norm from Discovery 12.2, not just geometrically.

(a) Can $\|\mathbf{v}\|$ ever be negative?

(b) What is $\|\mathbf{0}\|$? Is $\mathbf{0}$ the only vector that has this value for its norm?

(c) Complete the formulas:

- $\|2\mathbf{v}\| = \underline{\qquad} \|\mathbf{v}\|$
- $\|-2\mathbf{v}\| = \underline{\qquad} \|\mathbf{v}\|$
- $\|k\mathbf{v}\| = \underline{\qquad} \|\mathbf{v}\|$

Discovery 12.4 A **unit vector** is one whose norm is equal to 1.

(a) Verify that the standard basis vectors are all unit vectors, in all dimensions.

(b) Fill in the blanks with an appropriate scalar multiple.

- If $\|\mathbf{u}\| = 1/2$, then $\underline{\qquad}\,\mathbf{u}$ is a unit vector.
- If $\|\mathbf{w}\| = 2$, then $\underline{\qquad}\,\mathbf{w}$ is a unit vector.
- For every nonzero $\mathbf{v}$, $k\mathbf{v}$ is a unit vector for *both* $k = \underline{\qquad}$ and $k = \underline{\qquad}$.

Discovery 12.5 Plot points $P(1,3)$ and $Q(4,-1)$ in the xy-plane. Now draw in the vectors $\mathbf{u}$ and $\mathbf{v}$ that correspond to $\overrightarrow{OP}$ and $\overrightarrow{OQ}$. Complete the triangle by drawing a vector between P and Q. Do you remember how to express this vector as a combination of $\mathbf{u}$ and $\mathbf{v}$? Now compute the distance between P and Q by computing the norm of this third vector.

Recall that in math we measure angles in *radians*. Here are some common conversions:

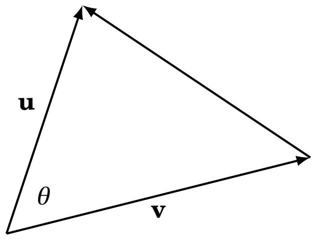
$$
\begin{aligned}
30^\circ &= \pi/6 \text{ rad}, & 90^\circ &= \pi/2 \text{ rad},\\
45^\circ &= \pi/4 \text{ rad}, & 180^\circ &= \pi \text{ rad}.\\
60^\circ &= \pi/3 \text{ rad}, & &
\end{aligned}
$$

Discovery 12.6

(a) In the xy-plane, what is the angle between $\mathbf{e}_1$ and $\mathbf{e}_2$? ... between $\mathbf{e}_1$ and $\mathbf{u} = (1,1)$? ... between $\mathbf{e}_1$ and $2\mathbf{e}_1$? ... between $\mathbf{e}_1$ and $-\mathbf{e}_2$? ... between $\mathbf{e}_1$ and $\mathbf{v} = (1,-1)$? ... between $\mathbf{e}_1$ and $-\mathbf{e}_1$?

(b) Fill in the blanks: an angle θ between a pair of two-dimensional vectors should satisfy $\underline{\qquad} \le \theta \le \underline{\qquad}$.

Discovery 12.7 In the diagram below, consider $\mathbf{u}$ and $\mathbf{v}$ to be two-dimensional vectors. Label the third vector with the appropriate combination of $\mathbf{u}$ and $\mathbf{v}$, just as you did in Discovery 12.5.



There is a version of Pythagoras that applies here even though $\theta \neq 90°$, called **the law of cosines**:

$$a^2 + b^2 - c^2 = 2ab\cos\theta,$$

where a is the length of $\mathbf{u}$, b is the length of $\mathbf{v}$, and c is the length of the "hypotenuse" across from θ. (If θ were 90°, the right-hand side of this equality would be zero and this law would "collapse" to the same equality as Pythagoras.)

Use the formulas from Discovery 12.2 to rewrite the left-hand side of the law of cosines in terms of the components of $\mathbf{u} = (u_1, u_2)$ and $\mathbf{v} = (v_1, v_2)$, then simplify until you get

$$2 \times (\text{simple formula}).$$

Using the new expression $2 \times (\text{simple formula})$ from Discovery 12.7 as the left-hand side in the law of cosines, and dividing both sides by $2ab$, we get

$$\cos\theta = \frac{(\text{simple formula})}{\|\mathbf{u}\|\,\|\mathbf{v}\|}.$$

(Remember that a and b are the lengths of $\mathbf{u}$ and $\mathbf{v}$, respectively.)

The "simple formula" part of this angle formula turns out to be an important one — it is called the **Euclidean inner product** or **standard inner product** (or just simply the **dot product**) of $\mathbf{u}$ and $\mathbf{v}$, and written $\mathbf{u}\cdot\mathbf{v}$.

Discovery 12.8 Let's extend the computational pattern from Discovery 12.7. In the two-dimensional case in Task a below, you should just enter the "simple formula" you discovered above. In the subsequent tasks in higher dimensions, use the pattern from the two-dimensional case to create a similar higher-dimensional formula.

(a) *In two dimensions.*

For $\mathbf{u} = (u_1, u_2)$, $\mathbf{v} = (v_1, v_2)$: $\mathbf{u}\cdot\mathbf{v} =$.

(b) *In three dimensions.*

For $\mathbf{u} = (u_1, u_2, u_3)$, $\mathbf{v} = (v_1, v_2, v_3)$: $\mathbf{u}\cdot\mathbf{v} =$.

(c) *In four dimensions.*

For $\mathbf{u} = (u_1, u_2, u_3, u_4)$, $\mathbf{v} = (v_1, v_2, v_3, v_4)$: $\mathbf{u}\cdot\mathbf{v} =$.

(d) *Arbitrary dimension.*

For $\mathbf{u} = (u_1, u_2, \dots, u_n)$, $\mathbf{v} = (v_1, v_2, \dots, v_n)$: $\mathbf{u}\cdot\mathbf{v} =$.

Discovery 12.9 What is the formula for the dot product of a vector with itself?

For $\mathbf{v} = (v_1, v_2, \dots, v_n)$, $\mathbf{v}\cdot\mathbf{v} =$.

Compare your answer with Discovery 12.2.

Discovery 12.10 Using the formula for the dot product for two-dimensional vectors, verify that it has the following properties.

Remember. When verifying an equality, make sure to use proper LHS versus RHS procedure!

(a) $\mathbf{v}\cdot\mathbf{u} = \mathbf{u}\cdot\mathbf{v}$.

(b) $\mathbf{u}\cdot(\mathbf{v} + \mathbf{w})$.

(c) $k(\mathbf{u}\cdot\mathbf{v}) = (k\mathbf{u})\cdot\mathbf{v} = \mathbf{u}\cdot(k\mathbf{v})$.

(d) $\mathbf{0}\cdot\mathbf{v} = 0$.

Do you think all these properties will still be true for higher-dimensional vectors?

Discovery 12.11

(a) For two-dimensional column vectors $\mathbf{u} = \left[\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right]$ and $\mathbf{v} = \left[\begin{smallmatrix} v_1 \\ v_2 \end{smallmatrix}\right]$, compute the matrix product $(\mathbf{u}^{\mathrm{T}})\mathbf{v}$.

What do you notice? Do you think the same will happen for higher-dimensional column vectors?

(b) Suppose $\mathbf{u}$ and $\mathbf{v}$ are n-dimensional column vectors and A is an $n \times n$ matrix. Use what you discovered in Task a to fill in the blank:

$$(A\mathbf{u}) \cdot \mathbf{v} = \mathbf{u} \cdot (\qquad \mathbf{v}).$$

12.2 Terminology and notation

norm (of a vector v)

the quantity $\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$; also called the **length** or **magnitude** of **v**

unit vector

a vector whose norm is equal to 1

normalization (of a vector v)

the unit vector $\dfrac{1}{\|\mathbf{v}\|}\mathbf{v}$

distance (between two vectors u and v)

the distance between the terminal points of the two vectors when their initial points are placed at the same point; can be computed as $\|\mathbf{u} - \mathbf{v}\|$ (or equivalently as $\|\mathbf{v} - \mathbf{u}\|$)

dot product (of two vectors u and v of the same dimension)

the quantity

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n;$$

also referred to as the **Euclidean inner product** or **standard inner product** of **u** and **v**

angle (between two vectors u and v of the same dimension)

the angle θ satisfying both

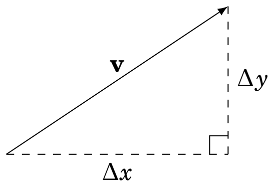
$$0 \le \theta \le \pi \qquad \text{and} \qquad \cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|}$$

12.3 Concepts

In this section.

12.3.1 Geometric length of a vector: the norm

We can easily determine the length of a vector in the plane from its components using the Pythagorean Theorem.



If we let ℓ represent the length of $\mathbf{v}$, then Pythagoras tells us that

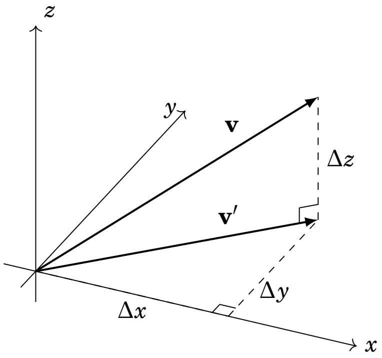
$$\ell^2 = (\Delta x)^2 + (\Delta y)^2.$$

Remember. There is no such operation as *squaring* a vector, so it would be incorrect to write $\mathbf{v}^2 = (\Delta x)^2 + (\Delta y)^2$.

We write $\|\mathbf{v}\|$ to mean the length of the vector $\mathbf{v}$ in the plane. Keep in mind in all that follows that $\|\mathbf{v}\|$ is always a *single number*, since it measures a length. If $\mathbf{v}$ has components $\mathbf{v} = (v_1, v_2)$ (where $v_1 = \Delta x$ and $v_2 = \Delta y$), then solving for ℓ in the Pythagorean equation above gives us

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2}.$$

For a vector $\mathbf{v} = (v_1, v_2, v_3)$ in $\mathbb{R}^3$, consider the vector $\mathbf{v}' = (v_1, v_2, 0)$ sitting in the xy-plane.



Applying the Pythagorean Theorem to the vertical triangle, we find

$$\|\mathbf{v}\|^2 = \left\|\mathbf{v}'\right\|^2 + (\Delta z)^2.$$

But $\mathbf{v}'$ lies flat in the xy-plane, and we have already analyzed that case above:

$$\left\|\mathbf{v}'\right\|^2 = (\Delta x)^2 + (\Delta y)^2.$$

Combining these, we get

$$\|\mathbf{v}\|^2 = \left((\Delta x)^2 + (\Delta y)^2\right) + (\Delta z)^2 = v_1^2 + v_2^2 + v_3^2,$$

so that

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + v_3^2}.$$

The word *length* ceases to have any meaning in $\mathbb{R}^4$, so in general we refer to $\|\mathbf{v}\|$ as the **norm** of $\mathbf{v}$ in any dimension. We imagine that if we were able to somehow measure length in $\mathbb{R}^n$ for $n \geq 4$, then the pattern where we used length in $\mathbb{R}^2$ to help us compute length in $\mathbb{R}^3$ would be repeated, and we would be able to use length in $\mathbb{R}^3$ to help us compute "length" in $\mathbb{R}^4$, and then we would be able to use "length" in $\mathbb{R}^4$ to help us compute "length" in $\mathbb{R}^5$, and so on. So it seems reasonable to define the **norm** of a vector $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ in $\mathbb{R}^n$ to be

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}.$$

Square roots are annoying to work with algebraically, so we often work with the square of a norm, for which we developed the formula

$$\|\mathbf{v}\|^2 = v_1^2 + v_2^2 + \cdots + v_n^2$$

in Discovery 12.2.

12.3.2 Properties of the norm

We explored some other basic properties of the norm in Discovery 12.3. First, when we take the square root of a nonzero number, we always take the *positive* square root, so a norm is never a negative number. This property agrees with our conception of norm as a length in $\mathbb{R}^2$ and $\mathbb{R}^3$, since in geometry we usually require lengths to be nonnegative.

Second, the zero vector $\mathbf{0} = (0,0,\dots,0)$ always has norm 0 in every dimension, since

$$\|\mathbf{0}\| = \sqrt{0^2 + 0^2 + \cdots + 0^2} = \sqrt{0} = 0.$$

And it is the *only* vector that has norm 0, since as soon as one of the components of a vector is nonzero, the sum of squares under the square root sign in the norm formula will be a positive number. There is no possibility of cancellation to zero under the square root, even if a vector has a mix of positive and negative components, because squaring the components will never have negative results.

Finally, we considered the effect of a scalar multiplication on norm. Geometrically, in $\mathbb{R}^2$ and $\mathbb{R}^3$ we think of scalar multiplication as scaling a vector's length by some scale factor k, so we should expect the numerical norm of a vector to be multiplied by the scale factor. And that is (almost) exactly what happens:

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}, \qquad \begin{aligned} \|k\mathbf{v}\| &= \sqrt{(kv_1)^2 + (kv_2)^2 + \cdots + (kv_n)^2} \\ &= \sqrt{k^2v_1^2 + k^2v_2^2 + \cdots + k^2v_n^2} \\ &= \sqrt{k^2(v_1^2 + v_2^2 + \cdots + v_n^2)} \\ &= \sqrt{k^2}\sqrt{v_1^2 + v_2^2 + \cdots + v_n^2} \\ &= \sqrt{k^2}\,\|\mathbf{v}\|. \end{aligned}$$

We need to be a little careful with the last step, because *it is not always true that* $\sqrt{k^2} = k$. In particular, the result of $\sqrt{k^2}$ is never negative, so if k is negative then it is impossible for $\sqrt{k^2}$ to be equal to k. The proper formula for all values of k is $\sqrt{k^2} = |k|$, so our norm formula becomes

$$\|k\mathbf{v}\| = |k|\,\|\mathbf{v}\|.$$

12.3.3 Unit vectors and normalization

In the plane or in space, a vector with length 1 is convenient geometrically because it can be used as a "meter stick" — every scalar multiple of that vector will have length equal to the (absolute value of) the scale factor. For example, if $\mathbf{u}$ has length 1, then both $3\mathbf{u}$ and $-3\mathbf{u}$ have length 3. The same pattern will hold in any dimension when we replace the word "length" with "norm." A vector with norm 1 is called a **unit vector**. One of the reasons the standard basis vectors are so special is that each of them is a unit vector, as we saw in Discovery 12.4. Thus each standard basis vector can be used as a “meter stick” along the corresponding axis.

We also explored how to scale a nonzero vector to a unit vector in Discovery 12.4. For example, if a vector has norm 1/2, then we can scale it up to a unit vector by multiplying it by 2 to double its norm. Conversely, if a vector has norm 2, we can scale it down to a unit vector by multiplying it by 1/2 to halve its norm. In general, we can scale any nonzero vector $\mathbf{v}$ in $\mathbb{R}^n$ up or down to a unit vector by multiplying it by scale factor $k = \frac{1}{\|\mathbf{v}\|}$, since then

$$\left\| \frac{1}{\|\mathbf{v}\|}\mathbf{v} \right\| = \left| \frac{1}{\|\mathbf{v}\|} \right| \|\mathbf{v}\| = \frac{1}{\|\mathbf{v}\|}\,\|\mathbf{v}\| = 1.$$

In the above, we have used the formula for the norm of a scalar multiple, $\|k\mathbf{v}\| = |k|\,\|\mathbf{v}\|$, with $k = \frac{1}{\|\mathbf{v}\|}$. The absolute value brackets on this particular scalar k can be removed because norms are never negative, and so $|k| = k$ in this case.

Remember. We should never divide by a vector because there is no such vector operation (see Warning 11.3.2). But $\|\mathbf{v}\|$ is a *number*, not a *vector*, so $k = \frac{1}{\|\mathbf{v}\|}$ is valid.

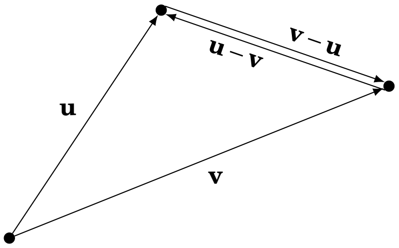
In fact, every nonzero vector $\mathbf{v}$ is parallel to exactly two corresponding unit vectors, because $k\mathbf{v}$ and $-k\mathbf{v}$ always have the same norm. So

$$\mathbf{u}_1 = \frac{1}{\|\mathbf{v}\|}\mathbf{v}, \qquad \text{and} \qquad \mathbf{u}_2 = -\frac{1}{\|\mathbf{v}\|}\mathbf{v}$$

are always unit vectors, as long as $\mathbf{v} \neq \mathbf{0}$.

12.3.4 Distance between vectors

As we saw in Subsection 11.3.4, if we position $\mathbf{u}$ and $\mathbf{v}$ to share the same initial points, then the difference vectors $\mathbf{u}-\mathbf{v}$ and $\mathbf{v}-\mathbf{u}$ run between the terminal points of $\mathbf{u}$ and $\mathbf{v}$.

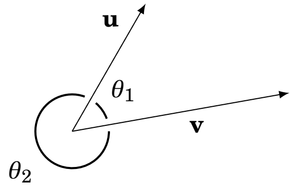


So we can measure the distance between the terminal points of $\mathbf{u}$ and $\mathbf{v}$ by computing $\|\mathbf{u}-\mathbf{v}\|$ or $\|\mathbf{v}-\mathbf{u}\|$, as we discovered in Discovery 12.5. This process is even more straightforward when the common initial point of $\mathbf{u}$ and $\mathbf{v}$ is chosen to be the origin, so that the components of $\mathbf{u}$ and $\mathbf{v}$ are the same as the coordinates of their respective terminal points.

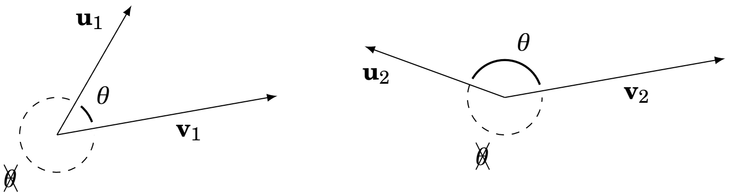
Remark 12.3.1 The analysis above illustrates a useful strategy to compute distances in the plane or in space: determine some vector that traverses the distance in question, and then compute the norm of that vector to obtain the desired distance. Combined with some of the vector geometry that we will develop in the next few chapters, this strategy is often easier than trying to determine the coordinates of the points at the endpoints of the desired distance. You should remember this strategy when we explore the geometry of lines and planes in Chapters 13–14.

12.3.5 Angle between vectors in the plane and in space

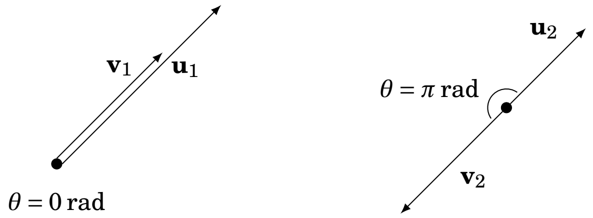
Two vectors in the plane, when given the same initial point, have two angles between them.



We only need to know one of these two angles, since the other can be computed from the knowledge that the sum of the two angles is 2π radians. We generally prefer to avoid ambiguity in math, so it would be nice to have a systematic way to choose one of the two angles between a pair of vectors that we can refer to as *the* angle between the vectors. We will not distinguish between clockwise and counterclockwise, because those terms will become meaningless when we move up a dimension. Instead we will always choose the smaller angle to be *the* angle between the two vectors.



Thus, the angle between two vectors in the plane will always be between 0 and π radians. Note that it is possible for the angle to be *exactly* 0 radians or *exactly* π radians, in the case the the two vectors are parallel.



How can we measure the angle between vectors in three-dimensional space?

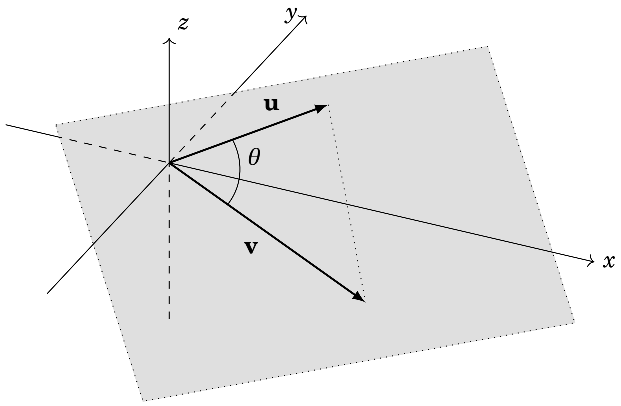


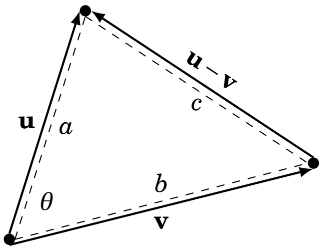
Figure 12.3.2 Diagram of the angle between vectors in space, embedded in a plane.

In space, two vectors that are positioned to share the same initial point can be completed to a triangle, and that triangle will lie in a plane. The angle between the two vectors can then be taken to be the smaller of the two angles between the two vectors in that shared plane.

Note. In Figure 12.3.2, you should imagine the shaded surface passing through the origin, with the two vectors $\mathbf{u}$ and $\mathbf{v}$ lying flat in it.

12.3.6 Dot product

In Discovery 12.7, we combined vector geometry with some high school geometry to determine a formula for the (cosine of the) angle between two plane vectors. Recall from Subsection 11.3.4 that a vector that runs between the terminal points of two vectors that share an initial point is a difference vector.



The lengths of the sides of the triangle formed by these three vectors are just the norms of the vectors:

$$a = \|\mathbf{u}\|, \qquad b = \|\mathbf{v}\|, \qquad c = \|\mathbf{u}-\mathbf{v}\|.$$

The Law of Cosines applied to this triangle says that $a^2 + b^2 - c^2 = 2ab\cos\theta$.

Careful. It would be *nonsense* to write this as $\mathbf{u}^2 + \mathbf{v}^2 - (\mathbf{u}-\mathbf{v})^2 = 2\mathbf{u}\mathbf{v}\cos\theta$, because there are no such operations as multiplying or squaring vectors.

Let's give our plane vectors some components so that we can work with this equality:

$$\mathbf{u} = (u_1, u_2), \qquad \mathbf{v} = (v_1, v_2), \qquad \mathbf{u}-\mathbf{v} = (u_1 - v_1, u_2 - v_2).$$

Now we have

$$\begin{aligned} a^2 &= \|\mathbf{u}\|^2 & b^2 &= \|\mathbf{v}\|^2 & c^2 &= \|\mathbf{u}-\mathbf{v}\|^2 \\ &= u_1^2 + u_2^2, & &= v_1^2 + v_2^2, & &= (u_1 - v_1)^2 + (u_2 - v_2)^2 \\ & & & & &= u_1^2 - 2u_1v_1 + v_1^2 + u_2^2 - 2u_2v_2 + v_2^2, \end{aligned}$$

and so after some cancelling we have

$$a^2 + b^2 - c^2 = 2u_1v_1 + 2u_2v_2.$$

Using the expression on the right above for the left-hand side of the equality $a^2 + b^2 - c^2 = 2ab\cos\theta$ for $\cos\theta$, solving for $\cos\theta$, and then substituting $a = \|\mathbf{u}\|$ and $b = \|\mathbf{v}\|$ leads to

$$\cos\theta = \frac{u_1v_1 + u_2v_2}{\|\mathbf{u}\|\,\|\mathbf{v}\|}. \tag{$*$}$$

The expression on the left and the denominator on the right are both familiar — we have the ordinary cosine function from trigonometry and we have some vector norms. However, before we worked through Discovery 12.7, the expression in the numerator on the right-hand side was unknown.

Earlier in this chapter, we mentioned how two vectors in space with their initial points at the origin lie inside a common flat plane (see Figure 12.3.2). If we repeated the above geometric analysis of vector angle in this flat surface inside space, we would come to a similar conclusion:

$$\cos\theta = \frac{u_1v_1 + u_2v_2 + u_3v_3}{\|\mathbf{u}\|\,\|\mathbf{v}\|}. \tag{$**$}$$

There is an obvious pattern to the numerators on the right-hand sides of equations (∗) and (∗∗). And it seems that the value that these numerator formulas compute is important, since it provides a link between the two most important quantities in geometry: length and angle. So we give it a name, the **dot product** (or the **Euclidean inner product**), and use the symbol • between two vectors to represent this quantity. The formula can obviously be extended to higher dimensions than just the plane $\mathbb{R}^2$ and space $\mathbb{R}^3$, so we will do just that:

$$\mathbf{u}\cdot\mathbf{v} = u_1v_1 + u_2v_2 + \cdots + u_nv_n.$$

Warning 12.3.3 The result of the computation $\mathbf{u}\cdot\mathbf{v}$ is a *number*, which is important to keep in mind if you are working algebraically with an expression containing a dot product. See Proposition 12.5.3 in Subsection 12.5.1 for algebraic rules involving the dot product.

12.3.7 Angle between vectors in $\mathbb{R}^n$

Even though we can't "see" geometry in $\mathbb{R}^n$ for $n > 3$, we have already seen that we can perform computations related to geometry in these spaces. We can attach the number $\|\mathbf{v}\|$ to a vector $\mathbf{v}$ in $\mathbb{R}^n$ that can be interpreted as its "length." And for two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$, we can compute the number $\mathbf{u}\cdot\mathbf{v}$ that is somehow related to the geometric relationship between $\mathbf{u}$ and $\mathbf{v}$. We have seen that in the plane and in space, $\mathbf{u}\cdot\mathbf{v}$ links the lengths of $\mathbf{u}$ and $\mathbf{v}$ to the angle between them. But do higher-dimensional vectors have angles between them? Is there some number that we can attach to $\mathbf{u}$ and $\mathbf{v}$ that "measures" the angle between them, even if we can't see or measure this angle directly?

The equalities in (∗) and (∗∗) suggest a pattern we can copy into $\mathbb{R}^n$ in general. We *define* the angle between $\mathbf{u}$ and $\mathbf{v}$ to be the unique angle θ, between 0 and π, that makes

$$\cos\theta \qquad \text{and} \qquad \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|} \tag{$***$}$$

equal.

Question 12.3.4

- For every pair of vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$, can we always determine a suitable angle θ in the domain $0 \le \theta \le \pi$ that works (i.e. that makes the two quantities in (∗∗∗) equal)?
- For some pair of vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$, might it be possible that there are *several* values of θ in the domain $0 \le \theta \le \pi$ that work?

□

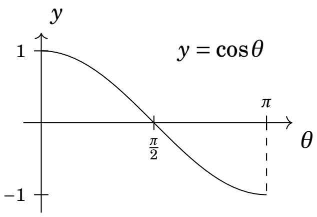
Fortunately, for a pair of (nonzero) plane vectors or space vectors, there is *exactly one* number (once we restrict to the domain $0 \le \theta \le \pi$) that gets to call itself *the* angle between the vectors. It would not bode well for the possibility of somehow doing geometry in higher-dimensional spaces if there were sometimes *two* numbers that could be reasonably called "the angle" between a pair of vectors, or sometimes *none* at all. Luckily neither of these is possible.

First, for a pair of nonzero vectors in $\mathbb{R}^n$, the formula

$$\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|}$$

can always be computed, and the result of the computation is always a single, definite number.

Second, looking at the provided graph of $y = \cos\theta$, there are no instances in the domain $0 \le \theta \le \pi$ where $\cos\theta$ computes to the *same* value for two *different* values of θ.



On this domain, we call the graph *one-to-one*. So a pair of vectors in $\mathbb{R}^n$ can never have *two* angles in the domain $0 \le \theta \le \pi$ between them, because there are never two solutions to the equation

$$\cos\theta = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|} \tag{†}$$

in that domain.

But is there always *some* solution to equation (†)? No matter what domain you work on, $\cos\theta$ never evaluates to a number greater than 1 or less than −1. Perhaps if we tried hard enough we could discover some unlucky pair of vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^{13}$ where

$$\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|}$$

computed to a number greater than 1 or to a number less than −1. In that case, it would be impossible for $\cos\theta$ to be equal to that number, and $\mathbf{u}$ and $\mathbf{v}$ would have *no* angle between them. It turns out that forming such an unlucky pair of vectors is impossible, and we know this courtesy of a couple of dead guys.

Theorem 12.3.5 The Cauchy-Schwarz inequality. *For every pair of vectors* $\mathbf{u}$ *and* $\mathbf{v}$ *in* $\mathbb{R}^n$*, the quantity*

$$\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|}$$

is always between −1 *and* 1 *(inclusive).*

Note. The "inequality" part of the Cauchy-Schwarz inequality is usually stated as $|\mathbf{u}\cdot\mathbf{v}| \le \|\mathbf{u}\|\,\|\mathbf{v}\|$, which for nonzero vectors $\mathbf{u}$ and $\mathbf{v}$ is equivalent to

$$-1 \le \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|} \le 1.$$

Since the graph $y = \cos\theta$ passes through every possible y-value in the range $-1 \le y \le 1$, and does so only once, equation (†) always has *one unique* solution for a pair of nonzero vectors.

12.3.8 Dot product versus norm

We have already seen that the dot product is intimately tied to the geometry of $\mathbb{R}^n$, acting as a link between norm (length) and angle. But as we discovered in Discovery 12.9, it is also directly linked to the norm by the observation

$$\|\mathbf{v}\|^2 = \left(\sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}\right)^2 \qquad \mathbf{v}\cdot\mathbf{v} = v_1v_1 + v_2v_2 + \cdots + v_nv_n$$

$$= v_1^2 + v_2^2 + \cdots + v_n^2, \qquad\qquad = v_1^2 + v_2^2 + \cdots + v_n^2.$$

So we obtain a very convenient formula: $\|\mathbf{v}\|^2 = \mathbf{v} \cdot \mathbf{v}$.

Remark 12.3.6 Really, this "new" link between dot product and norm is just the special case of equation (†) where **u** is taken to be equal to **v**, since in this case the angle between **u** and **v** (i.e. between **v** and itself) is zero, and $\cos 0 = 1$.

12.3.9 Dot product as matrix multiplication

The pattern in the formula for the dot product of two vectors should look vaguely familiar to you — it is a sum of products, which is exactly the pattern of the left-hand side of a linear equation, and so also the pattern in our "row-times-column" view of matrix multiplication in Subsection 4.3.7. In fact, the dot product can be defined in terms of matrix multiplication if we take our vectors to be *column vectors* and use the transpose to turn one of the columns into a row. Indeed, for

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}, \qquad \mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix},$$

we have

$$\begin{aligned} \mathbf{v}^{\mathrm{T}}\mathbf{u} &= \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \\ &= v_1u_1 + v_2u_2 + \cdots + v_nu_n \\ &= u_1v_1 + u_2v_2 + \cdots + u_nv_n \\ &= \mathbf{u} \cdot \mathbf{v}. \end{aligned}$$

So we obtain a matrix formula for dot product: $\mathbf{u} \cdot \mathbf{v} = \mathbf{v}^{\mathrm{T}}\mathbf{u}$.

Remark 12.3.7

- Technically, the result of multiplying the $1 \times n$ matrix $\mathbf{v}^{\mathrm{T}}$ and the $n \times 1$ matrix **u** should be a 1×1 matrix. But algebraically there is no difference between numbers and 1×1 matrices with respect to the operations of addition, subtraction, and multiplication, so it is common to think of a 1×1 matrix as just a number, as we did above.

- It might seem more natural to use

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^{\mathrm{T}}\mathbf{v} = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = u_1v_1 + u_2v_2 + \cdots + u_nv_n$$

  (as we did in Discovery 12.11), instead of the seemingly pointless reversal of order in the formula $\mathbf{u} \cdot \mathbf{v} = \mathbf{v}^{\mathrm{T}}\mathbf{u}$. However, if you continue on in your study of linear algebra beyond this course, you will discover that this reversal of order is necessary when studying *complex* vectors (that is, vectors where the components are *complex* numbers). Since this reversal of order is harmless here, we will start using it now so as to avoid confusion later.

12.4 Examples

In this section.

- Subsection 12.4.1 *The norm of a vector*
- Subsection 12.4.2 *Dot product and the angle between vectors*

12.4.1 The norm of a vector

Example 12.4.1 Basic computation examples. Here are a few examples of computing the norm of a vector, in various dimensions.

1. Consider $\mathbf{u} = (1,2)$ in $\mathbb{R}^2$. Then,
$$\|\mathbf{u}\| = \sqrt{1^2 + 2^2} = \sqrt{5}.$$
2. Consider $\mathbf{v} = (1,2,-1)$ in $\mathbb{R}^3$. Then,
$$\|\mathbf{v}\| = \sqrt{1^2 + 2^2 + (-1)^2} = \sqrt{6}.$$
3. Consider $\mathbf{w} = (1,2,-1,5)$ in $\mathbb{R}^4$. Then,
$$\|\mathbf{w}\| = \sqrt{1^2 + 2^2 + (-1)^2 + 5^2} = \sqrt{31}.$$

□

Example 12.4.2 Norms of the standard basis vectors. The standard basis vectors in $\mathbb{R}^n$ are always **unit vectors**:

$$\begin{aligned}
\|\mathbf{e}_1\| &= \sqrt{1^2 + 0^2 + \cdots + 0^2} = \sqrt{1} = 1,\\
\|\mathbf{e}_2\| &= \sqrt{0^2 + 1^2 + 0^2 + \cdots + 0^2} = \sqrt{1} = 1,\\
&\vdots\\
\|\mathbf{e}_n\| &= \sqrt{0^2 + \cdots + 0^2 + 1^2} = \sqrt{1} = 1.
\end{aligned}$$

□

Example 12.4.3 Normalizing vectors. We can scale any nonzero vector to a unit vector by dividing by its norm, and this **normalized** version of the vector will always be parallel to the original.

Let's carry this out for the vectors from Example 12.4.1 above.

1. We computed the norm of $\mathbf{u} = (1,2)$ to be $\|\mathbf{u}\| = \sqrt{5}$. Therefore, the scaled vector
$$\mathbf{u}' = \frac{1}{\sqrt{5}}\mathbf{u} = \left(\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}\right)$$
is a unit vector (i.e. $\left\|\mathbf{u}'\right\| = 1$).
2. We computed the norm of $\mathbf{v} = (1,2,-1)$ to be $\|\mathbf{v}\| = \sqrt{6}$. Therefore, the scaled vector
$$\frac{1}{\sqrt{6}}\mathbf{v} = \left(\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}, -\frac{1}{\sqrt{6}}\right)$$
is a unit vector.

3. We computed the norm of $\mathbf{w} = (1,2,-1,5)$ to be $\|\mathbf{w}\| = \sqrt{31}$. Therefore, the scaled vector
$$\frac{1}{\sqrt{31}}\mathbf{v} = \left(\frac{1}{\sqrt{31}}, \frac{2}{\sqrt{31}}, -\frac{1}{\sqrt{31}}, \frac{5}{\sqrt{31}},\right)$$
is a unit vector.

□

12.4.2 Dot product and the angle between vectors

Here is an example of using the dot product to determine the angle between vectors.

Example 12.4.4 Computing angle from dot product. What is the angle between vectors $\mathbf{u} = (1,2)$ and $\mathbf{v} = (-1,3)$ in $\mathbb{R}^2$?

From Discovery 12.7, we know that the angle θ between $\mathbf{u}$ and $\mathbf{v}$ satisfies

$$\cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}.$$

Also see. Subsection 12.3.7 and Corollary 12.5.5 in Subsection 12.5.2.

So compute

$$\mathbf{u} \cdot \mathbf{v} = 1 \cdot (-1) + 2 \cdot 3 = 5, \qquad \|\mathbf{u}\| = \sqrt{1^2 + 2^2} = \sqrt{5}, \qquad \|\mathbf{v}\| = \sqrt{(-1)^2 + 3^2} = \sqrt{10}.$$

Therefore,

$$\cos\theta = \frac{5}{\sqrt{5}\sqrt{10}} = \frac{1}{\sqrt{2}}.$$

The only angle in the domain $0 \le \theta \le \pi$ with this cosine value is $\theta = \pi/4$. □

12.5 Theory

In this section.

- Subsection 12.5.1 *Norm and dot product*
- Subsection 12.5.2 *Vector geometry inequalities and uniqueness of vector angles*

12.5.1 Norm and dot product

We'll begin with algebraic properties of norm and dot product.

Proposition 12.5.1 Properties of the norm. *The following are true for all vectors* $\mathbf{u}$ *and* $\mathbf{v}$ *in* $\mathbb{R}^n$ *and all scalars* k.

1. $\|\mathbf{v}\| \ge 0$*, and* $\|\mathbf{v}\| = 0$ *only for* $\mathbf{v} = 0$.
2. $\|-\mathbf{v}\| = \|\mathbf{v}\|$.
3. $\|k\mathbf{v}\| = |k| \, \|\mathbf{v}\|$.
4. $\|\mathbf{v} - \mathbf{u}\| = \|\mathbf{u} - \mathbf{v}\|$.

Warning 12.5.2 The norm is *not* additive; that is, it is ***not*** true in general that $\|\mathbf{u} + \mathbf{v}\|$ is equal to $\|\mathbf{u}\| + \|\mathbf{v}\|$. *Sometimes* the two quantities are equal, as you are asked to consider below, but the best we can say about the norm of a sum is contained in Theorem 12.5.6 below.

Proposition 12.5.3 Algebra rules of the dot product. *The following are true for all vectors* $\mathbf{u}$*,* $\mathbf{v}$*, and* $\mathbf{w}$ *in* $\mathbb{R}^n$*, and all scalars* k*.*

1. $\mathbf{v}\cdot\mathbf{u}=\mathbf{u}\cdot\mathbf{v}$.
2. $\mathbf{u}\cdot(\mathbf{v}+\mathbf{w})=\mathbf{u}\cdot\mathbf{v}+\mathbf{u}\cdot\mathbf{w}$.
3. $\mathbf{u}\cdot(\mathbf{v}-\mathbf{w})=\mathbf{u}\cdot\mathbf{v}-\mathbf{u}\cdot\mathbf{w}$.
4. $(\mathbf{u}+\mathbf{v})\cdot\mathbf{w}=\mathbf{u}\cdot\mathbf{w}+\mathbf{v}\cdot\mathbf{w}$.
5. $(\mathbf{u}-\mathbf{v})\cdot\mathbf{w}=\mathbf{u}\cdot\mathbf{w}-\mathbf{v}\cdot\mathbf{w}$.
6. $(k\mathbf{u})\cdot\mathbf{v}=k(\mathbf{u}\cdot\mathbf{v})$.
7. $\mathbf{u}\cdot(k\mathbf{v})=k(\mathbf{u}\cdot\mathbf{v})$.
8. $\mathbf{v}\cdot\mathbf{v}=\|\mathbf{v}\|^2$.
9. *Both* $\mathbf{v}\cdot\mathbf{0}=0$ *and* $\mathbf{0}\cdot\mathbf{v}=0$.

12.5.2 Vector geometry inequalities and uniqueness of vector angles

The Cauchy-Schwarz inequality. Here we will state the Cauchy-Schwarz inequality in its usual form. Note that this version applies to *every* pair of vectors, even if one is (or both are) the zero vector.

Theorem 12.5.4 The Cauchy-Schwarz inequality. *For every pair of vectors* $\mathbf{u}$ *and* $\mathbf{v}$ *in* $\mathbb{R}^n$*, we have* $|\mathbf{u}\cdot\mathbf{v}|\le\|\mathbf{u}\|\,\|\mathbf{v}\|$*.*

Proof. We will show that $(\mathbf{u}\cdot\mathbf{v})^2\le(\|\mathbf{u}\|\,\|\mathbf{v}\|)^2$. Once this is established, then for $\mathbf{u}\cdot\mathbf{v}$ to have a smaller square than $\|\mathbf{u}\|\,\|\mathbf{v}\|$, it must be smaller in magnitude. That is, $(\mathbf{u}\cdot\mathbf{v})^2\le(\|\mathbf{u}\|\,\|\mathbf{v}\|)^2$ can only be true if $|\mathbf{u}\cdot\mathbf{v}|\le\big|\|\mathbf{u}\|\,\|\mathbf{v}\|\big|$ is true. But since neither $\|\mathbf{u}\|$ nor $\|\mathbf{v}\|$ can be negative, we have $\big|\|\mathbf{u}\|\,\|\mathbf{v}\|\big|=\|\mathbf{u}\|\,\|\mathbf{v}\|$, and so

$$|\mathbf{u}\cdot\mathbf{v}|\le\|\mathbf{u}\|\,\|\mathbf{v}\|$$

will be established.

So, we will try to prove that $(\mathbf{u}\cdot\mathbf{v})^2\le(\|\mathbf{u}\|\,\|\mathbf{v}\|)^2$ is always true for every pair of vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$. We might as well assume that $\mathbf{v}$ is nonzero, since if it is zero then both $(\mathbf{u}\cdot\mathbf{v})^2$ and $(\|\mathbf{u}\|\,\|\mathbf{v}\|)^2$ are 0, and the required inequality is true. In the case that $\mathbf{v}$ is nonzero, then also $\|\mathbf{v}\|\neq0$ (Statement 1 of Proposition 12.5.1), and we can form the vector

$$\mathbf{w}=\mathbf{u}-a\mathbf{v},\qquad\qquad \text{where } a=\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{v}\|^2}$$

without worry that we've accidentally divided by zero. We will find that $\|\mathbf{w}\|^2$ is related to the inequality we are trying to prove, so compute

$$\begin{align}
\|\mathbf{w}\|^2 &= \mathbf{w}\cdot\mathbf{w} \tag{i}\\
&= (\mathbf{u}-a\mathbf{v})\cdot(\mathbf{u}-a\mathbf{v}) \tag{ii}\\
&= (\mathbf{u}-a\mathbf{v})\cdot\mathbf{u}-(\mathbf{u}-a\mathbf{v})\cdot(a\mathbf{v}) \tag{iii}\\
&= \mathbf{u}\cdot\mathbf{u}-(a\mathbf{v})\cdot\mathbf{u}-\big(\mathbf{u}\cdot(a\mathbf{v})-(a\mathbf{v})\cdot(a\mathbf{v})\big) \tag{iv}\\
&= \mathbf{u}\cdot\mathbf{u}-a(\mathbf{v}\cdot\mathbf{u})-a(\mathbf{u}\cdot\mathbf{v})+a\big(a(\mathbf{v}\cdot\mathbf{v})\big) \tag{v}\\
&= \mathbf{u}\cdot\mathbf{u}-a(\mathbf{u}\cdot\mathbf{v})-a(\mathbf{u}\cdot\mathbf{v})+a^2(\mathbf{v}\cdot\mathbf{v}) \tag{vi}\\
&= \|\mathbf{u}\|^2-2a(\mathbf{u}\cdot\mathbf{v})+a^2\,\|\mathbf{v}\|^2 \tag{vii}\\
&= \|\mathbf{u}\|^2-2\left(\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{v}\|^2}\right)(\mathbf{u}\cdot\mathbf{v})+\left(\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{v}\|^2}\right)^2\|\mathbf{v}\|^2 \tag{viii}\\
&= \|\mathbf{u}\|^2-2\frac{(\mathbf{u}\cdot\mathbf{v})^2}{\|\mathbf{v}\|^2}+\frac{(\mathbf{u}\cdot\mathbf{v})^2}{\|\mathbf{v}\|^2}, \notag\\
&= \|\mathbf{u}\|^2-\frac{(\mathbf{u}\cdot\mathbf{v})^2}{\|\mathbf{v}\|^2}, \notag
\end{align}$$

with justifications

(i) Rule 8 of Proposition 12.5.3;

(ii) using the definition of $\mathbf{w}$ above;

(iii) Rule 2 of Proposition 12.5.3;

(iv) Rule 4 of Proposition 12.5.3;

(v) Rule 6 and Rule 7 of Proposition 12.5.3;

(vi) Rule 1 of Proposition 12.5.3;

(vii) Rule 8 of Proposition 12.5.3; and

(viii) using the definition of a above.

Now, $\|\mathbf{w}\|^2$ cannot be negative, so we have

$$\begin{aligned} 0 &\le \|\mathbf{u}\|^2 - \frac{(\mathbf{u}\cdot\mathbf{v})^2}{\|\mathbf{v}\|^2} \\ \frac{(\mathbf{u}\cdot\mathbf{v})^2}{\|\mathbf{v}\|^2} &\le \|\mathbf{u}\|^2 \\ (\mathbf{u}\cdot\mathbf{v})^2 &\le \|\mathbf{u}\|^2\,\|\mathbf{v}\|^2, \end{aligned}$$

where multiplying both sides of the second inequality by the non-negative quantity $\|\mathbf{v}\|^2$ does not change the direction of the inequality.

Because $\mathbf{u}\cdot\mathbf{v}$ *could* be negative, we will change our last inequality above to

$$|\mathbf{u}\cdot\mathbf{v}|^2 \le (\|\mathbf{u}\|\,\|\mathbf{v}\|)^2.$$

Note. Squaring turns negative into positive anyway, so it doesn't matter if we introduce absolute value brackets to do that first.

In words, this inequality says that the square of one number is less than or equal to the square of another number. But when we square two numbers, the bigger number will always result in the bigger square (as long as neither number is negative). Since neither $|\mathbf{u}\cdot\mathbf{v}|$ nor $\|\mathbf{u}\|\,\|\mathbf{v}\|$ can be negative, the bigger number must be $\|\mathbf{u}\|\,\|\mathbf{v}\|$ to result in a bigger square (or the two numbers could be equal). That is,

$$|\mathbf{u}\cdot\mathbf{v}| \le \|\mathbf{u}\|\,\|\mathbf{v}\|.$$

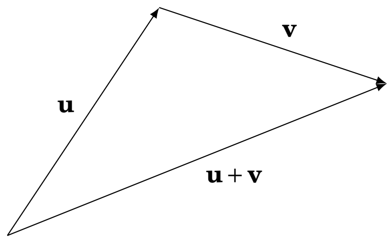
■

Corollary 12.5.5 Uniqueness of angle measures. *For every pair of nonzero vectors* $\mathbf{u}$ *and* $\mathbf{v}$ *in* $\mathbb{R}^n$*, there is one unique angle value* θ *in the domain* $0 \le \theta \le \pi$ *so that*

$$\cos\theta = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|}.$$

The triangle inequality. Here is another commonly used inequality. Remembering our view of sums of vectors as a chain of changes in position, it basically says that the shortest path between two points in $\mathbb{R}^n$ is the direct path.

Theorem 12.5.6 Triangle inequality. *For every pair of vectors* $\mathbf{u}$ *and* $\mathbf{v}$ *in* $\mathbb{R}^n$*, we have* $\|\mathbf{u}+\mathbf{v}\| \le \|\mathbf{u}\| + \|\mathbf{v}\|$.



Check your understanding. *Can you think of a situation where* $\|\mathbf{u}+\mathbf{v}\|$ *and* $\|\mathbf{u}\|+\|\mathbf{v}\|$ *are exactly equal? Is this the only way this would happen?*

Proof. As mentioned in this chapter, working with square roots algebraically is inconvenient, so we will work with the square of the norm $\|\mathbf{u}+\mathbf{v}\|$, and use Proposition 12.5.3 to avoid working directly with the components of our vectors.

We have

$$\begin{aligned}\|\mathbf{u}+\mathbf{v}\|^2 &= (\mathbf{u}+\mathbf{v})\cdot(\mathbf{u}+\mathbf{v}) && \text{(i)}\\ &= (\mathbf{u}+\mathbf{v})\cdot\mathbf{u}+(\mathbf{u}+\mathbf{v})\cdot\mathbf{v} && \text{(ii)}\\ &= \mathbf{u}\cdot\mathbf{u}+\mathbf{v}\cdot\mathbf{u}+\mathbf{u}\cdot\mathbf{v}+\mathbf{v}\cdot\mathbf{v} && \text{(iii)}\\ &= \mathbf{u}\cdot\mathbf{u}+2\mathbf{u}\cdot\mathbf{v}+\mathbf{v}\cdot\mathbf{v} && \text{(iv)}\\ &= \|\mathbf{u}\|^2+2\mathbf{u}\cdot\mathbf{v}+\|\mathbf{v}\|^2 && \text{(v),}\end{aligned}$$

with justifications

(i) Rule 8 of Proposition 12.5.3;

(ii) Rule 2 of Proposition 12.5.3;

(iii) Rule 4 of Proposition 12.5.3;

(iv) Rule 1 of Proposition 12.5.3; and

(v) Rule 8 of Proposition 12.5.3.

Now, keep in mind that $\mathbf{u}\cdot\mathbf{v}$ is a number, and it may be positive, negative, or zero. But every number x satisfies

$$x \le |x|\,, \qquad (*)$$

since if x is positive or zero then the two sides are equal, and if x is negative then obviously the negative number x must be less than the positive number $|x|$. Applying this for $x=\mathbf{u}\cdot\mathbf{v}$, we have $\mathbf{u}\cdot\mathbf{v}\le|\mathbf{u}\cdot\mathbf{v}|$, and so

$$\begin{aligned}\|\mathbf{u}+\mathbf{v}\|^2 &= \|\mathbf{u}\|^2+2\mathbf{u}\cdot\mathbf{v}+\|\mathbf{v}\|^2 && \text{(i)}\\ &\le \|\mathbf{u}\|^2+2\,|\mathbf{u}\cdot\mathbf{v}|+\|\mathbf{v}\|^2 && \text{(ii)}\\ &\le \|\mathbf{u}\|^2+2\,\|\mathbf{u}\|\,\|\mathbf{v}\|+\|\mathbf{v}\|^2 && \text{(iii)}\\ &= (\|\mathbf{u}\|+\|\mathbf{v}\|)^2 && \text{(iv),}\end{aligned}$$

with justifications

(i) continued from above;

(ii) rule (∗);

(iii) The Cauchy-Schwarz inequality; and

(iv) FOIL in reverse.

Following the chain of equalities and inequalities from beginning to end, we now have

$$\|\mathbf{u}+\mathbf{v}\|^2 \le (\|\mathbf{u}\| + \|\mathbf{v}\|)^2.$$

In words, this says that the square of one number is less than or equal to the square of another number. But when we square two numbers, the bigger number will always result in the bigger square (as long as neither number is negative). Since neither $\|\mathbf{u}+\mathbf{v}\|$ nor $\|\mathbf{u}\| + \|\mathbf{v}\|$ can be negative, the bigger number must be $\|\mathbf{u}\| + \|\mathbf{v}\|$ to result in a bigger square (or the two numbers could be equal). That is,

$$\|\mathbf{u}+\mathbf{v}\| \le \|\mathbf{u}\| + \|\mathbf{v}\|.$$

■

Chapter 13

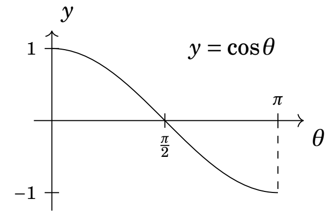
Orthogonal vectors

13.1 Discovery guide

Recall that for $\mathbf{u} = (u_1, u_2, \dots, u_n)$ and $\mathbf{v} = (v_1, v_2, \dots, v_n)$, the **dot product** of $\mathbf{u}$ and $\mathbf{v}$ is defined by the formula given below on the left. It is an important formula because if θ is the angle between two nonzero vectors $\mathbf{u}$ and $\mathbf{v}$, then θ satisfies both $0 \le \theta \le \pi$ and the formula given below on the right.

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n \qquad\qquad \cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}$$

Discovery 13.1 Based on the graph of $y = \cos\theta$ on domain $0 \le \theta \le \pi$ provided below, what can you say about $\mathbf{u} \cdot \mathbf{v}$ in the case that θ is acute? ... obtuse? ... right?



Extending the concept of **perpendicular** to higher dimensions, vectors $\mathbf{u}$ and $\mathbf{v}$ are called **orthogonal** if $\mathbf{u} \cdot \mathbf{v} = 0$.

Discovery 13.2

(a) Can you guess a vector $\mathbf{v} = (v_1, v_2)$ that is orthogonal to $\mathbf{u} = (1, -3)$ in the plane? Make sure your guess satisfies the definition of **orthogonal**: you need $\mathbf{u} \cdot \mathbf{v} = 0$.

(b) What relationship to your initial guess $\mathbf{v}$ will other vectors in the plane that are orthogonal to $\mathbf{u}$ have?

Hint. Draw a diagram of your vectors $\mathbf{u}$ and $\mathbf{v}$, both with initial points at the origin. On your diagram, how can you modify your intial guess $\mathbf{v}$ geometrically while still maintaining orthogonality with $\mathbf{u}$?

(c) Turn the pattern of your guess from Task a into a general pattern for vectors in the plane: if $\mathbf{u} = (a, b)$, then an example of a vector orthogonal to $\mathbf{u}$ is

$$\mathbf{v} = (\qquad , \qquad).$$

Discovery 13.3

(a) Draw the vector $\mathbf{a} = (3,1)$ in the xy-plane with its tail at the origin. Now imagine you were to also draw in every possible scalar multiple of $\mathbf{a}$ (positive, negative, zero, fractional, etc.). What geometric shape would these scalar multiples of $\mathbf{a}$ trace out? Draw this shape on your diagram.

(b) Plot the point $Q(4,4)$ on your diagram. On the line defined by $\mathbf{a}$ that you drew in the first part of this activity, draw in the point that you think is closest to Q. Label this point P. Now draw $\overrightarrow{PQ}$, and label this vector as $\mathbf{n}$.

What is the relationship between $\mathbf{n}$ and the line? What is the value of $\mathbf{n} \cdot \mathbf{a}$?

(c) Vector $\overrightarrow{OP}$ is parallel to $\mathbf{a}$, so $\overrightarrow{OP}$ is a scalar multiple of $\mathbf{a}$. Our goal is to determine the scalar k so that the head of $k\mathbf{a}$ lies at P. Complete the triangle in your diagram by drawing in the vector $\mathbf{u} = \overrightarrow{OQ}$.

Then express $\mathbf{n}$ as a combination of $\mathbf{u}$ and $k\mathbf{a}$.

Remember. $k\mathbf{a} = \overrightarrow{OP}$.

(d) Substitute your expression for $\mathbf{n}$ from Task c into your equation for $\mathbf{n} \cdot \mathbf{a}$ from Task b, and then solve for k as a formula in $\mathbf{u}$ and $\mathbf{a}$.

Now complete the general formula:

$$k\mathbf{a} = \left(\boxed{}\right)\mathbf{a}$$

(where in the brackets you should fill in a formula in the variable letters $\mathbf{u}$ and $\mathbf{a}$, *without* using their actual numerical components, that describes how to compute $k\mathbf{a}$ from $\mathbf{u}$ and $\mathbf{a}$).

The vector $k\mathbf{a}$ in Discovery 13.3 is called the **orthogonal projection of u onto a**, and we write $\operatorname{proj}_{\mathbf{a}} \mathbf{u}$ to mean this vector. It is also sometimes called the **vector component of u parallel to a**. The vector $\mathbf{n} = \mathbf{u} - \operatorname{proj}_{\mathbf{a}} \mathbf{u}$ is called the **vector component of u orthogonal to a**.

Note. The reason these vectors are called **components** of $\mathbf{u}$ is that the original vector $\mathbf{u}$ can be rebuilt out of these "components" by $\mathbf{u} = \operatorname{proj}_{\mathbf{a}} \mathbf{u} + (\mathbf{u} - \operatorname{proj}_{\mathbf{a}} \mathbf{u})$.

The same problem can be solved in higher dimensions by the same formula for $\operatorname{proj}_{\mathbf{a}} \mathbf{u}$.

Discovery 13.4

(a) Suppose $\mathbf{u}$ is orthogonal to $\mathbf{a}$. What is $\operatorname{proj}_{\mathbf{a}} \mathbf{u}$? What is the component of $\mathbf{u}$ orthogonal to $\mathbf{a}$?

(b) Answer the same two questions in the case that $\mathbf{u}$ is *parallel* to $\mathbf{a}$.

Discovery 13.5 If ℓ is the line through the origin and parallel to a vector $\mathbf{a}$, and $\mathbf{u}$ is some other vector, then our construction in Discovery 13.3 guarantees that $\operatorname{proj}_{\mathbf{a}} \mathbf{u}$ represents the closest point on ℓ to the terminal point of $\mathbf{u}$.

The distance between a point and a line is defined as the shortest (i.e. perpendicular) distance between the two. Use the orthogonal projection to come up with a procedure to determine the distance between the line $\ell\colon\ y = x/2$ and the point $Q(2,4)$.

Discovery 13.6 The homogeneous linear equation $2x + 3y = 0$ defines a line through the origin in $\mathbb{R}^2$ (i.e. the xy-plane).

(a) Recall that a point (x, y) lies on the line if and only if its coordinates satisfy the given equation. Let's consider such a point as the terminal point of the vector $\mathbf{x} = (x, y)$ with its initial point at the origin. Does the left-hand side of the equation for the line look like the formula for some quantity related to $\mathbf{x}$ and some other vector? Perhaps some quantity that we've been exploring in detail recently?

(b) In light of the first part of this activity, what does the right-hand side of the equation for the line say about the relationship between a vector $\mathbf{x} = (x, y)$ that lies along the line and the other special vector you identified in the previous part?

Terminology. This special vector for the line is called a **normal vector** for the line.

Discovery 13.7 The non-homogeneous linear equation $2x + 3y = 8$ defines a line through the point $P(1,2)$ in $\mathbb{R}^2$.

(a) Draw the line and label the point $P(1,2)$. Choose another arbitrary point on the line and label it $Q(x, y)$. Draw the vector $\mathbf{v} = \overrightarrow{PQ}$ along the line. Express the components of $\mathbf{v}$ as formulas in x and y.

(b) Draw the vector $\mathbf{n} = (2,3)$ (from the coefficients in the line equation, just as in Discovery 13.6) with its tail at P. What do you notice about the relationship between this normal vector and the vector $\mathbf{v}$ parallel to the line? Express this relationship in terms of the dot product, and then expand out this dot product.

Terminology. The equation involving the dot product that you obtain is called the **point-normal form** for the line.

The same sort of analysis can be carried out for a plane in space determined by algebraic equation $ax + by + cz = d$. The coefficients form a normal vector $\mathbf{n} = (a, b, c)$ that is perpendicular to the plane (i.e. orthogonal to every vector that is parallel to the plane), and given some specific point $\mathbf{x}_0 = (x_0, y_0, z_0)$ that lies on the plane, the plane can be described by the point-normal form $\mathbf{n} \cdot (\mathbf{x} - \mathbf{x}_0) = 0$.

Discovery 13.8 Consider the planes Π_1, Π_2, and Π_3 described algebraically below.

$$\Pi_1 : x - y + 2z = 2 \qquad \Pi_2 : 2x - 2y + 4z = 7 \qquad \Pi_3 : x - y + 3z = 2$$

Use the concept of normal vector to justify the claim that Π_1 and Π_2 are parallel, but that Π_3 is not parallel to either of Π_1 or Π_2.

Orthogonal projection onto a plane in space is a little more complicated, and is likely something you would learn about in a second course in linear algebra. But it's possible to use a different strategy to determine the distance between a point and a plane by using the fact that a plane has *one unique* normal "direction."

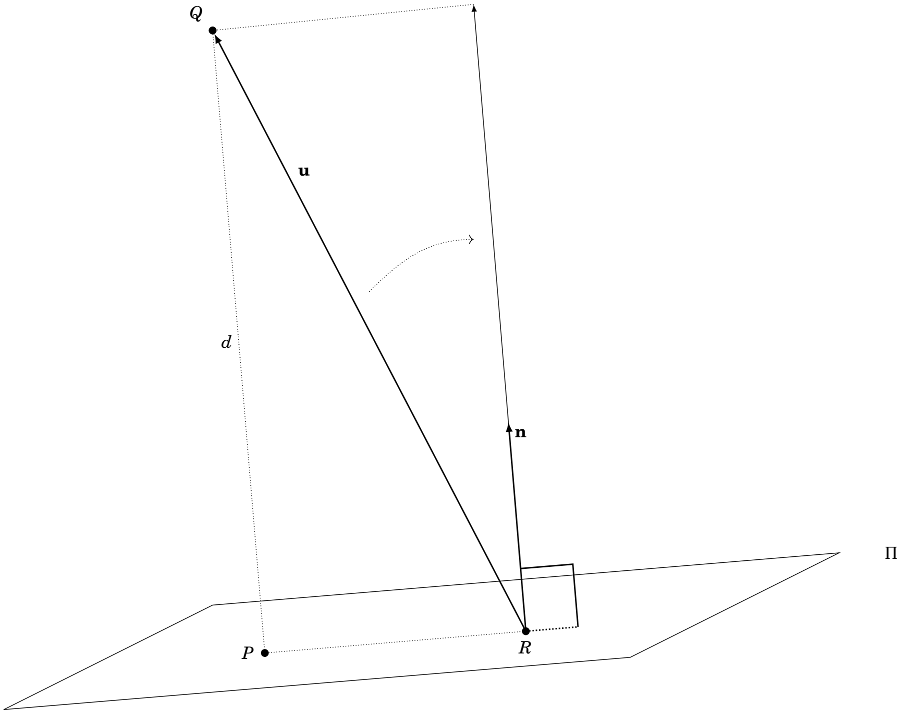
Discovery 13.9

(a) Using the diagram below as inspiration, come up with a procedure to determine the distance d between a point Q and a plane Π.

Hint. Determine a vector that represents an equivalent distance, and then d will be the norm of this vector.

(b) Come up with a procedure using vectors to determine the distance between parallel planes. Do not assume that either of the planes passes through the origin.

Hint. Find a way to reduce this problem to the problem in the first part of this activity.


Q
u
d
n
Π
P
R

13.2 Terminology and notation

orthogonal vectors
a pair of vectors whose dot product evaluates to 0

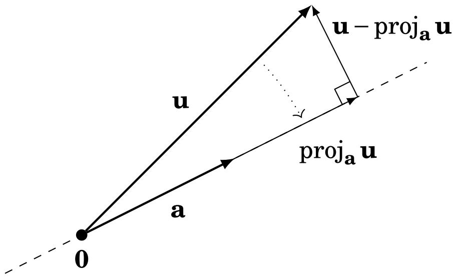
normal vector (to a line or a plane)
a vector that is orthogonal to the object of interest (i.e. the line or plane being considered)

orthogonal projection (of a vector u onto a second vector a)
the special scalar multiple of **a**,

$$\operatorname{proj}_{\mathbf{a}} \mathbf{u} = k\mathbf{a}, \qquad \text{where} \qquad k = \frac{\mathbf{u}\cdot\mathbf{a}}{\|\mathbf{a}\|^2};$$

sometimes called the **vector component of u parallel to a**

When the initial point of $\operatorname{proj}_{\mathbf{a}} \mathbf{u}$ is placed at the origin, the terminal point will be the point closest to **u** on the line passing through the origin and parallel to **a**.


$\mathbf{u} - \operatorname{proj}_{\mathbf{a}} \mathbf{u}$
$\mathbf{u}$
$\operatorname{proj}_{\mathbf{a}} \mathbf{u}$
$\mathbf{a}$
$\mathbf{0}$

vector component of a vector u orthogonal to a second vector a
the vector $\mathbf{u} - \operatorname{proj}_{\mathbf{a}} \mathbf{u}$

When the initial point of the vector $\mathbf{u} - \operatorname{proj}_{\mathbf{a}} \mathbf{u}$ is placed at the terminal point of $\operatorname{proj}_{\mathbf{a}} \mathbf{u}$, it points towards the terminal point of **u**, at a right angle to the line that passes through the origin and is parallel to **a**. (See the diagram above.)

point-normal form (of a line in $\mathbb{R}^2$)
the vector equation $\mathbf{n}\cdot(\mathbf{x}-\mathbf{x}_0) = 0$, where $\mathbf{x}_0$ is a vector from the origin to a known point on the line, **n** is a known normal vector for the line, and **x** is a variable vector representing an arbitrary point on the line (again as a vector from the origin)

point-normal form (of a plane in $\mathbb{R}^3$)
the vector equation $\mathbf{n}\cdot(\mathbf{x}-\mathbf{x}_0) = 0$, where $\mathbf{x}_0$ is a vector from the origin to a known point on the plane, **n** is a known normal vector for the plane, and **x** is a variable vector representing an arbitrary point on the plane (again as a vector from the origin)

cross product (of vectors u and v in $\mathbb{R}^3$)
a particular vector in R^3 that is orthogonal to both **u** and **v**; written $\mathbf{u}\times\mathbf{v}$

13.3 Concepts

In this section.

13.3.1 Values of $\mathbf{u} \cdot \mathbf{v}$

In Discovery 13.1, we compared the graph of the cosine function on the domain $0 \le \theta \le \pi$ with the formula

$$\cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}, \tag{$*$}$$

where θ is the angle between nonzero vectors $\mathbf{u}$ and $\mathbf{v}$. On the right of equation $(*)$, the denominator is always positive, so whether the whole fraction is positve, negative, or zero depends entirely on the dot product in the numerator. On the left, the cosine function is positive, negative, or zero precisely when the angle θ is acute, obtuse, or right. So we come to the following conclusions.

	θ	$\mathbf{u} \cdot \mathbf{v}$
acute:	$0 \le \theta < \pi/2$	positive
right:	$\theta = \pi/2$	zero
obtuse:	$\pi/2 < \theta \le \pi$	negative

Figure 13.3.1

13.3.2 Orthogonal vectors

Right angles are extremely important in geometry, and from Figure 13.3.1 we see that the dot product gives us a very convenient way to tell when the angle θ between two nonzero vectors $\mathbf{u}$ and $\mathbf{v}$ is right: ***we have*** $\theta = \pi/2$ ***precisely when*** $\mathbf{u} \cdot \mathbf{v} = 0$. In the plane or in space, $\mathbf{u}$ and $\mathbf{v}$ will be **perpendicular** when $\theta = \pi/2$ and $\mathbf{u} \cdot \mathbf{v} = 0$. Since we can't "see" right angles and perpendicular lines in higher dimensions, in general we say that $\mathbf{u}$ and $\mathbf{v}$ are **orthogonal** when $\mathbf{u} \cdot \mathbf{v} = 0$.

13.3.2.1 Orthogonal vectors in $\mathbb{R}^2$

In Discovery 13.2, we tried to find a pattern to the task of choosing some vector that is orthogonal to a given one in the plane. Rather than struggle with the geometry, we unleash the power of algebra: given vector $\mathbf{u} = (a, b)$, we are looking for a vector $\mathbf{v}$ so that $\mathbf{u} \cdot \mathbf{v} = 0$. Expanding out the dot product, we are looking to fill in the blanks in the following equation with components for $\mathbf{v}$:

$$a \cdot \boxed{} + b \cdot \boxed{} = 0.$$

Two numbers add to zero only if one is the negative of the other. We can make both terms in the sum the same number by entering b in the first blank and a in the second, so we can make the sum cancel to zero by also flipping the sign of one of those entries. For example,

$$a \cdot \boxed{b} + b \cdot \boxed{(-a)} = 0.$$

We have now answered the question in Discovery 13.2.c.

Pattern 13.3.2 Orthogonal vectors in the plane. Given vector $\mathbf{u} = (a,b)$ in the plane, two examples of vectors that are orthogonal to $\mathbf{u}$ are $\mathbf{v} = (b,-a)$ and $-\mathbf{v} = (-b,a)$, and every vector that is orthogonal to $\mathbf{u}$ is some scalar multiple of this example $\mathbf{v}$.

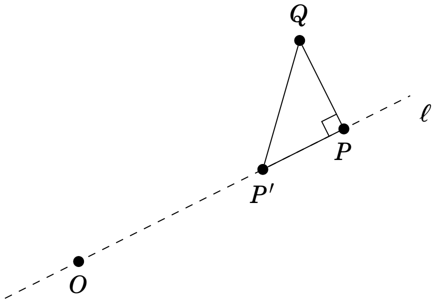
Note 13.3.3 For patterns of orthogonal vectors in $\mathbb{R}^3$, see Subsection 13.3.6.

13.3.3 Orthogonal projection

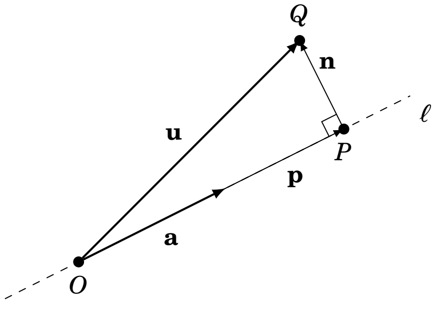
Orthogonal projection is a vector solution to a problem in geometry.

Question 13.3.4 Given a line through the origin in the plane, and a point not on the line, what point on the line is closest to the given point? □

In Question 13.3.4, write ℓ for the line through the origin and Q for the point not on that line. Consider the point P on ℓ at the foot of the perpendicular to ℓ from Q . Any other point P' on ℓ will form a right triangle with P and Q, making it farther from Q than P, since the distance $P'Q$ is the length of the hypotenuse in the right triangle.



All we know about P is that it is on line ℓ and it is at the vertex of a right angle with ℓ and Q. But if we introduce some vectors to help tackle this problem, then maybe we can use what we know about the dot product and right angles to help determine P.



In this diagram, $\mathbf{u}$ is the vector corresponding to directed line segment $\overrightarrow{OQ}$, and $\mathbf{p}$ is the vector corresponding to the directed line segment $\overrightarrow{OP}$, where P is our unknown closest point. Since $\mathbf{p}$ is placed with its tail at the origin, the components of $\mathbf{p}$ are precisely the coordinates of P. So determining $\mathbf{p}$ will solve the problem.

We are assuming that the line ℓ is known, and it would be nice to also have a vector means of describing it. But the vectors created by the points on this line (using the origin as a universal tail point) will all be parallel to each other, so (as we discovered in Discovery 13.3.a) line ℓ could be described as all scalar multiples of a particular vector $\mathbf{a}$. This vector can be arbitrarily chosen as any vector parallel to the line. Once we have chosen $\mathbf{a}$, we have reduced our problem from determining the *two* unknown components of the vector $\mathbf{p}$ to determining a *single* unknown scalar k so that $\mathbf{p} = k\mathbf{a}$.

As mentioned, since P is the closest point, the directed line segment $\overrightarrow{PQ}$ must be perpendicular to ℓ. On the diagram above, we have used the vector $\mathbf{n}$ to represent this direct line segment. As in Discovery 13.3.b, we know that $\mathbf{n}\cdot\mathbf{a}$ must be zero — this is the perpendicular condition. However, the vector $\mathbf{n}$ is unknown as well, since we don't know its initial point. But we can also use the triangle formed by $\mathbf{u}$, $\mathbf{n}$, and $\mathbf{p}$ to replace $\mathbf{n}$:

$$\mathbf{p}+\mathbf{n}=\mathbf{u} \qquad \Longrightarrow \qquad \mathbf{n}=\mathbf{u}-\mathbf{p}=\mathbf{u}-k\mathbf{a}$$

Replacing $\mathbf{n}$ by this expression in the condition $\mathbf{n}\cdot\mathbf{a}=0$ gives us an equation of numbers that we can solve for the unknown scale factor k, as we did in Discovery 13.3.d:

$$k = \frac{\mathbf{u}\cdot\mathbf{a}}{\|\mathbf{a}\|^2}.$$

This vector $\mathbf{p} = k\mathbf{a}$ pointing from the origin to the desired closest point P is called the **projection of u onto a** or sometimes the **vector component of u parallel to a**, and we write $\operatorname{proj}_{\mathbf{a}}\mathbf{u}$ to represent it.

Procedure 13.3.5 Closest point on a line (orthogonal projection). *Given a line ℓ through the origin and point Q that does not lie on ℓ, compute the point P on ℓ that is closest to Q as follows.*

1. *Choose any point P' on the line (excluding the origin), and form the parallel vector $\mathbf{a}=\overrightarrow{OP'}$.*
2. *Form the vector $\mathbf{u}=\overrightarrow{OQ}$.*
3. *Compute the projection vector*

$$\mathbf{p} = \operatorname{proj}_{\mathbf{a}}\mathbf{u} = \frac{\mathbf{u}\cdot\mathbf{a}}{\|\mathbf{a}\|^2}\,\mathbf{a}.$$

This projection vector will now point from the origin to the desired closest point P, parallel to the line ℓ*, so that* $\mathbf{p}=\overrightarrow{OP}$.

Remark 13.3.6 It is not actually necessary that Q be external to the line. If you were to carry out the procedure above in the case that Q lies on ℓ, the calculations would end up with $\mathbf{p}=\mathbf{u}$, confirming that Q was already the point on the line that is closest to itself.

The normal vector $\mathbf{n}$ in the diagram above is sometimes called the **vector component of u orthogonal to a**. Together, the projection vector and corresponding normal vector are called **components** of $\mathbf{u}$ (relative to $\mathbf{a}$) because they represent an **orthogonal decomposition of u**:

$$\mathbf{u}=\mathbf{p}+\mathbf{n},$$

where $\mathbf{p}$ is parallel to $\mathbf{a}$ and $\mathbf{n}$ is orthogonal to $\mathbf{a}$. While this decomposition is relative to $\mathbf{a}$, it is really only the *direction* of $\mathbf{a}$ that matters — if $\mathbf{a}'$ is parallel to $\mathbf{a}$ (even possibly opposite to $\mathbf{a}$), then both

$$\mathbf{p}=\operatorname{proj}_{\mathbf{a}}\mathbf{u}=\operatorname{proj}_{\mathbf{a}'}\mathbf{u}, \qquad \mathbf{n}=\mathbf{u}-\mathbf{p}=\mathbf{u}-\operatorname{proj}_{\mathbf{a}}\mathbf{u}=\mathbf{u}-\operatorname{proj}_{\mathbf{a}'}\mathbf{u}$$

will be true.

Procedure 13.3.7 Shortest distance to a line. *Given a line ℓ through the origin and point Q that does not lie on ℓ, compute the shortest distance from Q to the line as follows.*

1. *Compute the projection vector* $\mathbf{p} = \operatorname{proj}_{\mathbf{a}} \mathbf{u}$ *as in Procedure 13.3.5.*
2. *Compute the normal vector* $\mathbf{n} = \mathbf{u} - \mathbf{p}$.
3. *Compute the norm* $\|\mathbf{n}\|$.

The computed norm is the distance from the closest point P to the point Q.

Remark 13.3.8

1. These procedures and calculations can be easily modified to work for lines that do not pass through the origin: simply choose some arbitrary "initial" point R on the line to "act" as the origin.
2. All of these calculations can be performed in higher dimensions as well. In higher dimensions, it is true that there is no longer one unique perpendicular direction to a given vector $\mathbf{a}$, but the calculation of $\mathbf{n}$ as above will pick out the correction direction to extend from the line to the point Q at a right angle to the line.

13.3.4 Normal vectors of lines in the plane

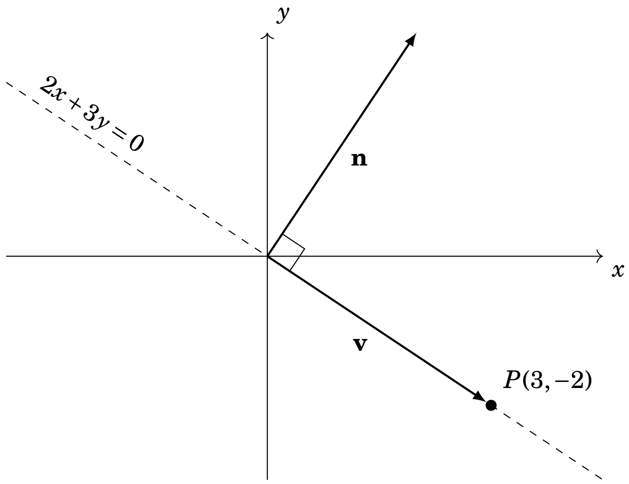
Consider the line $2x+3y=0$ that we investigated in Discovery 13.6. The point $(3,-2)$ is on this line, since

$$2 \cdot 3 + 3 \cdot (-2) = 0. \tag{$**$}$$

The left-hand side of this calculation looks a lot like a dot product — we could reinterpret equation $(**)$ as

$$(2,3) \bullet (3,-2) = 0.$$

So verifying that the point $(3,-2)$ is on the line is equivalent to checking that the corresponding vector $\mathbf{v} = (3,-2)$ (with its tail at the origin) is orthogonal to the vector $\mathbf{n} = (2,3)$ whose components are the *coefficients* from our line equation.

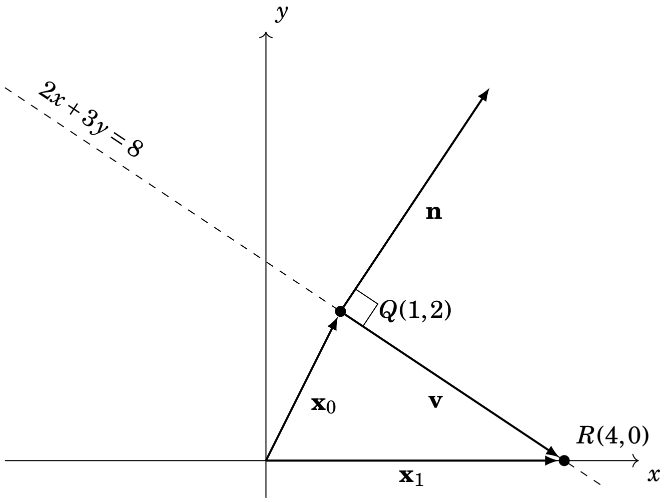


Every other point $\mathbf{x} = (x,y)$ on the line satisfies the same relationship, as the equation for the line could be rewritten in a vector form as

$$\mathbf{n} \cdot \mathbf{x} = 0. \tag{***}$$

The vector $\mathbf{n}$ is called a *normal* vector for the line. Note that normal vectors for a line are not unique — every nonzero scalar multiple of $\mathbf{n}$ will also be normal to the line, and this is equivalent to noting that we could multiply the equation $2x + 3y = 0$ by any nonzero factor to obtain a different equation that represents the same line in the plane.

In Discovery 13.7 we considered a line defined by a *non*homogeneous equation $2x + 3y = 8$. This line has the same slope as the line defined by $2x + 3y = 0$ that we investigate above, and so the vector $\mathbf{n} = (2,3)$ obtained from the coefficients on x and y in the equation must still be normal. The constant 8 just changes the y-intercept.

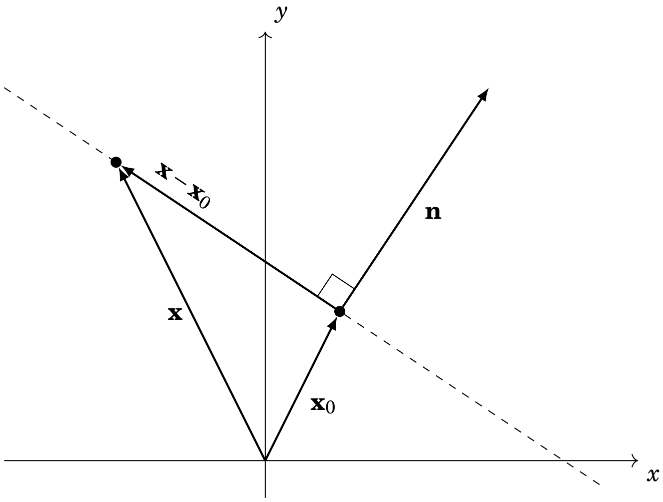


In the homogeneous case, vectors from the origin determined by a point on the line were also *parallel* to the line. Since things have shifted away from the origin in the nonhomogeneous case, to get a vector parallel to the line we need to consider *two* vectors from the origin to points on the line. Two convenient points for this the line are $Q(1,2)$ and $R(4,0)$, with corresponding vectors $\mathbf{x}_0 = (1,2)$ and $\mathbf{x}_1 = (4,0)$. Then the difference vector

$$\mathbf{v} = \mathbf{x}_1 - \mathbf{x}_0 = (3,-2)$$

is parallel to the line, as in the diagram above. In fact, this vector $\mathbf{v}$ is the same as previous vector $\mathbf{v}$ that appears parallel to the line through the origin in the diagram for the homogeneous case above, so we know it satisfies $\mathbf{n} \cdot \mathbf{v} = 0$.

Is there a way to use the normal vector $\mathbf{n}$ to create a vector condition by which we can tell if a vector $\mathbf{x}$ represents a point on the line, as we did with equation (***) in the homoegenous case? We need *two* points on the line to create a parallel difference vector, but we could compare the variable vector $\mathbf{x}$ with a arbitrarily chosen *fixed* vector representing a point on the line (like $\mathbf{x}_0$, say).



Every such difference vector $\mathbf{x}-\mathbf{x}_0$ is parallel to the line and hence orthogonal to the normal vector $\mathbf{n}$, so that we can describe the line as all points where the corresponding vector $\mathbf{x}$ satisfies

$$\mathbf{n}\cdot(\mathbf{x}-\mathbf{x}_0)=0. \tag{†}$$

This is called the **point-normal form** for the line, referring to the *point* on the line at the terminal point of $\mathbf{x}_0$ and the *normal* vector $\mathbf{n}$.

Pattern 13.3.9 Point-normal form for a line in $\mathbb{R}^2$. If (x_0,y_0) is a point on the line $\ell\colon ax+by=d$ (that is, $ax_0+by_0=d$ is true), then ℓ can alternatively be described as all points (x,y) that satisfy

$$(a,b)\cdot\big((x,y)-(x_0,y_0)\big)=0.$$

Remark 13.3.10 It may seem like the line parameter d has disappeared in converting from algebraic form $ax+by=d$ to point-normal form. But it has merely be replaced by the point (x_0,y_0), since $d=ax_0+by_0$. In fact, if we use the algebraic properties of the dot product to expand the left-hand side of the point-normal form equation, we can recover the original algebraic equation:

$$\begin{gathered}(a,b)\cdot\big((x,y)-(x_0,y_0)\big)=0\\(a,b)\cdot(x,y)-(a,b)\cdot(x_0,y_0)=0\\(ax+by)-(ax_0+by_0)=0\\(ax+by)-d=0\\ax+by=d.\end{gathered}$$

13.3.5 Normal vectors of planes in space

A similar analysis can be made for an equation $ax+by+cz=d$ describing a plane in space. The coefficients form a normal vector $\mathbf{n}=(a,b,c)$. For vectors $\mathbf{x}_0$ and $\mathbf{x}_1$ that both have initial point at the origin and terminal points on the plane, then the difference vector $\mathbf{x}_1-\mathbf{x}_0$ is parallel to the plane, hence normal to $\mathbf{n}$. If we keep a fixed choice of $\mathbf{x}_0$ but replace $\mathbf{x}_1$ by a variable vector $\mathbf{x}$, we can describe the plane as all points whose difference is orthogonal to $\mathbf{n}$, giving us a point-normal for a plane just as in equation (†).

Pattern 13.3.11 Point-normal form for a plane in $\mathbb{R}^3$. If (x_0,y_0,z_0) is a point on the plane $\Pi\colon ax+by+cz=d$ (that is, $ax_0+by_0+cz_0=d$ is true), then Π can alternatively be described as all points (x,y,z) that satisfy

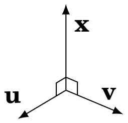
$$(a,b,c)\cdot\big((x,y,z)-(x_0,y_0,z_0)\big)=0.$$

Remark 13.3.12 A line in space does not have a point-normal form, because it does not have one unique normal "direction" like a line in the plane or a plane in space does. To describe a line in space in a similar fashion you would need *two* normal vectors. We will see several more convenient ways to describe a line in space in the next chapter.

13.3.6 The cross product

Seeing how the algebraic equation for a plane in $\mathbb{R}^3$ is connected to a normal vector to the plane, a basic problem is how to quickly obtain a normal vector. If we know two vectors that are parallel to the plane in question, the problem reduces to the following.

Question 13.3.13 Given two nonzero, nonparallel vectors in $\mathbb{R}^3$, determine a third vector that is orthogonal to each of the first two. □



So if $\mathbf{u}=(u_1,u_2,u_3)$ and $\mathbf{v}=(v_1,v_2,v_3)$ are our starting vectors, we would like to simultaneously solve the equations

$$\mathbf{u}\cdot\mathbf{x}=0, \qquad\qquad \mathbf{v}\cdot\mathbf{x}=0,$$

for the unknown vector $\mathbf{x}=(x,y,z)$. Expanding out the dot products, we get (surprise!) a system of linear equations:

$$\left\{\begin{array}{ccccccc} u_1x & + & u_2y & + & u_3z & = & 0, \\ v_1x & + & v_2y & + & v_3z & = & 0. \end{array}\right.$$

Specifically, we get a homogeneous system of two equations in the three unknown coordinates x,y,z. Now, since this system is homogeneous, it is consistent. But its general solution will also require at least one parameter, since its rank is at most 2, while we have three variables. In the diagram above, we can see what the "freedom" of a parameter corresponds to — we can make $\mathbf{x}$ longer or shorter, or turn it around to be opposite of the way it is pictured, and it will remain orthogonal to $\mathbf{u}$ and $\mathbf{v}$. Our end goal is a calculation formula and procedure that will compute one particular solution to this problem, so let's introduce a somewhat arbitrary additional equation to eliminate the need for a parameter in the solution.

$$\left\{\begin{array}{ccccccc} x & + & y & + & z & = & 1, \\ u_1x & + & u_2y & + & u_3z & = & 0, \\ v_1x & + & v_2y & + & v_3z & = & 0. \end{array}\right.$$

In matrix form, this system can be expressed as $A\mathbf{x}=\mathbf{b}$, with

$$A=\begin{bmatrix} 1 & 1 & 1 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{bmatrix}, \qquad\qquad \mathbf{b}=\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}. \qquad (\dagger\dagger)$$

Assuming that $\det A\neq 0$, Cramer's rule tells us the solution to this system.

$$x=\frac{1}{\det A}\begin{vmatrix} 1 & 1 & 1 \\ 0 & u_2 & u_3 \\ 0 & v_2 & v_3 \end{vmatrix} \qquad y=\frac{1}{\det A}\begin{vmatrix} 1 & 1 & 1 \\ u_1 & 0 & u_3 \\ v_1 & 0 & v_3 \end{vmatrix} \qquad z=\frac{1}{\det A}\begin{vmatrix} 1 & 1 & 1 \\ u_1 & u_2 & 0 \\ v_1 & v_2 & 0 \end{vmatrix}$$

$$= \frac{1}{\det A}\begin{vmatrix} u_2 & u_3 \\ v_2 & v_3 \end{vmatrix} \qquad = \frac{-1}{\det A}\begin{vmatrix} u_1 & u_3 \\ v_1 & v_3 \end{vmatrix} \qquad = \frac{1}{\det A}\begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix}$$

Now, each of x, y, z has a common factor of $1/\det A$, and all this common factor does is scale the length of our solution vector $\mathbf{x}$ without affecting orthogonality with $\mathbf{u}$ and $\mathbf{v}$. Even worse, $\det A$ depends on that extra equation we threw in, and we would like our solution to depend only on $\mathbf{u}$ and $\mathbf{v}$. So let's remove it and use solution

$$\mathbf{x} = \left(\begin{vmatrix} u_2 & u_3 \\ v_2 & v_3 \end{vmatrix}, -\begin{vmatrix} u_1 & u_3 \\ v_1 & v_3 \end{vmatrix}, \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix} \right).$$

We call this the **cross product of u and v**, and write $\mathbf{u} \times \mathbf{v}$ instead of $\mathbf{x}$. There is a trick to remembering how to compute the cross product: if we replace the top row of A by the standard basis vectors $\mathbf{i}, \mathbf{j}, \mathbf{k}$ in $\mathbb{R}^3$, then the cross product will be equal to its determinant expanded by cofactors along the first row. That is, setting

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} \qquad (\dagger\dagger\dagger)$$

and expanding the determinant along the first row yields

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} u_2 & u_3 \\ v_2 & v_3 \end{vmatrix} \mathbf{i} - \begin{vmatrix} u_1 & u_3 \\ v_1 & v_3 \end{vmatrix} \mathbf{j} + \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix} \mathbf{k},$$

as desired. See Example 13.4.4 in Subsection 13.4.3 for an example of using formula (†††) to compute cross products.

The cross product follows the **right-hand rule** — if you orient your right hand so that your fingers point in the direction of $\mathbf{u}$ and curl towards $\mathbf{v}$, then your thumb will point in the direction of $\mathbf{u} \times \mathbf{v}$.

Check your understanding. Compute the cross products of the standard basis vectors in the various combinations

$$\begin{array}{ll} \mathbf{i} \times \mathbf{j}, & \mathbf{j} \times \mathbf{i}, \\ \mathbf{j} \times \mathbf{k}, & \mathbf{k} \times \mathbf{j}, \\ \mathbf{i} \times \mathbf{k}, & \mathbf{k} \times \mathbf{i}, \end{array}$$

and verify that the right-hand rule holds in these cases.

Computing $\mathbf{v} \times \mathbf{u}$ instead of $\mathbf{u} \times \mathbf{v}$ should still produce a vector that is orthogonal to both $\mathbf{u}$ and $\mathbf{v}$, but the right-hand rule tells that the two should be opposite to each other. From equation (†††) we can be even more specific. Computing $\mathbf{v} \times \mathbf{u}$ would swap the second and third rows of the special matrix in equation (†††), and we know that the resulting determinant would be the negative of that for the matrix for computing $\mathbf{u} \times \mathbf{v}$, and so

$$\mathbf{v} \times \mathbf{u} = -\mathbf{u} \times \mathbf{v}.$$

See Proposition 13.5.5 in Subsection 13.5.3 for more properties of the cross product.

Remark 13.3.14 There is one more thing to say about our development of the cross product — Cramer's rule can only be applied if $\det A$ is not zero, where A is the matrix in (††). However, the coefficients in the extra equation we introduced did not figure into our final solution. So if $\det A$ ended up being zero for some particular vectors $\mathbf{u}$ and $\mathbf{v}$, we could just change the variable coefficients in that extra equation (but keep the 1 in the equals column) so that $\det A$ is not zero, and we would still come to the same formula for $\mathbf{u} \times \mathbf{v}$. And it follows from concepts we will learn in Chapter 20 that it is always possible to fill in the top row of this matrix A so that its determinant is nonzero, as long as we start with *nonparallel* vectors $\mathbf{u}$ and $\mathbf{v}$.

13.4 Examples

In this section.

- Subsection 13.4.1 *Orthogonal vectors*
- Subsection 13.4.2 *Orthogonal projection*
- Subsection 13.4.3 *Cross product*

13.4.1 Orthogonal vectors

Example 13.4.1 Testing for orthogonality. As in Discovery 13.2, and as discussed in Subsection 13.3.2, it's fairly easy to form orthogonal vectors in $\mathbb{R}^2$. And it's not that much more difficult in $\mathbb{R}^3$.

1. The vectors $\mathbf{u} = (3,7)$ and $\mathbf{v} = (-7,3)$ are orthogonal in $\mathbb{R}^2$, because

$$\mathbf{u} \cdot \mathbf{v} = 3 \cdot (-7) + 7 \cdot 3 = -21 + 21 = 0.$$

2. The vectors $\mathbf{u} = (3,7,1)$ and $\mathbf{v} = (-7,2,7)$ are orthogonal in $\mathbb{R}^3$, because

$$\mathbf{u} \cdot \mathbf{v} = 3 \cdot (-7) + 7 \cdot 2 + 1 \cdot 7 = -21 + 14 + 7 = 0.$$

□

Example 13.4.2 Orthogonality of the standard basis vectors. In $\mathbb{R}^n$, the standard basis vectors are always orthogonal to each other. When we compute $\mathbf{e}_i \cdot \mathbf{e}_j$ with $i \neq j$, the 1 in the i^{th} component of $\mathbf{e}_i$ won't line up with the 1 in the j^{th} component of $\mathbf{e}_j$, and we'll get a computation something like

$$\begin{aligned}\mathbf{e}_i \cdot \mathbf{e}_j &= 0 \cdot 0 + \cdots + 0 \cdot 0 + \overbrace{1 \cdot 0}^{i^{\text{th}} \text{ times } i^{\text{th}}} + 0 \cdot 0 + \cdots + 0 \cdot 0 + \overbrace{0 \cdot 1}^{j^{\text{th}} \text{ times } j^{\text{th}}} + 0 \cdot 0 + \cdots + 0 \cdot 0 \\ &= 0.\end{aligned}$$

□

13.4.2 Orthogonal projection

Let's complete the computations from Discovery 13.3.

Example 13.4.3 Using orthogonal projection to compute distance from a point to a line in $\mathbb{R}^2$. The line through the origin and parallel to $\mathbf{a} = (3,1)$ consists of all scalar multiples of $\mathbf{a}$. We would like to know the following.

- What is the point on this line closest to the point $Q(4,4)$?
- What is the distance from Q to the line?

We know that the point we are looking for is at the terminal point of $\operatorname{proj}_{\mathbf{a}} \mathbf{u}$, where $\mathbf{u} = \overrightarrow{OQ} = (4,4)$. So compute

$$\operatorname{proj}_{\mathbf{a}} \mathbf{u} = \frac{4 \cdot 3 + 4 \cdot 1}{3^2 + 1^2}(3,1) = \frac{8}{5}(3,1) = \left(\frac{24}{5}, \frac{8}{5}\right),$$

which tells us that the point on the line closest to Q is $P(24/5, 8/5)$. Now, the vector

$$\mathbf{n} = \overrightarrow{PQ} = (-4/5, 12/5)$$

will be a normal vector for the line, extending from P to Q, and so the norm of this vector represents the (perpendicular) distance between Q and the line:

$$d = \|\mathbf{n}\| = \sqrt{\left(-\frac{4}{5}\right)^2 + \left(\frac{12}{5}\right)^2} = \sqrt{\frac{160}{25}} = \frac{4\sqrt{10}}{5}.$$

□

13.4.3 Cross product

Here is an example of using the cross product to answer a geometry question in $\mathbb{R}^3$.

Example 13.4.4 Using cross product to determine the equation of a plane in $\mathbb{R}^3$. Suppose we would like to determine the equation of the plane in $\mathbb{R}^3$ that passes through the point $(3,3,3)$ and is parallel to the vectors $\mathbf{u} = (1,2,-3)$ and $\mathbf{v} = (0,2,5)$.

The equation we are looking for is of the form $ax + by + cz = d$. We know that a, b, c can be taken to be the components of any normal vector for the plane. A normal vector for the plane must be orthogonal to the plane, and hence must be orthogonal to each of $\mathbf{u}$ and $\mathbf{v}$. We can use the cross product to compute such a vector:

$$\begin{aligned} \mathbf{n} = \mathbf{u} \times \mathbf{v} &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 2 & -3 \\ 0 & 2 & 5 \end{vmatrix} \\ &= \mathbf{i}\big(2 \cdot 5 - (-3) \cdot 2\big) - \mathbf{j}\big(1 \cdot 5 - (-3) \cdot 0\big) + \mathbf{k}(1 \cdot 2 - 2 \cdot 0) \\ &= 16\mathbf{i} - 5\mathbf{j} + 2\mathbf{k}. \end{aligned}$$

So we can use $16x - 5y + 2z = d$ as the equation of the plane, for some as-yet-to-be-determined value of d. But we also know that the plane passes through the point $(3,3,3)$, so we must have

$$16 \cdot 3 - 5 \cdot 3 + 2 \cdot 3 = d \qquad \Longrightarrow \qquad d = 39.$$

Thus, the plane can be described algebraically by the equation $16x - 5y + 2z = 39$, or in point-normal form by the equation $\mathbf{n} \cdot (\mathbf{x} - \mathbf{x}_0) = 0$, where $\mathbf{n}$ is as computed above, $\mathbf{x}_0$ is the "base" point $(3,3,3)$, and $\mathbf{x} = (x,y,z)$ is a variable point. □

13.5 Theory

In this section.

13.5.1 Properties of orthogonal vectors and orthogonal projection

First we record a few properties of orthogonal vectors and orthogonal projection.

Proposition 13.5.1 Orthogonality versus vector operations. *The following apply to vectors in* $\mathbb{R}^n$.

1. *If* $\mathbf{u}$ *is orthogonal to* $\mathbf{v}$*, then it is orthogonal to every scalar multiple of* $\mathbf{v}$.
2. *If* $\mathbf{u}$ *is orthogonal to both* $\mathbf{v}$ *and* $\mathbf{w}$*, then it is also orthogonal to* $\mathbf{v}+\mathbf{w}$.
3. *If* $\mathbf{u}$ *is orthogonal to each of* $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$*, then* $\mathbf{u}$ *is also orthogonal to every linear combination of those vectors.*

Proof. These properties of orthogonal vectors follow directly from the definition of orthogonality (i.e. dot product equals 0) and from the algebraic properties of the dot product listed in Proposition 12.5.3, so we will omit detailed proofs. ■

Proposition 13.5.2 Properties of orthogonal projection. *Suppose* $\mathbf{u}$, $\mathbf{v}$*, and* $\mathbf{a}$ *are vectors in* $\mathbb{R}^n$*, with* $\mathbf{a} \neq \mathbf{0}$*, and* k *is a scalar. The the following hold.*

1. $\operatorname{proj}_{\mathbf{a}} \mathbf{0} = \mathbf{0}$.
2. $\operatorname{proj}_{\mathbf{a}}(k\mathbf{u}) = k(\operatorname{proj}_{\mathbf{a}} \mathbf{u})$.
3. $\operatorname{proj}_{\mathbf{a}}(\mathbf{u}+\mathbf{v}) = \operatorname{proj}_{\mathbf{a}} \mathbf{u} + \operatorname{proj}_{\mathbf{a}} \mathbf{v}$.
4. *For nonzero scalar* k, $\operatorname{proj}_{(k\mathbf{a})} \mathbf{u} = \operatorname{proj}_{\mathbf{a}} \mathbf{u}$.
5. *If* $\mathbf{u}$ *is parallel to* $\mathbf{a}$*, then* $\operatorname{proj}_{\mathbf{a}} \mathbf{u} = \mathbf{u}$.
6. *If* $\mathbf{u}$ *is orthogonal to* $\mathbf{a}$*, then* $\operatorname{proj}_{\mathbf{a}} \mathbf{u} = \mathbf{0}$.
7. $\left\lVert \operatorname{proj}_{\mathbf{a}} \mathbf{u} \right\rVert = \dfrac{|\mathbf{u} \cdot \mathbf{a}|}{\lVert \mathbf{a} \rVert}$.

Proof of Rule 4. Starting with the formula we determined for orthogonal projection, and using Rule 3 of Proposition 12.5.1 and Rule 7 of Proposition 12.5.3, we have

$$\begin{aligned}
\operatorname{proj}_{(k\mathbf{a})} \mathbf{u} &= \frac{\mathbf{u} \cdot (k\mathbf{a})}{\lVert k\mathbf{a} \rVert^2}\,(k\mathbf{a}) \\
&= \frac{k(\mathbf{u} \cdot \mathbf{a})}{|k|^2\,\lVert \mathbf{a} \rVert^2}\,(k\mathbf{a}) \\
&= \frac{\cancel{k^2}(\mathbf{u} \cdot \mathbf{a})}{\cancel{k^2}\,\lVert \mathbf{a} \rVert^2}\,\mathbf{a} \\
&= \frac{\mathbf{u} \cdot \mathbf{a}}{\lVert \mathbf{a} \rVert^2}\,\mathbf{a} \\
&= \operatorname{proj}_{\mathbf{a}} \mathbf{u}.
\end{aligned}$$

■

Proof of Rule 5. If $\mathbf{u}$ is parallel to $\mathbf{a}$, then it is a scalar multiple of $\mathbf{a}$: $\mathbf{u} = k\mathbf{a}$ for some scalar k. Then, using Rule 6 and Rule 8 of Proposition 12.5.3, we have

$$\begin{aligned}\operatorname{proj}_{\mathbf{a}} \mathbf{u} &= \frac{\mathbf{u} \cdot \mathbf{a}}{\|\mathbf{a}\|^2}\, \mathbf{a} \\ &= \frac{(k\mathbf{a}) \cdot \mathbf{a}}{\|\mathbf{a}\|^2}\, \mathbf{a} \\ &= \frac{k(\mathbf{a} \cdot \mathbf{a})}{\|\mathbf{a}\|^2}\, \mathbf{a} \\ &= k\,\frac{\|\mathbf{a}\|^2}{\|\mathbf{a}\|^2}\, \mathbf{a} \\ &= k\mathbf{a} \\ &= \mathbf{u}.\end{aligned}$$

■

Proofs of other rules. The rest of these properties of orthogonal projection follow from the properties of the dot product in Proposition 12.5.3 and from the formula

$$\operatorname{proj}_{\mathbf{a}} \mathbf{u} = \frac{\mathbf{u} \cdot \mathbf{a}}{\|\mathbf{a}\|^2}\, \mathbf{a},$$

so we will leave the remaining proofs to you, the reader. ■

13.5.2 Decomposition of a vector into orthogonal components

The following fact says that the decomposition of one vector into components (parallel and orthogonal) relative to another vector is unique.

Theorem 13.5.3 Uniqueness of orthogonal decomposition. *Suppose* $\mathbf{a}$ *is a nonzero vector in* $\mathbb{R}^n$*. Given another vector* $\mathbf{u}$ *in* $\mathbb{R}^n$*, there is one unique way to decompose* $\mathbf{u}$ *into a sum*

$$\mathbf{u} = \mathbf{p}_{\mathbf{a}} + \mathbf{n}_{\mathbf{a}},$$

where $\mathbf{p}_{\mathbf{a}}$ *is parallel to* $\mathbf{a}$ *and* $\mathbf{n}_{\mathbf{a}}$ *is normal (i.e. orthogonal) to* $\mathbf{a}$*.*

Proof. Clearly such a decomposition exists — see Remark 13.5.4 below. But suppose we have *two* such decompositions,

$$\mathbf{u} = \mathbf{p}_{\mathbf{a}} + \mathbf{n}_{\mathbf{a}}, \qquad \mathbf{u} = \mathbf{p}'_{\mathbf{a}} + \mathbf{n}'_{\mathbf{a}},$$

where both $\mathbf{p}_{\mathbf{a}}, \mathbf{p}'_{\mathbf{a}}$ are parallel to $\mathbf{a}$ and both $\mathbf{n}_{\mathbf{a}}, \mathbf{n}'_{\mathbf{a}}$ are orthogonal to $\mathbf{a}$. Then each of $\mathbf{n}_{\mathbf{a}}, \mathbf{n}'_{\mathbf{a}}$ are also orthogonal to each of $\mathbf{p}_{\mathbf{a}}, \mathbf{p}'_{\mathbf{a}}$ (Rule 1 of Proposition 13.5.1).

We can use the two decompositions to obtain two expressions for each of $\mathbf{p}_{\mathbf{a}} \cdot \mathbf{u}$ and $\mathbf{p}'_{\mathbf{a}} \cdot \mathbf{u}$:

$$\begin{aligned}\mathbf{p}_{\mathbf{a}} \cdot \mathbf{u} &= \mathbf{p}_{\mathbf{a}} \cdot (\mathbf{p}_{\mathbf{a}} + \mathbf{n}_{\mathbf{a}}) & \mathbf{p}'_{\mathbf{a}} \cdot \mathbf{u} &= \mathbf{p}'_{\mathbf{a}} \cdot (\mathbf{p}'_{\mathbf{a}} + \mathbf{n}'_{\mathbf{a}}) \\ &= \mathbf{p}_{\mathbf{a}} \cdot \mathbf{p}_{\mathbf{a}} + \mathbf{p}_{\mathbf{a}} \cdot \mathbf{n}_{\mathbf{a}} & &= \mathbf{p}'_{\mathbf{a}} \cdot \mathbf{p}'_{\mathbf{a}} + \mathbf{p}'_{\mathbf{a}} \cdot \mathbf{n}'_{\mathbf{a}} \\ &= \|\mathbf{p}_{\mathbf{a}}\|^2 + \mathbf{0} & &= \left\|\mathbf{p}'_{\mathbf{a}}\right\|^2 + \mathbf{0} \\ &= \|\mathbf{p}_{\mathbf{a}}\|^2, & &= \left\|\mathbf{p}'_{\mathbf{a}}\right\|^2,\end{aligned}$$

$$\begin{aligned}\mathbf{p}_{\mathbf{a}} \cdot \mathbf{u} &= \mathbf{p}_{\mathbf{a}} \cdot (\mathbf{p}'_{\mathbf{a}} + \mathbf{n}'_{\mathbf{a}}) & \mathbf{p}'_{\mathbf{a}} \cdot \mathbf{u} &= \mathbf{p}'_{\mathbf{a}} \cdot (\mathbf{p}_{\mathbf{a}} + \mathbf{n}_{\mathbf{a}}) \\ &= \mathbf{p}_{\mathbf{a}} \cdot \mathbf{p}'_{\mathbf{a}} + \mathbf{p}_{\mathbf{a}} \cdot \mathbf{n}'_{\mathbf{a}} & &= \mathbf{p}'_{\mathbf{a}} \cdot \mathbf{p}_{\mathbf{a}} + \mathbf{p}'_{\mathbf{a}} \cdot \mathbf{n}_{\mathbf{a}} \\ &= \mathbf{p}_{\mathbf{a}} \cdot \mathbf{p}'_{\mathbf{a}} + \mathbf{0} & &= \mathbf{p}_{\mathbf{a}} \cdot \mathbf{p}'_{\mathbf{a}} + \mathbf{0}\end{aligned}$$

$$= \mathbf{p_a} \cdot \mathbf{p}'_{\mathbf{a}}, \qquad\qquad = \mathbf{p_a} \cdot \mathbf{p}'_{\mathbf{a}}.$$

Since the bottom two calculations yield the same result, the quantities they begin with must be equal:

$$\mathbf{p_a} \cdot \mathbf{u} = \mathbf{p}'_{\mathbf{a}} \cdot \mathbf{u}.$$

But these are also the beginning quantities of the top two calculations, so those two calculations must have the same result,

$$\|\mathbf{p_a}\|^2 = \left\|\mathbf{p}'_{\mathbf{a}}\right\|^2.$$

Therefore, we can conclude that $\mathbf{p_a}$ and $\mathbf{p}'_{\mathbf{a}}$ are the same length. Since these two vectors are also parallel (because they are both parallel to **a**), we must have that either they are the same vector or are negatives of each other. However, if they were negatives of each other (i.e. $\mathbf{p}'_{\mathbf{a}} = -\mathbf{p_a}$), tracing through the two calculations of $\mathbf{p_a} \cdot \mathbf{u}$ above would tell us that

$$\|\mathbf{p_a}\|^2 = \mathbf{p_a} \cdot \mathbf{u} = \mathbf{p_a} \cdot \mathbf{p}'_{\mathbf{a}} = \mathbf{p_a} \cdot (-\mathbf{p_a}) = -(\mathbf{p_a} \cdot \mathbf{p_a}) = -\|\mathbf{p_a}\|^2,$$

which is only possible if $\|\mathbf{p_a}\| = 0$, in which case $\mathbf{p_a} = \mathbf{0}$, and then also $\mathbf{p}'_{\mathbf{a}} = -\mathbf{p_a} = \mathbf{0}$. Thus, in every case we have $\mathbf{p}'_{\mathbf{a}} = \mathbf{p_a}$. But then

$$\mathbf{n}'_{\mathbf{a}} = \mathbf{u} - \mathbf{p}'_{\mathbf{a}} = \mathbf{u} - \mathbf{p_a} = \mathbf{n_a}.$$

So the two decompositions we started with are actually the same decomposition, and it is not possible to have more than one such decomposition. ■

Remark 13.5.4 Clearly, in this decomposition we have $\mathbf{p_a} = \operatorname{proj}_{\mathbf{a}} \mathbf{u}$ and $\mathbf{n_a} = \mathbf{u} - \operatorname{proj}_{\mathbf{a}} \mathbf{u}$.

13.5.3 Properties of the cross product

Finally, we record a few properties of the cross product.

Remember. The cross product is only defined for vectors in $\mathbb{R}^3$.

Proposition 13.5.5 *Suppose* **u**, **v**, *and* **w** *are vectors in* $\mathbb{R}^3$, *and* k *is a scalar. Then the following hold.*

1. $\mathbf{u} \cdot (\mathbf{u} \times \mathbf{v}) = 0$.
2. $\mathbf{v} \cdot (\mathbf{u} \times \mathbf{v}) = 0$.
3. $\mathbf{u} \times \mathbf{0} = \mathbf{0}$.
4. $\mathbf{0} \times \mathbf{v} = \mathbf{0}$.
5. $\mathbf{v} \times \mathbf{u} = -\mathbf{u} \times \mathbf{v}$.
6. $(k\mathbf{u}) \times \mathbf{v} = k(\mathbf{u} \times \mathbf{v})$.
7. $\mathbf{u} \times (k\mathbf{v}) = k(\mathbf{u} \times \mathbf{v})$.
8. $(\mathbf{u} + \mathbf{v}) \times \mathbf{w} = \mathbf{u} \times \mathbf{w} + \mathbf{v} \times \mathbf{w}$.
9. $\mathbf{u} \times (\mathbf{v} + \mathbf{w}) = \mathbf{u} \times \mathbf{v} + \mathbf{u} \times \mathbf{w}$.
10. $\mathbf{u} \times \mathbf{u} = \mathbf{0}$.
11. *If* **u** *and* **v** *are parallel, then* $\mathbf{u} \times \mathbf{v} = \mathbf{0}$.

Proof idea. The first two statements just reflect the design goal in inventing the cross product: we were looking for a vector that was orthogonal to each of the two input vectors. The rest of the statements follow easily from the determinant formula (†††) for the cross product expressed in Subsection 13.3.6 combined with the properties of the determinant contained in Proposition 9.4.2. We leave detailed proofs to you, the reader. ■

CHAPTER 14

Geometry of linear systems

14.1 Discovery guide

Discovery 14.1 Begin with a set of xy-axes. Draw the vector $\mathbf{x}_0 = (3,0)$ with its tail at the origin, and then draw the vector $\mathbf{p} = (2,1)$ with its tail at the head of $\mathbf{x}_0$.

(a) Consider the expression $\mathbf{x} = \mathbf{x}_0 + t\mathbf{p}$ in the **parameter** t. Think of $\mathbf{x}$ as a variable vector: using different values of t, $\mathbf{x}$ evaluates to different vectors. Draw the vector $\mathbf{x}$ for $t = 1$ on your diagram with its tail at the origin and using a dashed line for the shaft of the arrow. Then do the same for $t = 2$, $t = -1$, $t = 1/2$, $t = -3$.

> **Note.** You shouldn't have to compute any coordinates to be able to draw these vectors, you should be able to just use your initial diagram of $\mathbf{x}_0$ and $\mathbf{p}$ to know where $\mathbf{x}$ ends up for these various values of t.

(b) Suppose you continued sketching in the different possible $\mathbf{x}$ vectors forever, using every possible value for the parameter t. What shape would be traced out by all of the *points* at the heads of the different versions of $\mathbf{x}$?

Discovery 14.2 The equation $x - 2y = 3$ defines a line ℓ in $\mathbb{R}^2$. We can also view this equation as a system of linear equations. Its solution requires one parameter.

(a) Set $y = t$ and then compute the parametric equation for x. Set $\mathbf{x}$ to be the variable vector $\mathbf{x} = \left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$. Fill in the vectors at below. Then compare with Discovery 14.1.

$$\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \\ t \end{bmatrix} = \begin{bmatrix} \\ \end{bmatrix} + t \begin{bmatrix} \\ \end{bmatrix}$$

(b) Use the line equation $x - 2y = 3$ to verify that the point $(4, 1/2)$ lies on ℓ. Then determine the value of the parameter t so that $\mathbf{x} = (4, 1/2)$.

Discovery 14.3 Consider the two planes

$$\Pi_1 : \; 2x - y + 5z = -5, \qquad\qquad \Pi_2 : \; x + 2y - 5z = 10$$

in $\mathbb{R}^3$.

(a) Verify that Π_1 and Π_2 are not parallel.

Hint. Compare their normal vectors.

(b) Two nonparallel planes must intersect in a line. Describe the line of intersection of Π_1 and Π_2 in the form $\mathbf{x} = \mathbf{x}_0 + t\mathbf{p}$.

Hint. Any point in the intersection must lie on both planes at once. That is, any point in the intersection must be a solution to the system of equations formed by the two plane equations.

Discovery 14.4 The equation $x - y + 5z = -5$ defines a plane in $\mathbb{R}^3$. We can also view this equation as a system of linear equations.

(a) Similarly to Discovery 14.2, determine vectors $\mathbf{x}_0$, $\mathbf{p}_1$, and $\mathbf{p}_2$ so that

$$\mathbf{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \mathbf{x}_0 + s\mathbf{p}_1 + t\mathbf{p}_2$$

describes all solutions to the equation (and hence all points on the plane).

(b) Use the plane's equation $x - y + 5z = -5$ to verify that the point $(1,1,-1)$ lies on the plane.

Then determine the values of the parameters s and t so that the formula

$$\mathbf{x} = \mathbf{x}_0 + s\mathbf{p}_1 + t\mathbf{p}_2$$

results in this point $\mathbf{x} = (1,1,-1)$.

Discovery 14.5 Draw a grid over the xy-plane, with a vertical line at each integer value of x and a horizontal line at each integer value of y. Then draw $\mathbf{e}_1$ and $\mathbf{e}_2$ on your diagram.

What does the decomposition $(3,2) = 3\mathbf{e}_1 + 2\mathbf{e_2}$ look like on your grid?
How about $(-1,2) = (-1)\mathbf{e}_1 + 2\mathbf{e}_2$?
How about $(3/2,-2) = (3/2)\mathbf{e}_1 + (-2)\mathbf{e}_2$?

Discovery 14.6 Draw a "grid" over the xy-plane as follows: at each integer value along the x-axis, draw both a vertical line and a slant line parallel to the line $y = x$. Then draw $\mathbf{u} = (1,1)$ and $\mathbf{e}_2$ on your diagram.

What does the decomposition $(3,2) = 3\mathbf{u} + (-1)\mathbf{e}_2$ look like on your grid?
How about $(-1,2) = (-1)\mathbf{u} + 3\mathbf{e}_2$?
How about $(3/2,-2) = (3/2)\mathbf{u} + (-7/2)\mathbf{e}_2$?

Discovery 14.7 The set of all solutions to the homogeneous equation $x - 2y + 3z = 0$ forms a plane in $\mathbb{R}^3$. We can solve this equation by assigning parameters $y = s$ and $z = t$, so that all solutions can be described parametrically by

$$(x,y,z) = s(2,1,0) + t(-3,0,1).$$

Discuss how the vectors $\mathbf{p}_1 = (2,1,0)$ and $\mathbf{p}_2 = (-3,0,1)$ create a "grid" on the plane defined by $x - 2y + 3z = 0$, similarly to the grids you worked with in Discovery 14.5 and Discovery 14.6.

Note. Since the plane equation

$$x - 2y + 3z = 0$$

is homogeneous, this plane passes through the origin.

Discovery 14.8 Determine the point of intersection of the line ℓ, described parametrically below left, and the plane Π, described algebraically below right.

$$\ell\colon\ \mathbf{x} = (2,0,3) + t(-1,1,1) \qquad\qquad \Pi\colon\ 2x + y - 3z = 7$$

Hint. The point of intersection is simultaneously on the line and on the plane.

Discovery 14.9 Set up a system of equations whose solution is the point of intersection of the line ℓ and the plane Π, described parametrically below.

$$\ell\colon\ \mathbf{x} = (2,0,3) + t(-1,1,1) \qquad \Pi\colon\ \mathbf{x} = (3,1,0) + r(1,1,1) + s(3,0,2)$$

Hint. The point of intersection is simultaneously on the line and on the plane.

14.2 Terminology and notation

point-parallel form (of a line in $\mathbb{R}^n$)
the vector equation $\mathbf{x} = \mathbf{x}_0 + t\mathbf{p}$, where $\mathbf{x}_0$ is a vector from the origin to a known point on the line, $\mathbf{p}$ is a known parallel vector for the line, $\mathbf{x}$ is a variable vector representing an arbitrary point on the line (again as a vector from the origin), and t is a scalar parameter that varies as the arbitrary vector $\mathbf{x}$ varies

point-parallel form (of a plane in $\mathbb{R}^n$)
the vector equation $\mathbf{x} = \mathbf{x}_0 + s\mathbf{p}_1 + t\mathbf{p}_2$, where $\mathbf{x}_0$ is a vector from the origin to a known point on the plane, $\mathbf{p}_1, \mathbf{p}_2$ are known parallel vectors for the plane that are not parallel to each other, $\mathbf{x}$ is a variable vector representing an arbitrary point on the plane (again as a vector from the origin), and s, t are scalar parameters that vary as the arbitrary vector $\mathbf{x}$ varies

14.3 Concepts

In this section.

14.3.1 Lines in the plane

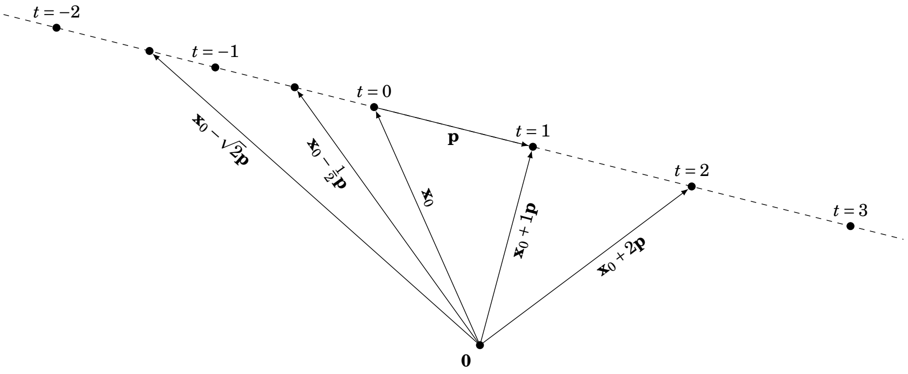
When we view a single linear equation in two variables as a (very simple) system of equations, we require a single parameter to solve. We've previously seen that we can use matrix algebra to express the general solution to a system of equations as a linear combination of column matrices, where the parameters appear as coefficients.

See. Examples 4.4.8–4.4.11 in Subsection 4.4.4. But we also reminded ourselves of this in Discovery 14.2.

When we interpret the column matrices in such a linear combination as vectors, we can investigate the geometry of the set of solutions, as we did in Discovery 14.1. For a general solution to $ax + by = c$ of the form

$$\mathbf{x} = \mathbf{x}_0 + t\mathbf{p},$$

the vector $\mathbf{x}_0$ corresponds to the particular solution for $t = 0$, and we can think of its terminal point as an "base" point on the line. When we vary the value of the parameter t, we get solutions that are vector sums of the base point $\mathbf{x}_0$ and scalar multiples of $\mathbf{p}$. Geometrically, these vector sums all involve tacking some scaled copy of $\mathbf{p}$ onto the end of $\mathbf{x}_0$.


$t = -2$
$t = -1$
$t = 0$
$t = 1$
$t = 2$
$t = 3$
$\mathbf{p}$
$\mathbf{x}_0 - \sqrt{2}\mathbf{p}$
$\mathbf{x}_0 - \frac{1}{2}\mathbf{p}$
$\mathbf{x}_0$
$\mathbf{x}_0 + 1\mathbf{p}$
$\mathbf{x}_0 + 2\mathbf{p}$
$\mathbf{0}$

The terminal points of all such vectors $\mathbf{x}$ trace out the line parallel to $\mathbf{p}$ that passes through the terminal point of $\mathbf{x}_0$. Since $\mathbf{p}$ is parallel to the line, we might think of $\mathbf{p}$ as a "direction" vector for the line.

14.3.2 Lines in space

Two nonparallel planes in space must intersect in a line, as in Discovery 14.3. If we have algebraic equations $a_1x + b_1y + c_1z = d_1$ and $a_2x + b_2y + c_2z = d_2$ for these planes, then solving for the points of intersection is the same as solving the linear system formed by these two equations. The assumption that the planes are not parallel guarantees that we will need one (and only one) parameter to solve the system, and then the general solution can be expressed in a vector form

$$\mathbf{x} = \mathbf{x}_0 + t\mathbf{p},$$

just as in the previous case of a line in the plane. To visualize, we can imagine the diagram in the previous subsection above as floating in space instead of lying in the plane.

14.3.3 Planes in space

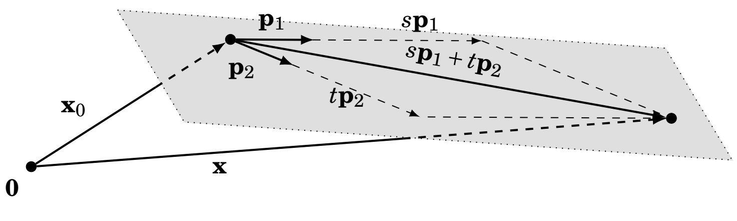
When we view a single linear equation in three variables as a system of equations, we require *two* parameters to solve. As before, we can use matrix algebra to express the general solution as a linear combination of column matrices, where the parameters appear as coefficients.

Recall. We explored this situation in Discovery 14.4.

Similarly to the vector description of a line, a parametric vector expression

$$\mathbf{x} = \mathbf{x}_0 + s\mathbf{p}_1 + t\mathbf{p}_2$$

can be interpreted as follows. The terminal point of the vector $\mathbf{x}_0$ is an "base" point on the plane, corresponding to parameter values $s = 0$ and $t = 0$. The vectors $\mathbf{p}_1$ and $\mathbf{p}_2$ are parallel to the plane. Similarly to the vector description of a line, as we vary the values of s and t we obtain other points on the plane by tacking on *linear combinations* of $\mathbf{p}_1$ and $\mathbf{p}_2$ to the end of $\mathbf{x}_0$.



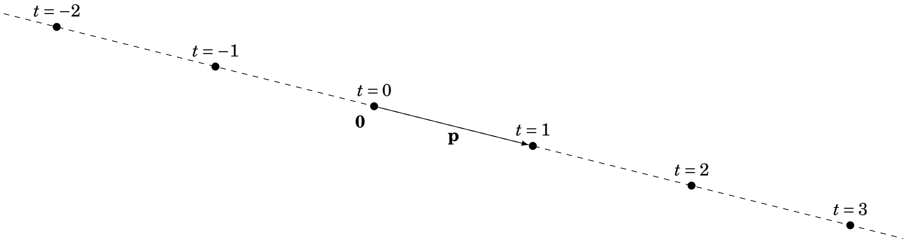
Note. In this diagram, all of the vectors lie flat in the shaded plane (and so, parallel to it) except for $\mathbf{x}_0$ and $\mathbf{x}$, which point from the origin into the plane.

14.3.4 Parallel vectors as a "basis" for lines and planes

In a vector description $\mathbf{x} = \mathbf{x}_0 + t\mathbf{p}$ for a line, the "base" point at the head of $\mathbf{x}_0$ gets you onto the line, and then one can get to any other point on the line by following a scalar multiple of the parallel vector $\mathbf{p}$. In this way, the parameter t effectively places a coordinate system on the line, where the integers are spaced apart by the length of $\mathbf{p}$.

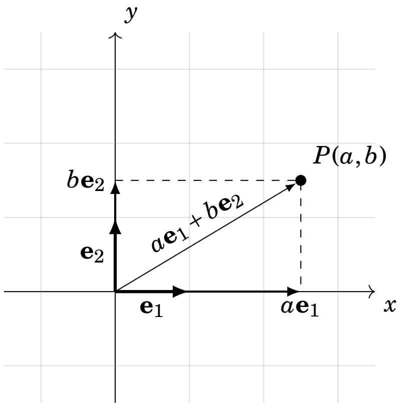
See. the line diagram earlier in Subsection 14.3.1.

Values of the parameter t are mapped to specific positions on the line, just as when we visualize the set of real numbers $\mathbb{R}$ along the real number line, where each real number represents a position on a line. This idea of a coordinate system along the line is more natural when the line passes through the origin, so that we can take $\mathbf{x}_0 = \mathbf{0}$. In this case we have $\mathbf{x} = t\mathbf{p}$, so that all points on the line correspond to scalar multiples of the parallel vector $\mathbf{p}$, and parameter value $t = 0$ corresponds to the origin. So the vector $\mathbf{p}$ tells us pretty much all we need to know about the line, and any other line that is parallel to this line could use the same parallel vector $\mathbf{p}$, it would just need a different "base point" vector $\mathbf{x}_0$.



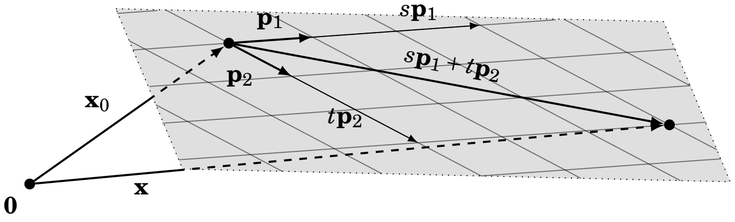
In the plane, the standard basis vectors $\mathbf{e}_1$,$\mathbf{e}_2$ play the same role *for the whole plane*, representing our xy-coordinate system and setting up a grid as in Discovery 14.5.

Recall. We have previously explored how vectors in the plane can be decomposed as linear combinations of the standard basis vectors: in Discovery 11.8, and further back in Subsection 11.3.7.



Notice. In this diagram, vertical grid lines appear at multiples of $\mathbf{e}_1$, and horizontal grid lines appear at multiples of $\mathbf{e}_2$.

When we have a vector description $\mathbf{x} = \mathbf{x}_0 + s\mathbf{p}_1 + t\mathbf{p}_2$ for a plane in space, scalar multiples of the vectors $\mathbf{p}_1$ and $\mathbf{p}_2$ form a grid on the plane in the same way (except that the grid lines will not be at right angles to each other if $\mathbf{p}_1$ and $\mathbf{p}_2$ are not).



The vectors $\mathbf{p}_1$ and $\mathbf{p}_2$ set up an st-coordinate system on the plane, where every point on the plane corresponds to a particular pair of parameter values, and vice versa, by adding the linear combination $s\mathbf{p}_1 + t\mathbf{p}_2$ onto $\mathbf{x}_0$. If the plane passes through the origin (as in Discovery 14.7), then we can take $\mathbf{x}_0$ to be the zero vector, so that the origin corresponds to $(s,t) = (0,0)$. Then every other point in the plane could be constructed as a linear combination of $\mathbf{p}_1$ and $\mathbf{p}_2$.

14.3.5 Summary

Combining with our knowledge of normal vectors from the previous chapter, we now have several ways to describe lines and planes in $\mathbb{R}^2$ and $\mathbb{R}^3$.

	Algebraic	Geometric	Vector
line in $\mathbb{R}^2$	$ax + by = c$	$\mathbf{n} \cdot (\mathbf{x} - \mathbf{x}_0) = 0$ where $\mathbf{n} = (a,b)$	$\mathbf{x} = \mathbf{x}_0 + t\mathbf{p}$
plane in $\mathbb{R}^3$	$ax + by + cz = d$	$\mathbf{n} \cdot (\mathbf{x} - \mathbf{x}_0) = 0$ where $\mathbf{n} = (a,b,c)$	$\mathbf{x} = \mathbf{x}_0 + s\mathbf{p}_1 + t\mathbf{p}_2$
line in $\mathbb{R}^3$	intersection of planes $a_1x + b_1y + c_1z = d_1$ and $a_2x + b_2y + c_2z = d_2$	common $\mathbf{x}$ that satisfy $\mathbf{n}_1 \cdot (\mathbf{x} - \mathbf{x}_0) = 0$ and $\mathbf{n}_2 \cdot (\mathbf{x} - \mathbf{x}_0) = 0$ where $\mathbf{n}_1 = (a_1, b_1, c_1)$, $\mathbf{n}_2 = (a_2, b_2, c_2)$	$\mathbf{x} = \mathbf{x}_0 + t\mathbf{p}$

Figure 14.3.1

Remark 14.3.2

- In both the Geometric and Vector columns, the vector $\mathbf{x}_0$ represents a fixed "base" point that is on the line or plane.

- In the Geometric column, the $\mathbf{n}$ vectors are normal vectors to the line or plane, and their components are precisely the coefficients from the corresponding entry in the Algebraic column. Note that in $\mathbb{R}^3$, there are 360° of normal directions to a line, so we need *two* normal vectors ($\mathbf{n}_1$ and $\mathbf{n}_2$) to be able to specify the direction of the line — and then the line is parallel to $\mathbf{n}_1 \times \mathbf{n}_2$. While the normal vector $\mathbf{n}$ for a line in $\mathbb{R}^2$ or a plane in $\mathbb{R}^3$ are essentially unique (for a specific line or plane, it can only be replaced by a nonzero scalar multiple), the pair of normal vectors for a line in $\mathbb{R}^3$ is not unique (there are many pairs of normal vectors that are not just scalar multiples of other pairs that would describe the same line). We can say something about $\mathbf{n}_1$ and $\mathbf{n}_2$ though — for a given line in $\mathbb{R}^3$, every such pair of normal vectors must be parallel to a plane that is normal to the line.

- In the Vector column, the $\mathbf{p}$ vectors are parallel to the line or plane. For a line in either $\mathbb{R}^2$ or $\mathbb{R}^3$, we would just need to know a second "base point" vector $\mathbf{x}_1$, and the we could take $\mathbf{p} = \mathbf{x}_1 - \mathbf{x}_0$. Or, for a line in $\mathbb{R}^3$, we could start with two known, nonparallel normal vectors $\mathbf{n}_1, \mathbf{n}_2$ for the line, and then we could take $\mathbf{p} = \mathbf{n}_1 \times \mathbf{n}_2$. For a plane in $\mathbb{R}^3$, we need *three* "known" points total, represented by some vectors $\mathbf{x}_0$, $\mathbf{x}_1$, $\mathbf{x}_2$. As long as these "known" points are not noncollinear, we can get the necessary vectors parallel to that plane by taking $\mathbf{p}_1 = \mathbf{x}_1 - \mathbf{x}_0$ and $\mathbf{p}_2 = \mathbf{x}_2 - \mathbf{x}_0$.

- We can realize similar geometric "shapes" in $\mathbb{R}^4$, $\mathbb{R}^5$, and higher dimensions, even though we can't "see" them. A line or plane in higher dimensions would have the same kind of vector description. The algebraic and geometric descriptions of lines in $\mathbb{R}^2$ and planes in $\mathbb{R}^3$, if adapted to be used in higher dimensions, would describe a *hyper*plane — some sort of "subspace" of n-dimensional space that is of one dimension lower. For example, similarly to how we might think of a plane in $\mathbb{R}^3$ as a "copy" of *the* plane ($\mathbb{R}^2$) sitting inside space ($\mathbb{R}^3$), we might imagine a hyperplane in $\mathbb{R}^4$ as a "copy" of $\mathbb{R}^3$ sitting inside $\mathbb{R}^4$.

14.4 Examples

In this section.

14.4.1 Describing lines and planes parametrically

First we will work out some of the activities from Discovery guide 14.1 that involve describing lines and planes parametrically.

Example 14.4.1 Parametrically describing a line in $\mathbb{R}^2$. In Discovery 14.2, we considered the equation $x - 2y = 3$ for a line in the plane. Setting parameter $y = t$ and isolating x in this equation leads to general solution

$$\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3+2t \\ t \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \end{bmatrix} + t \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

Geometrically, the vector $\mathbf{x}_0 = (3, 0)$ represents an "base" point on the line, and algebraically represents the particular solution to the system obtained from parameter value $t = 0$. The vector $\mathbf{p} = (2,1)$ represents a vector parallel to the line, and every other point on the line (i.e. every other solution to the system) can be obtained by adding a scalar multiple of $\mathbf{p}$ to $\mathbf{x}_0$. For example, the point $\mathbf{x} = (4, 1/2)$ lies on the line, as we can verify by checking

$$4 - 2 \cdot \frac{1}{2} = 4 - 1 = 3,$$

so that the coordinates of the point satisfy the line equation $x - 2y = 3$. We can solve for t in the vector equation

$$\begin{bmatrix} 4 \\ \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \end{bmatrix} + t \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

to see exactly how this point is a multiple of $\mathbf{p}$ away from $\mathbf{x}_0$:

$$\begin{bmatrix} 4 \\ \frac{1}{2} \end{bmatrix} - \begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 2t \\ t \end{bmatrix} \qquad \implies \qquad t = \frac{1}{2}. \tag{$*$}$$

Finally, we have

$$\begin{bmatrix} 4 \\ \frac{1}{2} \end{bmatrix} = \mathbf{x} + \frac{1}{2}\mathbf{p}.$$

□

Warning 14.4.2 In the previous example, we determined the value of the parameter t that corresponds to the point $(4, 1/2)$ on the line by solving vector equation $(*)$. When you are solving vector equations for parameters like this, make sure you check that your solution works in *every* coordinate! For example, in $(*)$ we see that $t = 1/2$ immediately from comparing the second coordinate on left and right. But it is important to check that this parameter value also works in the first coordinate (which it does).

Example 14.4.3 Parametrically describing the intersection of two planes in $\mathbb{R}^3$. In Discovery 14.3, we considered a system consisting of two equations in three variables,

$$\left\{ \begin{array}{rcrcrcr} 2x & - & y & + & 5z & = & -5, \\ x & + & 2y & - & 5z & = & 10. \end{array} \right.$$

Geometrically, each of the equations represents a plane in space, and solutions to the system represent points that are in common to both planes (that is, points in the intersection of the two planes). From the coefficients of the equations we may take $\mathbf{n}_1 = (2, -1, 5)$ and $\mathbf{n}_2 = (1, 2, -5)$ as normal vectors for the two planes, respectively. Since these normal vectors are not parallel, neither are the planes, and so they must intersect. Algebraically, this means that the coefficient parts of the two equations are not multiples of each other, so when we row reduce we will find two leading ones, representing the two independent equations with which we started. And so we will only require one parameter to express the general solution, which will then take the form of a line. The free variable in this system is z, so setting z to parameter t and solving we get

$$\mathbf{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -t \\ 5 + 3t \\ t \end{bmatrix} = \begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix} + t \begin{bmatrix} -1 \\ 3 \\ 1 \end{bmatrix}.$$

Here, the "base" point corresponding to $t = 0$ is $\mathbf{x}_0 = (0, 5, 0)$, and the vector $\mathbf{p} = (-1, 3, 1)$ is parallel to the line. □

Example 14.4.4 Parametrically describing a plane in $\mathbb{R}^3$. When we have just a single plane in space, as in Discovery 14.4, we can view its equation as a system of equations, just as we did with the line in Discovery 14.2. In that activity, we worked with equation $x - y + 5z = -5$. For this equation, we will need two parameters to express the general solution, and each parameter will provide us with a vector parallel to the plane. Setting $y = s$ and $z = t$, we can then use the plane equation to express x in terms of these parameters. This leads to general solution, in vector form:

$$\mathbf{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -5 + s - 5t \\ s \\ t \end{bmatrix} = \begin{bmatrix} -5 \\ 0 \\ 0 \end{bmatrix} + s \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -5 \\ 0 \\ 1 \end{bmatrix}.$$

Here, the "base" point on the plane is $\mathbf{x}_0 = (-5, 0, 0)$, which corresponds to parameter values $s = t = 0$. Every other point on the plane corresponds to other choices of parameter values. For example, as in the discovery activity, the point $(1, 1, -1)$ is on the plane. We can verify this by checking

$$1 - 1 + 5(-1) = -5,$$

so that the coordinates of the point satisfy the plane equation $x - y + 5z = -5$. We can also describe this point using the vector equation

$$\mathbf{x} = \mathbf{x}_0 + s\mathbf{p}_1 + t\mathbf{p}_2$$

as follows:

$$\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} -5 \\ 0 \\ 0 \end{bmatrix} + s \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -5 \\ 0 \\ 1 \end{bmatrix}.$$

Solving this vector equation for s and t leads to a ... system of linear equations!

$$\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} - \begin{bmatrix} -5 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} s \\ s \\ 0 \end{bmatrix} + \begin{bmatrix} -5t \\ 0 \\ t \end{bmatrix}$$
$$\begin{bmatrix} 6 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} s - 5t \\ s \\ t \end{bmatrix}.$$

From the second and third coordinates we immediately see $s = 1$ and $t = -1$. However, it's important to also verify that $s - 5t = 6$ for this choice of parameter values, to satisfy the equality of the two first coordinates on right and left. □

14.4.2 Determining points of intersection

When lines and/or planes are described using algebraic equations, determining points of intersection only requires solving the linear systems that those equations form together. Here we will demonstrate determining points of intersection when some or all of the lines and/or planes involved are described parametrically by working out some of the activities from Discovery guide 14.1.

Example 14.4.5 Intersection of a parametrically-described line and an algebraically-described plane in $\mathbb{R}^3$. In Discovery 14.8, we have a line described by a vector equation and a plane described algebraically, and would like to determine their point of intersection (if there is one). Any such point of intersection must be on the line, and so its coordinates can be described in terms of the parameter t:

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

If this point is also on the plane, its coordinates must satisfy the equation for the plane:

$$\begin{aligned} 2x + y - 3z &= 7 \\ 2(2-t) + t - 3(3+t) &= 7 \\ -5 - 4t &= 7 \\ t &= -3. \end{aligned}$$

Substituting this parameter value in our expression for $\mathbf{x}$ gives us the point of intersection:

$$\begin{bmatrix} 5 \\ -3 \\ 0 \end{bmatrix}.$$

Check your understanding. What if the line and plane had been parallel with no point of intersection? What would have happened when we tried to solve for t? Or, what if the line and plane had been parallel, but with the line lying inside the plane? How would this have become evident from the algebra of attempting to solve for t?

□

Example 14.4.6 Intersection of parametrically-described line and plane in $\mathbb{R}^3$. In Discovery 14.9, we again want to determine the point of intersection (if any) of a line and a plane, but this time both line and plane are described by vector equations. If a point lies on *both* line and plane, its coordinates must have a *simultaneous* description by both vector equations in terms of parameters:

$$\mathbf{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 3 \end{bmatrix} + t \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2-t \\ t \\ 3+t \end{bmatrix},$$

$$\mathbf{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \\ 0 \end{bmatrix} + r \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + s \begin{bmatrix} 3 \\ 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 3+r+3s \\ 1+r \\ r+2s \end{bmatrix}.$$

Now, this point can only have one set of coordinates, so these two descriptions must actually be the same. This lets us set up a ... system of linear equations!

$$\begin{aligned} x&: \quad 2-t = 3+r+3s, \\ y&: \quad t = 1+r, \\ z&: \quad 3+t = r+2s, \end{aligned} \qquad \Longrightarrow \qquad \left\{ \begin{array}{rcrcrcr} r & + & 3s & + & t & = & -1, \\ r & & & - & t & = & -1, \\ r & + & 2s & - & t & = & 3. \end{array} \right.$$

We could put this system in a matrix and row reduce, but we only really care about the value of parameter t in the solution, because knowing t allows us to determine $\mathbf{x}$ from the vector description for the line. So we can use Cramer's rule instead. Write A for the coefficient matrix of this system, and A_3 for the matrix obtained by replacing the third column of A by the vector of constants. Then,

$$\det A_3 = \begin{vmatrix} 1 & 3 & -1 \\ 1 & 0 & -1 \\ 1 & 2 & 3 \end{vmatrix} = -12,$$

$$\det A = \begin{vmatrix} 1 & 3 & 1 \\ 1 & 0 & -1 \\ 1 & 2 & -1 \end{vmatrix} = 4$$

$$\implies \quad t = \frac{\det A_3}{\det A} = -3.$$

Now that we have $t = -3$, we can determine the point of intersection:

$$\mathbf{x} = \begin{bmatrix} 2-t \\ t \\ 3+t \end{bmatrix} = \begin{bmatrix} 5 \\ -3 \\ 0 \end{bmatrix}.$$

Note. We got the same answer as in the previous example because the lines and planes in the two discovery activities are actually the same, but the plane is described in two different ways in the two examples.

□

CHAPTER 15

Abstract vector spaces

15.1 Discovery guide

Suppose you have a collection of mathematical objects. The objects in the collection may satisfy all/some/none of the following rules, depending on the objects. In the rule statements, bold variable letters represent arbitrary objects in the collection, and ordinary variable letters represent arbitrary **scalars** (i.e. numbers).

List 15.1.1 (A) Addition rules

1. The objects can be *added* (two at a time), and the resulting "sum" is always equal to another in the collection of objects.
2. Every $\mathbf{v}, \mathbf{w}$ satisfy
$$\mathbf{w} + \mathbf{v} = \mathbf{v} + \mathbf{w}.$$
3. Every $\mathbf{u}, \mathbf{v}, \mathbf{w}$ satisfy
$$\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}.$$
4. There is a special **zero object** in the collection, so that every $\mathbf{v}$ satisfies $\mathbf{v} + \mathbf{0} = \mathbf{v}$.
5. Every $\mathbf{v}$ has an **opposite** $\widetilde{\mathbf{v}}$ so that $\mathbf{v} + \widetilde{\mathbf{v}} = \mathbf{0}$.

List 15.1.2 (S) Scalar multiplication rules

1. The objects can be *scaled* by a numerical factor (called a **scalar**), and the resulting "scaled object" is always equal to another in the collection of objects.
2. Every $k, \mathbf{v}, \mathbf{w}$ satisfy
$$k(\mathbf{v} + \mathbf{w}) = k\mathbf{v} + k\mathbf{w}.$$
3. Every $k, m, \mathbf{v}$ satisfy
$$(k + m)\mathbf{v} = k\mathbf{v} + m\mathbf{v}.$$
4. Every $k, m, \mathbf{v}$ satisfy
$$k(m\mathbf{v}) = (km)\mathbf{v}.$$
5. Every $\mathbf{v}$ satisfies $1\mathbf{v} = \mathbf{v}$.

Discovery 15.1 Read and briefly discuss the rules in your group. In particular, make sure everyone in your group understands the differences between the LHS and RHS in each of Rule A 2, Rule A 3, Rule S 2, Rule S 3, and Rule S 5.

It may help to come up with expressions for these algebra rules in *plain*

English rather than letters and variables. For example, Rule A 2 states that order doesn't matter in adding objects.

Discovery 15.2 These rules are modelled on the properties of vectors in $\mathbb{R}^n$. Convince yourself that all the rules are true when the collection of mathematical objects considered is "all vectors in $\mathbb{R}^2$." In particular, make sure you know what the *zero object* is in the collection (Rule A 4), and how to determine an object's *opposite* (Rule A 5).

Discovery 15.3 For each of the following collections of objects, convince yourself that all the rules are true. In particular, make sure you know what the *zero object* is in the collection (Rule A 4), and how to determine an object's *opposite* (Rule A 5).

(a) All 2×2 matrices.

(b) All $m \times n$ matrices. (Here m and n are some specific but unknown numbers.)

(c) All polynomials in the variable x.

(d) All polynomials in the variable x of degree 2 or less (i.e. no x^3 or higher allowed).

(e) All real numbers.

> **Careful.** In the last example collection, both *objects* and *scalars* are numbers. Don't get mixed up!

Discovery 15.4 Suppose you have a collection of objects that satisfies all of the rules. (Don't pick a specific example collection, just think in the abstract.)

(a) For an object $\mathbf{v}$, is it necessarily always true that $\mathbf{0} + \mathbf{v} = \mathbf{v}$?

Hint. Look at Rule A 2 and Rule A 4.

(b) For an object $\mathbf{v}$ and its opposite $\widetilde{\mathbf{v}}$, is it necessarily always true that $\widetilde{\mathbf{v}} + \mathbf{v} = \mathbf{0}$?

Hint. Look at Rule A 2 and Rule A 5.

(c) By Rule A 5, every object has an opposite which itself is an object. What is the opposite of an opposite? Make sure you can justify that your answer satisfies the definition of **opposite** contained in Rule A 5.

(d) Suppose $\mathbf{v}$ is an object. What object do you think $0\mathbf{v}$ should be equal to? Do the rules provide *direct* evidence to support your guess?

(e) Here is a justification of your guess from Task d. (Assuming you guessed correctly!) Fill in the blanks with the identifier of the rule that justifies each step, working down the left-hand side first. Make sure you understand how and for what objects that rule is being applied.

$\mathbf{v} + \widetilde{\mathbf{v}} = \mathbf{0}$		$0\mathbf{v} + (1\mathbf{v} + \widetilde{\mathbf{v}}) = \mathbf{0}$	
$1\mathbf{v} + \widetilde{\mathbf{v}} = \mathbf{0}$		$0\mathbf{v} + (\mathbf{v} + \widetilde{\mathbf{v}}) = \mathbf{0}$	
$(0+1)\mathbf{v} + \widetilde{\mathbf{v}} = \mathbf{0}$	(arithmetic)	$0\mathbf{v} + \mathbf{0} = \mathbf{0}$	
$(0\mathbf{v} + 1\mathbf{v}) + \widetilde{\mathbf{v}} = \mathbf{0}$		$0\mathbf{v} = \mathbf{0}$	

(f) Use the rules to "simplify" the expression $\mathbf{v} + (-1)\mathbf{v}$. Make sure each step is justified by a specific rule, similarly to Task e.

> **Note.** As well as the rules from the top of this discovery guide, you may also use your newly justified rule from Task e. This is a useful pattern: every time we use existing rules to create a new rule, that new rule can be freely used to help create even more rules.

Hint. Start by using Rule S 5 backwards, as used to transform the first line to the second in Task e.

(g) Take $\mathbf{v}+(-1)\mathbf{v}=X$, where X is your final simplified expression from Task f. We can "cancel" the $\mathbf{v}$ from the LHS by adding $\widetilde{\mathbf{v}}$ to both sides of the equality. Based on the resulting equality after doing that, what do you think is a better name for $\widetilde{\mathbf{v}}$ than *opposite of* $\mathbf{v}$?

Discovery 15.5 Nominate one member of your group to become an object, and consider the collection of objects that consists of just *one* object (namely, the group member you nominated).

(a) Can you come up with some sort of addition so that Rule A 1 is true?

(b) Can you come up with some sort of scaling operation so that Rule S 1 is true?

(c) Check whether the other eight rules hold true with the operations you have devised in this activity.

15.2 Motivation

The rules of vector algebra listed in Proposition 11.5.1 are valid whether we take a geometric view (using directed line segments) or an an algebraic view (using column vectors) of vector addition and scalar multiplication. But all these rules have counterparts in matrix algebra in Proposition 4.5.1, which suggests that these algebra patterns might be more universal — are there other collections of algebraic objects that can be added and scaled and that follow the same rules of algebra when we do so?

If we observe similar algebraic patterns elsewhere (and we will), then it is worth the effort to *abstract* the concepts of vector and vector algebra — to disassociate them from any specific ideas of what they are, and deal with them as abstract concepts. This is ultimately where mathematics becomes most powerful: when it recognizes, describes, and analyzes patterns in familiar contexts that can then be recognized and exploited in new contexts.

The cycle of life of a mathematical idea is as follows.

- Extract common patterns from familiar *model* systems (on the left in Figure 15.2.1 below).
- Describe the core features of these common patterns and use them as the basis for an abstract system.
- Deduce new properties of the abstract system based on the aspects of the underlying patterns that describe it.
- Recognize the common patterns described by the abstract system in new systems (on the right in Figure 15.2.1 below).
- Interpret the new abstract properties back in the known systems, new and old, and apply these properties to solve problems.

Since the abstract properties are logically deduced from the underlying patterns that defines the abstract system, every specific system that follows these common patterns must have the same properties.

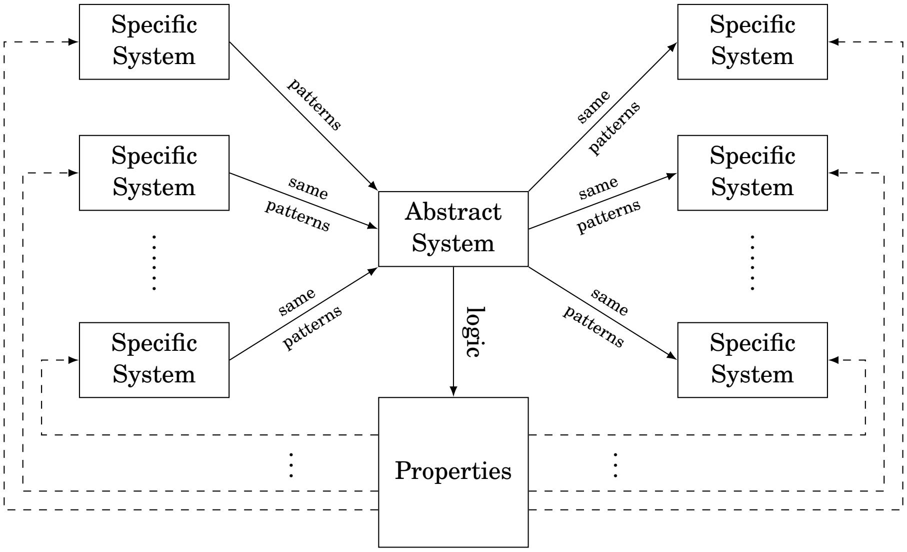


Figure 15.2.1 The cycle of abstract mathematical models.

Following this cycle for systems that follow the patterns of the rules of algebra for **vector addition** and **scalar multiplication** is our task for the next few chapters. In this chapter, we begin our study of the abstract system we can extract from our familiar model systems of vectors in $\mathbb{R}^n$ and $m \times n$ matrices, both of which satisfy the same rules of algebra with respect to addition and scalar multiplication.

15.3 Terminology and notation

vector addition
a rule for associating to a pair of objects $\mathbf{v}$ and $\mathbf{w}$ a third object $\mathbf{v} + \mathbf{w}$

scalar multiplication
a rule for associating to a number k and an object $\mathbf{v}$ another object $k\mathbf{v}$

vector space
a collection of mathematical objects, along with appropriate conceptions of vector addition and scalar multiplication, that satisfies the Vector space axioms

vector an object in a vector space

zero vector
the special vector $\mathbf{0}$ in a vector space that satisfies vector addition Axiom A 4

negative vector (of a vector v)
the special vector $-\mathbf{v}$ that satisfies vector addition Axiom A 5 relative to $\mathbf{v}$

vector subtraction
for vectors $\mathbf{v}$ and $\mathbf{w}$, write $\mathbf{v} - \mathbf{w}$ to mean $\mathbf{v} + (-\mathbf{w})$

trivial vector space
a vector space that consists of a single object, which then must be the zero vector in that space; also called the **zero vector space**

Here follows the notation we will use for some common vector space examples.

$\mathbb{R}^n$ the usual vector space of n-tuples of real numbers that we have been studying in Chapters 11–14

$\mathrm{M}_{m \times n}(\mathbb{R})$ the vector space of all $m \times n$ matrices with entries that are real numbers; when $m = n$ we sometimes just write $\mathrm{M}_n(\mathbb{R})$ to mean the vector space of all square $n \times n$ matrices

$\mathrm{P}(\mathbb{R})$ the vector space of all polynomials with real coefficients in a single variable

$\mathrm{P}_n(\mathbb{R})$ the vector space of all polynomials with real coefficients in a single variable that have degree n or less

$F(D)$ the vector space of all real-valued functions that are defined on the domain D

15.4 Concepts

In this section.

- Subsection 15.4.1 *The ten vector space axioms*
- Subsection 15.4.2 *Instances of vector spaces*

15.4.1 The ten vector space axioms

A vector space consists of a collection of objects, which are usually all of the same kind. For example, the collection of all vectors in $\mathbb{R}^2$, or the collection of all 3×5 matrices. To do the type of vector algebra we are familiar with, we need two operations that can be performed with these objects: some sort of **addition**, and some sort of **scalar multiplication**. So that algebra with these objects and operations works the way we expect, we demand that the operations always conform to the following rules, called **axioms**. Essentially, these rules consist of our "favourite" properties of algebra with vectors in $\mathbb{R}^n$ and of algebra with matrices, and we would like to explore whether similar algebraic systems can be found elsewhere.

Definition 15.4.1 Vector space axioms. A collection of objects is called a **vector space**, and the objects inside are then referred to as **vectors**, if the collection satisfies all ten of the following axioms. In the axiom statements, bold variable letters represent arbitrary objects in the collection, and ordinary variable letters represent arbitrary scalars (i.e. numbers).

List 15.4.2 (A) Addition axioms

1. The objects can be *added* (two at a time), and the resulting "sum" is always equal to another in the collection of objects.
2. Every $\mathbf{v}, \mathbf{w}$ satisfy
$$\mathbf{w} + \mathbf{v} = \mathbf{v} + \mathbf{w}.$$
3. Every $\mathbf{u}, \mathbf{v}, \mathbf{w}$ satisfy
$$\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}.$$
4. There is a special **zero object** in the collection, so that every $\mathbf{v}$ satisfies $\mathbf{v} + \mathbf{0} = \mathbf{v}$.
5. Every $\mathbf{v}$ has a **negative** $-\mathbf{v}$ so that $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$.

List 15.4.3 (S) Scalar multiplication axioms

1. The objects can be *scaled* by a numerical factor (called a **scalar**), and the resulting "scaled object" is always equal to another in the collection of objects.
2. Every $k, \mathbf{v}, \mathbf{w}$ satisfy
$$k(\mathbf{v} + \mathbf{w}) = k\mathbf{v} + k\mathbf{w}.$$
3. Every $k, m, \mathbf{v}$ satisfy
$$(k + m)\mathbf{v} = k\mathbf{v} + m\mathbf{v}.$$
4. Every $k, m, \mathbf{v}$ satisfy
$$k(m\mathbf{v}) = (km)\mathbf{v}.$$
5. Every $\mathbf{v}$ satisfies $1\mathbf{v} = \mathbf{v}$.

◊

Remark 15.4.4

- In Axiom A 5, the negative symbol ***does not*** mean that we are multiplying the vector by -1. It is literally just a negative *symbol*, and should be read as "the negative of." So for a vector $\mathbf{v}$ in a vector space, the symbols $-\mathbf{v}$ mean "the vector that is the negative of $\mathbf{v}$," defined by the property that it adds with $\mathbf{v}$ to the special zero vector.

  A look ahead. The algebra of vectors *will* lead to a connection between the negative of a vector with respect to addition, and the scalar multiple of a vector by the scalar -1. See Rule 2.e from Proposition 15.6.2 in Subsection 15.6.2.

- Many of these axioms describe two different ways of performing the operations, and state that the different ways always produce the same result.

  - For example, in Axiom A 3, the brackets on the left-hand side tell us to add vectors $\mathbf{v}$ and $\mathbf{w}$ first, in whatever way addition is defined in that space, and then to add that resulting sum vector to $\mathbf{u}$. On the right, the brackets tell us to add vectors $\mathbf{u}$ and $\mathbf{v}$ first, and then to add that resulting sum vector to $\mathbf{w}$. The equals sign in the middle means that we require the two different addition processes to always have the same result.
  - For another example, in Axiom S 2, the brackets on the left tell us to add $\mathbf{v}$ and $\mathbf{w}$, and then scale that sum vector by scalar k, whereas the brackets on the right tell us to scale each of $\mathbf{v}$ and $\mathbf{w}$ by k separately first, and then add those two scaled vectors together. The equals sign in the middle means that we require the add-then-scale process on the left to always have the same result as the scale-then-add process on the right.

  When we first encounter a new collection of objects for which we have some ideas of addition and scalar multiplication, we don't know that the two different orders of operations will always have the same result. Before we can call our new collection a **vector space**, we need to verify all of these sorts of things.

15.4.2 Instances of vector spaces

The vector space $\mathbb{R}^n$. One set of prototypical examples of vector spaces are the collections of vectors we have been studying in Chapters 11–14: $\mathbb{R}^2$, $\mathbb{R}^3$, and the higher-dimensional spaces $\mathbb{R}^n$, $n \geq 4$. In these spaces,

- adding vectors or scalar multiplying a vector results in a vector in the same space, satisfying Axiom A 1 and Axiom S 1;
- the zero vector is $\mathbf{0} = (0, 0, \dots, 0)$ as usual;
- the negative of a vector is the parallel vector of the same length in the opposite direction; and
- we know that the rest of the axioms hold true from our knowledge of vector algebra in these spaces (Proposition 11.5.1).

In Discovery 15.3.e, we discovered that even the collection of real numbers itself can be considered as a vector space. We might think of this space as $\mathbb{R}^1$, and visualize its vectors as directed line segments lying along the real number line.

The vector space $\mathrm{M}_{m\times n}(\mathbb{R})$. Another set of prototypical examples of vectors spaces are the collections of matrices of given dimensions, $\mathrm{M}_{m\times n}(\mathbb{R})$. But these matrix spaces represent our first expansion of the word **vector** to include other kinds of objects — since all ten axioms hold true here, we can justifiably refer to *any* matrix, of any size, as a *vector*. In these spaces,

- adding matrices or scalar multiplying a matrix does not change its dimensions, so these operations always result in a vector in the same space, satisfying Axiom A 1 and Axiom S 1;
- the zero vector is the zero matrix of the appropriate size;
- the negative of a vector is the matrix of the same dimensions where all the entries are the negatives of those of the original matrix; and
- we know that the rest of the axioms hold true from our knowledge of matrix algebra (Proposition 4.5.1).

Spaces of polynomials. In Discovery 15.3, we also explored some new examples of vector spaces consisting of polynomials as vectors. First, we considered the collection $\mathrm{P}(\mathbb{R})$ of polynomials with real coefficients of arbitrary degree in Discovery 15.3.c. Here are some observations on the vector space axioms for this space.

- We add polynomials algebraically, by adding like terms. For example,

  $$(5x^3+3x^2+2x-1)+(6x^{101}-3x^3+x+1)=6x^{101}+2x^3+3x^2+3x.$$

  Clearly, the result of adding polynomials is another polynomial, satisfying Axiom A 1.

- We scalar multiply a polynomial by distributing the scalar across the addition of the polynomials terms. For example,

  $$-2(6x^{101}-3x^3+x+1)=12x^{101}+6x^3-2x-2.$$

  The result of multiplying a polynomial by a scalar is another polynomial, satisfying Axiom S 1.

- The zero vector is the constant (i.e. degree zero) polynomial $p(x)=0$.
- The negative of a vector is the polynomial of the same degree where all the coefficients are the negatives of those of the the original polynomial.
- The rest of the axioms are familiar rules of algebra involving polynomial expressions.

Next, we considered the collection $\mathrm{P}_2(\mathbb{R})$ of polynomials with real coefficients of maximum degree 2 (Discovery 15.3.d). Everything here works the same as in $\mathrm{P}(\mathbb{R})$, except that we need to reconsider Axiom A 1 and Axiom S 1. We define vector addition and scalar multiplication for polynomials as before, but we need to make sure that the result of each of these operations is always equal to another in the collection of objects. But neither of the operations can *increase* the degree of a polynomial, so their results will always again be a polynomial of degree 2 or less.

The space of functions. This instance of a vector space is a generalization of the space of polynomials. We let $F(D)$ represent the collection of *all* functions, not just polynomials, defined on a domain D of real numbers. Our first task is define how the two operations will work.

To create a new “sum” function out of two old functions, or to create a new “scaled” version of an old function, we first need understand *how* to create new functions. To define a new function, we must *describe the input-output process*. If we have two functions, f and g, then these functions have an already defined input-output process, but addition must somehow take these two processes and create a new one that is the *sum* of the old. The natural thing to do would be to add outputs of the two old process. That is, we define the sum function $f+g$ by the input-output rule

$$(f+g)(x) = f(x)+g(x). \tag{$*$}$$

For example, if function f produces output 5 at input 3 (i.e. $f(3)=5$), and function g produces output 2 at input 3 (i.e. $g(3)=2$), then the sum function $f+g$ will produce output 7 at input 3:

$$(f+g)(3) = f(3)+g(3) = 5+2 = 7.$$

Similarly, to scale a function we should scale its outputs. That is,

$$(kf)(x) = kf(x). \tag{$**$}$$

For example, if function f produces output 5 at input 3 (i.e. $f(3)=5$), then the scaled function $\sqrt{2}f$ will produce output $5\sqrt{2}$ at input 3:

$$(\sqrt{2}f)(3) = \sqrt{2}\bigl(f(3)\bigr) = \sqrt{2}(5) = 5\sqrt{2}.$$

Both of these processes result in a new function with the same domain as the old, so Axiom A 1 and Axiom S 1 are satisfied.

In this space, the zero vector is the *zero function*, whose outputs are always zero: $\mathbf{0}(x)=0$ for all x. And the negative of a function is obtained by negating all of its outputs: $(-f)(x) = -f(x)$.

See Subsection 15.5.2 for examples of carrying out the verification of some of the vector axioms in this space.

The trivial vector space. In Discovery 15.5, we explored the possibility of a vector space with just one vector in it. But Axiom A 4 requires that every vector space have a zero vector, so that single vector inside must be it. And when we add this vector to itself or try to scale this vector, the condition that “the result is always equal to another in the collection” in both Axiom A 1 and Axiom S 1 requires that the result is actually always equal to that one vector, because there are no other vectors in the collection to choose from.

This simple vector space consisting of *just* a zero vector is called **the zero vector space** or **the trivial vector space**.

A weird instance of a vector space. To emphasize the fact that the words **vector**, **addition**, and the phrase **scalar multiplication** can potentially mean *anything*, let’s consider a weird example.

We’ll take our collection of objects to be the collection of positive numbers. (But our scalars can still be *any* number, whether positive, negative, or zero.) To help distinguish between a vector and a scalar, we’ll put brackets around a number if it is to mean a vector (as if we were considering $\mathbb{R}^1$). And to make sure we don't get mixed up with ordinary addition and multiplication of numbers, we'll use the symbols $\oplus$ and $\odot$ to mean vector addition and scalar multiplication, respectively.

To define **vector addition** in this space, we'll actually use ordinary *multiplication* of numbers. That is, for numbers $a, b > 0$, we will add the vectors $(a), (b)$ according to the rule

$$(a) \oplus (b) = (ab). \tag{†}$$

This definition satisfies Axiom A 1 because multiplying two positive numbers results in another positive number.

To define **scalar multiplication** in this space, we'll use exponentiation. That is, for number $a > 0$ and scalar k (also a number), we will scale the vector (a) by the scale factor k according to the rule

$$k \odot (a) = (a^k), \tag{††}$$

with the usual conventions that $a^0 = 1$ and $a^{-1} = 1/a$. This definition satisfies Axiom S 1 because a power of a positive number results in another positive number.

What is the zero vector in this space? For Axiom A 4 to hold true, we need a vector (z) so that

$$(a) \oplus (z) = (a)$$

for all other vectors (a). But we can't use $z = 0$, because the vectors in our space must all be *positive* numbers. Inserting the definition of $\oplus$ as ordinary multiplication, we need a positive number z so that

$$(az) = (a)$$

for all $a > 0$. But we only get $az = a$ for all $a > 0$ when $z = 1$. So in this weird space, the **zero vector** is the *number one*.

Comment. The identity of the zero vector in this space should actually not be that surprising — the number one is to multiplication what the number zero is to addition.

What is the negative of a vector in this space? Given positive $a > 0$, the negative of vector (a) can't be $(-a)$, because all our vectors have to be *positive* numbers. To repeat, *in this case,* $-(a)$ *is* not *equal to* $(-a)$. We know that every vector in this space is represented by a single positive number. That is, the negative of (a) must be equal to (b) for some positive number b. To satisfy Axiom A 5, this negative vector must satisfy

$$(a) \oplus \big(-(a)\big) = \mathbf{0}.$$

So we need $b > 0$ so that

$$(a) \oplus (b) = \mathbf{0}.$$

Inserting the definition of $\oplus$ as ordinary multiplication, and inserting $\mathbf{0} = (1)$ from above, we see we need

$$(ab) = (1).$$

That is, we need $ab = 1$, so that $b = 1/a$ (which is positive since a is positive). So we have

$$-(a) = \big(a^{-1}\big).$$

In this weird space, the negative (i.e. *additive* inverse) of a vector corresponds to the reciprocal (i.e. *multiplicative* inverse) of the positive number representing that vector.

In Subsection 15.5.1, we will verify some of the other vector space axioms for algebra in this space.

15.5 Examples

In this section.

15.5.1 Verifying axioms: the space of positive numbers

Here we will continue our "weird" example from the end of Subsection 15.4.2, and verify some of the other axioms for vectors in that space.

Let's start with Axiom A 2. Here we would like to verify that the vector equality $\mathbf{v}+\mathbf{w}=\mathbf{w}+\mathbf{v}$ is always true when the vectors are positive numbers and vector addition is defined to be ordinary multiplication, as defined in (†) in Subsection 15.4.2.

Verifying an equality. When verifying an equality, *we make sure to always consider the left- and right-hand expressions separately*.

For this space, vectors are positive numbers, so we should take $\mathbf{v}=(a)$ and $\mathbf{w}=(b)$ for *arbitrary*, unspecified positive numbers a and b (where again we use brackets to distinguish between numbers that are vectors and numbers that are scalars). Then,

$$\begin{aligned}
\text{LHS} &= \mathbf{v}+\mathbf{w} & \qquad \text{RHS} &= \mathbf{w}+\mathbf{v}\\
&= (a)\oplus(b) & &= (b)\oplus(a)\\
&= (ab), & &= (ba).
\end{aligned}$$

Now, we know that ordinary multiplication of numbers can be performed in either order, so $ba=ab$, and thus LHS = RHS as desired.

Verifying axioms. It's important that we verify axioms using *arbitrary* vectors and scalars, so that we know our verifications will be true regardless of the specific vectors and scalars considered. A vector axiom being *sometimes* true, for specific example vectors and scalars, is not good enough — we need the axioms to *always* be true, for *all possible* vectors in the collection, and *all possible* scalars.

We will leave the other addition axioms up to you, but let's verify one of the scalar multiplication axioms. Consider Axiom S 2. We need to verify that $k(\mathbf{v}+\mathbf{w})=k\mathbf{v}+k\mathbf{w}$ is always true for all scalars k and all vectors $\mathbf{v}$ and $\mathbf{w}$, where scalar multiplication is defined as in (††) in Subsection 15.4.2. When considering the left- and right-hand sides of this vector equality, we need to be sure to *pay attention to the order of operations on each side*. Again, take $\mathbf{v}=(a)$ and $\mathbf{w}=(b)$ for *arbitrary*, unspecified positive numbers a and b. Then,

$$\begin{aligned}
\text{LHS} &= k(\mathbf{v}+\mathbf{w}) & \qquad \text{RHS} &= k\mathbf{v}+k\mathbf{w}\\
&= k\odot\big((a)\oplus(b)\big) & &= \big(k\odot(a)\big)\oplus\big(k\odot(b)\big)\\
&= k\odot(ab) & &= \left(a^k\right)\oplus\left(b^k\right)\\
&= \left((ab)^k\right), & &= \left(a^k b^k\right).
\end{aligned}$$

We can now see that LHS = RHS as desired because of the exponent law $(ab)^k = a^k b^k$ from the algebra of ordinary numbers.

Check your understanding. Verify Axiom A 3 and Axioms S 3–5 for our "weird" example space, using a similar procedure as in this subsection.

15.5.2 Verifying axioms: the space of functions

Here we will verify some of the axioms for vectors in the space $F(D)$. We will be verifying equality of functions, so we need to make sure we know what it means for two functions to be equal.

Definition 15.5.1 Equality of functions. Two functions are **equal** when the input-output processes they represent always produce the same outputs. That is, functions f and g are equal if $f(x) = g(x)$ for *all possible* input x-values. ♢

Let's start with Axiom A 3. Here we would like to verify that the vector equality $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$ is always true when the vectors are functions and addition is defined in $F(D)$ as in (∗) in Subsection 15.4.2.

Let's take $\mathbf{u} = f$, $\mathbf{v} = g$, and $\mathbf{w} = h$, where f, g, h are *arbitrary*, unspecified functions that are all defined on the domain D. From Definition 15.5.1 above, we see that we need to verify that the sum functions $f + (g + h)$ and $(f + g) + h$ always produce the same output when fed the same input. So suppose x is an input value in the domain D. Then,

$$\begin{aligned} \text{LHS} &= \big(f + (g + h)\big)(x) \\ &= f(x) + (g + h)(x) && \text{(i)} \\ &= f(x) + \big(g(x) + h(x)\big), && \text{(ii)} \end{aligned}$$

$$\begin{aligned} \text{RHS} &= \big((f + g) + h\big)(x) \\ &= (f + g)(x) + h(x) && \text{(iii)} \\ &= \big(f(x) + g(x)\big) + h(x), && \text{(iv)} \end{aligned}$$

with justifications

(i) definition of the sum of f and $g + h$;

(ii) definition of the sum of g and h;

(iii) definition of the sum of $f + g$ and h; and

(iv) definition of the sum of f and g.

Now, $f(x), g(x), h(x)$ are just *numbers* — they are the output y-values produced by the functions from the input value x — and we know that we can group numbers with brackets in any combination when adding. So LHS = RHS as desired.

Now let's verify Axiom A 5, using the definition $(-f)(x) = -f(x)$ from Subsection 15.4.2 (where we have also defined $\mathbf{0}(x) = 0$). We must verify that the sum function $f + (-f)$ is the same as the zero function $\mathbf{0}$, which means we must verify that these functions always have the same outputs. So suppose x is an input value in the domain D. Then,

$$\begin{aligned} \text{LHS} &= \big(f + (-f)\big)(x) \\ &= f(x) + (-f)(x) && \text{(i)} \\ &= f(x) + \big(-f(x)\big) && \text{(ii)} \\ &= f(x) - f(x) && \text{(iii)} \\ &= 0 && \text{(iv)} \\ &= \mathbf{0}(x) && \text{(v)} \\ &= \text{RHS}, \end{aligned}$$

as desired, with justifications

(i) definition of the sum of f and $-f$;

(ii) definition of the negative of f;

(iii) algebra of numbers;

(iv) algebra of numbers; and

(v) definition of the zero function.

As a last example, let's verify Axiom S 3 in this space. We need to verify that $(k+m)\mathbf{v} = k\mathbf{v} + m\mathbf{v}$ is always true for all scalars k and m and all vectors $\mathbf{v}$ when the vectors are functions and the vector operations are defined in $F(D)$ as in $(*)$ and $(**)$ in Subsection 15.4.2. And when considering the left- and right-hand sides of this vector equality, we need to be sure to *pay attention to the order of operations on each side*.

Again, take $\mathbf{v} = f$ for *arbitrary*, unspecified function f. Then we actually need to verify that the function $(k+m)f$ always produces the same outputs as the function $kf + mf$. So suppose x is an input value in the domain D. Then,

$$\begin{aligned} \text{LHS} &= \big((k+m)f\big)(x) \\ &= (k+m)f(x), && \text{(i)} \end{aligned}$$

$$\begin{aligned} \text{RHS} &= (kf+mf)(x) \\ &= (kf)(x) + (mf)(x) && \text{(ii)} \\ &= kf(x) + mf(x), && \text{(iii)} \end{aligned}$$

with justifications

(i) definition of scalar multiplication of f by $k+m$;

(ii) definition of the sum of kf and mf; and

(iii) definition of scalar multiplication of f by k and by m.

Again, $f(x)$ is just a *number*, and k, m are also numbers, and we know that we can distribute the multiplication of $f(x)$ across the sum of k and m in the expression $(k+m)f(x)$. Therefore, LHS = RHS as desired.

Check your understanding. Verify Axiom A 2, Axiom S 2, Axiom S 4, and Axiom S 5, using a similar procedure as in the examples of this subsection. Also verify that in Subsection 15.4.2, we have chosen the correct **zero vector** and the correct concept of **negative vector** in this space (Axiom A 4 and Axiom A 5).

15.6 Theory

In this section.

15.6.1 Uniqueness of the zero vector and of negatives

Notice that Axiom A 4 and Axiom A 5 only say that there is *a* zero vector, and that every vector has *some* negative — they don't say that there is *only one* zero vector, or that every vector has *only one* negative. There is no need to make these axioms that strong — we can instead just logically deduce these properties from the weaker axioms we already have.

Proposition 15.6.1

1. *In a vector space, there is one unique zero vector.*
2. *A vector in a vector space has one unique negative vector.*

Proof of Statement 1. Suppose there were two vectors, $\mathbf{0}_1$ and $\mathbf{0}_2$, that could each fulfill the requirement of Axiom A 4. But then we would have both

$$\mathbf{0}_1 + \mathbf{0}_2 = \mathbf{0}_1 \qquad \text{(i),}$$

and

$$\begin{aligned} \mathbf{0}_1 + \mathbf{0}_2 &= \mathbf{0}_2 + \mathbf{0}_1 && \text{(ii)} \\ &= \mathbf{0}_2 && \text{(iii),} \end{aligned}$$

with justifications

(i) Axiom A 4 with $\mathbf{v} = \mathbf{0}_1$ and $\mathbf{0} = \mathbf{0}_2$;

(ii) Axiom A 2; and

(iii) Axiom A 4 with $\mathbf{v} = \mathbf{0}_2$ and $\mathbf{0} = \mathbf{0}_1$.

Since $\mathbf{0}_1 + \mathbf{0}_2$ equals both $\mathbf{0}_1$ and $\mathbf{0}_2$, we must have $\mathbf{0}_1 = \mathbf{0}_2$. So there can't really be more than one zero vector, since multiple zero vectors would end up having to be equal to each other. ■

Proof of Statement 2. Suppose a vector $\mathbf{v}$ could have two negatives, $(-\mathbf{v})_1$ and $(-\mathbf{v})_2$. But then,

$$\begin{aligned} (-\mathbf{v})_2 &= (-\mathbf{v})_2 + \mathbf{0} && \text{(i)} \\ &= (-\mathbf{v})_2 + \big(\mathbf{v} + (-\mathbf{v})_1\big) && \text{(ii)} \\ &= \big((-\mathbf{v})_2 + \mathbf{v}\big) + (-\mathbf{v})_1 && \text{(iii)} \\ &= \big(\mathbf{v} + (-\mathbf{v})_2\big) + (-\mathbf{v})_1 && \text{(iv)} \\ &= \mathbf{0} + (-\mathbf{v})_1 && \text{(v)} \\ &= (-\mathbf{v})_1 + \mathbf{0} && \text{(vi)} \\ &= (-\mathbf{v})_1 && \text{(vii),} \end{aligned}$$

with justifications

(i) Axiom A 4;

(ii) Axiom A 5 with $-\mathbf{v} = (-\mathbf{v})_1$;

(iii) Axiom A 3;

(iv) Axiom A 2;

(v) Axiom A 5 with $-\mathbf{v} = (-\mathbf{v})_2$;

(vi) Axiom A 2; and

(vii) Axiom A 4.

So $\mathbf{v}$ can't really have more than one negative vector, since multiple negative vectors would end up having to be equal to each other. ■

15.6.2 Basic vector algebra rules

There was also no need to include the condition $\mathbf{0}+\mathbf{v}=\mathbf{v}$ in Axiom A 4 or the condition $-\mathbf{v}+\mathbf{v}=\mathbf{0}$ in Axiom A 5, as these can be deduced from the axioms we have, as we did in Discovery 15.4.a and Discovery 15.4.b. Let's record these properties, and some others that can be deduced from the axioms.

Keeping it simple. We want the axioms to be as simple as possible, to reduce the amount of work it takes to verify that an example collection of objects is actually a vector space. The stronger we make the axioms, the more we have to check in examples. So, wherever possible, we leave conditions that seem "axiom-like" to be left as properties to be *logically deduced* from the axioms. This way, these extra properties become *automatically* true, *without checking*, in every collection that we have successfully checked the ten axioms.

Proposition 15.6.2 *Suppose that* $\mathbf{u},\mathbf{v},\mathbf{w}$ *are vectors in a vector space, and that* k *is a scalar. Then the following are always true.*

1. *Additional rules of the zero vector.*
   - *(a)* $\mathbf{0}+\mathbf{v}=\mathbf{v}$.
   - *(b)* $0\mathbf{v}=\mathbf{0}$.
   - *(c)* $-\mathbf{0}=\mathbf{0}$.
   - *(d)* $k\mathbf{0}=\mathbf{0}$.
   - *(e) If* $k\mathbf{v}=\mathbf{0}$*, then either* $k=0$ *or* $\mathbf{v}=\mathbf{0}$ *(or both).*
2. *Additional rules of vector negatives.*
   - *(a)* $-\mathbf{v}+\mathbf{v}=\mathbf{0}$.
   - *(b)* $-(\mathbf{u}+\mathbf{v})=(-\mathbf{u})+(-\mathbf{v})$.
   - *(c)* $-(-\mathbf{v})=\mathbf{v}$.
   - *(d)* $-(k\mathbf{v})=k(-\mathbf{v})$.
   - *(e)* $(-1)\mathbf{v}=-\mathbf{v}$.
   - *(f)* (Cancellation) *If* $\mathbf{u}+\mathbf{w}=\mathbf{v}+\mathbf{w}$*, then* $\mathbf{u}=\mathbf{v}$.

Proofs of Rules 1.a–1.b, Rule 2.c, and Rule 2.e. We have already considered these rules in Discovery 15.4. ■

Proof of Rule 1.c. The zero vector is a special vector, but it's still a vector so it must have a negative because of Axiom A 5. Now, Statement 2 of Proposition 15.6.1 with $\mathbf{v}=\mathbf{0}$ tells us that the *only* way to fill the blank in

$$\mathbf{0}+\boxed{}=\mathbf{0}$$

is with the negative $-\mathbf{0}$. But Axiom A 4 with $\mathbf{v}=\mathbf{0}$ says that we may also fill this blank with plain $\mathbf{0}$. Therefore, we must have $\mathbf{0}=-\mathbf{0}$, as desired. ■

Proof of Rule 1.d. We need to verify the equality $k\mathbf{0}=\mathbf{0}$:

$$\begin{aligned}
\text{LHS} &= k\mathbf{0} && \\
&= k\big(\mathbf{0}+(-\mathbf{0})\big) && \text{(i)}\\
&= k\big(\mathbf{0}+(-1)\mathbf{0}\big) && \text{(ii)}\\
&= k\mathbf{0}+k\big((-1)\mathbf{0}\big) && \text{(iii)}\\
&= k\mathbf{0}+(-k)\mathbf{0} && \text{(iv)}\\
&= \big(k+(-k)\big)\mathbf{0} && \text{(v)}\\
&= 0\mathbf{0} && \text{(vi)}
\end{aligned}$$

$$\begin{aligned} &= \mathbf{0} && \text{(vii)}\\ &= \text{RHS}, \end{aligned}$$

as desired, with justifications

(i) Axiom A 5 with $\mathbf{v} = \mathbf{0}$;

(ii) Rule 2.e;

(iii) Axiom S 2;

(iv) Axiom S 4;

(v) Axiom S 3;

(vi) algebra of numbers; and

(vii) Rule 1.b.

■

Proof of Rule 1.e. Suppose $k\mathbf{v} = \mathbf{0}$. Regardless of this starting assumption, either k is equal to 0 or it is not. If it is, then the desired conclusion "either $k = 0$ or $\mathbf{v} = \mathbf{0}$" is true, regardless of whether $\mathbf{v}$ is zero of not. On the other hand, if k is not equal to 0, then the reciprocal k^{-1} exists, and so we can use it to compute

$$\begin{aligned} \mathbf{v} &= 1\mathbf{v} && \text{(i)}\\ &= (k^{-1}k)\mathbf{v} && \text{(ii)}\\ &= k^{-1}(k\mathbf{v}) && \text{(iii)}\\ &= k^{-1}\mathbf{0} && \text{(iv)}\\ &= \mathbf{0}, && \text{(v)} \end{aligned}$$

with justifications

(i) Axiom S 5;

(ii) algebra of numbers;

(iii) Axiom S 4;

(iv) assumption $k\mathbf{v} = \mathbf{0}$; and

(v) Rule 1.d.

In this case, the desired conclusion "either $k = 0$ or $\mathbf{v} = \mathbf{0}$" is true again. ■

Proof of Rule 2.f. Suppose $\mathbf{u} + \mathbf{w} = \mathbf{v} + \mathbf{w}$. Starting with $\mathbf{u}$, we can use the axioms to work in a $\mathbf{w}$ and then convert to $\mathbf{v}$:

$$\begin{aligned} \mathbf{u} &= \mathbf{u} + \mathbf{0} && \text{(i)}\\ &= \mathbf{u} + \big(\mathbf{w} + (-\mathbf{w})\big) && \text{(ii)}\\ &= (\mathbf{u} + \mathbf{w}) + (-\mathbf{w}) && \text{(iii)}\\ &= (\mathbf{v} + \mathbf{w}) + (-\mathbf{w}) && \text{(iv)}\\ &= \mathbf{v} + \big(\mathbf{w}) + (-\mathbf{w})\big) && \text{(v)}\\ &= \mathbf{v} + \mathbf{0} && \text{(vi)}\\ &= \mathbf{v}, && \text{(vii)} \end{aligned}$$

as desired, with justifications

(i) Axiom A 4;

(ii) Axiom A 5;

(iii) Axiom A 3;

(iv) assumption $\mathbf{u}+\mathbf{w}=\mathbf{v}+\mathbf{w}$;

(v) Axiom A 3;

(vi) Axiom A 5; and

(vii) Axiom A 4.

■

Remark 15.6.3 Again, keep in mind the difference between the left- and right-hand sides in Rule 2.e in the proposition above. The left-hand side is the scalar multiple of $\mathbf{v}$ by the scalar -1, while the right-hand side is the special negative vector that adds with $\mathbf{v}$ to the zero vector. These are two different processes of obtaining a new vector from the old vector $\mathbf{v}$, and the point of the rule is to verify our intuition that these two processes should always return the same result. One of the advantages of this rule is that it eliminates any ambiguity in our definition of **vector subtraction**, since now it doesn't matter if we interpret $\mathbf{v}-\mathbf{w}$ to mean $\mathbf{v}+(-\mathbf{w})$ or $\mathbf{v}+(-1)\mathbf{w}$.

CHAPTER 16

Subspaces

16.1 Discovery guide

Recall that a **vector space** is a collection of objects (called **vectors**) that satisfies all of the axioms in Definition 15.4.1.

Discovery 16.1 Sometimes you have a subcollection of vectors inside a larger vector space, and would like to know whether the subcollection is also a vector space, all on its own.

Definition. A **subcollection** is a collection of objects, each of which is a member of some "larger" collection. For example, the collection of even numbers is a subcollection of the collection of whole numbers.

(a) In the large vector space, you would already *know* (from having checked) that *all* ten axioms are true. Because all the vectors in the subcollection also "live" in the large vector space, *six* of the axioms will automatically be true for the subcollection (and the remaining four may or may not be true). Identify these six axioms that are automatically true.

Hint. It is easier to identify the six that are *definitely* true rather than the four that *might* be false.

(b) Using $\mathbb{R}^2$ as the large vector space, for each of the following subcollections, which of those four remaining axioms are true and which are false? (Consider all vectors as positioned with initial point at the origin.)

(i) All points on the line $y = x$.

(ii) All points on the line $y = x + 1$.

(iii) All points on the circle of radius 1 centred at the origin.

In Proposition 16.5.1, we will prove that the task of checking the four "possibly false" axioms you identified in Discovery 16.1 for a particular subcollection can be refined to the following test.

The Subspace Test.

(i) *Nonempty.*

The subcollection contains at least one vector.

(ii) *Closed under vector addititition.*

The sum of two vectors in the subcollection is always equal to another vector *in the subcollection*.

(iii) *Closed under scalar multiplication.*

A scalar multiple of a vector in the subcollection is always equal to another vector *in the subcollection*.

Discovery 16.2 In each of the following, check each part of the Subspace Test for subcollection W inside vector space V.

- If you think a part of the Subspace Test is *true*, ***justify it without resorting to examples***.
- If you think a part of the Subspace Test is *false*, ***provide an explicit example that demonstrates it***.

Logic 101. To demonstrate that a general statement is true, we work in the abstract with *arbitrary* objects, so that our justification is valid no matter what objects one considers. But to demonstrate that a general statement is false, all we have to do is demonstrate a specific **counterexample**, because one exception is all that is needed to prove the general rule to be false.

(a) $V = \mathbb{R}^3$; $W =$ the xy-plane.

(b) $V = \mathbb{R}^3$; $W =$ the plane parallel to the xy-plane at height $z = 1$.

(c) $V =$ all 10×10 matrices; $W =$ diagonal 10×10 matrices.

(d) $V =$ all 12×12 matrices; $W =$ those 12×12 matrices with a 7 in the $(1,1)$ entry.

(e) $V =$ all 6×4 matrices; $W =$ those 6×4 matrices with 0 in each of the four corner entries.

(f) $V =$ all polynomials; $W =$ those polynomials of degree 2 or less.

(g) $V =$ all polynomials; $W =$ those polynomials of degree exactly 2.

(h) $V =$ all polynomials; $W =$ those polynomials with constant term equal to 0.

(i) $V = \mathbb{R}^3$; $W =$ all column vectors $\mathbf{x}$ that satisfy the matrix equation $A\mathbf{x} = \mathbf{0}$, where A is some fixed 2×3 matrix.

Hint. You don't need to know the entries of the matrix A to carry out the Subspace Test — use matrix algebra instead to test a sum or scalar multiple in the equation $A\mathbf{x} = \mathbf{0}$.

(j) $V = \mathbb{R}^3$; $W =$ all possible linear combinations of vectors $\mathbf{u} = (1,1,1)$ and $\mathbf{v} = (3,2,-1)$.

As we will see from Proposition 16.5.5 in Subsection 16.5.2, the pattern in Discovery 16.2.j always works: if V is a vector space and S is a set of vectors in V, then the subcollection W of all possible linear combinations of vectors from S is a subspace of V, called the **span of** S, and we write $W = \operatorname{Span} S$.

Discovery 16.3 In each of the following, determine if the given vector $\mathbf{v}$ is a member of $\operatorname{Span} S$. That is, determine if $\mathbf{v}$ can be expressed as a linear combination of the vectors in S.

Hint. Don't guess at it, set up equations and solve! The unknowns in your equations will be the scalars in the linear combination of the S-vectors to try to make the vector $\mathbf{v}$. Start with a vector equation

$$\mathbf{v} = \text{linear combination of } S\text{-vectors with variables as scalars.}$$

This should somehow lead to a (gasp!) system of linear equations in your unknown scalars.

(a) $V = \mathbb{R}^3$, $S = \big\{(1,0,1),(2,1,-1)\big\}$, $\mathbf{v} = (1,-1,4)$.

(b) V = all 2×3 matrices, $S = \big\{\left[\begin{smallmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right]\big\}$, $\mathbf{v} = \left[\begin{smallmatrix} 0 & 2 & 2 \\ 3 & -3 & -3 \end{smallmatrix}\right]$.

(c) V = all polynomials, $S = \{1, 1+x, 1+x^2\}$, $\mathbf{v} = 2 - x + 3x^2$.

Discovery 16.4 In each of the following, try to convince yourself that $V = \operatorname{Span} S$. That is, convince yourself that *every* vector in V can be expressed as a linear combination of the vectors in S.

Remember. You can't rely on specific examples to verify a general statement!

(a) $V = \mathbb{R}^5$, $S = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5\}$.

(b) V = all 2×2 matrices, $S = \big\{\left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right]\big\}$.

(c) V = all polynomials of degree 3 or less, $S = \{1, x, x^2, x^3\}$.

16.2 Terminology and notation

subspace a subset of vectors in a vector space that itself is a vector space under the same addition and scalar multiplication operations as the parent vector space

trivial subspace
the subspace of a vector space consisting of only the zero vector

linear combination (of a collection of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$)
a vector that can be expressed as

$$k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + \cdots + k_m\mathbf{v}_m$$

for some collection of scalars $k_1, k_2, \ldots, k_m$

subspace generated by a set of vectors S
the subspace of a vector space consisting of all possible linear combinations of vectors in S; also called the **span of S**, and written $\operatorname{Span} S$

spanning set (for a vector space)
a set of vectors in a vector space (or subspace of a vector space) where the subspace generated by the set is in fact the whole space; could also be called a **generating set** of vectors for the space

solution space of homogeneous system $A\mathbf{x} = 0$
the subspace of $\mathbb{R}^n$ (where n is the number of columns of A) consisting of all solutions to the system

16.3 Concepts

In this section.

When faced with any big problem, mathematical or otherwise, it is often a good idea to break the big problem up into smaller parts. In a vector space, the “smaller parts” are smaller vector spaces inside the larger space, called **subspaces**.

16.3.1 Recognizing subspaces

How can we recognize subspaces? To be a subspace, a subcollection of vectors must satisfy all ten vector space axioms on its own. But Axiom A 2, Axiom A 3, and Axioms S 2–5 all concern the *algebra* of vectors, and don’t really take into account where the vectors are considered to “live”. Since these algebra axioms are true about *all* vectors in the larger space, they are automatically true about the vectors in the subcollection. So that leaves Axiom A 1, Axiom A 4, Axiom A 5, and Axiom S 1.

For axioms Axiom A 1 and Axiom S 1, we already have addition and scalar multiplication of vectors defined as in the large space. But when we are considering whether the smaller collection is a vector space all on its own, the vectors *not* in this collection are no longer relevant. So the parts of these axioms that say "the result is always equal to another in the collection of objects" now refer to the *subcollection* of vectors under consideration. That is, we need to make sure that when vectors in the subcollection are added or scalar multiplied, then the result is again in the subcollection, not somewhere else in the vector space at large. We call this property being **closed** under the vector operations.

Similarly, for Axiom A 5, we already know that every vector in the large space has a negative, hence so does every vector in the subcollection. But we need to check that the subcollection is *closed* under taking negatives — that the negative of a vector in the subcollection is again in the subcollection. And we know that there is a zero vector in the large space, but the subcollection needs a zero vector too, to satisfy Axiom A 4. The zero vector from the larger space already satisfies the property that $\mathbf{v}+\mathbf{0}=\mathbf{v}$ for all vectors, but again *we need the zero vector to be in the subcollection*.

As we will prove in Proposition 16.5.1, we really only need to check a subcollection for closure under addition and scalar multiplication in order to verify that it is also a vector space.

Procedure 16.3.1 Subspace Test. *To test whether a subcollection of vectors in a vector space is a sub*space *(that is, a vector space on its own), check whether all three of the following conditions are met.*

(i) Nonempty.

The subcollection contains at least one vector.

(ii) Closed under vector addition.

Given vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ in the subcollection, the sum $\mathbf{w}_1+\mathbf{w}_2$ is also in the subcollection.

(iii) Closed under scalar multiplication.

Given vector $\mathbf{w}$ in the subcollection and scalar k, the scaled vector $k\mathbf{w}$ is also in the subcollection.

Remark 16.3.2

- The first condition might seem unnecessary. But in math it is possible to accidentally be considering some collection of objects that in reality contains no objects. For example, consider $\mathrm{M}_1(\mathbb{R})$, the vector space of all 1×1 matrices. We could try to determine whether the subcollection of all 1×1 matrices whose square is equal to $\begin{bmatrix}-1\end{bmatrix}$ is a subspace of of $\mathrm{M}_1(\mathbb{R})$, but we'd be wasting our time because there are no such matrices.

- The logic of the test works in reverse as well: every subspace satisfies the three statements of the test because it is a vector space all on its own and thus satisfies the ten vector space axioms. (This is the "and only if" part of Proposition 16.5.1.) So a subspace is always nonempty because it must contain a zero vector (Axiom A 4), and it is always closed under the vector operations (Axiom A 1 and Axiom S 1).

See Section 16.4 for examples of applying the Subspace Test to verify that certain subcollections of vectors in vector spaces form subspaces.

16.3.2 Building subspaces

Suppose we are studying a problem for which certain vectors in a certain vector space are important. We would like to do linear algebra with these certain special vectors, so the fact that they are part of a vector space is essential, but perhaps not all of the vector space in which they live is relevant to the problem. Can we form a smaller subspace which contains these important vectors? Even better, can we determine the *smallest* subspace which contains these important vectors?

As stated above, every subspace must satisfy the three parts of the Subspace Test. So if a subspace contains our special vectors, then it must also contain all scalar multiples of those vectors. And it must also be closed under vector addition, so it must also contain all sums of scalar multiples of the special vectors. Therefore, it must contain every *linear combination* of the special vectors. (In other words, subspaces are also *closed under taking linear combinations*.)

As we noted in Discovery guide 16.1, and will verify in Subsection 16.5.2 (Proposition 16.5.5), the subcollection of a vector space consisting of all linear combinations of a set of vectors S is always a subspace, called the **span of** S and written $\operatorname{Span} S$. So the process of taking *all possible* linear combinations of a set of vectors can be used to build subspaces. And sometimes, as in Discovery 16.4, the space that span builds ends up being the *whole* larger space.

A look ahead. In Subsection 16.5.2, we will see that *every* vector space (whether a subspace of a larger space or not) can be described as the span of some set of vectors (Statement 3 of Proposition 16.5.5). And in Chapters 17–19 we will study optimal ways for doing so.

16.3.3 The subspaces of $\mathbb{R}^n$

We saw in Subsection 14.3.2 that a line through the origin in $\mathbb{R}^n$ can be described in vector form as $\mathbf{x} = t\mathbf{p}$ for some vector $\mathbf{p}$ that is parallel to the line (and taking "initial" point $\mathbf{x}_0 = \mathbf{0}$, since the line passes through the origin). Similarly, we saw that a plane through the origin in $\mathbb{R}^3$ can be described in vector form as $\mathbf{x} = s\mathbf{p}_1 + t\mathbf{p}_2$ for some vectors $\mathbf{p}_1$ and $\mathbf{p}_2$ that are parallel to the plane but not parallel to each other. With our new, more sophisticated view of vector spaces and subspaces, we can now recognize a line $\mathbf{x} = t\mathbf{p}$ as the subspace $\operatorname{Span}\{\mathbf{p}\}$, and a plane $\mathbf{x} = s\mathbf{p}_1 + t\mathbf{p}_2$ as the subspace $\operatorname{Span}\{\mathbf{p}_1, \mathbf{p}_2\}$.

In fact, *every* (nontrivial) subspace of $\mathbb{R}^n$ has a geometric interpretation as a line or a plane or some sort of higher-dimensional hyperplane. In particular,

- the subspaces of $\mathbb{R}^2$ are precisely
  - the zero space $\{\mathbf{0}\}$,
  - $\operatorname{Span}\{\mathbf{p}\}$ for a nonzero vector $\mathbf{p}$, which builds a line through the origin, and
  - $\operatorname{Span}\{\mathbf{p}_1, \mathbf{p}_2\}$ for two nonzero, nonparallel vectors $\mathbf{p}_1$ and $\mathbf{p}_2$, which builds the whole plane $\mathbb{R}^2$;
- the subspaces of $\mathbb{R}^3$ are precisely
  - the zero space $\{\mathbf{0}\}$,
  - $\operatorname{Span}\{\mathbf{p}\}$ for a nonzero vector $\mathbf{p}$, which builds a line through the origin,
  - $\operatorname{Span}\{\mathbf{p}_1, \mathbf{p}_2\}$ for two nonzero, nonparallel vectors $\mathbf{p}_1$ and $\mathbf{p}_2$, which builds a plane through the origin, and
  - $\operatorname{Span}\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3\}$ for three nonzero, non-coplanar vectors $\mathbf{p}_1$, $\mathbf{p}_2$, and $\mathbf{p}_3$, which builds all of space $\mathbb{R}^3$;

- the subspaces of $\mathbb{R}^4$ are precisely
  - the zero space $\{\mathbf{0}\}$,
  - $\operatorname{Span}\{\mathbf{p}\}$ for a nonzero vector $\mathbf{p}$, which builds a line through the origin,
  - $\operatorname{Span}\{\mathbf{p}_1, \mathbf{p}_2\}$ for two nonzero, nonparallel vectors $\mathbf{p}_1$ and $\mathbf{p}_2$, which builds a plane through the origin,
  - $\operatorname{Span}\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3\}$ for three nonzero, non-coplanar vectors $\mathbf{p}_1$, $\mathbf{p}_2$, and $\mathbf{p}_3$, which builds a hyperplane through the origin, and
  - $\operatorname{Span}\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4\}$ for four nonzero, non-cohyperplanar vectors $\mathbf{p}_1$, $\mathbf{p}_2$, $\mathbf{p}_3$, and $\mathbf{p}_4$, which builds all of four-dimensional space $\mathbb{R}^4$;
- etc.

16.3.4 Recognizing when two subspaces are the same

One of the goals of the next few chapters is to determine how to describe vector spaces using spanning sets in an optimal fashion. In this endeavour, we will want to know when a refinement of a spanning set still spans the space we are trying to describe. So, in particular, we will need to know when two spanning sets generate the same subspace. Since spans are defined by linear combinations, and subspaces are closed under linear combinations, this will not be as difficult as it sounds, and we provide a test for this situation as Proposition 16.5.6 in Subsection 16.5.3.

16.4 Examples

In this section.

16.4.1 The Subspace Test

First, let's practise applying the Subspace Test.

Remark 16.4.1 Since every vector space must have a zero vector (Axiom A 4), so too must a subspace. But since the vector operations of a subspace are the same as the operations of the larger space, it will turn out that the zero vector in a subspace must always be the same as the zero vector in the larger space (see Proposition 16.5.2). So often the best way to check the Nonempty clause of the Subspace Test is to verify that it contains the zero vector.

Here are some examples from Discovery guide 16.1.

Example 16.4.2 A plane through the origin in $\mathbb{R}^3$. In Discovery 16.2.a, we considered the subcollection of vectors in $\mathbb{R}^3$ consisting of all vectors from the origin with terminal point in the xy-plane. Note that any such vector must have a 0 as its z-component.

Let's apply the Subspace Test.

Nonempty. We know that the xy-plane is nonempty; in particular, it contains the zero vector since it contains the origin.

Closed under vector addition. If vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ are both in the xy-plane, then their z-components are both zero. So we can write these vectors as

$$\mathbf{u}_1 = (x_1, y_1, 0), \qquad \mathbf{u}_2 = (x_2, y_2, 0).$$

Then,

$$\mathbf{u}_1 + \mathbf{u}_2 = (x_1 + x_2, y_1 + y_2, 0).$$

Since this sum vector also has zero z-component, it is again in the xy-plane, as required.

Closed under scalar multiplication. If vector $\mathbf{u}$ is in the xy-plane, then its z-component is zero, so we can write it as

$$\mathbf{u} = (x, y, 0).$$

Then for every scalar k, we have

$$k\mathbf{u} = (kx, ky, 0).$$

Since this scaled vector also has zero z-component, it is again in the xy-plane, as required.

Conclusion. Since all parts of the Subspace Test pass, the xy-plane is a subspace of $\mathbb{R}^3$. □

Example 16.4.3 A subspace of matrices. In Discovery 16.2.d, we considered the subcollection of $\mathrm{M}_{12}(\mathbb{R})$ consisting of all those 12×12 matrices that have a 7 in the $(1,1)$ entry.

Let's apply the Subspace Test.

Nonempty. This collection is clearly not empty, since the 12×12 matrix with 7 in *every* entry is in the collection.

Note. The zero matrix is clearly *not* in the collection, so we could conclude right now that this subcollection is not a subspace. But since the Subspace Test itself has not yet failed, we will continue.

Closed under vector addition. If matrices A_1 and A_2 are both in the subcollection, then they each have a 7 in the $(1,1)$ entry. But then $A_1 + A_2$ has 14 in the $(1,1)$ entry, not 7. So the sum vector is *not* in the subcollection.

Conclusion. There is no need to check the third clause of the Subspace Test, since the second has already failed to pass. But it shouldn't be too difficult to see that scalar multiples of such a matrix will also fail to remain in the subcollection (except when the scalar is 1). □

Example 16.4.4 Restricting degree creates a subspace of polynomials. In Discovery 16.2.f, we considered the subcollection of $\mathrm{P}(\mathbb{R})$ consisting of those polynomials that have degree 2 or less.

Let's apply the Subspace Test.

Nonempty. Clearly this subcollection is nonempty, as any constant polynomial has degree 0, which is less than 2. In particular, the zero polynomial $\mathbf{0}(x) = 0$ (which is the zero vector in this space) also has degree less than 2.

Closed under vector addition. Suppose p_1 and p_2 are two polynomials in this subcollection. Then the degree of each of these polynomials is 2 or less, so we can write

$$p_1(x) = a_1x^2 + b_1x + c_1, \qquad p_2(x) = a_2x^2 + b_2x + c_2.$$

Note. Even though each expression has an x^2 term, the degree could still be less than 2 because the leading coefficient a_i could be zero.

Then,

$$\begin{aligned} p_1(x) + p_2(x) &= (a_1x^2 + b_1x + c_1) + (a_2x^2 + b_2x + c_2) \\ &= (a_1 + a_2)x^2 + (b_1 + b_2)x + (c_1 + c_2). \end{aligned}$$

Since this sum polynomial also has degree 2 (or less, since a_1 and a_2 could cancel or could both be zero), it is again in the subcollection, as required.

Closed under scalar multiplication. Suppose p is a polynomial in this subcollection. Then the degree of this polynomial is 2 or less, so we can write

$$p(x) = ax^2 + bx + c.$$

Then for every scalar k, we have

$$kp(x) = kax^2 + kbx + kc.$$

Since this scaled polynomial also has degree 2 (or less, since either k or a could be zero), it is again in the subcollection, as required.

Conclusion. Since all parts of the Subspace Test pass, the collection of all polynomials of degree 2 or less is a subspace of P(ℝ). □

Remark 16.4.5 Similar to this last example, the Subspace Test can be used to verify that $\mathrm{P}_n(\mathbb{R})$, the subcollection of P(ℝ) consisting of all polynomials with degree n or less, is a subspace for every fixed value of positive integer n.

16.4.2 Important subspace examples

Here are a few more important examples of subspaces.

Example 16.4.6 The trivial subspace. Consider the subcollection in a vector space consisting of *just* the zero vector. Since we already know that the zero vector space is, indeed, a vector space, there is no need for the Subspace Test. ***In every vector space, the zero space $\{\mathbf{0}\}$ is always a subspace.*** □

Example 16.4.7 The full subspace. Consider the subcollection in a vector space consisting of *every* vector. (This may not seem like a *sub*collection, but every vector in this subcollection is in the original vector space.) Since it is obviously true that the collection of all vectors in a vector space forms a vector space, we have that ***every vector space is a subspace of itself***. □

Example 16.4.8 The solution space of a homogeneous system. Suppose A is a fixed $m \times n$ matrix. Solutions to the homogeneous system $A\mathbf{x} = \mathbf{0}$ can be considered as (column) vectors in $\mathbb{R}^n$, so the solution set to this system is a subcollection of a vector space. Is it a subspace? Let's apply the Subspace Test, similarly to Discovery 16.2.i.

Nonempty. Since a homogeneous system is always consistent, the solution set is nonempty. In particular, the solution set contains the zero vector, since this is the vector corresponding to the trivial solution.

Closed under vector addition. Suppose $\mathbf{x}_1$ and $\mathbf{x}_2$ are two solutions to this system. Then both

$$A\mathbf{x}_1 = \mathbf{0} \qquad \text{and} \qquad A\mathbf{x}_2 = \mathbf{0}.$$

To check if the sum vector is also in the solution set, we need to check whether $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$ satisfies the matrix equation $A\mathbf{x} = \mathbf{0}$:

$$A(\mathbf{x}_1 + \mathbf{x}_2) = A\mathbf{x}_1 + A\mathbf{x}_2 = \mathbf{0} + \mathbf{0} = \mathbf{0}.$$

So the sum vector is indeed in the solution set.

Closed under scalar multiplication. Suppose $\mathbf{x}_0$ is a solution to this system. Then $A\mathbf{x}_0 = \mathbf{0}$. For a scalar k, to check whether the scaled vector $k\mathbf{x}_0$ is also in the solution set, we need to check whether $\mathbf{x} = k\mathbf{x}_0$ satisfies the matrix equation $A\mathbf{x} = \mathbf{0}$:

$$A(k\mathbf{x}_0) = kA\mathbf{x}_0 = k\mathbf{0} = \mathbf{0}.$$

So the scaled vector is indeed in the solution set.

Conclusion. Since all parts of the Subspace Test pass, the solution set of the homogeneous system $A\mathbf{x} = \mathbf{0}$ is a subspace of $\mathbb{R}^n$. □

Remark 16.4.9 Every subspace is somehow defined by a homogeneous condition or a set of homogeneous conditions. In the solution space example above, this was explicit — the subcollection was directly defined as the solution set of a homogeneous matrix equation $A\mathbf{x} = \mathbf{0}$. On the other hand, it's easy to see that the solution set of a nonhomogeneous system $A\mathbf{x} = \mathbf{b}$ would *not* be a subspace of $\mathbb{R}^n$, since it would not contain the zero vector.

Let's reconsider some of the examples of Discovery 16.2 from this perspective.

- In Discovery 16.2.a, the xy-plane is a subspace of $\mathbb{R}^3$, and it corresponds to the homogeneous condition $z = 0$. However, in Discovery 16.2.b, the plane parallel to the xy-plane in $\mathbb{R}^3$ but shifted one unit upward is *not* a subspace, and it corresponds to the *non*homogeneous condition $z = 1$.

- In Discovery 16.2.c, the collection of 10×10 diagonal matrices is a subspace of $\mathrm{M}_{10}(\mathbb{R})$, and it corresponds to the homogeneous conditions that the off-diagonal entries be zero. However, in Discovery 16.2.d, the collection of those 12×12 matrices with a 7 in the $(1,1)$ entry is *not* a subspace of $\mathrm{M}_{12}(\mathbb{R})$, and this collection corresponds to the *non*homogeneous condition of requiring the top-left entry be 7.

- In Discovery 16.2.f, the collection $\mathrm{P}_2(\mathbb{R})$ of polynomials of degree 2 or less is a subspace of $\mathrm{P}(\mathbb{R})$, and it corresponds to the homogeneous conditions of requiring the coefficient on every power x^n, $n \geq 3$, be zero. However, in Discovery 16.2.g, the collection of polynomials of degree *exactly* 2 is *not* a subspace. While this subcollection requires the same homogeneous conditions as those defining P_2, it also requires the *non*homogeneous condition that the coefficient on x^2 be *non*zero.

16.4.3 Determining if a vector is in a span

In Discovery 16.3, we explored the question of determining whether a given vector was in the subspace generated by a specified spanning set. For this to be true, that vector must be a linear combination of vectors in the spanning set.

Example 16.4.10 A span of $\mathbb{R}^3$-vectors. This example corresponds to Discovery 16.3.a. Working in $\mathbb{R}^3$, we would like to determine if $\mathbf{v} = (1,-1,4)$ is in $\operatorname{Span} S$ for $S = \{(1,0,1),(2,1,-1)\}$. Let's try to express $\mathbf{v}$ as a linear combination of vectors in the spanning set, and see if it works out. Set

$$(1,-1,4) = s(1,0,1) + t(2,1,-1),$$

for unknown scalars s,t. Combining the linear combination on the right into a single vector and comparing components on each side, we get (surprise!) a linear system of equations:

$$\left\{\begin{array}{rcrcr} 1 & = & s & + & 2t, \\ -1 & = & & & t, \\ 4 & = & s & - & t. \end{array}\right.$$

Now, we don't actually care about the solution to this system — we only care if the system is *consistent* or not, because if it's consistent then there *is* a way to express $\mathbf{v}$ as a linear combination of the spanning vectors, so that $\mathbf{v}$ is in $\operatorname{Span} S$. And, as you can check for yourself, this system is consistent.

There is an interesting pattern to note if we actually convert the system above into an augmented matrix:

$$\left[\begin{array}{rr|r} 1 & 2 & 1 \\ 0 & 1 & -1 \\ 1 & -1 & 4 \end{array}\right].$$

Notice that ***the columns in the coefficient matrix are precisely the vectors in the spanning set, and the column of constants is precisely the vector that we are testing as being in*** $\operatorname{Span} S$ ***or not***. □

Example 16.4.11 A span of matrices. This example corresponds to Discovery 16.3.b. Working in $M_{2\times 3}(\mathbb{R})$, we would like to determine if $\mathbf{v} = \left[\begin{smallmatrix} 0 & 2 & 2 \\ 3 & -3 & -3 \end{smallmatrix}\right]$ is in $\operatorname{Span} S$, for

$$S = \left\{ \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right\}.$$

Here, we can see more directly that $\mathbf{v}$ is *not* in $\operatorname{Span} S$. Notice that the nonzero entries of the matrices in S do not overlap. From this, we can see that every linear combination of these spanning matrices will have the first two entries in the second row equal to each other. But the entries of $\mathbf{v}$ do not have this property.

If we didn't notice this, we could carry out a similar procedure as in the previous example, beginning with the vector equation

$$\begin{bmatrix} 0 & 2 & 2 \\ 3 & -3 & -3 \end{bmatrix} = r \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} + s \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} + t \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

in the unknown scalars r,s,t. Combining the linear combination on the right into one matrix, and then comparing entries on each side, we get a linear system with augmented matrix

$$\left[\begin{array}{ccc|r} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 \\ 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 3 \\ 0 & 1 & 0 & -3 \\ 0 & 0 & 1 & -3 \end{array}\right].$$

Notice. The pattern in the columns versus the vectors in our problem again!

In this matrix, the inconsistency is obvious in the fourth and fifth rows. □

Example 16.4.12 A span of polynomials. This example corresponds to Discovery 16.3.c. Working in $\mathrm{P}(\mathbb{R})$, we would like to determine if the vector $\mathbf{v} = 2 - x + 3x^2$ is in $\operatorname{Span} S$ for $S = \{1, 1+x, 1+x^2\}$. Again, we set up a vector equation expressing $\mathbf{v}$ as a linear combination in the vectors of S with unknown scalars:

$$2 - x + 3x^2 = r \cdot 1 + s(1+x) + t(1+x^2).$$

Two polynomials are only equal if they have the same degree and all the same coefficients. From this, we get the following linear system:

$$\begin{array}{rrcrcrcr} \text{constant term:} & 2 & = & r & + & s & + & t, \\ x \text{ term:} & -1 & = & & & s, & & \\ x^2 \text{ term:} & 3 & = & & & & & t, \end{array}$$

which can be converted into augmented matrix

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

Notice. the pattern in the columns versus the vectors in our problem again!

This system is consistent, so $\mathbf{v}$ is indeed in $\operatorname{Span} S$. □

16.4.4 Determining if a spanning set generates the whole vector space

In Discovery 16.4, we attempted to determine whether a given spanning set generated the entire vector space. In other words, we attempted to answer: is *every* vector in the vector space somehow a linear combination of spanning set vectors? In the three examples of that discovery activity, the answer was very clearly *yes*.

Example 16.4.13 A spanning set for $\mathbb{R}^5$. In Discovery 16.4.a, clearly every vector in $\mathbb{R}^5$ can be decomposed as a linear combination of the standard basis vectors:

$$(a,b,c,d,e) = a\mathbf{e}_1 + b\mathbf{e}_2 + c\mathbf{e}_3 + d\mathbf{e}_4 + e\mathbf{e}_5.$$

□

Example 16.4.14 A spanning set for $\mathrm{M}_2(\mathbb{R})$. In Discovery 16.4.b, every vector in $\mathrm{M}_2(\mathbb{R})$ can be decomposed as a linear combination of the provided spanning set vectors:

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

□

Example 16.4.15 A spanning set for $\mathrm{P}_3(\mathbb{R})$. In Discovery 16.4.c, every vector $a + bx + cx^2 + dx^3$ in $\mathrm{P}_3(\mathbb{R})$ is *naturally* expressed as a linear combination of 1 and the powers of x up to x^3, where the scalars in the linear combination are just the coefficients of the polynomial. □

Remark 16.4.16 In each of the spaces in the examples above, there are analogues for other "dimensions" of vectors.

1. In $\mathbb{R}^n$, the standard basis vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ always form a spanning set for the entire vector space.

2. In $\mathrm{M}_{m\times n}(\mathbb{R})$, the set of "standard basis vectors," consisting of those matrices that have all zero entries except for a single 1 in one specific entry, is always a spanning set for the entire vector space.

3. When we write a polynomial, we naturally write it as a *linear combination of the constant polynomial* 1 *and powers of* x. So in $\mathrm{P}_n(\mathbb{R})$, the "standard basis vectors" $1, x, x^2, \dots, x^n$ form a spanning set for the entire vector space.

In more complicated examples, the question "Is $\operatorname{Span} S$ equal to the whole space?" may be more difficult to answer with the concepts we have accumulated so far. We will make this question more manageable through a deeper study of the relationships of vectors to one another with respect to linear combinations, and by attaching a notion of "size" to subspaces. In the meantime, see Example 16.6.2 in Section 16.6 for a preliminary method of answering this question.

16.5 Theory

In this section.

16.5.1 The Subspace Test

First we formally state the Subspace Test, and provide a proof.

Proposition 16.5.1 Subspace Test. *A subcollection of vectors in a vector space is a subspace if and only if all three of the following conditions are met.*

(i) The subcollection is nonempty. That is, it contains at least one vector.

(ii) The subcollection is ***closed under vector addition****. That is, given vectors* $\mathbf{w}_1$ *and* $\mathbf{w}_2$ *in the subcollection, the sum* $\mathbf{w}_1 + \mathbf{w}_2$ *is also in the subcollection.*

(iii) The subcollection is ***closed under scalar multiplication****. That is, given vector* $\mathbf{w}$ *in the subcollection and scalar* k*, the scaled vector* $k\mathbf{w}$ *is also in the subcollection.*

Proof. Suppose we have a subcollection of vectors in a vector space that satisfies the three conditions of the Subspace Test. We would like to verify that this subcollection is a vector space all on its own, using the same vector addition and scalar multiplication operations as the larger space. But as explored in Discovery 16.1 and discussed in Subsection 16.3.1, we don't need to verify all ten vector space axioms — we get the six algebra axioms for free from knowing that is how vector algebra works in the larger space. The remaining four axioms are Axiom A 1, Axiom A 4, Axiom A 5, and Axiom S 1, so we really only need to verify that the subcollection contains the zero vector, and is closed under vector addition, scalar multiplication, and taking negatives. Furthermore, from Condition ii and Condition iii of the test, we already know that the subcollection is closed under addition and scalar multiplication. So we are down to checking the zero vector and negatives.

To show that the subcollection must contain the zero vector, consider that Condition i of the test guarantees that the subcollection contains *some* vector $\mathbf{v}$. But then Condition iii of the test tells us that the subcollection must also contain every scalar multiple of $\mathbf{v}$. In particular, by applying Rule 1.b of Proposition 15.6.2 we may say that the subcollection must contain $0\mathbf{v} = \mathbf{0}$, as desired.

To show that the subcollection must be closed under taking negatives, consider a vector $\mathbf{w}$ in the subcollection. Again, Condition i of the test says that the subcollection must also contain every scalar multiple of $\mathbf{w}$. In particular, by applying Rule 2.e of Proposition 15.6.2 we may say that the subcollection must contain $(-1)\mathbf{w} = -\mathbf{w}$, as desired.

Finally, we will consider the "only if" part of the statement. Suppose we have a subcollection of vectors in a vector space that we already know is a subspace. A subspace is itself a vector space, so it must be nonempty (since it at least contains *some* zero vector by Axiom A 4), it must be closed under vector addition (Axiom A 1), and it must be closed under scalar multiplication (Axiom S 1). In other words, it must pass the Subspace Test. ■

As per the proposition above, every subspace satisfies the conditions of the Subspace Test. But we can go a little further.

Proposition 16.5.2 Properties of subspaces. *Every subspace of a vector space contains the zero vector of that space, and is closed under vector addition, scalar multiplication, taking negatives, and taking linear combinations.*

Proof. We have already established that a subspace is closed under the vector operations. Verifying that it also contains the zero vector and is closed under taking negatives is exactly as in the proof of Proposition 16.5.1 above, since we know that a subspace always passes the Subspace Test.

It remains to show that a subspace is closed under linear combinations. So suppose that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_\ell$ are vectors in the subspace. Since the subspace is closed under scalar multiplication, the vectors $k_1\mathbf{v}_1, k_2\mathbf{v}_2, \ldots, k_\ell\mathbf{v}_\ell$ are all also in the subspace. And then, since the subspace is also closed under addition, the linear combination $k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + \cdots + k_\ell\mathbf{v}_\ell$ is also in the subspace. ■

From the first property of subspaces listed above, we can deduce our observation about the best way to verify Condition i of the Subspace Test.

Corollary 16.5.3 Subspaces must contain the zero vector. *If a subcollection of a vector space does* not *contain the zero vector of the larger space, then it cannot be a subspace.*

16.5.2 Universal examples of subspaces

Here we recognize examples of subspaces that occur in every vector space.

Proposition 16.5.4 The trivial and full subspaces. *In every vector space, both the zero space* $\{\mathbf{0}\}$ *and the whole space are subspaces.*

Proof. We verified these examples of subspaces in Example 16.4.6 and Example 16.4.7 of Subsection 16.4.2. ■

Proposition 16.5.5 Creating subspaces via spanning sets.

1. *If S is a nonempty collection of vectors in a vector space, then* $\operatorname{Span} S$ *is a subspace of that vector space, and it contains every vector in S.*

2. *The subspace* $\operatorname{Span} S$ *is the* smallest *subspace that contains every vector in S in the following sense: every other subspace that contains the vectors of S must also contain* $\operatorname{Span} S$ *as a subspace.*

3. *Every vector space (and hence, every subspace of a vector space) has a spanning set.*

Proof of Statement 1. Recall that $\operatorname{Span} S$ is the collection of all possible linear combinations of vectors in S. First we verify that $\operatorname{Span} S$ contains every vector in S. Indeed, if $\mathbf{v}$ is a vector in S, then it is trivially a linear combination of vectors in S by $\mathbf{v} = 1\mathbf{v}$.

Let V represent the vector space from which the collection of vectors S is taken. First, we know that every vector in $\operatorname{Span} S$ is a vector in V, because the vectors in $\operatorname{Span} S$ are linear combinations of the vectors in S, and V is closed under taking linear combinations (Proposition 16.5.2, where V is considered as a subspace of itself using Proposition 16.5.4). So $\operatorname{Span} S$ is a subcollection of V.

Now let's apply the Subspace Test to $\operatorname{Span} S$.

Nonempty. We know $\operatorname{Span} S$ is nonempty because it contains each of the vectors of S.

Closed under vector addition. Suppose $\mathbf{u}$ and $\mathbf{v}$ are vectors in $\operatorname{Span} S$. Then each is a linear combination of vectors in S, say

$$\begin{aligned}\mathbf{u} &= k_1\mathbf{u}_1 + k_2\mathbf{u}_2 + \cdots + k_s\mathbf{u}_s,\\ \mathbf{v} &= m_1\mathbf{v}_1 + m_2\mathbf{v}_2 + \cdots + m_t\mathbf{v}_t,\end{aligned}$$

where each of $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_s$ and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_t$ are vectors in S. Then,

$$\begin{aligned}\mathbf{u}+\mathbf{v} = k_1\mathbf{u}_1 &+ k_2\mathbf{u}_2 + \cdots + k_s\mathbf{u}_s\\ &+ m_1\mathbf{v}_1 + m_2\mathbf{v}_2 + \cdots + m_t\mathbf{v}_t,\end{aligned}$$

which is again a linear combination of vectors in S, so $\mathbf{u}+\mathbf{v}$ is also in $\operatorname{Span} S$. This shows that $\operatorname{Span} S$ is closed under vector addition.

Closed under scalar multiplication. Suppose $\mathbf{v}$ is a vector in $\operatorname{Span} S$. Then it is a linear combination of vectors in S, say

$$\mathbf{v} = m_1\mathbf{v}_1 + m_2\mathbf{v}_2 + \cdots + m_\ell\mathbf{v}_\ell,$$

where each of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_\ell$ are vectors in S. Then for every scalar k,

$$\begin{aligned}k\mathbf{v} &= k(m_1\mathbf{v}_1 + m_2\mathbf{v}_2 + \cdots + m_t\mathbf{v}_\ell)\\ &= km_1\mathbf{v}_1 + km_2\mathbf{v}_2 + \cdots + km_\ell\mathbf{v}_\ell,\end{aligned}$$

which is again a linear combination of vectors in S, so $k\mathbf{v}$ is always also in $\operatorname{Span} S$. This shows that $\operatorname{Span} S$ is closed under scalar multiplication.

Conclusion. Since $\operatorname{Span} S$ passes the Subspace Test, it is a subspace of V. ■

Proof of Statement 2. We wish to show that every other subspace that contains the vectors of S must also contain $\operatorname{Span} S$ as a subspace. So suppose we have another subspace that contains the vectors of S. Then it must contain every linear combination of the vectors in S, since subspaces are closed under taking linear combinations (Proposition 16.5.2). That is, if a subspace contains all of the vectors in S, then it must also contain all of the vectors in $\operatorname{Span} S$.

Note. There is no need to use the Subspace Test to prove that $\operatorname{Span} S$ is a subspace of this other subspace — we already know from Statement 1 that $\operatorname{Span} S$ is a subspace of V, the vector space from which the vectors S are taken. So $\operatorname{Span} S$ is a vector space all on its own, hence will be a subspace of *any* space that contains all of its vectors. (See the definition of **subspace** in Section 16.2.)

■

Proof of Statement 3. A vector space V always has an obvious spanning set — itself! That is, we claim that $V = \operatorname{Span} V$ is always true. To verify this, we must demonstrate that each vector in the collection V is also in the collection $\operatorname{Span} V$, and vice versa, so that they are exactly the same collection of vectors. However, by applying Statement 1 we can immediately say that $\operatorname{Span} V$ is a subspace of V (implying every vector in $\operatorname{Span} V$ is in V) that contains every vector of the spanning set V (i.e. every vector in V is in $\operatorname{Span} V$). ■

16.5.3 Equality of subspaces created via spanning sets

Finally, we provide a way to determine when two spanning sets generate the same subspace.

Proposition 16.5.6 Comparing spanned spaces. *Suppose S and S' are two sets of vectors in a vector space.*

1. *If each vector in S can be expressed as a linear combination of the vectors in S', then* $\operatorname{Span} S$ *is a subspace of* $\operatorname{Span} S'$.

2. *If each vector in S can be expressed as a linear combination of the vectors in S',* and *each vector in S' can be expressed as a linear combination of the vectors in S, then* $\operatorname{Span} S$ *and* $\operatorname{Span} S'$ *are the same space.*

Proof of Statement 1. Recall that $\operatorname{Span} S'$ is the collection of all possible linear combinations of the vectors in S'. So assuming that each vector in S can be expressed as a linear combination of the vectors in S' is the same as assuming that each vector in S is in $\operatorname{Span} S'$. But Statement 2 of Proposition 16.5.5 tells us that $\operatorname{Span} S$ is the *smallest* subspace that contains all the vectors in S, and that $\operatorname{Span} S$ must therefore be a subspace of $\operatorname{Span} S'$. ■

Proof of Statement 2. Now we assume both that each vector in S can be expressed as a linear combination of the vectors in S' and that each vector in S' can be expressed as a linear combination of the vectors in S. Then we can apply Statement 1 of this proposition twice, first to conclude that $\operatorname{Span} S$ is a subspace of $\operatorname{Span} S'$, and second to conclude that $\operatorname{Span} S'$ is a subspace of $\operatorname{Span} S$. But then $\operatorname{Span} S$ is a subcollection of the vectors in $\operatorname{Span} S'$, and also vice versa. This can only happen if they are in fact the same collection of vectors. ■

16.6 More examples

Before concluding this chapter, we'll illustrate the uses of Proposition 16.5.6 with two examples.

Example 16.6.1 Recognizing when two subspaces are the same. Consider the sets of vectors $S = \{(1,0,0),(0,1,0)\}$ and $S' = \{(1,1,0),(1,0,0),(1,-1,0)\}$ in $\mathbb{R}^3$. It should be clear that $\operatorname{Span} S$ is the xy-plane in $\mathbb{R}^3$. Does $\operatorname{Span} S'$ generate the same subspace?

To answer this question, we use Statement 2 of Proposition 16.5.6, which gives us two new questions to answer.

- Can each vector in S be expressed as a linear combination of the vectors in S'? Yes, because

$$\begin{bmatrix}1\\0\\0\end{bmatrix} = 0\begin{bmatrix}1\\1\\0\end{bmatrix} + 1\begin{bmatrix}1\\0\\0\end{bmatrix} + 0\begin{bmatrix}1\\-1\\0\end{bmatrix},$$

$$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + 0\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \left(-\frac{1}{2}\right)\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

- Can each vector in S' be expressed as a linear combination of the vectors in S? Yes, because

$$\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = 1\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 1\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix},$$

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = 1\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 0\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix},$$

$$\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = 1\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + (-1)\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

Since both questions have been answered in the affirmative, Statement 2 of Proposition 16.5.6, tells us that $\operatorname{Span} S$ and $\operatorname{Span} S'$ are the same space. □

Example 16.6.2 Determining if a spanning set generates the whole vector space. Consider the set of vectors $S = \{A_1, A_2, A_3, A_4\}$ in $M_2(\mathbb{R})$, where

$$A_1 = \begin{bmatrix} 0 & -1 \\ 2 & 1 \end{bmatrix}, \qquad A_3 = \begin{bmatrix} 0 & 1 \\ -2 & 0 \end{bmatrix},$$

$$A_2 = \begin{bmatrix} 1 & 2 \\ 4 & -1 \end{bmatrix}, \qquad A_4 = \begin{bmatrix} 0 & 0 \\ 1 & -2 \end{bmatrix}.$$

Is this set a spanning set for *all* of $M_2(\mathbb{R})$? That is, is $M_2(\mathbb{R}) = \operatorname{Span} S$? We already know a spanning set for $M_2(\mathbb{R})$ — the set of standard basis vectors $\mathcal{B} = \{E_{11}, E_{12}, E_{21}, E_{22}\}$, where

$$E_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad E_{12} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

$$E_{21} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \qquad E_{22} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

That is, we already know that $M_2(\mathbb{R}) = \operatorname{Span} \mathcal{B}$. So we can turn our question into: is $\operatorname{Span} S = \operatorname{Span} \mathcal{B}$? With this new version of our problem, we can use the same method as in the previous example. However, we don't need to explicitly verify that each vector in S can be expressed as a linear combination of the vectors in $\mathcal{B}$. Besides being obvious, this fact is already implied by our assertion that $M_2(\mathbb{R}) = \operatorname{Span} \mathcal{B}$, since clearly each vector in S is a vector in $M_2(\mathbb{R})$. So it just remains to verify that each vector in $\mathcal{B}$ can be expressed as a linear combination of the vectors in S. Let's begin with vector E_{11}. We use the same strategy as in the examples in Subsection 16.4.3: express E_{11} as a linear combination of the vectors in S with unknown scalar coefficients, set up equations in those unknown scalars, and determine whether the resulting linear system is consistent.

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = k_1\begin{bmatrix} 0 & -1 \\ 2 & 1 \end{bmatrix} + k_2\begin{bmatrix} 1 & 2 \\ 4 & -1 \end{bmatrix} + k_3\begin{bmatrix} 0 & 1 \\ -2 & 0 \end{bmatrix} + k_4\begin{bmatrix} 0 & 0 \\ 1 & -2 \end{bmatrix}$$

$$= \begin{bmatrix} k_2 & -k_1 + 2k_2 + k_3 \\ 2k_1 + 4k_2 - 2k_3 + k_4 & k_1 - k_2 - 2k_4 \end{bmatrix}$$

Comparing entries on left and right sides leads to the system of equations

$$\left\{ \begin{array}{rcrcrcrcl} & & k_2 & & & & & = & 1, \\ -k_1 & + & 2k_2 & + & k_3 & & & = & 0, \\ 2k_1 & + & 4k_2 & - & 2k_3 & + & k_4 & = & 0, \\ k_1 & - & k_2 & & & - & 2k_4 & = & 0, \end{array} \right.$$

which can be put in an augmented matrix and reduced.

$$\left[\begin{array}{rrrr|r} 0 & 1 & 0 & 0 & 1 \\ -1 & 2 & 1 & 0 & 0 \\ 2 & 4 & -2 & 1 & 0 \\ 1 & -1 & 0 & -2 & 0 \end{array} \right] \xrightarrow[\text{reduce}]{\text{row}} \left[\begin{array}{rrrr|r} 1 & 0 & 0 & 0 & -15 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & -17 \\ 0 & 0 & 0 & 1 & -8 \end{array} \right]$$

Notice. Once again, we have a pattern in the columns of the initial augmented matrix versus the vectors involved.

The reduced augmented matrix above tells us that

$$k_1 = -15, \qquad k_2 = 1, \qquad k_3 = -17, \qquad k_4 = -8,$$

and so

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = -15 \begin{bmatrix} 0 & -1 \\ 2 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 4 & -1 \end{bmatrix} - 17 \begin{bmatrix} 0 & 1 \\ -2 & 0 \end{bmatrix} - 8 \begin{bmatrix} 0 & 0 \\ 1 & -2 \end{bmatrix},$$

though we only cared about the *existence* of a solution, not the actual solution itself.

In a similar manner, one can calculate that

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = 4 \begin{bmatrix} 0 & -1 \\ 2 & 1 \end{bmatrix} + 5 \begin{bmatrix} 0 & 1 \\ -2 & 0 \end{bmatrix} + 2 \begin{bmatrix} 0 & 0 \\ 1 & -2 \end{bmatrix},$$

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

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

We have now verified that each vector in $\mathcal{B}$ can be expressed as a linear combination of the vectors in S. As discussed above, we already knew that each vector in S can be expressed as a linear combination of the vectors in $\mathcal{B}$. Therefore, Statement 2 of Proposition 16.5.6 tells us that $\operatorname{Span} S = \operatorname{Span} \mathcal{B}$. Since we already knew that $\operatorname{Span} \mathcal{B}$ is equal to the entire space $M_2(\mathbb{R})$, we must also have $\operatorname{Span} S$ equal to this entire space. □

Chapter 17

Linear independence

17.1 Discovery guide

Discovery 17.1 Consider the vectors $\mathbf{v}_1 = (1,0,1)$, $\mathbf{v}_2 = (1,1,2)$, and $\mathbf{v}_3 = (1,-1,0)$.

(a) Do you remember what Span means? Explain why the vector

$$\mathbf{x} = 3\mathbf{v}_1 + 2\mathbf{v}_2 - \mathbf{v}_3$$

is in $\operatorname{Span}\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$.

Note. You can compute $\mathbf{x}$ if you like, but it is not necessary.

(b) Actually, $\mathbf{v}_2$ can be expressed as a linear combination of $\mathbf{v}_1$ and $\mathbf{v}_3$ — do you see how?

Use this and the expression for $\mathbf{x}$ in Task a to express $\mathbf{x}$ as a linear combination of *just* $\mathbf{v}_1$ and $\mathbf{v}_3$.

(c) Task b shows that $\mathbf{x}$ is in $\operatorname{Span}\{\mathbf{v}_1, \mathbf{v}_3\}$. Do you think that similar calculations and the same reasoning can be carried out for every vector in $\operatorname{Span}\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$?

What does this say about $\operatorname{Span}\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ versus $\operatorname{Span}\{\mathbf{v}_1, \mathbf{v}_3\}$?

Discovery 17.1 demonstrates a common pattern: when one of the vectors in a spanning set can be expressed as a linear combination of the others, that vector becomes *redundant*, and a smaller spanning set can be used in place of the original one. We'll give this situation a name: a set of vectors is called **linearly dependent** if (at least) one of the vectors in the set can be written as a linear combination of other vectors in the set; otherwise the set of vectors is called **linearly independent**. However, it can be tedious to check each vector in a set one-by-one to see if it is a linear combination of others. Luckily, for a finite set of vectors, there is a way to check all of them all at once.

Test for Linear Dependence/Independence. To test whether vectors

$$\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$$

are linearly dependent or independent, set up the vector equation

$$k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + \cdots + k_m\mathbf{v}_m = \mathbf{0}, \qquad (*)$$

where the coefficients $k_1, k_2, \dots, k_m$ are (scalar) variables.

- If vector equation $(*)$ has a nontrivial solution in the variables $k_1, k_2, \dots, k_m$, then the vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$ are **linearly dependent**.

- Otherwise, if vector equation (*) has *only* the trivial solution $k_1 = 0, k_2 = 0, \dots, k_m = 0$, then the vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$ are **linearly independent**.

Check your understanding. Do you see why equation (*) always has *at least* the trivial solution?

Discovery 17.2

(a) Use the test to verify that $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ from Discovery 17.1 are linearly dependent.

Note. Forming vector equation (*) using the three vectors of Discovery 17.2.a should lead to a homogeneous linear system in variables k_1, k_2, k_3. Look at the columns in your matrix for this homogeneous ystem — what pattern do you notice?

(b) Use the test to verify that $\mathbf{v}_1, \mathbf{v}_3$ from Discovery 17.1 are linearly independent.

The next discovery activity will help you understand the Test for Linear Independence/Dependence. To keep it simple, we'll consider just three vectors at a time.

Discovery 17.3

(a) Consider abstract vectors $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$, and suppose the vector equation

$$k_1\mathbf{u}_1 + k_2\mathbf{u}_2 + k_3\mathbf{u}_3 = \mathbf{0} \tag{**}$$

has a nontrivial solution. This means that there are values for the scalars k_1, k_2, k_3, at least one of which is not zero, so that equation (**) is true.

Use some algebra to manipulate equation (**) to demonstrate that one of the vectors can be expressed as a linear combination of the others (and hence, *by definition*, the vectors $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ are linearly dependent).

Careful. Make sure you don't accidentally divide by zero!

(b) Consider abstract vectors $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$, and suppose the vector equation

$$k_1\mathbf{w}_1 + k_2\mathbf{w}_2 + k_3\mathbf{w}_3 = \mathbf{0} \tag{***}$$

has *only* the trivial solution. We would like to see why this means that $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$ are linearly independent.

Suppose they weren't: for example, suppose $\mathbf{w}_3 = c_1\mathbf{w}_1 + c_2\mathbf{w}_2$ were true for some scalars c_1, c_2. Manipulate this expression for $\mathbf{w}_3$ until is says something about equation (***). Do you see now why $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$ *cannot* satisfy the *definition* of linearly dependence, and hence must be linearly independent?

Discovery 17.4 In each of the following vector spaces, practise using the Test for Linear Dependence/Independence of the given set of vectors.

(a) $V = \mathrm{M}_2(\mathbb{R})$, $S = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\}$.

(b) $V = \mathrm{M}_2(\mathbb{R})$, $S = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix} \right\}$.

(c) $V = \mathrm{P}(\mathbb{R})$, $S = \{1 + x, 1 + x^2, 2 - x + 3x^2\}$.

Hint. After setting up the vector equation from the test for linear dependence/independence, you are solving for the scalars k_1, k_2, k_3, not for x.

On the right-hand side, the zero represents the *zero vector*, which in this space is the *zero polynomial*. What are the coefficients on powers of x in the zero polynomial? The left-hand side, being equal, must have the same coefficients.

(d) $V = \mathrm{P}(\mathbb{R}) S = \{1, x, x^2, x^3\}$

Discovery 17.5

(a) Do you think it's possible to have a set of three linearly independent vectors in $\mathbb{R}^2$? Why or why not?

(b) Do you think it's possible to have a set of four linearly independent vectors in $\mathbb{R}^3$? Why or why not?

Discovery 17.6

(a) What does the definition of linear dependence say in the case of just two vectors?

(b) If the test for linear dependence/independence is to remain true in the case of a "set" of vectors consisting of just *one* vector, how should we define linear dependence/independence for such a set?

17.2 Terminology and notation

linearly dependent set of vectors

a set of vectors in which there is at least one example of a vector that can be expressed as a linear combination of other vectors in the set; we also consider the set of vectors consisting of *only* the zero vector as linearly dependent

linearly independent set of vectors

a set of vectors that is not linearly dependent; that is, a set of vectors in which *no* vector can be expressed as a linear combination of other vectors in the set

17.3 Concepts

In this section.

Statement 3 of Proposition 16.5.5 guarantees that every vector space has a spanning set. To prove this statement, we verified that every vector space is trivially generated by itself, i.e. $V = \operatorname{Span} V$.

But this doesn't give us a useful description of the vector space V, it basically just says "to build all the vectors in V out of some vectors in V, use all the vectors in V." The point of a spanning set is to give you a set of building-block vectors that can be used to construct every other vector in the space through linear combinations. To make an analogy to chemistry, the vectors in a spanning set act like *atoms*, and linear combinations are then like *molecules* built out of different combinations of different quantities of those atoms. So we would like our set of building blocks to be as simple as possible — that is, we would like to get down to some sort of *optimal* description of a vector space in terms of a spanning set. Linear dependence and independence are precisely the concepts we will need in order to judge whether we have such an optimal spanning set.

17.3.1 Reducing spanning sets

Suppose we have a spanning set S for a space V, so that $V = \operatorname{Span} S$. This equality of spaces says that every vector in the space V can somehow be expressed as a linear combination of vectors in S.

Suppose further that one of the vectors in S can be expressed as a linear combination of some of the others, say

$$\mathbf{w} = k_1\mathbf{u}_1 + k_2\mathbf{u}_2 + \cdots + k_m\mathbf{u}_m, \tag{$*$}$$

where each of $\mathbf{w}, \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ is a vector in S.

Note. This was basically the situation for Discovery 17.1, where $\mathbf{v}_2$ played the role of $\mathbf{w}$.

Then **w** *is not actually needed when expressing vectors in V as linear combinations of vectors in S.*

Indeed, consider a vector **x** in V expressed as a linear combination of vectors in S, including **w**, say

$$\mathbf{x} = c\mathbf{w} + c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + k_m\mathbf{v}_n,$$

where each of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is a vector in S. But then we could substitute the expression in $(*)$ for **w** to obtain

$$\begin{aligned}\mathbf{x} &= c(k_1\mathbf{u}_1 + k_2\mathbf{u}_2 + \cdots + k_m\mathbf{u}_m) + c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + k_m\mathbf{v}_n \\ &= ck_1\mathbf{u}_1 + ck_2\mathbf{u}_2 + \cdots + ck_m\mathbf{u}_m + c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + k_m\mathbf{v}_n.\end{aligned}$$

Here, each of the $\mathbf{u}_i$ vectors and the $\mathbf{v}_j$ vectors are in S, making this an expression for **x** as a linear combination of vectors in S *but not including* **w**.

If **w** isn't needed to express vectors in V as linear combinations of vectors in S, then we should have $\operatorname{Span} S = \operatorname{Span} S'$, where S' is the set of all vectors in S *except* **w**. That is, we can discard **w** from the spanning set for V, and *still* have a spanning set.

See. Lemma 17.5.4 in Subsection 17.5.2.

This pattern will help us reduce down to some sort of *optimal* spanning set: we can keep removing these redundant spanning vectors that are linear combinations of other spanning vectors until there are none left.

17.3.2 Linear dependence and independence

A set of vectors is called **linearly dependent** precisely when it is non-optimal as a spanning set in the way described above: when one of the vectors in the set is a linear combination of others in the set. However, it is usually not obvious that some vector is redundant in this way — checking each vector in turn is tedious, and also would not be a very convenient way to reason with the concept of linear dependence in the abstract.

However, having an expression for a vector **w** as a linear combination of other vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$, such as

$$\mathbf{w} = k_1\mathbf{u}_1 + k_2\mathbf{u}_2 + \cdots + k_m\mathbf{u}_m, \qquad (**)$$

is equivalent to having a nontrivial linear combination of these vectors equal to the zero vector:

$$k_1\mathbf{u}_1 + k_2\mathbf{u}_2 + \cdots + k_m\mathbf{u}_m + (-1)\mathbf{w} = \mathbf{0}. \qquad (***)$$

And vice versa, since if we have a nontrivial linear combination of these vectors that results in the zero vector, say

$$k_1\mathbf{u}_1 + k_2\mathbf{u}_2 + \cdots + k_m\mathbf{u}_m + k\mathbf{w} = \mathbf{0},$$

and the coefficient k on **w** is nonzero, then we can rearrange to get

$$\mathbf{w} = -\frac{k_1}{k}\mathbf{u}_1 + \left(-\frac{k_2}{k}\right)\mathbf{u}_2 + \cdots + \left(-\frac{k_m}{k}\right)\mathbf{u}_m,$$

an expression for **w** as a linear combination of the others.

Again, the advantage of checking for linear combinations equal to the zero vector is that in a collection of vectors S, we usually don't know ahead of time which one will be the odd one out (that is, which one will play the role of **w** as above). In a nontrivial linear combination equalling $\mathbf{0}$, we can take as the redundant vector $\mathbf{w}$ any of the vectors whose coefficient is nonzero (which is required to perform the step of dividing by k in the algebra isolating $\mathbf{w}$ above).

Now, we can always take the *trivial* linear combination, where all coefficients are 0, to get a result of $\mathbf{0}$. But if this is the *only* linear combination of a set of vectors by which we can get $\mathbf{0}$ as a result, then *none* of the vectors can act as the redundant $\mathbf{w}$ as above, because an expression like (∗∗) *always* leads to an equality like (∗∗∗), involving a nontrivial linear combination.

This logic leads to the Test for Linear Dependence/Independence.

Procedure 17.3.1 Test for Linear Dependence/Independence. *To test whether vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ are linearly dependent / independent, solve the homogeneous vector equation*

$$k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + \cdots + k_m\mathbf{v}_m = \mathbf{0}$$

in the (scalar) variables $k_1, k_2, \ldots, k_m$.

If this vector equation has a nontrivial solution for these coefficient variables, then the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ are linearly dependent.

Otherwise, if this vector equation has only *the trivial solution $k_1 = 0, k_2 = 0, \ldots, k_m = 0$, then the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ are linearly independent.*

17.3.3 Linear dependence and independence of just one or two vectors

For a pair $\mathbf{u}, \mathbf{v}$ of vectors to be linearly dependent, one must be a linear combination of the other. But a linear combination of one vector is just a scalar multiple, and so *a pair of vectors is linearly dependent if one is a scalar multiple of the other*. If both vectors are nonzero, that scalar must also be nonzero and so could be shifted to the other side as its reciprocal:

$$\mathbf{u} = k\mathbf{v} \qquad \iff \qquad \frac{1}{k}\mathbf{u} = \mathbf{v} \qquad \text{(for } k \neq 0\text{)}.$$

So nonzero vectors $\mathbf{u}, \mathbf{v}$ *are linearly dependent if and only if each is a scalar multiple of the other*. In $\mathbb{R}^n$, we would have called such vectors **parallel**.

What about a set containing a single vector? Our motivation for creating the concept of linear dependence/independence was to measure whether a spanning set contained redundent information or whether it was somehow “optimal.” A spanning set consisting of a single nonzero vector cannot be reduced to a smaller spanning set, so it is already optimal and we should refer to that spanning set as linearly independent. This coincides with the result of the test for linear dependence/independence for a single vector $\mathbf{v}$: if $\mathbf{v}$ is nonzero, then there are no nontrivial solutions to the vector equation $k\mathbf{v} = \mathbf{0}$.

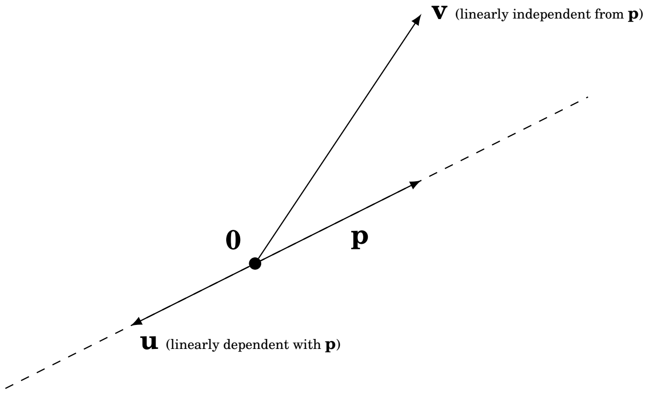
See. Rule 1.e of Proposition 15.6.2.

But what about the zero vector by itself? Scalar multiples of $\mathbf{0}$ remain $\mathbf{0}$, so Span$\{\mathbf{0}\}$ is the trivial vector space consisting of just the zero vector. Is $\{\mathbf{0}\}$ an optimal spanning set for the trivial space, or can it be reduced further? By convention, we also consider Span$\{\}$ to be the trivial vector space (where $\{\}$ represents a set containing *no* vectors, called the **empty set**), because we always want the Span process to return a subspace of the vector space in which we’re working. So the spanning set $\{\mathbf{0}\}$ *can* be reduced to the empty set, and still span the same space. This line of reasoning leads us to consider the set of vectors containing only the zero vector to be linearly dependent.

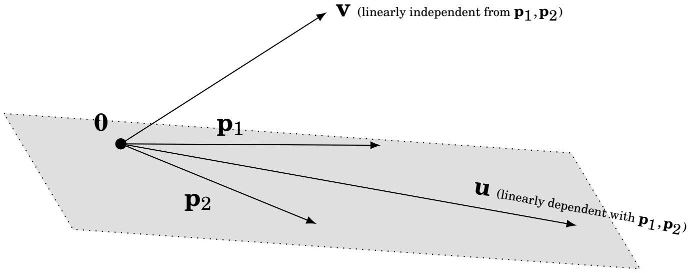
And. We should consider the empty set to be linearly independent, since it cannot be reduced.

17.3.4 Linear dependence and independence in $\mathbb{R}^n$

Independent directions. In Subsection 16.3.3, we discussed how a nonzero vector in $\mathbb{R}^n$ spans a line through the origin. If a second vector is linearly dependent with the vector spanning the line, then as discussed in the previous subsection (Subsection 17.3.3), that second vector must be parallel with the first, hence parallel with the line. To get something linearly independent, we need to branch off in a new direction from the line.


$\mathbf{v}$ (linearly independent from $\mathbf{p}$)
$\mathbf{0}$
$\mathbf{p}$
$\mathbf{u}$ (linearly dependent with $\mathbf{p}$)

In Subsection 16.3.3, we also discussed how a pair of nonzero, nonparallel vectors in $\mathbb{R}^n$ span a plane through the origin. If a third vector is linearly dependent with those two spanning vectors, it is somehow a linear combination of them and so lies in that plane. To get something linearly independent, we need to branch off in a new direction from that plane.


$\mathbf{v}$ (linearly independent from $\mathbf{p}_1, \mathbf{p}_2$)
$\mathbf{0}$
$\mathbf{p}_1$
$\mathbf{p}_2$
$\mathbf{u}$ (linearly dependent with $\mathbf{p}_1, \mathbf{p}_2$)

This idea that we can enlarge an independent set by including a new vector in a new direction works in abstract vector spaces as well, as we will see in Proposition 17.5.6 in Subsection 17.5.2.

Maximum number of linearly independent vectors. In Discovery 17.5, we considered the possible sizes of linearly independent sets in $\mathbb{R}^2$ and $\mathbb{R}^3$. We can certainly have two linearly independent vectors in $\mathbb{R}^2$, since clearly the standard basis vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ form a linearly independent set in $\mathbb{R}^2$. Could we have three? Two linearly independent vectors would have to be nonparallel, and so they would have to span a plane — i.e. they would have to span the *entire* plane. Geometrically, a third linearly independent vector would have to branch off in a "new direction," as in our discussion above. But in $\mathbb{R}^2$ there is no new third direction in which to head — we can't have a vector breaking up out of the plane "into the third dimension," because there is no third dimension available in $\mathbb{R}^2$. Algebraically, if we had three vectors $\mathbf{u} = (u_1, u_2)$, $\mathbf{v} = (v_1, v_2)$, $\mathbf{w} = (w_1, w_2)$ in $\mathbb{R}^2$

and attempted the test for dependence/independence, we would start by setting up the vector equation

$$k_1\mathbf{u} + k_2\mathbf{v} + k_3\mathbf{w} = \mathbf{0}.$$

Combining the linear combination on the left back into one vector, and comparing coordinates on either side, we would obtain linear system

$$\begin{cases} u_1k_1 + v_1k_2 + w_1k_3 = 0, \\ u_2k_1 + v_2k_2 + w_2k_3 = 0, \end{cases}$$

in the unknown coefficients k_1, k_2, k_3. Because there are only two equations, the reduced form for the coefficient matrix for this system can have no more than two leading ones, so it requires at least one parameter to solve, which means there are nontrivial solutions. So three vectors in $\mathbb{R}^2$ can never by linearly independent.

We come to a similar conclusion in $\mathbb{R}^3$ using both geometric and algebraic reasoning — three independent vectors in $\mathbb{R}^3$ is certainly possible (for example, the standard basis vectors), but a set of four vectors in $\mathbb{R}^3$ can never be linearly independent. Geometrically, once you have three independent vectors pointing in three "independent directions," there is no new direction in $\mathbb{R}^3$ for a fourth independent vector to take. Algebraically, we could set up the test for independence with four vectors in $\mathbb{R}^3$ and it would lead to a homogeneous system of three equations (one for each coordinate) in four variables (one unknown coefficient for each vector). Since the system would have more variables than equations, it would require parameters to solve, leading to nontrivial solutions.

And the pattern continues in higher dimensions — a collection of more than four vectors in $\mathbb{R}^4$ must be linearly dependent, a collection of more than five vectors in $\mathbb{R}^5$ must be linearly dependent, and so on. In fact, this pattern can also be found in abstract vectors spaces — see Lemma 17.5.7 in Subsection 17.5.2. And this pattern will help us transplant the idea of **dimension** from $\mathbb{R}^n$ to abstract spaces.

17.4 Examples

In this section.

17.4.1 Testing dependence/independence

Here we will carry out examples of applying the Test for Linear Dependence/ Independence.

Example 17.4.1 Testing dependence/independence in $\mathbb{R}^n$. Are the vectors $(1,0,0,1), (1,1,0,-1), (2,1,0,0), (5,1,0,5)$ in $\mathbb{R}^4$ linearly dependent or independent? Set up the test:

$$k_1(1,0,0,1) + k_2(1,1,0,-1) + k_3(2,1,0,0) + k_4(5,1,0,5) = (0,0,0,0).$$

Notice how we have used the proper zero vector in this space on the right-hand side. On the left-hand side, we want to combine the expression into one vector so that we can compare with the zero vector.

$$(k_1,0,0,k_1) + (k_2,k_2,0,-k_2) + (2k_3,k_3,0,0) + (5k_4,k_4,0,5k_4) = (0,0,0,0)$$

$$(k_1+k_2+2k_3+5k_4, k_2+k_3+k_4, 0, k_1-k_2+5k_4) = (0,0,0,0)$$

Comparing components on either side, we obtain a system of four equations in the unknown scalars from the linear combination:

$$\left\{\begin{array}{rcrcrcrcl} k_1 & + & k_2 & + & 2k_3 & + & 5k_4 & = & 0, \\ & & k_2 & + & k_3 & + & k_4 & = & 0, \\ & & & & & & 0 & = & 0, \\ k_1 & - & k_2 & & & + & 5k_4 & = & 0. \end{array}\right.$$

Now we'll solve this homogeneous system by row reducing its coefficient matrix.

$$\begin{bmatrix} 1 & 1 & 2 & 5 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 5 \end{bmatrix} \xrightarrow[\text{reduce}]{\text{row}} \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad (*)$$

Note that here it was not necessary to reduce all the way to RREF, as we are not actually interested in the solutions to this system — we only need to know whether there exist nontrivial solutions. From the reduced matrix, we can see that k_3 is a free variable and would be assigned a parameter in the general solution. The necessity of a parameter means there are an infinite number of solutions, which in particular means there are nontrivial solutions. Therefore, this collection of vectors is *linearly dependent*. □

Remark 17.4.2 Notice how the vectors from $\mathbb{R}^4$ that we were testing in the previous example ended up as columns in the coefficient matrix in $(*)$ — we saw a similar pattern in Example 16.4.10 (and in the other examples in Subsection 16.4.3), where we tested whether a particular vector was in the span of some collection of vectors.

Example 17.4.3 Testing dependence/independence in $\mathrm{M}_{m\times n}(\mathbb{R})$.

1. Consider the matrices in Discovery 17.4.a. First we set up the Test for Linear Dependence/Independence. Again, we use the proper zero vector on the right-hand side, and then we combine the expression on the left-hand side into one vector so that we may compare against the zero vector.

$$k_1\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + k_2\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + k_3\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} k_1 & 0 \\ 0 & k_1 \end{bmatrix} + \begin{bmatrix} 0 & k_2 \\ k_2 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & k_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} k_1 & k_2 \\ k_2 & k_1+k_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

   There is no need to set up a system of equations here — we can see from comparing the top rows on either side that $k_1 = 0$ and $k_2 = 0$. Then, from the $(2,2)$ entries, we see that $k_1 + k_3 = 0$. But since we already have $k_1 = 0$, we get $k_3 = 0$ as well. So there is only the trivial solution, and these vectors are *linearly independent*.

2. Consider the matrices in Discovery 17.4.b. Again, we start by setting up the Test for Linear Dependence/Independence using the appropriate zero vector.

$$k_1\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + k_2\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} + k_3\begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

As before, this will lead to a homogeneous system of equations in the unknown scalars k_1, k_2, k_3, and the coefficient matrix of this system will have the entries of the three vectors as columns:

$$\begin{bmatrix} 1 & 1 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & -1 & -2 \end{bmatrix} \xrightarrow[\text{reduce}]{\text{row}} \begin{bmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 5/2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

From the reduced matrix, we see that k_3 is a free variable and will be assigned a parameter in the general solution. The necessity of a parameter implies nontrivial solutions, so these vectors are *linearly dependent.*

The reduced matrix can also be used to tell us exactly how these vectors are linearly dependent. Since k_3 is free, we obtain a solution to the system for every possible value we assign to it. To get a simple nontrivial solution, let's set $k_3 = 1$. Then solving the equations represented by the nonzero rows of the reduced matrix gives us $k_1 = -1/2$ and $k_2 = -5/2$. Putting these back into the vector equation from when we first set up the Test for Linear Dependence/Independence, we get

$$\left(-\frac{1}{2}\right)\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \left(-\frac{5}{2}\right)\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} + \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\implies \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \frac{5}{2}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

From this we see exactly how one of the vectors in our collection can be expressed as a linear combination of others in the collection.

□

Example 17.4.4 Testing dependence/independence in $\mathrm{P}_n(\mathbb{R})$**.** Consider the polynomials from Discovery 17.4.c. Are they linearly dependent or independent? Set up the test, using the zero polynomial as the zero vector on the right-hand side:

$$k_1(1+x) + k_2(1+x^2) + k_3(2-x+3x^2) = 0.$$

As usual, we simplify the linear combination on the left-hand side into one vector. Here, this means collecting like terms.

$$(k_1 + k_2 + 2k_3) + (k_1 - k_3)x + (k_2 + 3k_3)x^2 = 0.$$

The polynomial on the left can only be equal to the zero polynomial if all its coefficients are zero, leading to the following system of equations:

$$\left\{\begin{array}{rcrcrcl} k_1 & + & k_2 & + & 2k_3 & = & 0, \\ k_1 & & & - & k_3 & = & 0, \\ & & k_2 & + & 3k_3 & = & 0. \end{array}\right.$$

Once again, we reduce the coefficient matrix to determine if there are nontrivial solutions:

$$\begin{bmatrix} 1 & 1 & 2 \\ 1 & 0 & -1 \\ 0 & 1 & 3 \end{bmatrix} \xrightarrow[\text{reduce}]{\text{row}} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix}.$$

Compare. The columns of the initial coefficient matrix with the vectors being tested.

Since variable k_3 is free, there exist nontrivial solutions and the vectors are *linearly dependent*. □

Example 17.4.5 Testing dependence/independence in $F(D)$. Let's do an example in a function space. Consider vectors $f(x) = x$, $g(x) = \sin(\pi x/2)$ and $h(x) = \cos(\pi x/2)$ in $F(\mathbb{R})$, the space of functions defined on the whole real number line. Are these functions linearly dependent or independent? Let's start the test by setting up the vector equation

$$k_1 x + k_2 \sin(\pi x/2) + k_3 \cos(\pi x/2) = 0.$$

Here, there is no algebraic way we can simplify the expression on the left-hand side. However, remember that the 0 on the right-hand side represents the zero *function*, and that functions are only equal when they always produce the same output given the same input (Definition 15.5.1). So let's try substituting some input x-values into the functions on either side of our vector equation above:

$$\begin{aligned} x=0: &\qquad k_1\cdot 0 + k_2\cdot 0 + k_3\cdot 1 = 0,\\ x=1: &\qquad k_1\cdot 1 + k_2\cdot 1 + k_3\cdot 0 = 0,\\ x=2: &\qquad k_1\cdot 2 + k_2\cdot 0 + k_3\cdot(-1) = 0. \end{aligned}$$

From the first equation we see $k_3 = 0$. Combining this with the third equation we also get $k_1 = 0$. Then combining that with the second equation we finally get $k_2 = 0$. Since only the trivial solution is possible, these vectors are *linearly independent*. □

17.4.2 Linear independence of "standard" spanning sets

Finally, let's check the "standard" spanning sets of our favourite example vector spaces.

Example 17.4.6 Independence of the standard basis vectors in $\mathbb{R}^n$. The standard basis vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ form a spanning set for $\mathbb{R}^n$, and they are also linearly independent, as we see if we apply the test:

$$k_1\mathbf{e}_1 + k_2\mathbf{e}_1 + \cdots + k_n\mathbf{e}_n = \mathbf{0} \qquad \Longrightarrow \qquad (k_1, k_2, \dots, k_n) = (0, 0, \dots, 0).$$

So clearly each scalar k_j must be zero, which means there is only the trivial solution. □

Example 17.4.7 Independence of the standard spanning vectors in $\mathrm{M}_{m\times n}(\mathbb{R})$. In Remark 16.4.16, we noted that there is also a "standard" set of spanning vectors in $\mathrm{M}_{m\times n}(\mathbb{R})$, consisting of those matrices that have all zero entries except for a single 1 in one specific entry. We might call these "standard basis vectors" for $\mathrm{M}_{m\times n}(\mathbb{R})$. Write E_{ij} for the matrix of this type with a 1 in the $(i,j)^{\text{th}}$ entry. These spanning vectors are also linearly independent. Here, when we apply the Test for Linear Dependence/Independence, it is best if we enumerate our scalars with the same scheme as the vectors:

$$k_{11}E_{11} + k_{12}E_{12} + \cdots + k_{mn}E_{mn} = \mathbf{0} \qquad \Longrightarrow \qquad [k_{ij}] = \mathbf{0}.$$

Again, we immediately see that only the trivial solution is possible. □

Example 17.4.8 Independence of the standard spanning vectors in $\mathrm{P}_n(\mathbb{R})$. Also in Remark 16.4.16, we noted that the powers of x (along with the constant polynomial 1) form a spanning set for $\mathrm{P}_n(\mathbb{R})$. We might call these "standard basis vectors” for $\mathrm{P}_n(\mathbb{R})$. Are they linearly independent? Apply the Test:

$$k_0 \cdot 1 + k_1 x + k_2 x^2 + \cdots + k_n x^n = 0.$$

If any of these coefficients are nonzero, the polynomial on the left-hand side will be nonzero, so only the trivial solution is possible. Therefore, powers of x are always linearly independent in $\mathrm{P}_n(\mathbb{R})$. □

17.5 Theory

In this section.

17.5.1 Basic facts about linear dependence and independence

First we’ll formally record our test, but we will let our discussion in Subsection 17.3.2 serve as a proof.

Proposition 17.5.1 Test for Linear Dependence/Independence. *Vectors* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ *are linearly dependent if the vector equation*

$$k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + \cdots + k_m\mathbf{v}_m = \mathbf{0}$$

has a nontrivial solution in the (scalar) variables $k_1, k_2, \ldots, k_m$*. Otherwise, if this vector equation has* only *the trivial solution* $k_1 = 0, k_2 = 0, \ldots, k_m = 0$*, then the vectors* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ *are linearly independent.*

We will further explore the connection between linear independence and spanning sets in the next subsection below, but for now recall that we introduced these new concepts to help us determine when a spanning set could be reduced. The next statement reflects the fact that the zero vector does not help span anything other than itself, so it is not useful as a member of a spanning set.

Proposition 17.5.2 Zero is linearly dependent. *Any set of vectors that contains the zero vector is linearly dependent.*

Proof. Suppose S is a set of vectors containing the zero vector. We’ll break into cases depending on what *else* is in S besides $\mathbf{0}$.

S consists of only *the zero vector.* Then S is linearly dependent by definition.

S contains at least one *nonzero vector* $\mathbf{v}$. But then $\mathbf{0}$ can be expressed as the linear combination $\mathbf{0} = 0\mathbf{v}$. Since we have found a vector in S that can be expressed as a linear combination of another vector in S, the set of vectors is linearly dependent.

Note. This does not violate Proposition 17.5.1, since in the equality $\mathbf{0} = 0\mathbf{v}$, vector $\mathbf{0}$ is acting as a vector in S. This equality is equivalent to $1 \cdot \mathbf{0} + 0 \cdot \mathbf{v} = \mathbf{0}$, which is an equality of a *non*trivial linear combination of vectors from S and the zero vector, as required by Proposition 17.5.1.

■

Here are some facts about how linear dependence and independence behave when enlarging/reducing collections of vectors.

Proposition 17.5.3 Dependence/independence versus subcollections.

1. *A collection of vectors that contains a subcollection that is linearly dependent is itself linearly dependent.*
2. *In a linearly independent collection of vectors, every subcollection is also linearly independent.*

Proof of Statement 1. Suppose S is a collection of vectors in a vector space, and S' is a linearly dependent subcollection of S. Then some vector in S' can be expressed as a linear combination of other vectors in S'. But because S' is a subcollection of S, all these vectors in S' are also vectors in S. So we can also say that some vector in S can be expressed as a linear combination of other vectors in S, making S a linearly dependent set. ■

Proof of Statement 2. Suppose S is a linearly independent collection of vectors in a vector space. Then no subcollection of S can be linearly dependent, because if S contained such a linearly dependent subcollection then Statement 1 of this proposition would imply that S itself is linearly dependent, which we assume it is not. So every subcollection of S must be linearly independent. ■

17.5.2 Linear dependence and independence of spanning sets

First we'll record our observation about preserving spans when reducing spanning sets. Then, in the following proposition, we'll take this idea to its logical conclusion.

Lemma 17.5.4 Dependent spanning sets can be reduced. *Suppose S is a set of vectors in a vector space and $\mathbf{w}$ is both a vector in S and expressible as a linear combination of vectors in S besides itself. Then* $\operatorname{Span} S = \operatorname{Span} S'$*, where S' is the one-smaller set of vectors obtained by removing $\mathbf{w}$ from S.*

Proof. Using Statement 2 of Proposition 16.5.6, we just need to show that every vector in S can be expressed as a linear combination of vectors in S', and vice versa. However, S and S' are the same set of vectors *except* that S contains $\mathbf{w}$ while S' does not. So from the trivial expression $\mathbf{v} = 1\mathbf{v}$, we immediately have that every vector $\mathbf{v}$ in S (other than $\mathbf{w}$) is a linear combination of itself (which is a vector in S'), and vice versa. And we have also assumed that $\mathbf{w}$ is expressible as a linear combination of vectors in S besides itself. Since the vectors in such a linear combination are also in S', we know that $\mathbf{w}$ is expressible as a linear combination of vectors in S'. ■

Proposition 17.5.5 Fully reducing finite spanning sets. *Every finite spanning set can be reduced to a linearly independent spanning set. That is, if S is a spanning set for a vector space and contains a finite number of vectors, then some subcollection of vectors in S will both span that vector space and be linearly independent.*

Clarification. *In this proposition, we consider the hypothetical "can be reduced" to allow the possibility of* not *reducing the set at all, in case the initial spanning set is already linearly independent.*

Proof. If S is already linearly independent, then we have our desired linearly independent spanning set. Otherwise, there is some vector $\mathbf{w}$ in S that is a linear combination of the other vectors in S. If we set S' to be the subcollection of S

consisting of every vector *except* $\mathbf{w}$, then from Lemma 17.5.4 we know that S' will remain a spanning set for the vector space. If S' is linearly independent, then we have our desired linearly independent spanning set. Otherwise, we can continue removing linearly dependent vectors in this way until we end up with a linearly independent spanning set. And since we assumed there were a finite number of vectors in S, this one-by-one removal process must indeed come to an end at some point. ■

Just as a vector that points up out of a plane in $\mathbb{R}^3$ must be linearly independent from vectors parallel to the plane, in any vector space we can enlarge a linearly independent set by including new vectors that are not linear combinations of the old. The next statement encapsulates this idea, and will help us in the next chapter to develop a "bottom-up" approach to building an optimal spanning set for a vector space, as opposed to the "top-down" approach made possible by Lemma 17.5.4 and Proposition 17.5.5.

Proposition 17.5.6 Enlarging independent sets. *If S is a linearly independent set of vectors and vector $\mathbf{v}$ is* not *in* Span$S$*, then the set of vectors containing both $\mathbf{v}$ and every vector in S is still linearly independent.*

Proof. Write S' for the set of vectors containing both the vector $\mathbf{v}$ and every vector in S. The set S' will be linearly independent if none of its vectors can be expressed as a linear combination of other vectors in the set.

So suppose $\mathbf{w}$ is a vector in S'. There are two cases to consider.

Case $\mathbf{w} = \mathbf{v}$. In this case, we already know that $\mathbf{w}$ cannot be a linear combination of other vectors in S', because the other vectors in S' are the vectors in S, and we assumed that $\mathbf{v}$ is not in Span$S$.

Case $\mathbf{w} \neq \mathbf{v}$. In this case, $\mathbf{w}$ is in the set S. Since S is assumed to be linearly independent, we know that $\mathbf{w}$ cannot be a linear combination of other vectors from just S. Could it be a linear combination involving other vectors in S *and* $\mathbf{v}$? Suppose we had

$$\mathbf{w} = k_1\mathbf{u}_1 + k_2\mathbf{u}_2 + \cdots + k_m\mathbf{u}_m + k\mathbf{v},$$

for some vectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m$ in S and scalars $k_1, k_2, \dots, k_m, k$ (assuming $k \neq 0$ so that $\mathbf{v}$ is indeed involved in the linear combination). But then we could isolate $\mathbf{v}$ as

$$\mathbf{v} = \frac{1}{k}\mathbf{w} - \frac{k_1}{k}\mathbf{u}_1 - \frac{k_2}{k}\mathbf{u}_2 - \cdots - \frac{k_m}{k}\mathbf{u}_m,$$

a linear combination of vectors in S, which is not possible because we have assumed that $\mathbf{v}$ is not in Span$S$. ■

The final fact below records our observation in Discovery 17.5 and Subsection 17.3.4 that after a certain number, a collection of vectors can be too numerous to be linearly independent.

Lemma 17.5.7 Too-large sets must be dependent. *If a vector space can be spanned by n vectors, then every collection containing* more *than n vectors must be linearly dependent.*

Proof. Suppose S is a set of n vectors in a vector space so that S is a spanning set for the space. By Proposition 17.5.5, there are vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{n'}$ in S that are both linearly independent and also a spanning set for the vectors space. Since this is a subcollection of S, we must have $n' \leq n$. We'll refer to this subcollection of S as S'.

Now further suppose we have a collection of vectors $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m$ in the vector space, with $m > n$. Since S' is a spanning set, we can express each $\mathbf{w}_i$ as a linear combination of the vectors in S:

$$\begin{aligned}
\mathbf{w}_1 &= a_{11}\mathbf{v}_1 + a_{21}\mathbf{v}_2 + \cdots + a_{n'1}\mathbf{v}_{n'},\\
\mathbf{w}_2 &= a_{12}\mathbf{v}_1 + a_{22}\mathbf{v}_2 + \cdots + a_{n'2}\mathbf{v}_{n'},\\
&\vdots\\
\mathbf{w}_m &= a_{1m}\mathbf{v}_1 + a_{2m}\mathbf{v}_2 + \cdots + a_{n'm}\mathbf{v}_{n'}.
\end{aligned}$$

Let's apply the Test for Linear Dependence/Independence to $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m$: suppose there are scalars $k_1, k_2, \ldots, k_m$ so that

$$k_1\mathbf{w}_1 + k_2\mathbf{w}_2 + \cdots + k_m\mathbf{w}_m = \mathbf{0}.$$

If we substitute in the above expressions for each $\mathbf{w}_i$ in terms of the vectors in S' and collect like terms, we get

$$\begin{gathered}
(a_{11}k_1 + a_{12}k_2 + \cdots + a_{1m}k_m)\mathbf{v}_1 + (a_{21}k_1 + a_{22}k_2 + \cdots + a_{2m}k_m)\mathbf{v}_2\\
+ \cdots + (a_{n'1}k_1 + a_{n'2}k_2 + \cdots + a_{n'm}k_m)\mathbf{v}_{n'} = \mathbf{0}.
\end{gathered}$$

If we set c_1 to be the coefficient expression on $\mathbf{v}_1$ in the expression above, and c_2 to be the coefficient expression on $\mathbf{v}_2$, and so on, then we obtain a vector equality

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_{n'}\mathbf{v}_{n'} = \mathbf{0}.$$

But the vectors in this linear combination are the vectors in S', and we have assumed that S' is a linearly independent set. So this vector equality can only be true for the trivial solution where each $c_i = 0$. This leads to homogeneous system

$$\left\{\begin{array}{ccccccccc}
a_{11}k_1 & + & a_{12}k_2 & + & \cdots & + & a_{1m}k_m & = & 0,\\
a_{21}k_1 & + & a_{22}k_2 & + & \cdots & + & a_{2m}k_m & = & 0,\\
 & & & & \vdots & & & & \\
a_{n'1}k_1 & + & a_{n'2}k_2 & + & \cdots & + & a_{n'm}k_m & = & 0,
\end{array}\right.$$

in the variables $k_1, k_2, \ldots, k_m$. Now, we have assumed $m > n \geq n'$, so we have more variables than equations in the homogeneous system above. But then the solution will require parameters, which means there are nontrivial solutions. Thus, the Test for Linear Dependence/Independence tells us that the collection $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m$ is linearly dependent. ■

Chapter 18

Basis and Coordinates

18.1 Discovery guide

Suppose V is a vector space and S is a finite spanning set for V (i.e. $V = \operatorname{Span} S$). In the previous chapter, we saw that if S is linearly dependent, then (at least) one vector can be removed from S, and the resulting smaller set will still be a spanning set. You can imagine repeating this process until finally you are left with a spanning set that is linearly independent.

See. Lemma 17.5.4 and Proposition 17.5.5.

This leads to the following definition.

basis for a vector space
a linearly independent spanning set for the space

Discovery 18.1 In each of the following, determine whether S is a basis for V. If it is not a basis, make sure you know which property S violates, independence or spanning.

Note. A specific example could violate both, but we only need to know it violates one of the two properties to know it's not a basis.

(a) $V = \mathbb{R}^3$, $S = \{(1,0,0),(1,1,0),(1,1,1),(0,0,2)\}$.

(b) $V = \mathbb{R}^3$, $S = \{(1,0,0),(1,1,0),(0,0,2)\}$.

(c) $V = \mathrm{M}_2(\mathbb{R})$, $S = \left\{ \begin{bmatrix} 2 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \right\}$.

(d) $V =$ the space of 2×2 upper triangular matrices,
$S = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix} \right\}$.

(e) $V =$ the space of 3×3 lower triangular matrices,
$S = \left\{ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right\}$.

(f) $V = \mathrm{P}_3(\mathbb{R})$, the space of all polynomials of degree 3 or less, $S = \{1, x, x^2\}$.

(g) $V = \mathrm{P}_3(\mathbb{R})$, $S = \{1, x, x^2, x^3\}$.

As discussed in the introduction to this discovery guide above, a spanning set that is not linearly independent contains redundant information in the form of vectors that are not actually needed to form a spanning set. This redundancy manifests itself in other ways, as the next discovery activity will demonstrate.

Discovery 18.2 Consider the set $S = \{(1,0),(1,1),(1,-1)\}$ of vectors in $\mathbb{R}^2$. This set spans $\mathbb{R}^2$ but is not linearly independent.

(a) Since S spans $\mathbb{R}^2$, it is possible to express vector $(3,-3)$ as a linear combination of the vectors in S.

Demonstrate a way to do this:

$$(3,-3) = \boxed{}(1,0) + \boxed{}(1,1) + \boxed{}(1,-1).$$

(b) Here is the redundant part. Demonstrate a *different* way to express $(3,-3)$ as a linear combination of the vectors in S:

$$(3,-3) = \boxed{}(1,0) + \boxed{}(1,1) + \boxed{}(1,-1).$$

(c) How many different ways do you think there are to do this?

The next discovery activity will demonstrate that the redundancies of Discovery 18.2 cannot happen for a *basis*.

Discovery 18.3 Suppose V is a vector space, $S = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is a basis for V, and $\mathbf{w}$ is a vector in V.

Since S is a spanning set, there is a way to express $\mathbf{w}$ as linear combinations of the vectors in S:

$$\mathbf{w} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + a_3\mathbf{v}_3.$$

Suppose there were a *different* such expression:

$$\mathbf{w} = b_1\mathbf{v}_1 + b_2\mathbf{v}_2 + b_3\mathbf{v}_3.$$

Use the vector identity

$$\mathbf{w} - \mathbf{w} = \mathbf{0}$$

and the two different expressions for $\mathbf{w}$ above to show that having these two different expressions violates the linear independence of S.

Discovery 18.3 shows that when we have a basis $S = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ for a vector space V, each vector in V has *one unique* expression as a linear combination of the vectors in S. For $\mathbf{w} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n$, the coefficients $c_1, c_2, \dots, c_n$ are called the **coordinates of w relative to** S. Since these coordinates consist of n coefficients, we sometimes relate $\mathbf{w}$ to a vector in $\mathbb{R}^n$ by collecting its coordinates into an n-tuple:

$$(\mathbf{w})_S = (c_1, c_2, \dots, c_n).$$

This is called the **coordinate vector of w relative to** S.

Discovery 18.4 In each of the following, determine the coordinate vector of $\mathbf{w}$ relative to the provided basis S for V.

(a) $V = \mathrm{M}_2(\mathbb{R})$, $S = \left\{ \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right] \right\}$, $\mathbf{w} = \left[\begin{smallmatrix} -1 & 5 \\ 3 & -2 \end{smallmatrix}\right]$.

(b) $V = \mathrm{M}_2(\mathbb{R})$, $S = \left\{ \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 & 0 \\ 1 & 1 \end{smallmatrix}\right] \right\}$, $\mathbf{w} = \left[\begin{smallmatrix} -1 & 5 \\ 3 & -2 \end{smallmatrix}\right]$.

(c) $V = \mathrm{P}_3(\mathbb{R})$, $S = \{1, x, x^2, x^3\}$, $\mathbf{w} = 3 + 4x - 5x^3$.

(d) $V = \mathbb{R}^3$, $S = \{(-1,0,1),(0,2,0),(1,1,0)\}$, $\mathbf{w} = (1,1,1)$.

(e) $V = \mathbb{R}^3$, $S = \{(1,0,0),(0,1,0),(0,0,1)\}$, $\mathbf{w} = (-2,3,-5)$.

Discovery 18.5 In each of the following, determine which vector $\mathbf{w}$ in V has the given coordinate vector $(\mathbf{w})_S$.

(a) $V = \mathrm{M}_2(\mathbb{R})$, $S = \left\{ \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right] \right\}$, $(\mathbf{w})_S = (3,-5,1,1)$.

(b) $V = \mathrm{M}_2(\mathbb{R})$, $S = \left\{ \left[\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right], \left[\begin{smallmatrix}1&1\\0&0\end{smallmatrix}\right], \left[\begin{smallmatrix}0&0\\1&0\end{smallmatrix}\right], \left[\begin{smallmatrix}0&0\\1&1\end{smallmatrix}\right] \right\}$, $(\mathbf{w})_S = (3,-5,1,1)$.

(c) $V = P_3$, $S = \{1, x, x^2, x^3\}$, $(\mathbf{w})_S = (-3,1,0,3)$.

(d) $V = \mathbb{R}^3$, $S = \{(-1,0,1),(0,2,0),(1,1,0)\}$, $(\mathbf{w})_S = (1,1,1)$.

(e) $V = \mathbb{R}^3$, $S = \{(1,0,0),(0,1,0),(0,0,1)\}$, $(\mathbf{w})_S = (-2,3,-5)$.

Discovery 18.6 Coordinate vectors let us transfer vector algebra in a space V to the familiar space $\mathbb{R}^n$.

For example, consider the basis

$$S = \left\{ \begin{bmatrix}1&0\\0&0\end{bmatrix}, \begin{bmatrix}1&1\\0&0\end{bmatrix}, \begin{bmatrix}0&0\\1&0\end{bmatrix}, \begin{bmatrix}0&0\\1&1\end{bmatrix} \right\}$$

for the space $\mathrm{M}_2(\mathbb{R})$ from Task 18.5.b.

(a) In Task 18.5.b you have already determined the vector $\mathbf{w}$ in $\mathrm{M}_2(\mathbb{R})$ that has coordinate vector $(\mathbf{w})_S = (3,-5,1,1)$. Now do the same to determine the vector $\mathbf{v}$ in $\mathrm{M}_2(\mathbb{R})$ that has coordinate vector $(\mathbf{v})_S = (-1,2,0,3)$.

(b) *Do some algebra in* $\mathrm{M}_2(\mathbb{R})$*:*

Using your vectors $\mathbf{v}$ from Task a, and $\mathbf{w}$ from Task 18.5.b compute the linear combination $2\mathbf{v} + \mathbf{w}$.

Note: Vectors $\mathbf{v}$ and $\mathbf{w}$ “live” in the space $\mathrm{M}_2(\mathbb{R})$, so your computation in this task should involve 2×2 matrices, and should also result in a 2×2 matrix.

(c) *Do the same algebra in* $\mathbb{R}^4$*:*

Compute $2(\mathbf{v})_S + (\mathbf{w})_S$, using the coordinate vectors $(\mathbf{v})_S$ and $(\mathbf{w})_S$ provided to you in Task 18.6.a.

Note: These coordinate vectors “live” in the space $\mathbb{R}^4$, so your computation in this task should involve four-dimensional vectors, and should also result in a four-dimensional vector.

(d) *Compare your results:*

Consider your four-dimensional result vector from Task c as a coordinate vector for some vector in $\mathrm{M}_2(\mathbb{R})$ relative to S. Similarly to your computations in Task 18.5.b and Task a, determine the matrix in $\mathrm{M}_2(\mathbb{R})$ that has coordinate vector equal to your result vector from Task c. Then compare with your result matrix from Task 18.6.b. Surprised?

18.2 Terminology and notation

basis for a vector space
a linearly independent spanning set

ordered basis
a basis where the basis vectors are always written in a particular order, and linear combinations of the basis vectors are always expressed in that order

coordinates of a vector $\mathbf{w}$ relative to a basis $\mathcal{B} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$
the unique set of scalars $c_1, c_2, \ldots, c_n$ so that $\mathbf{w} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n$

coordinate vector associated to a vector $\mathbf{w}$ relative to a basis $\mathcal{B}$
the vector $(c_1, c_2, \ldots, c_n)$ in $\mathbb{R}^n$ formed by the coordinates of $\mathbf{w}$ relative to $\mathcal{B}$

$(\mathbf{w})_{\mathcal{B}}$ notation to mean the coordinate vector $(c_1, c_2, \ldots, c_n)$ in $\mathbb{R}^n$ for the vector $\mathbf{w}$, relative to the basis $\mathcal{B}$ for the vector space that contains $\mathbf{w}$

$[\mathbf{w}]_{\mathcal{B}}$ notation to mean the coordinate vector in $\mathbb{R}^n$ for the vector $\mathbf{w}$, relative to the basis $\mathcal{B}$ for the vector space that contains $\mathbf{w}$, *realized as a column vector* (i.e. as a column matrix)

18.3 Concepts

In this section.

18.3.1 Basis as a minimal spanning set

The purpose of a spanning set for a vector space is to be able to describe *every* vector in the space systematically in terms of linear combinations of certain specific vectors. But to be able to do this as simply as possible, we would like our spanning set to be "optimal" for this purpose. We have seen that spanning sets can contain redundant information — when a spanning set is linearly dependent, then one of its vectors can be expressed as a linear combination of others, and so that particular vector is not needed for the purpose of describing *every* vector in the vector space in terms of linear combinations of spanning vectors. Even worse, we saw in Discovery 18.2 that a linearly dependent spanning set allows for other types of redundancy. In particular, if a spanning set is linearly dependent, then every vector in the vector space will have an *infinite* number of different descriptions as linear combinations of spanning vectors. Clearly such a situation is not "optimal."

However, Lemma 17.5.4 and Proposition 17.5.5 tell us that we can remove this redundancy while still keeping a spanning set. By eliminating linearly dependent vectors from a spanning set one at a time, we can eventually reduce to a linearly independent spanning set — a **basis** for the space. As we saw in Discovery 18.3, a basis will no longer exhibit the second kind of redundancy discussed above, so that in terms of a basis, every vector in the space has *one unique* description as a linear combination (where we do not consider reorderings of the linear combination expression, or insertion or removal of basis vectors with a zero coefficient, as different expressions). And since a basis is linearly independent, it seems that it cannot contain any of the first kind of redundancy discussed above, because none of its vectors can be expressed as a linear combination of others. So it would be reasonable to guess that a basis is *minimal* in the sense that it cannot be reduced any further while still remaining a spanning set. And this is exactly true, as we will see in Statement 1.a of Theorem 18.5.2 in Subsection 18.5.2.

18.3.2 Basis as a maximal linearly independent set

As above, a spanning set that is not linearly independent can be reduced to one that is, making it a basis, and a basis cannot be reduced any further while still remaining a spanning set. But perhaps we can also work the other way. A linearly independent set that does not span the whole vector space can be enlarged using Proposition 17.5.6; perhaps we could continue to enlarge the set until it *does* span the whole vector space, at which point it would become a basis.

A look ahead. We will pursue this idea of enlarging a linearly independent set to a basis further in Chapter 19.

But we know from Lemma 17.5.7 that a collection of vectors that is larger than a known (finite) spanning set must be linearly dependent. Since a basis is, by definition, a special kind of spanning set, a basis is also *maximal* in the sense that it cannot be enlarged any further while still remaining linearly independent.

18.3.3 Basis is not unique

It is important to note that a vector space will not have just *one* basis. In fact, except for the trivial vector space, every vector space has an *infinite* number of different possible bases. But often spaces have an obvious, preferred basis, called the **standard basis** for the space. We will see examples of standard bases for various spaces in Subsection 18.4.2.

18.3.4 Ordered versus unordered basis

In mathematics, usually a collection or set of objects is considered to be **unordered** — all that matters is the inclusion of the members of the collection, not the order in which those members are written down. For example, if $V = \operatorname{Span}\{\mathbf{u},\mathbf{v},\mathbf{w}\}$ for some collection of vectors $\mathbf{u},\mathbf{v},\mathbf{w}$ in V, saying that $\{\mathbf{w},\mathbf{v},\mathbf{u}\}$ is a spanning set for V is the same as saying that $\{\mathbf{u},\mathbf{v},\mathbf{w}\}$ is a spanning set for V. However, we usually prefer one uniform way to describe vectors in V as linear combinations of spanning vectors. It would be inconsistent to write

$$\mathbf{x} = a_1\mathbf{u} + a_2\mathbf{v} + a_3\mathbf{w}$$

for some vector $\mathbf{x}$ in V, and then to write

$$\mathbf{y} = b_1\mathbf{w} + b_2\mathbf{v} + b_3\mathbf{u}$$

for some other vector $\mathbf{y}$ in V. So usually we take a spanning set to have a particular order, and to always express linear combinations in that order, especially when our spanning set is a basis. To emphasize that a basis has such a preferred ordering, we might refer to it as an **ordered basis**. But you should assume from this point forward that every basis is an **ordered** one.

18.3.5 Coordinates of a vector

18.3.5.1 Basic concept of coordinates relative to a basis

Suppose we have a basis for a vector space. Since a basis is a spanning set, every vector in the space has a decomposition as a linear combination of these basis vectors. But, as we saw in Discovery 18.3, a vector in the space cannot have *more* than one such decomposition. That is, every vector $\mathbf{w}$ in the vector space has *one unique* expression as a linear combination of the basis vectors. Because of this, we can consider the coefficients that go into such an expression as a "signature" or "code" that uniquely identifies $\mathbf{w}$. For example, if $V = \operatorname{Span}\mathcal{B}$ for some basis $\mathcal{B} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$, and we have a vector $\mathbf{w}$ in V for which

$$\mathbf{w} = 3\mathbf{v}_1 + 5\mathbf{v}_2 + (-1)\mathbf{v}_3, \tag{$*$}$$

then the numbers $3, 5, -1$ (in that order) uniquely identify the vector $\mathbf{w}$ relative to the (ordered) basis, and no other triple of numbers identify $\mathbf{w}$. These coefficients are called the **coordinates** of $\mathbf{w}$ relative to the basis $\mathcal{B}$. Now, we already have a concept that collects together triples of numbers in particular orders — vectors in $\mathbb{R}^3$. So, in this example, to every vector in the space V (which may be a space of matrices or a space of functions or etc.) we can associate one unique vector in $\mathbb{R}^3$ whose components are the coordinates of the vector relative to the basis $\mathcal{B}$. For our example vector $\mathbf{w}$ above, we can collect the three coefficients from the linear combination in ($*$) either into a triple of coordinates (x, y, z) or into a column vector:

$$(\mathbf{w})_{\mathcal{B}} = (3, 5, -1), \qquad [\mathbf{w}]_{\mathcal{B}} = \begin{bmatrix} 3 \\ 5 \\ -1 \end{bmatrix}.$$

The equivalent $\mathbb{R}^3$-vectors $(\mathbf{w})_{\mathcal{B}}$ and $[\mathbf{w}]_{\mathcal{B}}$ are each called the **coordinate vector** of $\mathbf{w}$ relative to $\mathcal{B}$, the only difference between the two being the presentation.

To repeat, since $\mathcal{B}$ is a spanning set, every vector in the space can be expressed as a linear combination of the vectors in $\mathcal{B}$, and so every vector has an associated coordinate vector. And since $\mathcal{B}$ is linearly independent, it contains no redundancy as a spanning set, and so each vector can only have *one unique* coordinate vector associated to it. Which also means that every vector in $\mathbb{R}^3$ can be interpreted as a coordinate vector relative to $\mathcal{B}$, and can be traced to one particular vector in V.

In general, the number of coordinates required is the same as the number of vectors in the basis being used. So if $V = \operatorname{Span}\mathcal{B}$ for basis $\mathcal{B} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$, then the coordinate vector for each vector in V needs to be a vector in $\mathbb{R}^n$. See Subsection 18.4.3 for examples of computing coordinate vectors and of interpreting vectors in $\mathbb{R}^n$ as coordinate vectors.

Warning 18.3.1 Order matters in coordinate vectors. Because of Axiom A 2, reordering a linear combination of vectors does not produce a different vector as the end result. However, when extracting coefficients from a linear combination to form a coordinate vector, order definitely does matter, since we have decided to consider every basis as an **ordered basis**.

For example, if $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is a basis for a space V, then the vector $\mathbf{v} = \mathbf{u}_1 + 2\mathbf{u}_2 + 3\mathbf{u}_3$ has coordinate vector

$$(\mathbf{v})_{\mathcal{B}} = (1, 2, 3).$$

If we rearrange the linear combination to $\mathbf{v} = 2\mathbf{u}_2 + \mathbf{u}_1 + 3\mathbf{u}_3$, we are obviously not forming a different vector $\mathbf{v}$, but we *are* changing our *point of view* to a different **ordered basis**, $\mathcal{B}' = \{\mathbf{u}_2, \mathbf{u}_1, \mathbf{u}_3\}$, creating a different *coordinate* vector for $\mathbf{v}$:

$$(\mathbf{v})_{\mathcal{B}'} = (2,1,3).$$

18.3.5.2 Linearity of coordinates

In Discovery 18.6, we discovered that performing a computation $2\mathbf{v} + \mathbf{w}$ in a vector space V and performing the corresponding calculation $2(\mathbf{v})_{\mathcal{B}} + (\mathbf{w})_{\mathcal{B}}$ with the corresponding coordinate vectors in $\mathbb{R}^n$ relative to some basis $\mathcal{B}$ of V would essentially yield the same result. (That is, the result of combining coordinate vectors ends up being the coordinate vector of the result of combining the original vectors.)

To consider why this works out, let's consider the operations involved in a linear combination (vector addition and scalar multiplication) separately. For the remainder of this discussion, suppose $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ is a basis for a particular vector space V.

Addition of coordinate vectors. If you have two vectors in V expressed uniquely as linear combinations of the basis vectors,

$$\begin{array}{ccccccccc} \mathbf{v} & = & a_1\mathbf{u}_1 & + & a_2\mathbf{u}_2 & + & \dots & + & a_n\mathbf{u}_n, \\ \mathbf{w} & = & b_1\mathbf{u}_1 & + & b_2\mathbf{u}_2 & + & \dots & + & b_n\mathbf{u}_n, \end{array}$$

then adding the vectors can be accomplished by adding the linear combinations. Algebraically, we can add linear combinations by collecting like terms, and when we do so we will be adding the corresponding coefficients on each basis vector. But coefficients on basis vectors are where components of coordinate vectors come from, and so we can say that ***the coordinate vector of a sum is the sum of the coordinate vectors***.

Scalar multiplication of a coordinate vector. If you have a vectors in V expressed uniquely as linear combinations of the basis vectors,

$$\mathbf{v} = a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \ldots + a_n\mathbf{u}_n,$$

then multiplying this vector by a scalar can be accomplished by scalar multiplying the linear combination. Algebraically, we can scalar multiply a linear combination by distributing the scalar through the vector sum, and when we do so we will be multiplying the coefficient on each basis vector by the scalar. But coefficients on basis vectors are where components of coordinate vectors come from, and so we can say that ***the coordinate vector of a scalar multiple is the scalar multiple of the coordinate vector***.

18.4 Examples

In this section.

18.4.1 Checking a basis

Let's start by working through Discovery 18.1, where we were asked to determine whether a collection of vectors forms a basis for a vector space. In each case we are looking to check two properties: that the collection is **linearly independent**, and that it forms a **spanning set** for the whole vector space.

Example 18.4.1 A collection of vectors too large to be a basis. In Discovery 18.1.a, we considered a set S of four vectors in $V = \mathbb{R}^3$.

We already know that $\mathbb{R}^3$ can be spanned by the *three* standard basis vectors, and so Lemma 17.5.7 tells that any set of *more* than three vectors in $\mathbb{R}^3$ must be linearly dependent. Set S contains four vectors, so it can't be a basis because it is linearly dependent. However, S is a spanning set — can you see how? □

Example 18.4.2 A nonstandard basis for $\mathbb{R}^3$. In Discovery 18.1.b, we considered a set S of three vectors in $V = \mathbb{R}^3$.

This set S is linearly independent, which can be verified using the Test for Linear Dependence/Independence. As we saw in many examples in Section 17.4, the vector equation

$$k_1(1,0,0) + k_2(1,1,0) + k_3(0,0,2) = (0,0,0)$$

that we use to begin the Test for Linear Dependence/Independence leads to a homogeneous system. In this case, that system has coefficient matrix

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix},$$

where the vectors in S appear as columns. This matrix can be reduced to I in two operations, and so only the trivial solution is possible.

The set S is also a spanning set for V. To check this, we need to make sure that *every* vector in $\mathbb{R}^3$ can be expressed as a linear combination of the vectors in S. That is, we need to check that if (x,y,x) is an arbitrary vector in $\mathbb{R}^3$, then we can always determine scalars a,b,c so that

$$a(1,0,0) + b(1,1,0) + c(0,0,2) = (x,y,z).$$

Similar to the Test for Linear Dependence/Independence, the above vector equation leads to a system of equations with augmented matrix

$$\left[\begin{array}{ccc|c} 1 & 1 & 0 & x \\ 0 & 1 & 0 & y \\ 0 & 0 & 2 & z \end{array}\right].$$

The same two operations as before will reduce the coefficient part of this matrix to I, so that a solution always exists, regardless of the values of x,y,z. But it's also possible to determine a solution directly by inspection of the vector equation above, as clearly

$$(x-y)(1,0,0) + y(1,1,0) + \frac{z}{2}(0,0,2) = (x,y,z)$$

will be a solution.

Because this set is both linearly independent and a spanning set, it is a basis for the space. □

Example 18.4.3 An independent set that does not span. In Discovery 18.1.c, we considered a set S of three vectors in $V = \mathrm{M}_2(\mathbb{R})$.

This set S is linearly independent (check using the test!), but it is not a spanning set. We can see a linear combination of these vectors will never have a nonzero entry in the $(1,2)$ entry. In particular, the vector

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

is not in $\operatorname{Span} S$. Since S does not span the entire space, it is not a basis for V.

Note. We have determined that S is not a basis for the whole space V. However, since S is linearly independent, it *is* a basis for the *subspace* of V that it spans (i.e. the subspace $\operatorname{Span} S$).

□

Example 18.4.4 A set that neither spans nor is independent. In Discovery 18.1.d, we considered a set S of four vectors in the space V of all 2×2 upper triangular matrices.

This set of vectors is not a basis because it is *neither* a spanning set nor linearly independent.

It can't be a spanning set for the space V because a linear combination of these vectors will always have the same number in both diagonal entries. In particular, the vector

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

is upper triangular, so it is in V, but it is not in $\operatorname{Span} S$.

Also, we could use the test to determine that these vectors are linearly dependent, but we can see directly that one of these vectors is a linear combination of others:

$$\begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

□

Example 18.4.5 The standard basis for the space of 3×3 lower triangular matrices. In Discovery 18.1.e, we considered a set S of six vectors in the space V of all 3×3 lower triangular matrices.

We might call these matrices the "standard basis vectors" for the space of 3×3 lower triangular matrices, since when we simplify a linear combination of them, such as

$$\begin{aligned} & k_{11} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + k_{21} \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + k_{22} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\ & + k_{31} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} + k_{32} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} + k_{33} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad (18.4.1) \\ & = \begin{bmatrix} k_{11} & 0 & 0 \\ k_{21} & k_{22} & 0 \\ k_{31} & k_{32} & k_{33} \end{bmatrix}, \qquad (*) \end{aligned}$$

we see that the coefficients in the linear combination on the left correspond directly to the entries in the resulting sum matrix on the right, just as with other "standard" bases that we've encountered.

The set S is a spanning set for V, since we can clearly achieve every possible vector in this space using linear combinations of vectors in S by varying the coefficients in the general linear combination $(*)$ above.

The left-hand side of $(*)$ is also the left-hand side of the vector equation that we use in the Test for Linear Dependence/Independence, and from the right-hand side of $(*)$ we can see that if we set this linear combination to equal the zero vector (which is the 3×3 zero matrix here), the only solution is the trivial one.

Since S is both linearly independent and a spanning set, it is a basis for V. □

Example 18.4.6 Another independent set that does not span. In Discovery 18.1.f, we considered a set S of three vectors in the space $V = \mathrm{P}_3(\mathbb{R})$.

We have already seen in Subsection 17.4.2 that powers of x are always linearly independent in a space of polynomials. But this set of polynomials cannot be a spanning set for $\mathrm{P}_3(\mathbb{R})$ because no linear combination of $1, x, x^2$ will ever produce a polynomial of degree 3. So S is not a basis. □

Example 18.4.7 The standard basis for $\mathrm{P}_3(\mathbb{R})$. In Discovery 18.1.g, we considered a set S of four vectors in the space $V = \mathrm{P}_3(\mathbb{R})$.

Again, we know that powers of x are linearly independent in a space of polynomials. However, this time S *is* also a spanning set, since we naturally write polynomials of degree 3 as linear combinations of powers of x:

$$a_0 \cdot 1 + a_1 x + a_2 x^2 + a_3 x^3.$$

Such linear combinations can also be used to produce polynomials of degree less than 3, by setting the coefficients on the higher powers to 0. Since S is both independent and a spanning set, it is a basis for $\mathrm{P}_3(\mathbb{R})$. □

Remark 18.4.8 After we study the concept of **dimension** in the next chapter, the process of determining whether a set of vectors is a basis will become simpler. It is fairly straightforward to check the linear independence condition, since this usually reduces to solving a homogeneous system of linear equations, but checking the spanning condition directly is more tedious. In Chapter 19, we will see that if we know the correct *number* of vectors required in a basis, we only need to check *one* of the two conditions in the definition of **basis** (Corollary 19.5.6). And, as mentioned, usually it is the linear independence condition that is easier to verify.

18.4.2 Standard bases

In Subsection 17.4.2, we checked that certain "standard" spanning sets for our main examples of vector spaces were also linearly independent. Since they both span and are linearly independent, that makes each of them a basis for the space that contains them. We'll list them again here.

Terminology. In particular, verifying that these "standard" spanning sets are in fact bases will justify our use of the phrase **standard basis** to describe some of them in previous chapters.

Example 18.4.9 The standard basis of $\mathbb{R}^n$. The standard basis vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ form a basis for $\mathbb{R}^n$, justifying the word "basis" in our description "standard basis vectors" for these vectors. □

Example 18.4.10 The standard basis of $\mathrm{M}_{m\times n}(\mathbb{R})$. The space $\mathrm{M}_{m\times n}(\mathbb{R})$ of $m \times n$ matrices also has a standard basis: the collection of matrices E_{ij} that have all entries equal to 0 except for a single 1 in the $(i,j)^{\text{th}}$ entry. □

Example 18.4.11 The two standard bases of $\mathrm{P}_n(\mathbb{R})$. A space of polynomials $\mathrm{P}_n(\mathbb{R})$ also has a standard basis: the collection $1, x, x^2, x^3, \ldots, x^n$ of powers of x.

As an **ordered basis**, we have two reasonable choices here: the order already presented, and the reverse order $x^n, x^{n-1}, \ldots, x^2, x, 1$. We will stick with the order of increasing powers of x, so that when we index the coefficients in a linear combination, as in

$$a_0 \cdot 1 + a_1 x + a_2 x^2 + \ldots + a_n x^n,$$

then their indices are increasing with the exponents on x. □

18.4.3 Coordinate vectors

Finally, we'll do some computations with coordinate vectors, by working Discovery 18.4 and Discovery 18.5.

First, from Discovery 18.4.

Example 18.4.12 Determining a coordinate vector.

1. In Discovery 18.4.a, we considered a vector $\mathbf{w}$ in $\mathrm{M}_2(\mathbb{R})$ relative to the standard basis.

   First, decompose $\mathbf{w}$ as a linear composition of the vectors in S. Since S is the standard basis for $\mathrm{M}_2(\mathbb{R})$, this can be done by inspection:

   $$\begin{bmatrix} -1 & 5 \\ 3 & -2 \end{bmatrix} = (-1)\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + 5\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + 3\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + (-2)\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

   To get the coordinate vector, we wrap the four coefficients up (in order) in an $\mathbb{R}^4$ vector:

   $$(\mathbf{w})_S = (-1, 5, 3, -2).$$

2. In Discovery 18.4.b, we considered the same vector from $\mathrm{M}_2(\mathbb{R})$ as in the previous example, but relative to a nonstandard basis.

   We could probably also decompose $\mathbf{w}$ by inspection here, but instead we'll demonstrate the general method. Write $\mathbf{w}$ as an unknown linear combination of the basis vectors, and then simplify the linear combination:

   $$\begin{aligned} \begin{bmatrix} -1 & 5 \\ 3 & -2 \end{bmatrix} &= k_1\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + k_2\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} + k_3\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + k_4\begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} k_1 + k_2 & k_2 \\ k_3 + k_4 & k_4 \end{bmatrix}. \end{aligned}$$

   Comparing entries on left- and right-hand sides, we obtain a system of equations:

   $$\left\{ \begin{array}{rcrcr} k_1 & + & k_2 & = & -1, \\ & & k_2 & = & 5, \\ k_3 & + & k_4 & = & 3, \\ & & k_4 & = & -2. \end{array} \right.$$

   If we had more complicated basis vectors, we would have a more complicated system, which we could solve by forming an augmented matrix and row reducing. As it is, we can solve by inspection:

   $$k_1 = -6, \qquad k_2 = 5, \qquad k_3 = 5, \qquad k_4 = -2.$$

   We collect these four coefficients (in order) in an $\mathbb{R}^4$ vector:

   $$(\mathbf{w})_S = (-6, 5, 5, -2).$$

Even though we were working with the *same* vector as in the previous example, *we ended up with a different coordinate vector* because it is relative to a different basis.

3. In Discovery 18.4.c, we considered a vector $\mathbf{w}$ in $P_3(\mathbb{R})$ relative to the standard basis.

   The standard basis of $P_3(\mathbb{R})$ consists of powers of x (along with the constant polynomial 1), and our polynomial $\mathbf{w}$ is naturally written as a linear combination of powers of x. However, there is no x^2 term, so we need to insert one:
   $$\mathbf{w} = 3 \cdot 1 + 4x + 0x^2 - 5x^3.$$
   Once again, we wrap up these four coefficients (in order) in an $\mathbb{R}^4$ vector:
   $$(\mathbf{w})_S = (1, 4, 0, -5).$$

4. In Discovery 18.4.d, we considered a vector $\mathbf{w}$ in $\mathbb{R}^3$ relative to a nonstandard basis.

   Rather than try to guess, we should set up equations and solve. Start by writing $\mathbf{w}$ as an unknown combination of the basis vectors and combine into a single vector expression:
   $$\begin{aligned}(1,1,1) &= k_1(-1,0,1) + k_2(0,2,0) + k_3(1,1,0)\\ &= (-k_1 + k_3, 2k_2 + k_3, k_1).\end{aligned}$$
   This leads to a system of equations:
   $$\left\{\begin{array}{rcrcrcr} -k_1 & & & + & k_3 & = & 1,\\ & & 2k_2 & + & k_3 & = & 1,\\ k_1 & & & & & = & 1.\end{array}\right.$$
   We could probably solve by inspection again, but let's form an augmented matrix and reduce:
   $$\left[\begin{array}{rrr|r} -1 & 0 & 1 & 1\\ 0 & 2 & 1 & 1\\ 1 & 0 & 0 & 1\end{array}\right] \xrightarrow[\text{reduce}]{\text{row}} \left[\begin{array}{rrr|r} 1 & 0 & 0 & 1\\ 0 & 1 & 0 & -1/2\\ 0 & 0 & 1 & 2\end{array}\right]$$
   Notice again how the columns in the initial augmented matrix, including the column of constants, are the vectors involved. The column of constants in the final reduced matrix is our coordinate vector:
   $$(\mathbf{w})_S = (1, -1/2, 2).$$

5. In Discovery 18.4.e, we considered a vector $\mathbf{w}$ in $\mathbb{R}^3$ relative to the standard basis.

   This is similar to the first example — we have the standard basis for $\mathbb{R}^3$, so it is simple to decompose $\mathbf{w}$ as a linear combination of the vectors in the basis:
   $$\mathbf{w} = -2\mathbf{e}_1 + 3\mathbf{e}_2 + (-5)\mathbf{e}_3.$$
   Collect these coefficients together into an $\mathbb{R}^3$ vector:
   $$(\mathbf{w})_S = (-2, 3, -5).$$

□

Remark 18.4.13 The last two parts of the example above might seem kind of weird, but the point is all about *point of view*. Relative to the *standard basis*, a vector in $\mathbb{R}^n$ is equal to its own coordinate vector. In other words, the standard basis is standard because it corresponds to the natural way that we think of vectors in $\mathbb{R}^3$ — in terms of its x-, y-, and z-coordinates. This is similar to how the standard basis for a polynomial space leads to coordinate vectors that just record the coefficients of polynomials, or how the standard basis for a matrix space leads to coordinate vectors that just record the entries of the matrices.

But if we *change* our point of view and use a nonstandard basis for $\mathbb{R}^n$, then coordinate vectors allow us to use vectors in $\mathbb{R}^n$ to represent other vectors in $\mathbb{R}^n$, where everything is "tuned" to the perspective of the nonstandard basis. And similarly if we use nonstandard bases in other spaces.

Now we'll work through Discovery 18.5. This activity is the same as the previous, but in reverse — we are given a coordinate vector from $\mathbb{R}^n$, and we can use its components as the coefficients in a linear combination of the basis vectors. We'll complete two of the examples from this discovery activity, and leave the rest to you.

Example 18.4.14 Determining a vector from its coordinate vector.

1. This is Discovery 18.5.a.

   Just compute the linear combination using the coordinate vector components as coefficients, in the proper order:

$$\begin{aligned}\mathbf{w} &= 3\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} + (-5)\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix} + 1\begin{bmatrix}0 & 0\\ 1 & 0\end{bmatrix} + 1\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}\\ &= \begin{bmatrix}3 & -5\\ 1 & 1\end{bmatrix}.\end{aligned}$$

   This result should not be surprising, as both a 2×2 matrix and a vector in $\mathbb{R}^4$ are just a collection of four numbers.

2. This is Discovery 18.5.b.

   Again, just compute the linear combination using the coordinate vector components as coefficients, in the proper order:

$$\begin{aligned}\mathbf{w} &= 3\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} + (-5)\begin{bmatrix}1 & 1\\ 0 & 0\end{bmatrix} + 1\begin{bmatrix}0 & 0\\ 1 & 0\end{bmatrix} + 1\begin{bmatrix}0 & 0\\ 1 & 1\end{bmatrix}\\ &= \begin{bmatrix}-2 & -5\\ 2 & 1\end{bmatrix}.\end{aligned}$$

   Even though we were working with the *same* coordinate vector as in the previous example, *we ended up with a different matrix result* because it is relative to a different basis.

3. This is Discovery 18.5.c.

   Use the same process here as in the previous two examples above:

$$\mathbf{w} = -3 \cdot 1 + 1x + 0x^2 + 3x^3 = -3 + x + 3x^3.$$

□

18.5 Theory

In this section.

- Subsection 18.5.1 *Reducing to a basis*
- Subsection 18.5.2 *Basis as optimal spanning set*

18.5.1 Reducing to a basis

First we will restate Proposition 17.5.5 in the language of our new concept of **basis**.

Proposition 18.5.1 *Every finite spanning set can be reduced to a basis. That is, if S is a spanning set for a vector space and contains a finite number of vectors, then some subcollection of vectors in S will be a basis for the vector space.*

Clarification. *Again, we consider the hypothetical "can be reduced" to allow the possibility of* not *reducing the spanning set at all, in case it is already a basis.*

Proof. Proposition 17.5.5 states that every finite spanning set can be reduced to a linearly independent spanning set. But that's exactly what a basis is — a linearly independent spanning set. ■

A look ahead. In the next chapter, we will also extend Proposition 17.5.6 to obtain a counterpart to the above proposition, where we *build up* to a basis instead of *reducing* to one: every linearly independent set can be extended to a basis. (See Proposition 19.5.4 in Subsection 19.5.2.)

18.5.2 Basis as optimal spanning set

The remaining facts establish that a basis is the answer to our quest for an *optimal* spanning set — no unnecessary spanning vectors, and no multiple ways of expressing vectors in the space as linear combinations of the spanning vectors.

Theorem 18.5.2 Basis is optimal.

1. (a) *A basis is a* minimal *spanning set, in the sense that no proper subcollection of vectors from the basis could still be a spanning set for the vector space.*

   (b) *A finite collection of vectors in a vector space that forms a minimal spanning set (in the same sense as in Statement 1.a) must be a basis for that space.*

2. (a) *A finite basis is a* maximal *linearly independent set, in the sense that it cannot be a proper subcollection of some linearly independent set of vectors.*

   (b) *A collection of vectors in a vector space that forms a maximal linearly independent set (in the same sense as in Statement 2.b) must be a basis for that space.*

Proof of Statement 1.a. Suppose $\mathcal{B}$ is a basis for a vector space. That is, suppose $\mathcal{B}$ is a linearly independent spanning set. If we remove even one vector $\mathbf{v}$ from $\mathcal{B}$, the remaining vectors cannot still form a spanning set for the space. Because if they could, then $\mathbf{v}$, being a vector in that vector space, could be expressed as a linear combination of some number of the remaining vectors in $\mathcal{B}$. In other words, some vector in $\mathcal{B}$ would be expressible as a linear combination of others, which would violate the assumption that $\mathcal{B}$ is linearly independent. ■

Proof of Statement 1.b. Suppose S is a spanning set for a vector space, and that S contains a finite number of vectors. Further suppose that S is *minimal* in the sense that no proper subcollection of S also forms a spanning set for the space. We would like to prove that this forces S to be a basis. We already assuming one-half of the definition of **basis**, so we only need to consider the other half: must S also be linearly independent? If it were linearly dependent instead, then it could be reduced to subcollection that forms a linearly independent spanning set (Proposition 17.5.5). But S doesn't have any subcollections that form spanning sets for the vector space, let alone any linearly independent ones. So S cannot be linearly dependent, forcing it to be linearly independent, as required. ■

Proof of Statement 2.a. Suppose $\mathcal{B}$ is a basis for a vector space, and that $\mathcal{B}$ contains a finite number of vectors. Then by definition of **basis**, $\mathcal{B}$ is a spanning set for the vector space, and so any collection of vectors that contains more vectors than $\mathcal{B}$ must be linearly dependent (Lemma 17.5.7). In particular, if some collection of vectors contains $\mathcal{B}$ as a proper subcollection, then that larger collection must be linearly dependent. ■

Proof of Statement 2.b. Suppose S is a linearly independent collection of vectors in a vector space, and that S is maximal in the sense that no other linearly independent collection of vectors can contain S as a proper subcollection. We would like to prove that this forces S to be a basis for the space. We already assuming one-half of the definition of **basis**, so we only need to consider the other half: must S also be a spanning set for the space? If it were *not* a spanning set, then $\operatorname{Span} S$ would merely be a proper subspace, and there would be other vectors in the full vector space that are *not* in that subspace. Let $\mathbf{v}$ be one such vector. Then Proposition 17.5.6 tells us that the collection of vectors containing both $\mathbf{v}$ and every vector in S must be linearly independent. But this is not possible, since this new, "larger" linearly independent collection would contain S as a proper subcollection, and we have assumed that S is a maximal linearly independent set of vectors. So S must also be a spanning set for the vector space, as required. ■

Theorem 18.5.3 Uniqueness of coordinate vectors. *Given a basis for a vector space, every vector in the space has* one unique *expression as a linear combination of the basis vectors.*

Proof. We will prove that two different linear combination expressions involving basis vectors must compute to two different vectors, which will imply that one single vector in the vector space cannot have two different expressions as linear combinations of basis vectors. So suppose we have two different linear combination expressions involving basis vectors. Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$ be a complete list of the basis vectors involved in both expressions. By attaching a zero coefficient to missing vectors, we can assume that *both* linear combination expressions involve *all* of these basis vectors. Let $a_1, a_2, \dots, a_m$ represent the corresponding coefficients in one of these linear combination expressions, and let $b_1, b_2, \dots, b_m$ represent the corresponding coefficients in the other. Note that we must have at least one instance of $a_j \neq b_j$ in these collections of coefficients, because we have assumed that these linear combination expressions are different. Now, these two linear combination expressions compute to two vectors in the vector space,

$$\mathbf{v} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \dots + a_m\mathbf{v}_m, \qquad \mathbf{w} = b_1\mathbf{v}_1 + b_2\mathbf{v}_2 + \dots + b_m\mathbf{v}_m.$$

We would like to prove that $\mathbf{v} \neq \mathbf{w}$. Equivalently, we would like to prove that $\mathbf{v} - \mathbf{w} \neq \mathbf{0}$. By collecting like terms, this difference vector can also be expressed as a linear combination as

$$\mathbf{v} - \mathbf{w} = (a_1 - b_1)\mathbf{v}_1 + (a_2 - b_2)\mathbf{v}_2 + \cdots + (a_m - b_m)\mathbf{v}_m.$$

Since we have at least one instance of $a_j \neq b_j$, we have at least one nonzero coefficient in the expression above, and so the linear combination above is nontrivial. And our basis vectors are linearly independent, so a nontrivial linear combination of basis vectors cannot equal the zero vector. Therefore, $\mathbf{v} - \mathbf{w} \neq \mathbf{0}$ as desired. ■

Remark 18.5.4 In the theorem above, for the purposes of the *uniqueness* of an expression as a linear combination of basis vectors, we do not consider reordering a linear combination, or including or removing a term with a 0 coefficient, as producing different linear combinations. (However, recall that for the purposes of forming coordinate vectors, order in a linear combination does definitely matter, as described in Warning 18.3.1.)

CHAPTER 19

Dimension

19.1 Discovery guide

Recall.

A **basis** for a vector space is a linearly independent spanning set.

Discovery 19.1 Answer each of the following assuming *nonzero* vectors in $\mathbb{R}^3$.

(a) What geometric shape is the span of one nonzero vector?

(b) **(i)** What is the definition of **linearly dependent** for a set of two vectors?

(ii) What does this mean geometrically?

(iii) What is the shape of the span of two nonzero linearly dependent vectors?

(c) **(i)** What does **linearly independent** mean geometrically for a set of two vectors?

(ii) What is the shape of the span of two linearly independent vectors?

(d) Based on your answers so far, do you think a set of *two* vectors can be a basis for $\mathbb{R}^3$?

(e) **(i)** What is the definition of **linearly dependent** for a set of three vectors?

(ii) What does this mean geometrically?

(iii) What is the shape of the span of three nonzero linearly dependent vectors? (There are actually two possibilities here.)

(f) **(i)** What does **linearly independent** mean geometrically for a set of three vectors?

(ii) What is the "shape" of the span of three linearly independent vectors?

(g) Do you think a set of *four* vectors can be a basis for $\mathbb{R}^3$?

(h) Determine the "dimension" of each of the following subspaces of $\mathbb{R}^3$. In each case, how does the number you come up with correspond with the answers you've given throughout this activity?

(i) A line through the origin.

(ii) A plane through the origin.

(iii) All of $\mathbb{R}^3$.

(iv) The trivial subspace (i.e. just the origin).

We've been using the word "dimension" informally throughout our developlment of the concepts of vectors (e.g. calling vectors in $\mathbb{R}^2$ *two-dimensional* vectors), but finally we can match our intuition about the "dimension" of the various types of subspaces of $\mathbb{R}^3$ with the theoretical concepts of **linear independence** and **spanning** to make the following definition.

dimension of a vector space
the number of vectors required in a basis for that space

One way to obtain a basis for a space (and hence to determine its dimension) is to *assign parameters* — then each *independent* parameter corresponds to a basis vector.

For example, in $\mathbb{R}^2$ we have natural parameters associated to the x- and y-coordinates: $\mathbf{x} = (x, y)$. Expanding this into a linear combination, we get $\mathbf{x} = x(1,0) + y(0,1)$. Parameter x corresponds to vector $(1,0)$ and parameter y corresponds to vector $(0,1)$, and together the two corresponding vectors form a basis $\{(1,0),(0,1)\}$ for $\mathbb{R}^2$. (In fact, the **standard basis** for $\mathbb{R}^2$!). Since there were *two* independent parameters required to described an arbitrary vector in the space, this led to *two* basis vectors, and so the dimension of $\mathbb{R}^2$ is (surprise!) 2.

Step-by-step procedure.

(a) Express arbitrary elements in the space in terms of parameters.

(b) Use any extra conditions to reduce to the minimum number of *independent* parameters (if necessary).

(c) Split up your parametric vector description into a linear combination based on the remaining parameters.

(d) Extract the basis vector attached to each parameter.

(e) Count the basis vectors to determine the dimension of the space (which should also correspond to the number of independent parameters required).

Discovery 19.2 In each of the following, determine a basis for the given space using the parameter method outlined above, similarly to the provided $\mathbb{R}^2$ example. Then count the dimension of the space.

(a) $\mathbb{R}^3$.

(b) The subspace of $\mathbb{R}^3$ consisting of vectors whose second coordinate is zero.

(c) The subspace of $\mathbb{R}^3$ consisting of vectors whose first and third coordinates are equal.

(d) $M_2(\mathbb{R})$, i.e. the space of 2×2 matrices.

(e) The subspace of $M_2(\mathbb{R})$ consisting of upper-triangular matrices.

(f) The subspace of $M_2(\mathbb{R})$ consisting of upper-triangular matrices whose diagonal entries add to zero.

(g) The subspace of $M_2(\mathbb{R})$ consisting of matrices whose entries sum to zero.

(h) $P_5(\mathbb{R})$, i.e. the space of polynomials of degree 5 or less.

(i) The subspace of $\mathrm{P}_5(\mathbb{R})$ consisting of polynomials with constant term equal to zero.

(j) The subspace of $\mathrm{P}_5(\mathbb{R})$ consisting of *odd* polynomials, i.e. those involving only odd powers of x (and no constant term).

(k) The subspace of $\mathrm{P}_5(\mathbb{R})$ consisting of *even* polynomials, i.e. those involving only even powers of x (and a constant term).

A vector space is called **finite-dimensional** if it can be spanned by a finite set; otherwise, it is called **infinite-dimensional**. For example, $\mathbb{R}^n$ is finite-dimensional for each value of n, because it can be spanned by the finite set of standard basis vectors $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$.

Discovery 19.3 Is the vector space of all polynomials is finite- or infinite-dimensional?

Hint. If S is a finite set of polynomials, what are the possible degrees of the polynomials in $\operatorname{Span} S$?

We've already seen that a linearly dependent spanning set can be reduced to a basis (Proposition 18.5.1). Working the other way, we will use Proposition 17.5.6 to argue in Subsection 19.5.2 that a linearly independent set that is not a spanning set can be *built up* to a basis by including additional vectors (Proposition 19.5.4). Proposition 17.5.6 tells us exactly how to do this: to ensure linear independence at each step, the new vector to be included should not be in the span of the old (i.e. the new should not be any linear combination of the old).

Discovery 19.4 In each of the following, enlarge the provided linearly independent set into a basis for the space.

Hint. Since we now know the dimensions of these spaces, we know how many linearly independent vectors are required to form a basis. Just guess *simple* new vectors to include in the given set, one at a time, and for each make sure your new vector is not a linear combination of the vectors you already have. (You can check this by trying to solve an appropriate system of linear equations.)

(a) $V = \mathbb{R}^3$, $S = \{(1,1,0), (1,0,1)\}$.

(b) $V = \mathrm{M}_2(\mathbb{R})$, $S = \left\{ \left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right], \left[\begin{smallmatrix} 1 & 0 \\ 1 & -1 \end{smallmatrix}\right] \right\}$.

Discovery 19.5 Suppose V is a finite-dimensional vector space, and W is a subspace of V.

(a) What is the relationship between $\dim W$ and $\dim V$? Justify your answer in terms of the *definition* of dimension.

Hint. The pattern of the previous exercise, where a linearly independent set can be built up into a basis, might help in articulating your justification.

(b) Is it possible for $\dim W = \dim V$ to be true?

19.2 Terminology and notation

finite-dimensional vector space
a vector space for which there exists a finite spanning set

infinite-dimensional vector space
a vector space for which there does not exists a finite spanning set

dimension of a finite-dimensional vector space
the number of vectors required in a basis for the space

$\dim V$ notation for the dimension of a finite-dimensional vector space V

Remark 19.2.1 In the case of an infinite-dimensional space V, we might write $\dim V = \infty$ to indicate this property. Similarly, we might write $\dim V < \infty$ to mean that a space V is finite-dimensional.

19.3 Concepts

In this section.

19.3.1 The "just-right" number of vectors in a spanning set

In Discovery 19.1, we reminded ourselves of the geometric interpretation of linear dependence and independence in $\mathbb{R}^3$.

Also see. Subsection 17.3.4

That discovery activity ties these new, abstract concepts back to our previous descriptions of lines and planes in Chapter 14. In that chapter, we described a line via an "initial" vector and a parallel vector, and we described a plane via an "initial" vector and *two* parallel vectors that are not parallel to each other. Recall that for a line or plane in $\mathbb{R}^3$ to be a *subspace*, it must contain the zero vector (i.e. it must pass through the origin). In this case, we can (and always will) take the "initial" point to be the origin.

So a line through the origin can be described by the vector equation $\mathbf{x} = t\mathbf{p}$, where $\mathbf{p}$ is a nonzero vector parallel to the line. With our new concept of span, we can instead write $L = \operatorname{Span}\{\mathbf{p}\}$ to represent the line L through the origin that is parallel to the vector $\mathbf{p}$. One vector is the "just-right" size for the spanning set for a line. If we had a spanning set for L consisting of two vectors, then because L goes through the origin, and because spanning vectors are always part of the space they span, both vectors would have to be parallel to the line and so would be parallel to each other. That is, the two spanning vectors would be scalar multiples of each other, and the spanning set would be linearly dependent.

Similarly, a plane P through the origin described by the vector equation $\mathbf{x} = s\mathbf{p}_1 + t\mathbf{p}_2$, where $\mathbf{p}_1, \mathbf{p}_2$ are nonzero vectors parallel to the plane but not to each other, can also be represented as $P = \operatorname{Span}\{\mathbf{p}_1, \mathbf{p}_2\}$. Two is the "just-right" size for the spanning set for a plane — one vector would only span a line, and three vectors that are all parallel to the plane would have to be linearly dependent.

When we consider all of $\mathbb{R}^3$, the "just-right" size for a spanning set is three — two vectors would only span a plane (or just a line if the two vectors are parallel to each other), and four vectors would be linearly dependent.

So it seems that there is always a "just-right" size for a spanning set to be a basis — if it's too small it spans only a subspace and not the whole space, and if it's too large it will be linearly dependent. We call this "just-right" size the **dimension** of the space.

Checking that a proposed spanning set actually *does* span the whole space can be difficult, as we noted at the end of Subsection 16.4.4. In Subsection 19.5.2, we will find that the concept of dimension gives us a powerful way to sidestep this task if we already know the dimension of the space. If we have the "just-right" number of vectors, and those vectors are linearly independent, then the subspace they span will have the same "size" (i.e. dimension) as the whole space, which will force that subspace to in fact be the whole space.

See. Proposition 19.5.5.

19.3.2 Dimension as geometric "degrees of freedom"

Again thinking of our tasks and results in Discovery 19.1, we can make the geometric interpretation of **dimension** more explicit.

Lines have dimension 1. Imagine standing on a line; how many "degrees of freedom" of movement do you have while staying on the line? You can only move forwards or backwards, and backwards is just the opposite (i.e. negative) of forwards. So you only have one "degree of freedom" on a line, and this is reflected in the fact that a basis for a line requires only *one* vector — that vector will represent the forward direction, and its negative will represent the backward direction. One "degree of freedom" on a line, and the dimension of a line is 1.

Planes have dimension 2. On a plane, you have two "degrees of freedom" of movement: you can move forwards/backwards (one direction and its opposite), or you can move side-to-side (a second direction and its opposite). So a basis for a plane requires exactly *two* vectors that do not represent the same direction, and the dimension of a plane is 2.

Space has dimension 3. When we consider all of space, we add a third dimension representing a third "degree of freedom," since you can now move upwards or downwards in addition to the previous forward/backward and side-to-side directions.

19.3.3 Dimension as algebraic "degrees of freedom"

There is an algebraic interpretation of the "degrees of freedom" point of view discussed above that we can transplant from $\mathbb{R}^n$ to other vector spaces. Consider again a plane in $\mathbb{R}^3$ described via a vector equation $\mathbf{x} = s\mathbf{p}_1 + t\mathbf{p}_2$. Each of the vectors $\mathbf{p}_1, \mathbf{p}_2$ represents an independent direction of movement along the plane, providing us with our two geometric degrees of freedom on this plane of dimension 2. Algebraically, these two degrees of freedom are provided by the parameters s and t. To convert the *general* formula $\mathbf{x} = s\mathbf{p}_1 + t\mathbf{p}_2$ representing *all* vectors in the plane to a *specific* formula representing *one* vector on the plane, we need to choose a specific value for s (related to how far to move in the direction $\mathbf{p}_1$) and a specific value for t (related to how far to move in the direction $\mathbf{p}_2$). These two values can be chosen *independently* — that is, choosing a value for t does not depend on what value is chosen for s, and vice versa. So two independent parameters in a general description of every vector, representing two "degrees of freedom," corresponds to the dimension value of 2 for the plane.

In Discovery 19.2, we practised using this point of view to not only determine the dimension of a space, but to extract a basis for the space from a general parametric description of the vectors in the space. Below is a general procedure for the process. See Subsection 19.4.1 for examples of using this procedure.

Procedure 19.3.1 Obtaining a basis from paramaters. *To determine a basis for a subspace U of a vector space V, when subspace U is not already described in terms of a spanning set:*

1. *Determine a general, parametric expression capable of expressing* all *vectors in V.*
2. *Use the defining conditions of the subspace U to reduce your general expression from the previous step to the minimum number of independent parameters possible.*
3. *Expand the reduced parametric expression from the previous step to a linear combination of the form*

   $$\mathbf{x} = (parameter) \cdot (vector) + (parameter) \cdot (vector) + \cdots + (parameter) \cdot (vector),$$

   where there is one term in the linear combination for each independent parameter, and the vectors involved are specific *vectors in the space, not involving parameters.*
4. *The collection of specific vectors in the general linear combination expression from the previous step,* without parameters*, should now form a basis for U.*

Remark 19.3.2

- This procedure can still be used in the case U is equal to the whole space V, but likely Step 2 will not be needed. In this case, the procedure is likely to produce a **standard basis** for V.
- In Remark 16.4.9, we claimed that that every subspace is somehow defined by one or more homogeneous conditions. Typically, in Step 2 of the procedure, you will be using such homogeneous conditions to express relationships between the parameters, in which some parameters can be solved for and then eliminated by substituting for them in the general parametric vector expression from Step 1.
- This procedure was actually one of the first things we learned, back in Chapter 2! Except back then we called the procedure **row reduction**. When we solve a homogeneous system of equations with $m \times n$ coefficient matrix A, we are attempting to determine all vectors $\mathbf{x}$ that satisfy the homogeneous condition $A\mathbf{x} = \mathbf{0}$. We could have started this process by assigning parameters $x_1 = t_1$, $x_2 = t_2$, and so on, at the beginning of the process, but this was not necessary because the matrix-reduction process doesn't involve any variable/parameter letters. By row reducing, we simplify the original homogeneous conditions (i.e. the original equations in the system) so that it becomes obvious how we can isolate certain of the variables and express them in terms of the others (or determine that they

are always zero and so can be eliminated completely). We then assign the *minimum* number of parameters necessary, leading to a general, parametric expression for all vectors in the solution space. See Subsection 19.4.1 for an example of using this procedure to determine a basis for the solution space of a homogeneous system.

19.3.4 The dimension of a subspace

In Discovery 19.5, we considered how the dimension of a subspace compares to the dimension of the whole space. The dimension of a space is defined to be the number of vectors required in a basis (i.e. a linearly independent spanning set) for the space. We know what **spanning set** means for a subspace — a set of vectors is a spanning set when the collection of all possible linear combinations of the spanning set vectors is the same as the collection of all vectors in the subspace. But the definition of **linearly independent** does not seem to be relative to the space that the vectors are in, except for the use of the vector operations for that space, which are always the same in a subspace as they are in the whole space.

In more detail, the definitions of linear dependence and independence involve only the zero vector and the concept of linear combination, and every subspace contains the zero vector and is closed under taking linear combinations (Proposition 16.5.2). So if we have a set of vectors in a subspace of a larger vector space, and we would like to determine whether that set is linearly dependent or independent, it is irrelevant whether we consider those vectors as being a part of the subspace or as being a part of the large space — the answer will be the same regardless of our point of view on where these vectors "live."

It seems like a spanning set for a subspace should require fewer vectors than a spanning set for the larger space. This was our experience in Discovery 19.2, where eliminating dependent parameters using the subspace conditions led to a smaller basis. And the concepts of **linear dependence/independence** are independent of the concept of **subspace**. So our intuition is that the dimension of a subspace should be less than the dimension of the whole space, and that is exactly what we will see in Subsection 19.5.3.

19.3.5 The dimension of the trivial vector space

What should the dimension of the trivial vector space $\{\mathbf{0}\}$ be? If this were the subspace of $\mathbb{R}^n$ consisting of just the origin, we would have zero "degrees of freedom" of movement, as we couldn't move at all without leaving the subspace. And if we want a general algebraic expression describing all vectors in this space, zero parameters are needed since we can simply write $\mathbf{x} = \mathbf{0}$. So both our geometric and our algebraic conceptions of **dimension** suggest that $\dim\{\mathbf{0}\}$ should be 0.

Furthermore, in the previous subsection we decided that the dimension of a subspace should be smaller than the dimension of the whole space. The trivial vector space is always a subspace of every vector space, even 1-dimensional spaces. But clearly the trivial space is not the same "size" as a 1-dimensional space, so its dimension should be *strictly* smaller than 1, which only leaves dimension-0 as a possibility.

But what about the technical definition of dimension? How many vectors are required in a basis for the trivial space? A basis for $\{\mathbf{0}\}$ cannot contain a nonzero vector, because a span always contains its spanning vectors and this space does not contain anything nonzero. But while the *collection of vectors* consisting of just the zero vector *is* a spanning set for the *space of vectors* consisting of just a zero vector, we decided in Chapter 17 that the zero vector all by itself should be considered a linearly dependent set. However, the collection of vectors that contains *no vectors at all* (i.e. the empty set) is linearly independent, because it does not contain an example of a vector that can be expressed as a linear combination of other vectors in the set (since it contains nothing at all). So if we just decide that Span{} should result in the trivial vector space, then we can consider the empty set of vectors {} as a basis for the trivial space $\{\mathbf{0}\}$, and this basis contains 0 vectors.

For all of these reasons, it seems correct to consider $\dim\{\mathbf{0}\}$ to be 0.

19.4 Examples

In this section.

- Subsection 19.4.1 *Determining a basis from a parametric expression*
- Subsection 19.4.2 *An infinite-dimensional example*
- Subsection 19.4.3 *Enlarging a linearly independent set to a basis*

19.4.1 Determining a basis from a parametric expression

Example 19.4.1 From the discovery guide. First, let's carry out some of the examples from Discovery 19.2.

1. In Discovery 19.2.c, we considered a certain subspace of $\mathbb{R}^3$.

   An arbitrary vector in $\mathbb{R}^3$ requires three parameters to describe its three components: $\mathbf{x} = (a, b, c)$. If we restrict to just those vectors whose first and third components are equal, we can replace c by a, to get

   $$\mathbf{x} = (a, b, a) = (a, 0, a) + (0, b, 0) = a(1, 0, 1) + b(0, 1, 0).$$

   So a basis for this subspace of $\mathbb{R}^3$ is $\mathcal{B} = \{(1, 0, 1), (0, 1, 0)\}$, and the dimension is 2.

2. In Discovery 19.2.g, we considered a certain subspace of $M_2(\mathbb{R})$.

   An arbitrary matrix in $M_2(\mathbb{R})$ requires four parameters to describe its four entries:

   $$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

   If we restrict to those matrices whose entries sum to zero, so that $a + b + c + d = 0$, then we can isolate $d = -a - b - c$ and substitute that into the matrix:

   $$\begin{aligned} A &= \begin{bmatrix} a & b \\ c & -a-b-c \end{bmatrix} \\ &= \begin{bmatrix} a & 0 \\ 0 & -a \end{bmatrix} + \begin{bmatrix} 0 & b \\ 0 & -b \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ c & -c \end{bmatrix} \\ &= a \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} + b \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} + c \begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix}. \end{aligned}$$

   So this subspace of $M_2(\mathbb{R})$ has dimension 3, with basis

   $$\mathcal{B} = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix} \right\}.$$

3. In Discovery 19.2.j, we considered a certain subspace of $\mathrm{P}_5(\mathbb{R})$.

   An arbitrary polynomial in $\mathrm{P}_5(\mathbb{R})$ requires *six* parameters, one for each power of x, along with a parameter for the constant term:

   $$p(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5.$$

   If we restrict to only odd polynomials, we need to eliminate the constant term and the even powers of x:

   $$p(x) = a_1 x + a_3 x^3 + a_5 x^5.$$

   (Equivalently, we have applied the homogeneous conditions $a_0 = 0$, $a_2 = 0$, and $a_4 = 0$.) So this subspace of $\mathrm{P}_5(\mathbb{R})$ has dimension 3, with basis $\mathcal{B} = \{x, x^3, x^5\}$.

□

Example 19.4.2 Dimensions of familiar spaces via parameters. Now let's examine how the dimensions of our favourite example spaces relate to our parametric point of view. We considered specific examples of these in parts of Discovery 19.2, but here we'll work more generally.

1. An arbitrary vector in $\mathbb{R}^n$ requires n parameters, one for each component:

   $$\mathbf{x} = (x_1, x_2, \ldots, x_n).$$

   If we expanded this into a linear combination, each parameter would be attached to a standard basis vector $\mathbf{e}_j$. Since we've got n parameters and a corresponding n standard basis vectors, we have

   $$\dim \mathbb{R}^n = n.$$

2. An arbitrary $m \times n$ matrix in $\mathrm{M}_{m\times n}(\mathbb{R})$ requires mn parameters, one for each entry:

   $$A = [a_{ij}], \qquad 1 \le i \le m,\ 1 \le j \le n.$$

   If we expanded this into a linear combination, each parameter would be attached to a standard basis matrix E_{ij}, with zeros in all entries except for a single 1 in the $(i,j)^{\text{th}}$ entry. Since we've got mn parameters and a corresponding mn standard basis matrices, we have

   $$\dim \mathrm{M}_{m\times n}(\mathbb{R}) = mn.$$

3. An arbitrary polynomial in $\mathrm{P}_n(\mathbb{R})$, the space of polynomials of degree n or less, requires $n+1$ parameters, one for each power of x plus an extra one for the constant term:

   $$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n.$$

   This is already naturally expressed as a linear combination, and each parameter is attached to a polynomial from the standard basis

   $$\mathcal{B} = \{1, x, x^2, \ldots, x^n\}.$$

   Since we've got $n+1$ parameters and a corresponding $n+1$ standard basis polynomials, we have

   $$\dim \mathrm{P}_n(\mathbb{R}) = n + 1.$$

□

Example 19.4.3 The solution space of a homogeneous system. In Remark 19.3.2, we noted how assigning parameters after row reducing a homogeneous system corresponded directly to a parameter-based procedure for determine the basis for a space. Let's illustrate this correspondence with an example.

Consider the homogeneous system in Discovery 2.4, which we solved in Example 2.4.4. In Example 16.4.8, we used the Subspace Test to verify that the solution set of a homogeneous system with an $m \times n$ coefficient matrix is a subspace of $\mathbb{R}^n$. The system from Discovery 2.4 has a 4×4 coefficient matrix that we reduced:

$$\begin{bmatrix} 3 & 6 & -8 & 13 \\ 1 & 2 & -2 & 3 \\ 2 & 4 & -5 & 8 \end{bmatrix} \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \begin{bmatrix} 1 & 2 & 0 & -1 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Assigning parameters to free variables x_2, x_4, we obtained the general solution in parametric form:

$$x_1 = -2s + t, \qquad x_2 = s, \qquad x_3 = 2t, \qquad x_4 = t.$$

We can use these expressions as components in a general solution vector, and expand it out to a linear combination, just as in the previous examples in this subsection:

$$\mathbf{x} = \begin{bmatrix} -2s+t \\ s \\ 2t \\ t \end{bmatrix} = \begin{bmatrix} -2s \\ s \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} t \\ 0 \\ 2t \\ t \end{bmatrix} = s \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} 1 \\ 0 \\ 2 \\ 1 \end{bmatrix}$$

Since two parameters are needed to describe the solution vectors for this system, the solution space has dimension 2, and a basis for this subspace is

$$\mathcal{B} = \left\{ \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 2 \\ 1 \end{bmatrix} \right\}.$$

A look ahead. We will study solution spaces of homogeneous systems further in Chapter 20.

□

19.4.2 An infinite-dimensional example

All of the examples in the previous subsection involved **finite-dimensional** spaces. Here's an example of an **infinite-dimensional** space.

In Discovery 19.3, we considered the space of *all* polynomials. This space cannot be spanned by any finite collection of polynomials, because such a collection would have a polynomial of largest degree, and then every linear combination of those polynomials would have that degree or smaller. So the span of those polynomials could never include polynomials of larger degree. Thus,

$$\dim \mathrm{P}(\mathbb{R}) = \infty.$$

We can still come up with a basis for this space, but it will contain an infinite number of vectors:

$$\mathrm{P}(\mathbb{R}) = \operatorname{Span}\{1, x, x^2, x^3, \ldots\}.$$

This equality says that every polynomial is a linear combination of a finite number of powers of x. This spanning set is also linearly independent because no power of x can be expressed as a linear combination of other powers of x.

19.4.3 Enlarging a linearly independent set to a basis

In Discovery 19.4.b, we are given a linearly independent set S of vectors in $V = \mathrm{M}_2(\mathbb{R})$, and we would like to enlarge it to a basis for the whole space. Since S is linearly independent, it is a basis for the subspace $\operatorname{Span} S$. Since we know that $\dim \mathrm{M}_2(\mathbb{R}) = 4$, we need to add two more linearly independent vectors to get up to a basis for V. To do this, we can use Proposition 17.5.6, which says that to enlarge a linearly independent set, we need to add a vector from outside the span of the vectors we already have.

An obvious source for candidate vectors to use to enlarge S is the standard basis $\mathcal{B} = \{E_{11}, E_{12}, E_{21}, E_{22}\}$. We know that $\operatorname{Span} S$ can't contain *all four* of these vectors, because then Statement 1 of Proposition 16.5.6 would imply that all of $V = \operatorname{Span}\mathcal{B}$ would be contained in $\operatorname{Span} S$, which is not possible because $\dim(\operatorname{Span} S)$ is just 2. So let's start by checking whether E_{11} is in $\operatorname{Span} S$. The vector equation

$$k_1 \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + k_2 \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

leads to a system of equations with augmented matrix as on the left below, which we can reduce:

$$\left[\begin{array}{rr|r} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & -1 & 0 \end{array}\right] \xrightarrow[\text{reduce}]{\text{row}} \left[\begin{array}{rr|r} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{array}\right].$$

The one in the last column indicates that the system is inconsistent, which is what we want — there is no solution, so E_{11} is not in $\operatorname{Span} S$, and so we can enlarge S by including E_{11} and it will remain linearly independent. Call the enlarged set S'.

Now let's check if E_{12} is in the span of these *three* linearly independent vectors that we have already. Our vector equation is now,

$$k_1 \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + k_2 \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix} + k_3 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

which leads to a system with augmented and reduced augmented matrices

$$\left[\begin{array}{rrr|r} 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{array}\right] \xrightarrow[\text{reduce}]{\text{row}} \left[\begin{array}{rrr|r} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right].$$

Again, there is no solution, so E_{12} is not in $\operatorname{Span} S'$. We are now up to four linearly independent vectors, which must form a basis for the 4-dimensional space $\mathrm{M}_2(\mathbb{R})$.

See. Corollary 19.5.6 in Subsection 19.5.2.

Our final basis is

$$\left\{ \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \right\}.$$

A look ahead. In the example above, we could have augmented our initial spanning set vectors with *all* standard basis vectors, and checked all of them all at once:

$$\left[\begin{array}{rr|rrrr} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 & 0 & 1 \end{array}\right] \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \left[\begin{array}{rr|rrrr} 1 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 1 & 0 & 0 & 1/2 & -1/2 \\ 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1/2 & -1/2 \end{array}\right].$$

By changing the position of the vertical line that indicates the separation of the coefficients from the column of constants one column at a time, we can change our point of view on each augmented column from representing a vector to be achieved as a linear combination of the spanning vectors to a vector *included* as a spanning vector.

A look ahead. We will see in Chapter 20 how the leading ones in the reduced matrix tell us exactly which of the original six vectors are linearly independent.

19.5 Theory

In this section.

- Subsection 19.5.1 *Dimension as size of a basis*
- Subsection 19.5.2 *Consequences for the theory of linear dependence / independence and spanning*
- Subsection 19.5.3 *Dimension of subspaces*

19.5.1 Dimension as size of a basis

Since **dimension** is defined in terms of **basis**, it is important to know that we can always form a basis. The following fact is true for all vector spaces, but we will state and prove it only for finite-dimensional spaces. It is essentially just a restatement of Proposition 18.5.1 (which itself is a restatement of Proposition 17.5.5).

Theorem 19.5.1 *Every finite-dimensional vector space has a basis.*

Proof. By definition, a vector space is finite-dimensional when it has a finite spanning set. Proposition 18.5.1 states that every finite spanning set can be reduced to a basis. So if a finite spanning set exists for a space, so does a basis. ■

The next two facts allow us to attach a *single* number to a vector space as *the* dimension of the space.

Lemma 19.5.2 *A basis for a finite-dimensional vector space must contain a finite number of vectors.*

Proof. By definition, a finite-dimensional vector space has at least one example of a spanning set that contains a finite number of vectors. By Lemma 17.5.7, any other set of vectors from this space that contains *more* vectors than this example spanning set must be linearly dependent. But a basis is always linearly independent, and so cannot have more vectors than the finite number in this example spanning set. ■

Theorem 19.5.3 Uniformity of dimension. *Every basis for a finite-dimensional vector space has the same number of vectors.*

Proof. Suppose $\mathcal{B}_1$ and $\mathcal{B}_2$ are two different bases for a finite-dimensional vector space V. First, both $\mathcal{B}_1$ and $\mathcal{B}_2$ must contain a finite number of vectors, by Lemma 19.5.2. Now, $\mathcal{B}_1$ is a basis, so it is a spanning set, and so by Lemma 17.5.7 any set that contains more vectors than $\mathcal{B}_1$ must be linearly dependent. But $\mathcal{B}_2$ is also a basis, so it is linearly independent. Therefore, $\mathcal{B}_2$ cannot contain *more* vectors than there are in $\mathcal{B}_1$.

The same reasoning works the other way: $\mathcal{B}_1$ cannot contain *more* vectors than there are in the spanning set $\mathcal{B}_2$, otherwise it would be linearly dependent. Since neither set of vectors can contain more vectors than the other, the two sets must contain *exactly the same* number of vectors. ■

19.5.2 Consequences for the theory of linear dependence/independence and spanning

Now we extend Proposition 17.5.6 to establish a "building-up" counterpart to Proposition 18.5.1.

Proposition 19.5.4 Enlarging an independent set to a basis. *In a finite-dimensional vector space, every linearly independent set of vectors can be enlarged to a basis. That is, if S is a linearly independent set of vectors in a finite-dimensional vector space, then there exists a basis for the space that contains S as a subcollection.*

Clarification. *In this proposition, we consider the hypothetical "can be enlarged" to allow the possibility of not enlarging the set at all, in case the linearly independent set is already a basis.*

Proof. Suppose S is a linearly independent set of vectors in a finite-dimensional vector space. If it is also a spanning set, then it is already a basis and does not need to be enlarged. If it is not a spanning set, then there are vectors in the space that are not in $\operatorname{Span} S$. Choose a vector $\mathbf{v}$ *not* in $\operatorname{Span} S$, and let S' be the set that contains all the vectors of S as well as $\mathbf{v}$. By Proposition 17.5.6, the set S' is still linearly independent. If S' is also a spanning set, then it is a basis and we have the desired enlargement from S. Otherwise, we could again enlarge S' by some vector that is *not* in $\operatorname{Span} S'$ and still have a linearly independent set. We can continue in this fashion, but we will have to reach a point where we will not be able to enlarge our set any further without it becoming linearly dependent, since we know that in a finite-dimensional space, once a set of vectors gets too large it can no longer be linearly independent (Lemma 17.5.7). At this point, our enlarged linearly independent set *must* also be a spanning set (and hence a basis), since if it weren't we *would* be able to enlarge it again as before, with the enlarged set remaining independent. ■

The concept of dimension gives us another way to know whether a set of vectors is a basis, since it is the "just-right" size for a set of vectors to be a basis.

Proposition 19.5.5 Using dimension to help test basis. *Suppose S is a set of vectors in a finite-dimensional vector space, and the number of vectors in S is exactly equal to the dimension of the vector space.*

1. *If S is linearly independent, then we can conclude that S is also a spanning set without checking.*

2. *If S is a spanning set, then we can conclude that S is also linearly independent without checking.*

Proof of Statement 1. Assume that S is linearly independent. By Proposition 19.5.4, S can be enlarged to a basis for the vector space. But every basis for that space contains the same number of vectors (Theorem 19.5.3), and we have assumed that S already contains that number of vectors. So S must not need to be enlarged to become a basis — it must already be a basis itself, and so must be a spanning set. ■

Proof of Statement 2. Assume that S is a spanning set. By Proposition 18.5.1, S can be reduced to a basis for the vector space. But every basis for that space contains the same number of vectors (Theorem 19.5.3), and we have assumed that S already contains that number of vectors. So S must not need to be reduced to become a basis — it must already be a basis itself, and so must be linearly independent. ■

Corollary 19.5.6 *Suppose S is a set of vectors in a finite-dimensional vector space, and the number of vectors in S is exactly equal to the dimension of the vector space. If S is* either *known to be linearly independent* or *known to be a spanning set, then S must also have the other property, and hence must be a basis for the vector space.*

Remark 19.5.7 In a space whose dimension is known, the above corollary effectively reduces the amount of work required to check whether a set of vectors is a basis in half, since if we start with the right number of vectors in a basis-candidate set then we only need to check one of the requirements in the definition of **basis**. In practice, it is usually easier to carry out the Test for Linear Dependence/Independence than it is to check for spanning.

19.5.3 Dimension of subspaces

As discussed in Subsection 19.3.4, a set of linearly independent vectors in a subspace is still linearly independent when considered as a set of vectors in the larger space. So we can use Proposition 19.5.4 to relate a basis for a subspace to a basis for the whole space, and then also the dimension of the subspace to the dimension of the whole space.

Proposition 19.5.8 *Suppose U is a subspace of a finite-dimensional vector space V. Then the following all hold true.*

1. *Every basis for U can be enlarged to a basis for V.*
2. *We have* $\dim U \leq \dim V$.
3. *It is the case that* $\dim U = \dim V$ *only if U is actually the whole space V.*

Proof of Statement 1. Since U is a subspace of V, each vector of U is also a vector of V. So a basis for U will be a linearly independent set of vectors in V, which Proposition 19.5.4 tells us can be enlarged to a basis for V. ■

Proof of Statement 2. Recall that the dimenion of a vector space (whether a subspace of another space or not) is defined to be the number of vectors in a basis for the space. Since every basis for U can be enlarged to a basis for V, the number of vectors in a basis for U cannot be larger than the number of vectors in a basis for V. ■

Proof of Statement 3. Let $\mathcal{B}$ be a basis for U, so that $U = \operatorname{Span}\mathcal{B}$. If we have $\dim U = \dim V$, then the number of vectors in $\mathcal{B}$ is exactly equal to the dimension of V. But $\mathcal{B}$ is also linearly independent in V, so by Statement 1 of Proposition 19.5.5, it must also be a spanning set for V. Thus, $U = \operatorname{Span}\mathcal{B} = V$. ■

Chapter 20

Column, row, and null spaces

20.1 Discovery guide

In this discovery guide.

20.1.1 Column space

Take a minute to remind yourself of the column-wise view of matrix multiplication from (∗∗∗) in Subsection 4.3.7. In words, this matrix multiplication pattern says that in a matrix product AB,

- the first column of AB is the result of multiplying matrix A against the first column of B,
- the second column of AB is the result of multiplying matrix A against the second column of B,
- and so on.

In the first discovery activity, we'll use this pattern to obtain another important pattern involving the standard basis vectors.

Discovery 20.1 Notice that the columns of the identity matrix are precisely the standard basis vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ of $\mathbb{R}^n$. Use this observation, the matrix multiplication pattern described above, and the matrix identity $AI = A$ to complete the following.

- Product $A\mathbf{e}_1$ is equal to .
- Product $A\mathbf{e}_2$ is equal to .
- Product $A\mathbf{e}_j$ is equal to .

Discovery 20.2 Think of an $m \times 3$ matrix A as being made out of three column vectors from $\mathbb{R}^m$:

$$A = \begin{bmatrix} | & | & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 \\ | & | & | \end{bmatrix}.$$

(a) Suppose we want to compute $A\mathbf{x}$, where $\mathbf{x} = (5,3,-1)$ (but as a column vector). Use the pattern you discovered in Discovery 20.1 to fill in the following.

Since

$$\begin{bmatrix} 5 \\ 3 \\ -1 \end{bmatrix} = 5\mathbf{e}_1 + 3\mathbf{e}_2 + (-1)\mathbf{e}_3,$$

then

$$A\begin{bmatrix} 5 \\ 3 \\ -1 \end{bmatrix} = A(5\mathbf{e}_1 + 3\mathbf{e}_2 + (-1)\mathbf{e}_3) = 5\qquad + 3\qquad + (-1)\qquad .$$

From this, we see that the column vector $A\mathbf{x}$ is in the span of

.

(b) Convince yourself that the details/conclusion of Task a would be the same for *every* $\mathbf{x}$, not just the example $\mathbf{x}$ we used.

(c) Now consider system $A\mathbf{x} = \mathbf{b}$. If this system is consistent (i.e. has at least one solution), then our final conclusion from Task a would also be true about the column vector $\mathbf{b}$, since $\mathbf{b} = A\mathbf{x}$ for at least one $\mathbf{x}$.

So system $A\mathbf{x} = \mathbf{b}$ can only be consistent if $\mathbf{b}$ is in the span of

.

For $m \times n$ matrix A, from Discovery 20.2 it appears that the subspace of $\mathbb{R}^m$ obtained by taking the span of the columns of A is important when considering consistency of the system $A\mathbf{x} = \mathbf{b}$. Call this subspace the **column space of** A. Let's explore how to reduce our spanning set (the columns of A) down to a basis. For this task we'll need a fact about how multiplication by a matrix affects the linear independence of column vectors that we will state as Statement 1 of Proposition 20.5.1 in Subsection 20.5.1. You should read this statement before proceeding.

Discovery 20.3 The following matrix is in RREF:

$$B = \begin{bmatrix} 1 & 2 & 0 & 3 & 0 & 5 \\ 0 & 0 & 1 & 4 & 0 & 6 \\ 0 & 0 & 0 & 0 & 1 & 7 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

(a) Build a linearly independent set of column vectors from B by working from left to right, and either including or discarding each column based on whether it is linearly independent from the vectors you have already accumulated. (You should, of course, begin by "including" the first column.) What do you notice about your final set of linearly independent columns, relative to the reduced form of B?

(b) Suppose A is a matrix that can be reduced to B by a single elementary operation. Then there is an elementary matrix E so that

$$B = EA = \begin{bmatrix} \vert & \vert & \vert & \vert & \vert & \vert \\ E\mathbf{a}_1 & E\mathbf{a}_2 & E\mathbf{a}_3 & E\mathbf{a}_4 & E\mathbf{a}_5 & E\mathbf{a}_6 \\ \vert & \vert & \vert & \vert & \vert & \vert \end{bmatrix},$$

where the $\mathbf{a}_j$ are the columns of A. Use your answer to Task a along with the above-referenced Statement 1 to determine which columns of A form a linearly independent set.

(c) Now suppose A is a matrix that can be reduced to B by *two* elementary operations. Then there are elementary matrices E_1, E_2 so that $B = E_2E_1A$. Similarly to Task b, from $B = E_2(E_1A)$, decide which columns of E_1A are linearly independent. Then from the above-referenced Statement 1 and

$$E_1A = \begin{bmatrix} \vert & \vert & \vert & \vert & \vert & \vert \\ E_1\mathbf{a}_1 & E_1\mathbf{a}_2 & E_1\mathbf{a}_3 & E_1\mathbf{a}_4 & E_1\mathbf{a}_5 & E_1\mathbf{a}_6 \\ \vert & \vert & \vert & \vert & \vert & \vert \end{bmatrix}$$

(where the $\mathbf{a}_j$ are the columns of A), decide which columns of A are linearly independent.

(d) Now extrapolate to any number of row operations to complete the following statement: to create a linearly independent set of column vectors from a matrix A, row reduce A to RREF, and then take those columns of A that correspond to in RREF(A).

Discovery 20.4

(a) Use the procedure you've developed in Discovery 20.3.d to develop a reinterpretation of the Test for Linear Dependence/Independence for vectors in $\mathbb{R}^m$: if $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are vectors in $\mathbb{R}^m$, write these vectors as columns in a matrix, row reduce, and then you will know the original vectors are linearly independent if .

(b) Recall that a square matrix is invertible if and only if it can be row reduced to I. Use the procedure for testing linear independence that you've developed in Task a to create another condition that is equivalent to invertibility: a square matrix is invertible if and only if its columns .

(c) Let's go full circle. Combine Task a and Task b to complete the following condition: a collection of n vectors in $\mathbb{R}^n$ is a basis if and only if the square matrix formed by using the vectors as columns has determinant .

20.1.2 Row space

Why let the columns of a matrix have all the fun? Let's now explore the subspace of $\mathbb{R}^n$ formed by the span of the rows in an $m \times n$ matrix, called the **row space** of the matrix.

In the next discovery activity, we'll need to recall Statement 2 of Proposition 16.5.6 that gives us a way to determine when two spans are the same. You should re-read that statement before proceeding.

Discovery 20.5 Assume $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$ to be vectors in some vector space V.

(a) What does the above-referenced Statement 2 say about Span$\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$

and Span$\{\mathbf{v}_1, \mathbf{v}_4, \mathbf{v}_3, \mathbf{v}_2\}$?

(b) Complete the statement: if matrix A' is obtained from A by swapping two rows, then the row spaces of A' and of A are .

(c) What does the above-referenced Statement 2 say about Span$\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$ and Span$\{\mathbf{v}_1, \mathbf{v}_2, -7\mathbf{v}_3, \mathbf{v}_4\}$?

(d) Complete the statement: if matrix A' is obtained from A by multiplying some row by a nonzero constant, then the row spaces of A' and of A are .

(e) What does the above-referenced Statement 2 say about Span$\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$ and Span$\{\mathbf{v}_1, \mathbf{v}_2 + 3\mathbf{v}_1, \mathbf{v}_3, \mathbf{v}_4\}$?

(f) Complete the statement: if matrix A' is obtained from A by adding a multiple of one row to another, then the row spaces of A' and of A are .

Discovery 20.6

(a) Based on Discovery 20.5, the row spaces of a matrix and of its RREF are .

(b) Determine a basis for the row space of a matrix A for which

$$\mathrm{RREF}(A) = \begin{bmatrix} 1 & 0 & 2 & 0 & 3 \\ 0 & 1 & 4 & 0 & 5 \\ 0 & 0 & 0 & 1 & 6 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Discovery 20.7 If you have a collection of vectors in $\mathbb{R}^n$ and you want to obtain a basis for the subspace that the collection spans, you now have two options: either use those vectors as the *columns* in a matrix and row reduce to determine a basis for its column space, or use those vectors as the *rows* in a matrix and row reduce to determine a basis for its row space. Can you think of a reason you might choose to use column space instead of row space? And a reason you might choose to use row space instead of column space?

20.1.3 Null space

There is one more subspace of $\mathbb{R}^n$ associated to a matrix A — the solution space of the homogeneous system $A\mathbf{x} = \mathbf{0}$. Instead of **solution space**, from this point forward we will refer to it as the **null space** of A.

Recall. We have previously used the Subspace Test to show that the solution set of a homogeneous system with a $m \times n$ coefficient matrix is a subspace of $\mathbb{R}^n$ — see Example 16.4.8.

Discovery 20.8 Suppose A is a matrix whose RREF is as given below. Use the “independent parameter” method to determine a basis for the null space of A.

$$\mathrm{RREF}(A) = \begin{bmatrix} 1 & -1 & 0 & 2 & 3 \\ 0 & 0 & 1 & 2 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

20.1.4 Relationship between the three spaces

Discovery 20.9

(a) How can you determine the dimensions of the column/row/null spaces of a matrix from its RREF?

(b) For an $m \times n$ matrix A, what is the relationship between the dimension of its column space, the dimension of its null space, and its size?

20.2 Terminology and notation

The following definitions all apply to an $m \times n$ matrix.

column space
the subspace of $\mathbb{R}^m$ formed by the span of the columns of the matrix

row space
the subspace of $\mathbb{R}^n$ formed by the span of the rows of the matrix

null space
the subspace of $\mathbb{R}^n$ formed by the solution set of the homogeneous system with that matrix as the coefficient matrix

nullity the dimension of the null space of the matrix

20.3 Concepts

In this section.

We have already seen in Example 16.4.8 that the solution set of a homogeneous system $A\mathbf{x} = \mathbf{0}$ with $m \times n$ coefficient matrix A is a subspace of $\mathbb{R}^n$. We will return to this special subspace at the end of this section, but first we will discuss a special subspace of $\mathbb{R}^m$ related to *non*homogeneous systems with coefficient matrix A.

20.3.1 Column space

The "consistent space" of a coefficient matrix. The solution set of a nonhomogeneous system $A\mathbf{x} = \mathbf{b}$ with $m \times n$ coefficient matrix A cannot be a subspace of $\mathbb{R}^n$ because it can never contain the zero vector. Even worse, if the system is inconsistent, then the solution set does not contain any vectors at all.

Question 20.3.1 Amongst *all* systems with coefficient matrix A, which are consistent? □

We know that the homogeneous system $A\mathbf{x} = \mathbf{0}$ is consistent because it has at least the trivial solution. But for what other vectors of $\mathbf{b}$ besides $\mathbf{b} = \mathbf{0}$ is the system $A\mathbf{x} = \mathbf{b}$ consistent? It is possible to verify directly that the collection of all such $\mathbf{b}$ vectors is a subspace of $\mathbb{R}^m$.

Check your understanding. Apply the Subspace Test to verify that for a given $m \times n$ matrix A, the collection of all vectors $\mathbf{b}$ in $\mathbb{R}^m$ for which the system $A\mathbf{x} = \mathbf{b}$ is consistent forms a subspace of $\mathbb{R}^m$.

Until we know more about it, for now let's refer to this subspace of $\mathbb{R}^m$ as the **consistent space** of A.

Consistent space versus column space. To better understand this so-called consistent space, we should relate it back to the matrix A as we did in Discovery 20.2, because A is the only thing common to all the $\mathbf{b}$ vectors in this space. Let's again think of A as being made up of column vectors in $\mathbb{R}^m$:

$$A = \begin{bmatrix} | & | & & | \\ \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_n \\ | & | & & | \end{bmatrix}.$$

In Discovery 20.1, we found that the result of computing $A\mathbf{e}_j$ is $\mathbf{c}_j$, the j^{th} column of A (where $\mathbf{e}_j$ is the j^{th} standard basis vector in $\mathbb{R}^n$, as usual). But this says that each system $A\mathbf{x} = \mathbf{c}_j$ is consistent, since there is at least one solution $\mathbf{x} = \mathbf{e}_j$. Therefore, each of the columns of A is in the consistent space of A. And because the span of these columns is the *smallest* subspace that contains each of them, we can conclude that every vector in the column space $\operatorname{Span}\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n\}$ of A is also in the consistent space of A.

What other vectors could be in this space? If $A\mathbf{x} = \mathbf{b}$ is consistent, then it has at least one solution

$$\mathbf{x} = \mathbf{x}_0 = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n.$$

But then

$$\begin{aligned} \mathbf{b} &= A\mathbf{x}_0 \\ &= A(x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n) \\ &= x_1 A\mathbf{e}_1 + x_2 A\mathbf{e}_2 + \cdots + x_n A\mathbf{e}_n \\ &= x_1\mathbf{c}_1 + x_2\mathbf{c}_2 + \cdots + x_n\mathbf{c}_n. \end{aligned} \tag{$*$}$$

So whenever $A\mathbf{x} = \mathbf{b}$ is consistent, we find that $\mathbf{b}$ is equal to some linear combination of the columns of A (with coefficients taken from the components of a solution vector). In other words, every vector in the consistent space of A is also in the column space of A. So the two spaces are equal: ***the system*** $A\mathbf{x} = \mathbf{b}$ ***is consistent when, and only when, the vector of constants*** $\mathbf{b}$ ***is in the column space of*** A.

Determining a basis for a column space. Since the columns of A are, by definition, a spanning set for the column space of A, we can reduce it to a basis. Once again, we can apply row reduction to this task. Row reducing is equivalent to multiplying on the left by elementary matrices, and when we defined matrix multiplication in Subsection 4.3.7 we did so column-by-column:

$$EA = \begin{bmatrix} | & | & & | \\ E\mathbf{c}_1 & E\mathbf{c}_2 & \cdots & E\mathbf{c}_n \\ | & | & & | \end{bmatrix}.$$

Because matrix multiplication distributes over linear combinations, ***multiplying a collection of column vectors by a common matrix cannot create independence out of dependence***. Even better, the process of row reducing can be reversed (i.e. we are multiplying by *invertible* matrices), so it follows that ***multiplying a collection of column vectors by an*** **invertible** ***common matrix cannot create dependence out of independence***.

See. Proposition 20.5.1 in Subsection 20.5.1.

As we partially reasoned in Discovery 20.3, this means that we can recognize independence/dependence relationships amongst the columns of A from the independence/dependence relationships amongst the columns of reduced forms of A, leading to the following procedure.

Procedure 20.3.2 To determine a basis for the column space of matrix A.

1. *Reduce to at least REF.*
2. *Extract from A all those columns in positions corresponding to the positions of the leading ones in the reduced matrix.*

These extracted columns will form a basis for the column space of A.

Remark 20.3.3 It is important that you ***take the basis vectors from the columns of*** A, not from the columns of the reduced matrix — row operations do not change independence/dependence relationships amongst the columns, but they *do* change the column space.

Using column space to determine linear dependence/independence. In Discovery 20.4.a, we used this new procedure to create a reinterpretation of the Test for Linear Dependence/Independence for vectors in $\mathbb{R}^m$.

Procedure 20.3.4 To use row reduction to test linear dependence/ independence in $\mathbb{R}^m$. *To determine whether a collection of n vectors in* $\mathbb{R}^m$ *is linearly dependent or independent, form an* $m \times n$ *matrix by using the vectors as* columns, *and then row reduce to determine the rank of the matrix. If the rank is equal to n (i.e. there is a leading one in every column of the reduced matrix), then the vectors are linearly independent. If the rank is less that n (i.e. at least one column of the reduced matrix does not contain a leading one), then the vectors are linearly dependent.*

Note that this isn't really a new version of the Test for Linear Dependence/ Independence, it's just a shortcut — if we were to use the full test, the column vectors we are testing would appear as the columns of the coefficient matrix for the homogeneous system created by the test. (See Example 17.4.1, and the other examples in Subsection 17.4.1.)

In Discovery 20.4.b and Discovery 20.4.c, we also used this new procedure to connect column space to invertibility for a square matrix. We will summarize these new facts in Subsection 20.5.3.

20.3.2 Row space

Analyzing the row space of a matrix is considerably easier. As we discovered in Discovery 20.5 and Discovery 20.6, elementary row operations do not change the row space of a matrix, so the row spaces of a matrix and each of its REFs are the same space. Clearly we do not need the zero rows from an REF to span this space. But the pattern of leading ones guarantees that the nonzero rows in an REF are linearly independent.

See. Corollary 20.5.4 in Subsection 20.5.2.

Procedure 20.3.5 To determine a basis for the row space of matrix A.

1. *Reduce to at least REF.*
2. *Extract the nonzero rows from the REF you have computed.*

These extracted rows will form a basis for the row space of A.

Remark 20.3.6 Note the difference from the column space procedure — in this procedure we get the basis vectors from the *reduced* matrix, not from the original matrix.

We can also use the row space procedure to test vectors for linear independence.

Procedure 20.3.7 A second way to use row reduction to test linear dependence/independence in $\mathbb{R}^m$. *To determine whether a collection of m vectors in $\mathbb{R}^n$ is linearly dependent or independent, form an $m \times n$ matrix by using the vectors as* rows*, and then row reduce to determine the rank of the matrix. If the rank is equal to m (i.e. no zero rows can be produced by reducing), then the vectors are linearly independent. If the rank is less that n (i.e. reducing produces at least one zero row), then the vectors are linearly dependent.*

20.3.3 Column space versus row space

Question 20.3.8 Which procedure — column space or row space — should we use? □

When testing vectors from $\mathbb{R}^n$ for linear independence, we clearly have a choice of whether to form a matrix using those vectors as columns or as rows. But we also have a choice when computing a basis for either type of space, because the column space of a matrix is the same as the row space of the transpose, and the row space of a matrix is the same as the column space of the transpose.

In Discovery 20.7, you were asked to think about this question. You might have considered the end results of the two procedures to determine the pros and cons of one procedure over the other.

Column space

Produces a basis involving vectors from the original collection.

Row space

Produces a "simplified" basis.

In the column space procedure, we always go back to the original matrix to pick out certain columns. So, this procedure effectively performs the task of reducing a spanning set down to a basis, a task that we knew *could* be done (Proposition 18.5.1) but didn't have a systematic means of carrying out. In the row space procedure, we take our basis vectors from the simplified nonzero rows of an REF for the matrix. Because the leading one in each row is in a different position, expressing other vectors in the space as linear combinations of these basis vectors is much more straightforward than it is in general. In fact, if you have taken a basis for the row space from the RREF, expressing other vectors in the space as linear combinations of these basis vectors can be done by inspection.

20.3.4 Null space and the dimensions of the three spaces

We have already seen through examples in Subsection 19.4.1 how to extract a basis for the solution space for a homogeneous system $A\mathbf{x} = \mathbf{0}$ (now called the **null space** of A) from the parameters assigned after row reducing.

The null space of A doesn't just represent the set of solutions to the homogeneous system — Lemma 4.5.4 tells us that it represents most of the data we need in order to know the solution set of every system that has A as a coefficient matrix. If we know one specific solution $\mathbf{x} = \mathbf{x}_1$ to nonhomogeneous system $A\mathbf{x} = \mathbf{b}$, then every other solution can be obtained by adding to $\mathbf{x}_1$ a vector from the null space. Geometrically, this represents a translation of the null space away from the origin, like a plane that is translated away from the origin by an “initial” point $\mathbf{x}_1$.

All three spaces — column, row, and null — are connected through the RREF of the matrix. For column space, we get a basis vector for each leading one in the RREF. For row space, we get a basis vector for each nonzero row in the RREF, and a row in the RREF is nonzero precisely when it contains a leading one. So even though column space is a subspace of $\mathbb{R}^m$ and row space is a subspace of $\mathbb{R}^n$ (where A is an $m \times n$ matrix), ***the column space and the row space of A have the same dimension***, and this dimension is equal to the rank of A. On the other hand, for the null space we get one basis vector for each parameter required to solve the homogeneous system. Parameters are assigned to free variables, and free variables are those whose columns *do not* contain a leading one. So the dimension of the null space is equal to the difference between the number of columns of A and the rank of A, which is just a more sophisticated way to state Proposition 2.5.8.

20.4 Examples

In this section.

- Subsection 20.4.1 *The three spaces*
- Subsection 20.4.2 *Enlarging a linearly independent set*

20.4.1 The three spaces

We will do an example column space, row space, and null space, all in one example.

Consider the 4 × 5 matrix

$$A = \begin{bmatrix} -8 & 9 & 11 & 7 & 5 \\ 1 & -1 & -1 & -3 & 6 \\ -2 & 2 & 2 & 5 & -9 \\ 1 & -1 & -1 & 1 & -6 \end{bmatrix}.$$

Row reduce, as usual:

$$\begin{bmatrix} -8 & 9 & 11 & 7 & 5 \\ 1 & -1 & -1 & -3 & 6 \\ -2 & 2 & 2 & 5 & -9 \\ 1 & -1 & -1 & 1 & -6 \end{bmatrix} \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \begin{bmatrix} 1 & 0 & 2 & 0 & -1 \\ 0 & 1 & 3 & 0 & 2 \\ 0 & 0 & 0 & 1 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Column space of A. From the positions of the leading ones in the reduced matrix, we see that the first, second, and fourth columns of A are linearly independent, so a basis for the column space of A is

$$\mathcal{B}_{\text{col}} = \left\{ \begin{bmatrix} -8 \\ 1 \\ -2 \\ 1 \end{bmatrix}, \begin{bmatrix} 9 \\ -1 \\ 2 \\ -1 \end{bmatrix}, \begin{bmatrix} 7 \\ -3 \\ 5 \\ 1 \end{bmatrix} \right\},$$

and the dimension of the column space of A is 3.

We can also see from the reduced matrix the exact dependence relationships between the columns of A. In the reduced matrix, the leading-one columns are the first three standard basis vectors, and we can easily see how the third and fifth columns can be decomposed as linear combinations of these standard basis vectors. In A, the third and fifth columns can be decomposed in the exact same way as linear combinations of the vectors in $\mathcal{B}_{\text{col}}$. If we label the columns of A as $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4, \mathbf{c}_5$, then we have

$$\mathbf{c}_3 = 2\mathbf{c}_1 + 3\mathbf{c}_2, \qquad\qquad \mathbf{c}_5 = (-1)\mathbf{c}_1 + 2\mathbf{c}_2 + (-3)\mathbf{c}_4.$$

Row space of A. The leading ones guarantee that the nonzero rows in the reduced matrix are linearly independent. Since row reducing does not change the row space, we get our basis for the row space of A from the reduced matrix:

$$\mathcal{B}_{\text{row}} = \{\begin{bmatrix} 1 & 0 & 2 & 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 3 & 0 & 2 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 1 & -3 \end{bmatrix}\}.$$

The dimension of the row space of A is again 3.

Null space of A. Finally, for the null space of A we solve the homogeneous system as usual. The third and fifth columns represent free variables, so we set parameters $x_3 = s$ and $x_5 = t$. Solving for the remaining variables leads to a general solution in parametric form

$$x_1 = -2s + t, \qquad x_2 = -3s - 2t, \qquad x_3 = s, \qquad x_4 = 3t, \qquad x_5 = t.$$

In vector form, we have

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -2s+t \\ -3s-2t \\ s \\ 3t \\ t \end{bmatrix} = \begin{bmatrix} -2s \\ -3s \\ s \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} t \\ 2t \\ 0 \\ 3t \\ t \end{bmatrix} = s \begin{bmatrix} -2 \\ -3 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} 1 \\ 2 \\ 0 \\ 3 \\ 1 \end{bmatrix}.$$

So a basis for the null space of A is

$$\mathcal{B}_{\text{null}} = \left\{ \begin{bmatrix} -2 \\ -3 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 0 \\ 3 \\ 1 \end{bmatrix} \right\},$$

and the dimension of the null space is 2.

20.4.2 Enlarging a linearly independent set

Row space is also a convenient tool for enlarging a linearly independent set into a basis. Here are two examples of carrying out this task, one using vectors in $\mathbb{R}^n$, and one using vectors in another space, where we use the associated coordinate vectors in $\mathbb{R}^n$ to assist us.

Example 20.4.1 Using row space to enlarge a linearly independent set in $\mathbb{R}^4$. Suppose we would like to take the linearly independent set

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

of vectors in $\mathbb{R}^4$ and enlarge it into a basis for all of $\mathbb{R}^4$. Since $\dim \mathbb{R}^4 = 4$, we need two more vectors.

Using Proposition 17.5.6, we can start by determining a vector that is not in the subspace $U = \operatorname{Span}\{\mathbf{v}_1, \mathbf{v}_2\}$, where $\mathbf{v}_1, \mathbf{v}_2$ are the two given vectors. However, guess-and-check is not a very efficient method for doing this. Instead, let's set up a matrix with $\mathbf{v}_1$ and $\mathbf{v}_2$ as rows, so that U is precisely the row space of that matrix. We can then use row reduction to determine a simpler basis for U:

$$\begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & 6 & 1 & 1 \end{bmatrix} \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \begin{bmatrix} 1 & 3 & 0 & \frac{2}{3} \\ 0 & 0 & 1 & -\frac{1}{3} \end{bmatrix}.$$

We can see from the pattern of leading ones in the reduced matrix that to span all of $\mathbb{R}^4$, we need to introduce some "independence" in the second and fourth coordinates. So let's try enlarging our initial set of vectors by the second and fourth standard basis vectors:

$$\begin{bmatrix} 1 & 3 & 2 & 0 \\ 2 & 6 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The rows of the reduced matrix are the four standard basis vectors for $\mathbb{R}^4$, hence the row space of the reduced matrix is all of $\mathbb{R}^4$. We know that row operations do not change row space, so the rows of the initial matrix must also span all of $\mathbb{R}^4$. Since we have a spanning set for a dimension-4 space consisting of four vectors, those four vectors must for a basis for the space. □

Example 20.4.2 Using row space to enlarge a linearly independent set in $\mathrm{M}_2(\mathbb{R})$**.** Suppose we would like to take the linearly independent set

$$\left\{ \begin{bmatrix} 1 & 3 \\ 2 & 0 \end{bmatrix}, \begin{bmatrix} 2 & 6 \\ 1 & 1 \end{bmatrix} \right\}$$

of vectors in $\mathrm{M}_2(\mathbb{R})$ and enlarge it into a basis for all of $\mathrm{M}_2(\mathbb{R})$. Since $\dim \mathrm{M}_2(\mathbb{R}) = 4$, we need two more vectors. Now, *we cannot row reduce the given matrices* — that would be meaningless, as these matrices are not made of row vectors or column vectors, they are *themselves* vectors. However, we can get back to the land of row vectors by using coordinate vectors relative to the standard basis $\mathcal{S}$ for $\mathrm{M}_2(\mathbb{R})$:

$$(\mathbf{v}_1)_{\mathcal{S}} = (1,3,2,0), \qquad\qquad (\mathbf{v}_2)_{\mathcal{S}} = (2,6,1,1),$$

where $\mathbf{v}_1, \mathbf{v}_2$ are the two given vectors. These coordinate vectors are precisely the vectors from Example 20.4.1 above, so using those results we expect that we should be able to enlarge our basis using vectors $\mathbf{v}_3$ and $\mathbf{v}_4$ that have coordinate vectors

$$(\mathbf{v}_3)_{\mathcal{S}} = (0,1,0,0), \qquad\qquad (\mathbf{v}_4)_{\mathcal{S}} = (0,0,0,1).$$

Thus, we can enlarge the initial set of vectors to the basis

$$\left\{ \begin{bmatrix} 1 & 3 \\ 2 & 0 \end{bmatrix}, \begin{bmatrix} 2 & 6 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\}$$

for $\mathrm{M}_2(\mathbb{R})$. □

20.5 Theory

In this section.

20.5.1 Column space

First we'll record two facts concerning how multiplying each in a set of column vectors in $\mathbb{R}^n$ by a common matrix affects linear dependence and independence, leading to our conclusion about how to determine a basis for the column space of a matrix from examining its RREF.

Proposition 20.5.1 Dependence/independence versus matrix transformation. *Suppose $\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_\ell$ are column vectors in $\mathbb{R}^n$ and E is an $m \times n$ matrix.*

1. *If $\{E\mathbf{v}_1,E\mathbf{v}_2,\dots,E\mathbf{v}_\ell\}$ is a linearly independent set of vectors then so too is $\{\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_\ell\}$.*
2. *If E is square and invertible and $\{\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_\ell\}$ is a linearly independent set, then so too is $\{E\mathbf{v}_1,E\mathbf{v}_2,\dots,E\mathbf{v}_\ell\}$.*
3. *If E is square and invertible and $\mathbf{w}$ is another column vector in $\mathbb{R}^n$ so that vector $E\mathbf{w}$ is linearly dependent with the vectors $E\mathbf{v}_1,E\mathbf{v}_2,\dots,E\mathbf{v}_\ell$ by the dependence relation*

   $$E\mathbf{w} = k_1E\mathbf{v}_1 + k_2E\mathbf{v}_2 + \cdots + k_\ell E\mathbf{v}_\ell,$$

   then $\mathbf{w}$ is linearly dependent with $\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_\ell$ by a dependence relation involving the same scalars,

   $$\mathbf{w} = k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + \cdots + k_\ell\mathbf{v}_\ell.$$

Proof of Statement 1. Let's apply the Test for Linear Dependence/Independence to the vectors $\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_\ell$: suppose that $k_1,k_2,\dots,k_\ell$ are scalars so that

$$k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + \ldots + k_\ell\mathbf{v}_\ell = \mathbf{0}. \tag{$*$}$$

Multiplying both sides of this equation by the matrix E, and using some matrix algebra, we get

$$k_1E\mathbf{v}_1 + k_2E\mathbf{v}_2 + \ldots + k_\ell E\mathbf{v}_\ell = E\mathbf{0} = \mathbf{0}.$$

But we have assumed that the vectors $E\mathbf{v}_1,E\mathbf{v}_2,\dots,E\mathbf{v}_\ell$ are linearly independent, so the only way this linear combination could equal the zero vector is if all the scalars $k_1,k_2,\dots,k_\ell$ are zero. Thus, the linear combination in $(*)$ must be the trivial one, and so the vectors $\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_\ell$ are linearly independent. ■

Proof of Statement 2. We assume $\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_\ell$ are linearly independent. Since we also assume E to be invertible, we can restate this as saying that vectors

$$E^{-1}(E\mathbf{v}_1),E^{-1}(E\mathbf{v}_2),\dots,E^{-1}(E\mathbf{v}_\ell)$$

are linearly independent. Now we can apply Statement 1 with E replaced by E^{-1} and $\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_\ell$ replaced by $E\mathbf{v}_1,E\mathbf{v}_2,\dots,E\mathbf{v}_\ell$. ■

Proof of Statement 3. Simply apply the inverse E^{-1} to both sides of

$$E\mathbf{w} = k_1 E\mathbf{v}_1 + k_2 E\mathbf{v}_2 + \cdots + k_\ell E\mathbf{v}_\ell$$

to obtain

$$\mathbf{w} = k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + \cdots + k_\ell\mathbf{v}_\ell.$$

■

Corollary 20.5.2 Column space basis and dimension.

1. *A basis for the column space of an $m \times n$ matrix A can be formed from those columns of A (as column vectors in $\mathbb{R}^m$) in positions corresponding to the locations of the leading ones in the RREF of A.*

2. *The dimension of the column space of a matrix is equal to the rank of the matrix.*

Proof of Statement 1. By definition, the columns of A are a spanning set for the column space of A. By Proposition 18.5.1, this spanning set can be reduced to a basis; it's a matter of determining the largest possible linearly independent set of these spanning column vectors.

Let $E = E_t E_{t-1} \cdots E_1$ be the product of elementary matrices corresponding to some sequence of row operations that reduces A to its RREF. Because of the nature of RREF, each column of RREF(A) that contains a leading one will be a standard basis vector in $\mathbb{R}^m$, no two such leading-one columns will be the same standard basis vector, and each column that does not contain a leading one will be a linear combination of those leading-one columns that appear to its left. Therefore, the leading-one columns represent the largest set of linearly independent vectors that can be formed from the columns of RREF(A). Since E is invertible, the two statements of Proposition 20.5.1 tell us that the columns of A will have the same relationships: those columns in A that are in positions where the leading ones occur in RREF(A) will be linearly independent, and that will be the largest possible collection of linearly independent columns of A, because each of the other columns will be linearly dependent with the leading-one-position columns of A to its left.

Thus, we can reduce the spanning set made up of all columns of A to a basis for its column space by discarding the linearly dependent columns and keeping only those columns in positions corresponding to the locations of the leading ones occur in RREF(A). ■

Proof of Statement 2. Since we obtain a basis for column space by taking those columns in the matrix in positions corresponding to the leading ones in a reduced form for the matrix, the number of basis vectors is equal to the number of leading ones. ■

20.5.2 Row space

Next we'll record our observations concerning how the row operations affect the row space of a matrix, leading to our conclusion about how to obtain a basis for the row space of a matrix from its RREF.

Proposition 20.5.3 Row space versus row operations. *If an elementary row operation is applied to a matrix, then the row space of the new matrix is the same as the row space of the old matrix.*

Proof. Consider an $m \times n$ matrix A as a collection of row vectors in $\mathbb{R}^n$:

$$A = \begin{bmatrix} — & \mathbf{a}_1 & — \\ — & \mathbf{a}_2 & — \\ & \vdots & \\ — & \mathbf{a}_m & — \end{bmatrix}.$$

Then the row space of A is, by definition, the subspace $\operatorname{Span}\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m\}$ of $\mathbb{R}^m$.

As we did in Discovery 20.5, we will make repeated use of Statement 2 of Proposition 16.5.6, which tells us how to determine when two spanning sets generate the same subspace.

Let's consider each type of elementary row operation in turn.

(i) Suppose we swap two rows in A:

$$A = \begin{bmatrix} — & \mathbf{a}_1 & — \\ & \vdots & \\ — & \mathbf{a}_i & — \\ & \vdots & \\ — & \mathbf{a}_j & — \\ & \vdots & \\ — & \mathbf{a}_m & — \end{bmatrix} \to A' = \begin{bmatrix} — & \mathbf{a}_1 & — \\ & \vdots & \\ — & \mathbf{a}_j & — \\ & \vdots & \\ — & \mathbf{a}_i & — \\ & \vdots & \\ — & \mathbf{a}_m & — \end{bmatrix}.$$

The row space of the new matrix, A', is the span of its row vectors. But every row vector in A' is equal to one of the row vectors in A, and vice versa. So clearly the conditions of the above-referenced Statement 2 are satisfied, and the rowspaces of the two matrices are the same space.

(ii) Suppose we multiply one of the rows in A by a nonzero constant k:

$$A = \begin{bmatrix} — & \mathbf{a}_1 & — \\ & \vdots & \\ — & \mathbf{a}_i & — \\ & \vdots & \\ — & \mathbf{a}_m & — \end{bmatrix} \to A'' = \begin{bmatrix} — & \mathbf{a}_1 & — \\ & \vdots & \\ — & k\mathbf{a}_i & — \\ & \vdots & \\ — & \mathbf{a}_m & — \end{bmatrix}.$$

Again, most of the row vectors in the new matrix A'' are equal to one of the row vectors in A, and vice versa. So to fully satisfy the conditions of the above-referenced Statement 2, we need to verify that $k\mathbf{a}_i$ is somehow a linear combination of row vectors from A and that $\mathbf{a}_i$ is somehow a linear combination of row vectors from A''. But $k\mathbf{a}_i$ is already expressed as a scalar multiple of a row vector from A, and since k is nonzero we can also write

$$\mathbf{a}_i = \frac{1}{k} \cdot (k\mathbf{a}_i),$$

so that $\mathbf{a}_i$ is also a scalar multiple of a row vector from A''.

With the conditions of the above-referenced Statement 2 now fully satisfied, we can conclude that the rowspaces of the two matrices are the same space.

(iii) Suppose we replace one row vector in A by the sum of that row and a scalar multiple of another:

$$A = \begin{bmatrix} — & \mathbf{a}_1 & — \\ & \vdots & \\ — & \mathbf{a}_i & — \\ & \vdots & \\ — & \mathbf{a}_m & — \end{bmatrix} \to A''' = \begin{bmatrix} — & \mathbf{a}_1 & — \\ & \vdots & \\ — & \mathbf{a}_i + k\mathbf{a}_j & — \\ & \vdots & \\ — & \mathbf{a}_m & — \end{bmatrix}.$$

Once again, most of the row vectors in the new matrix A''' are equal to one of the row vectors in A, and vice versa. So to fully satisfy the conditions of the above-referenced Statement 2, we need to verify that $\mathbf{a}_i + k\mathbf{a}_j$ is somehow a linear combination of row vectors from A and that $\mathbf{a}_i$ is somehow a linear combination of row vectors from A'''. But $\mathbf{a}_i + k\mathbf{a}_j$ is already expressed as a linear combination of row vectors from A''', and for $\mathbf{a}_i$ we can write

$$\mathbf{a}_i = 1(\mathbf{a}_i + k\mathbf{a}_j) + (-k)\mathbf{a}_j,$$

a linear combination of row vectors from A'''.

Note. Row vector $\mathbf{a}_j$ has not been modified in the row operation, and so is a row vector for both A and A'''.

With the conditions of the above-referenced Statement 2 now fully satisfied, we can conclude that the rowspaces of the two matrices are the same space.

■

Corollary 20.5.4 Row space basis and dimension. *Let A represent a matrix.*

1. *If E is an invertible square matrix of compatible size, then A and EA have the same row space.*
2. *The row space of each REF for A (including the RREF of A) is always the same as that of A.*
3. *The nonzero rows of each REF for A form a basis for the row space of A.*
4. *The dimension of the row space of A is equal to the rank of A.*

Proof of Statement 1. Since E is invertible, it can be expressed as a product of elementary matrices (Theorem 6.5.2), and the product EA has the same result as applying to A the sequence of row operations represented by those elementary matrices. But Proposition 20.5.3 tells us that applying those operations does not change the row space. ■

Proof of Statement 2. Let F be an REF for A, and let $E_1, E_2, \ldots, E_\ell$ be elementary matrices corresponding to some sequence of row operations that reduces A to F. Set $E = E_\ell \cdots E_2 E_1$. Then E is an invertible matrix and $F = EA$. Therefore, F has the same row space as A by Statement 1 of this corollary. ■

Proof of Statement 3. Let F be an REF for A. By Statement 2 of this corollary, the rows of F are a spanning set for the row space of A. Clearly we can discard any zero rows from this spanning set, so it just remains to verify that the nonzero rows of F are linearly independent. For this, we will use Proposition 17.5.6, building up our linearly independent spanning set one vector at a time. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_\ell$ represent the nonzero rows of F, from top to bottom. Start with $\mathbf{v}_\ell$; all by itself, this one nonzero vector is linearly independent. Now, $\mathbf{v}_{\ell-1}$ cannot be in $\operatorname{Span}\{\mathbf{v}_\ell\}$, because the leading one in $\mathbf{v}_{\ell-1}$ appears to the left of the leading one in $\mathbf{v}_\ell$, and so no scalar multiple of $\mathbf{v}_\ell$ will have a nonzero entry in the component where $\mathbf{v}_{\ell-1}$ has its leading one. From this, Proposition 17.5.6 tells us that $\{\mathbf{v}_{\ell-1}, \mathbf{v}_\ell\}$ is linearly independent. Moving on, $\mathbf{v}_{\ell-2}$ cannot be in $\operatorname{Span}\{\mathbf{v}_{\ell-1}, \mathbf{v}_\ell\}$, because the leading one in $\mathbf{v}_{\ell-2}$ appears to the left of both the leading one in $\mathbf{v}_{\ell-1}$ and in $\mathbf{v}_\ell$, and so no linear combination of those two vectors will have a nonzero entry in the component where $\mathbf{v}_{\ell-2}$ has its leading one. From this, Proposition 17.5.6 tells us that $\{\mathbf{v}_{\ell-2}, \mathbf{v}_{\ell-1}, \mathbf{v}_\ell\}$ is linearly independent. Repeating this argument as we move up the rows of F, we see that the nonzero rows of F are linearly independent when taken altogether. ■

Proof of Statement 4. Applying Statement 3 of this corollary to the RREF for A, the nonzero rows of RREF(A) form a basis for the row space of A. But the nonzero rows of RREF(A) must all contain leading ones, so the number of vectors in a basis for the row space of A is equal to the number of leading ones in RREF(A), as desired. ■

20.5.3 Column and row spaces versus rank and invertibility

As discovered in Discovery 20.4, we can use our observations recorded in Proposition 20.5.1 to connect column space to invertibility. We can similarly use Corollary 20.5.4 to also connect row space to invertibility.

First, we will extend the list of properties that are equivalent to invertibility of a square matrix, first started in Theorem 6.5.2, and then continued in Theorem 10.5.3.

Theorem 20.5.5 Characterizations of invertibility. *For a square matrix A, the following are equivalent.*

1. *Matrix A is invertible.*
2. *Every linear system that has A as a coefficient matrix has one unique solution.*
3. *The homogeneous system $A\mathbf{x} = \mathbf{0}$ has only the trivial solution.*
4. *There is some linear system that has A as a coefficient matrix and has one unique solution.*
5. *The rank of A is equal to the size of A.*
6. *The RREF of A is the identity.*
7. *Matrix A can be expressed as a product of some number of elementary matrices.*
8. *The determinant of A is nonzero.*
9. *The null space of A consists of only the zero vector.*
10. *The columns of A are linearly independent.*
11. *The columns of A form a basis for $\mathbb{R}^n$, where n is the size of A.*
12. *The rows of A are linearly independent.*
13. *The rows of A form a basis for $\mathbb{R}^n$, where n is the size of A.*

In particular, an $n \times n$ matrix is invertible if and only if its columns form a basis for $\mathbb{R}^n$.

Proof. We have previously encountered the equivalence of many of these statements, most recently in Theorem 10.5.3. So currently we only need to concern ourselves with the new statements. For each of these, if we can establish equivalence of the new statement to *one* of the old, then the new statement must be equivalent to *all* of the old, by the transitivity of logical equivalence.

Statement 9. This is just restatement of Statement 3 using the concept of **null space**.

Statement 10. From our reinterpretation of Proposition 17.5.1, stated in Procedure 20.3.4, we know that *all* of the columns of A will be linearly independent if and only if every column of RREF(A) has a leading one. Therefore, this statement is equivalent to Statement 5.

Statement 11. This statement is equivalent to Statement 10, since Proposition 19.5.5 tells us that we need exactly n linearly independent vectors to form a basis for $\mathbb{R}^n$.

Statement 12. From the row space version of the Test for Linear Dependence/ Independence stated in Procedure 20.3.7, we know that *all* of the rows of A will be linearly independent if and only if every row of RREF(A) is nonzero. Therefore, this statement is also equivalent to Statement 5.

Statement 13. This statement is equivalent to Statement 12, again since Proposition 19.5.5 tells us that we need exactly n linearly independent vectors to form a basis for $\mathbb{R}^n$. ■

Finally, we'll record an observation from Discovery 20.9, which is just a reframing of Proposition 2.5.8.

Theorem 20.5.6 Rank-Nullity Theorem. *If A is an $m \times n$ matrix, then*

$$n = \operatorname{rank}(A) + \operatorname{nullity}(A).$$

That is,

$$\dim \mathbb{R}^n = \dim(\textit{column space of } A) + \dim(\textit{null space of } A).$$

Note. *The two spaces referenced in this theorem are connected through the matrix A, but may be subspaces of* different *vector spaces — the column space of A is a subspace of $\mathbb{R}^m$, while the null space is a subspace of $\mathbb{R}^n$.*

Proof. The dimension of the column space of A is equal to the number of leading ones in its RREF, while the dimension of the null space of A is equal to the number of free variables, which is equal to the number of columns in the RREF that do *not* have a leading one. These two numbers must add up to the total number of columns in A. ■

Part III

Introduction to Matrix Forms

CHAPTER 21

Eigenvalues and eigenvectors

21.1 Discovery guide

In Chapter 20, we began to see how the interaction between a matrix and column vectors can be used to understand the matrix. Here we will find that for each square matrix there are certain column vectors that are particularly well-suited to the task.

Discovery 21.1 Consider the matrix and column vectors

$$A = \begin{bmatrix} 7 & 8 \\ -4 & -5 \end{bmatrix}, \qquad \mathbf{u} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \qquad \mathbf{v} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}.$$

(a) Compute $A\mathbf{u}$. Carefully compare vectors $\mathbf{u}$ and $A\mathbf{u}$ — what do you notice? Now repeat for $\mathbf{v}$ and $A\mathbf{v}$.

(b) Verify that $\{\mathbf{u}, \mathbf{v}\}$ is a basis for $\mathbb{R}^2$.

Hint. Corollary 19.5.6.

(c) Because these vectors form a basis for $\mathbb{R}^2$, every vector in $\mathbb{R}^2$ can be expressed in one unique way as a linear combination of these basis vectors. We can use this fact, along with some matrix algebra and the patterns you noticed in Task a, to develop a simple way to compute products $A\mathbf{x}$ without actually performing matrix multiplication:

$$\mathbf{x} = a\mathbf{u} + b\mathbf{v} \qquad \Longrightarrow \qquad A\mathbf{x} = \underline{\hspace{10em}}.$$

From Discovery 21.1, it seems that pairs consisting of a scalar λ and (nonzero) vector $\mathbf{x}$ such that $A\mathbf{x} = \lambda\mathbf{x}$ are important to understanding how matrix A "operates" on *all* vectors by multiplication. For such a pair, the scalar λ is called an **eigenvalue** of A, and the corresponding vector $\mathbf{x}$ is called an **eigenvector** for A.

Notation and terminology. The symbol λ is the Greek letter *lambda*. The prefix *eigen* is German for specific/particular/"one's own."

It turns out that it is easier to determine potential eigenvalues for a matrix first, and to look for corresponding eigenvectors afterwards. In the next discovery activity we will develop a method to determine all eigenvalues of a matrix, independently of determining eigenvectors.

Discovery 21.2 For λ to be an eigenvalue for A, there must be at least one *nontrivial* solution $\mathbf{x}$ to the matrix equation $A\mathbf{x} = \lambda\mathbf{x}$.

(a) Use matrix algebra to turn the equation $A\mathbf{x} = \lambda\mathbf{x}$ into a homogeneous condition: $(\qquad)\,\mathbf{x} = \mathbf{0}$.

Careful. Make sure what you have in the brackets represents a *matrix*!

(b) We want *nontrivial* solutions to exist. Combine some knowledge from Chapter 6 and Chapter 10 to complete the statement below.

The homogeneous system from Task a has nontrivial solutions if and only if $\det(\qquad)$ is $\quad$.

Hint. Theorem 10.5.3.

We will see that the computation of the determinant you identified in Discovery 21.2.b always results in a degree n polynomial in the variable λ, where n is the size of the matrix. We will call this polynomial the **characteristic polynomial** of A. The eigenvalues of A are then precisely the roots of its characteristic polynomial.

Discovery 21.3 For each of the following matrices, compute its characteristic polynomial, and then use it to determine the eigenvalues of each matrix. Make sure to write your eigenvalue answers down, you will need them in Discovery 21.6.

Algebra help. When we solve for the roots of a polynomial by hand, our main method is factoring. So when computing a characteristic polynomial, keep it in factored form as much as possible — do not expand brackets unless you need to in order to be able to collect like terms and then factor further.

(a) $\begin{bmatrix} 7 & 8 \\ -4 & -5 \end{bmatrix}$

(b) $\begin{bmatrix} 2 & -4 & 4 \\ 0 & -6 & 8 \\ 0 & -6 & 8 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$

(d) $\begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -1 \end{bmatrix}$

Compare. Check your answer for the eigenvalues of the first matrix in Discovery 21.3 with your observations in Discovery 21.1.

Discovery 21.4 Complete each statement for the special type of matrix involved.

- The eigenvalues of a diagonal matrix are .
- The eigenvalues of an upper triangular matrix are .
- The eigenvalues of a lower triangular matrix are .

Hint. Proposition 7.5.1, and Statement 1 of Proposition 8.5.2.

Once we have determined the eigenvalues of a matrix, the next step is to determine corresponding eigenvectors. We do this for one eigenvalue at a time. Fortunately, we will ultimately find ourselves in familiar territory when we go looking for eigenvectors.

Discovery 21.5 For an eigenvalue λ of a matrix A, the corresponding eigenvectors are the nonzero solutions to the homogeneous system . Therefore, if we include the zero vector in with the collection of all eigenvectors for A that correspond to a particular eigenvalue λ, this collection is a subspace of $\mathbb{R}^n$ because it is equal to the space of matrix .

Hint. Task 21.2.a.

For an eigenvalue λ of a matrix A, the subspace of $\mathbb{R}^n$ consisting of all eigenvectors of A that correspond to λ (along with the zero vector) is called the **eigenspace of A corresponding to λ**.

Discovery 21.6 For each of the matrices in Discovery 21.3, determine a basis for each eigenspace by row reducing the matrix $\lambda I - A$, assigning parameters, and extracting null space basis vectors from the general parametric solution as usual.

Note. Substitute the actual eigenvalue in for variable λ *before* row reducing — do not row reduce with the variable λ still in there.

Discovery 21.7 From the initial definition of eigenvalue/eigenvector in the paragraph following Discovery 21.1, a matrix A has $\lambda = 0$ as an eigenvalue if and only if there are nonzero solutions to $A\mathbf{x} =$.

So from our previous study of matrices, we can conclude that A has $\lambda = 0$ as an eigenvalue precisely when A is .

Hint. Theorem 6.5.2.

21.2 Terminology and notation

The following definitions are relative to a given $n \times n$ matrix A.

eigenvector
a nonzero vector $\mathbf{x}$ in $\mathbb{R}^n$ such that $A\mathbf{x}$ is a scalar multiple of $\mathbf{x}$

eigenvalue
a scalar for which there exists an eigenvector $\mathbf{x}$ of A with $A\mathbf{x} = \lambda\mathbf{x}$

Note 21.2.1 Eigenvectors and eigenvalues go together in pairs, with the connection between the two provided by the equality $A\mathbf{x} = \lambda\mathbf{x}$. In this situation, we say that the two objects **correspond** to each other. So we might say that $\mathbf{x}$ is an eigenvector of A that **corresponds** to the eigenvalue λ. Equivalently, we might say that λ is an eigenvalue of A that **corresponds** to the eigenvector $\mathbf{x}$. However, note that an eigenvalue can correspond to many eigenvectors (in fact, an infinite number of them), an eigenvector must correspond to exactly one eigenvalue.

eigenspace
the subspace of $\mathbb{R}^n$ consisting of all eigenvectors of A that correspond to a specific eigenvalue λ, along with the zero vector

$E_\lambda(A)$ notation for the eigenspace of matrix A corresponding to the eigenvalue λ

Note 21.2.2 In other resources you may seem the terms **characteristic vector**, **characteristic value**, and **characteristic space** used in place of the terminology introduced above.

characteristic polynomial
the degree-n polynomial in the variable λ obtained by computing $\det(\lambda I - A)$

$c_A(\lambda)$ notation for the characteristic polynomial of matrix A

characteristic equation
the polynomial equation $\det(\lambda I - A) = 0$

21.3 Motivation

We have seen that when considering a specific matrix A, looking for patterns in the process of computing matrix-times-column-vector helps us to understand the matrix. In turn, this helps us understand all of the various systems $A\mathbf{x} = \mathbf{b}$ with common coefficient matrix A, since obviously the left-hand side of the matrix version of the system has matrix-times-column-vector form.

When we compute $A\mathbf{e}_j$ for a standard basis vector $\mathbf{e}_j$, the result is the j^{th} column of A. So if we computed each of $A\mathbf{e}_1, A\mathbf{e}_2, \dots, A\mathbf{e}_n$, we would have all of the columns of A as the results, which contain all of the data contained in A. These computations certainly let us *know* the matrix A, but they don't necessarily help us *understand* what A is really like as a matrix. In short, the standard basis for $\mathbb{R}^n$ is a great basis for understanding the vector space $\mathbb{R}^n$, but it is not so great for helping understand matrix products $A\mathbf{x}$ for a particular matrix A.

In Discovery 21.1, we discovered that for an $n \times n$ matrix A, if we can build a basis for $\mathbb{R}^n$ consisting of eigenvectors of A, then every matrix product $A\mathbf{x}$ becomes simple to compute once $\mathbf{x}$ is decomposed as a linear combination of these basis vectors. Indeed, if $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ is a basis for $\mathbb{R}^n$, and we have

$$A\mathbf{u}_1 = \lambda_1\mathbf{u}_1, \qquad A\mathbf{u}_2 = \lambda_2\mathbf{u}_2, \qquad \dots, \qquad A\mathbf{u}_n = \lambda_n\mathbf{u}_n,$$

then multiplication by A can be achieved by scalar multiplication:

$$\mathbf{x} = k_1\mathbf{u}_1 + k_2\mathbf{u}_2 + \cdots + k_n\mathbf{u}_n$$

$$\implies \qquad \begin{aligned} A\mathbf{x} &= k_1 A\mathbf{u}_1 + k_2 A\mathbf{u}_2 + \cdots + k_n A\mathbf{u}_n \\ &= k_1\lambda_1\mathbf{u}_1 + k_2\lambda_2\mathbf{u}_2 + \cdots + k_n\lambda_n\mathbf{u}_n. \end{aligned}$$

A complete study of how the concepts of eigenvalues and eigenvectors unlock all the mysteries of a matrix is too involved to carry out in full at this point, but we will get a glimpse of how it all works for a certain kind of square matrix in the next chapter. For the remainder of this chapter, we will be more concerned with how to calculate eigenvalues and eigenvectors.

21.4 Concepts

In this section.

21.4.1 Determining eigenvalues

To determine eigenvectors and their corresponding eigenvalues for a specific matrix A, we need to solve the matrix equation $A\mathbf{x} = \lambda\mathbf{x}$ for *both* the unknown eigenvector $\mathbf{x}$ and the unknown eigenvalue λ. This is not like any matrix equation we've tried to solve before — the right-hand side involves *unknown times unknown*, making the equation *nonlinear*. However, as in Discovery 21.2, we can use some matrix algebra to turn this equation into something more familiar:

$$\begin{aligned} A\mathbf{x} &= \lambda\mathbf{x} \\ \mathbf{0} &= \lambda I\mathbf{x} - A\mathbf{x} \\ \mathbf{0} &= (\lambda I - A)\mathbf{x}. \end{aligned}$$

A particular scalar λ will be an eigenvalue of A if and only if the above homogeneous system has nontrivial solutions.

Note. The "solution" $A\mathbf{0} = \lambda\mathbf{0}$ to the original equation $A\mathbf{x} = \lambda\mathbf{x}$ is not interesting because it works for *all* values of λ.

A homogeneous system with square coefficient matrix has nontrivial solutions precisely when that coefficient matrix is *not* invertible, which is the case precisely when the determinant of that coefficient matrix is equal to zero (Theorem 10.5.3). So ***there will exist eigenvectors of*** A ***corresponding to a particular scalar*** λ ***precisely when*** λ ***is a root of the characteristic equation*** $\det(\lambda I - A) = 0$.

Procedure 21.4.1 To determine all eigenvalues of a square matrix A. *Determine the roots of the characteristic equation* $\det(\lambda I - A) = 0$.

Remark 21.4.2 Because calculating $\det(\lambda I - A)$ only involves multiplication, addition, and subtraction, its result *is* always a polynomial in the variable λ. In fact, this polynomial will always be a **monic** polynomial of degree n (where A is $n \times n$).

Terminology. A polynomial is **monic** when the coefficient on the highest power of the variable is 1.

This is the reason we moved $A\mathbf{x}$ to the right-hand side to obtain $(\lambda I - A)\mathbf{x} = \mathbf{0}$ in our algebraic manipulations above, instead of moving $\lambda\mathbf{x}$ to the left-hand side to obtain $(A - \lambda I)\mathbf{x} = \mathbf{0}$ — if we had chosen this second option, the characteristic polynomial would have a leading coefficient of ±1 depending on whether n was even or odd.

21.4.2 Eigenvalues for special forms of matrices

In Discovery 21.4, we considered the eigenvalue procedure for diagonal and triangular matrices. Suppose A is such a matrix, with values $d_1, d_2, \ldots, d_n$ down its main diagonal. Then $\lambda I - A$ is of the same special form as A (diagonal or triangular), with entries $\lambda - d_1, \lambda - d_2, \ldots, \lambda - d_n$ down its main diagonal. Since we know that the determinant of a diagonal or triangular matrix is equal to the product of its diagonal entries (Statement 1 of Proposition 8.5.2), the characteristic polynomial for A will be

$$\det(\lambda I - A) = (\lambda - d_1)(\lambda - d_2)\cdots(\lambda - d_n),$$

and so the eigenvalues of A will be precisely its diagonal entries.

21.4.3 Determining eigenvectors

Once we know all possible eigenvalues of a square matrix A, we can substitute those values into the matrix equation $A\mathbf{x} = \lambda\mathbf{x}$ one at a time. With a value for λ substituted in, this matrix equation is no longer nonlinear and can be solved for all corresponding eigenvectors $\mathbf{x}$. But the homogeneous version $(\lambda I - A)\mathbf{x} = \mathbf{0}$ is more convenient to work with, since to solve this system we just need to row reduce the coefficient matrix $\lambda I - A$.

Procedure 21.4.3 To determine all eigenvectors of a square matrix A that correspond to a specific eigenvalue λ. *Compute the matrix $C = \lambda I - A$. Then the eigenvectors corresponding to λ are precisely the nontrivial solutions of the homogeneous system $C\mathbf{x} = \mathbf{0}$, which can be solved by row reducing as usual.*

21.4.4 Eigenspaces

Determining eigenvectors is the same as solving the homogeneous system $(\lambda I - A)\mathbf{x} = \mathbf{0}$, so the eigenvectors of A corresponding to a specific eigenvalue λ are precisely the nonzero vectors in the null space of $\lambda I - A$. In particular, since a null space is a subspace of $\mathbb{R}^n$, we see that the collection of all eigenvectors of A that correspond to a specific eigenvalue λ creates a subspace of $\mathbb{R}^n$, once we also include the zero vector in the collection. This subspace is called the **eigenspace** of A for eigenvalue λ, and we write $E_\lambda(A)$ for it.

Remark 21.4.4 Since determining eigenvectors is the same as determining a null space, the typical result of carrying out Procedure 21.4.3 for a particular eigenvalue of a matrix will be to obtain a basis for the corresponding eigenspace, by row reducing, assigning parameters, and then extracting basis vectors from the general parametric solution as usual.

21.4.5 Connection to invertibility

Recall that we do not call the zero vector an eigenvector of a square matrix A, because it would not correspond to *one* specific eigenvalue — the equality $A\mathbf{0} = \lambda\mathbf{0}$ is true for *all* scalars λ. However, the *scalar* $\lambda = 0$ *can* (possibly) be an eigenvalue for a matrix A, and we explored this possibility in Discovery 21.7.

In the case of $\lambda = 0$, the matrix equation $A\mathbf{x} = \lambda\mathbf{x}$ turns into the homogeneous system $A\mathbf{x} = \mathbf{0}$. And for $\lambda = 0$ to actually be an eigenvalue of A, there needs to be nontrivial solutions to this equation — which we know will occur precisely when A is *not invertible* (Theorem 6.5.2).

21.4.6 The geometry of eigenvectors

Multiplication of column vectors by a particular matrix can be thought of as a sort of **function**, i.e. an input-output process. But unlike the types of functions you are probably used to encountering, where the input is a number x and the output is a number y, this matrix-multiplication sort of function has a *column vector* $\mathbf{x}$ as input and a *column vector* $\mathbf{y}$ as output.

When the particular matrix used to form such a function is square, then the input and output vectors live in the same space (i.e. $\mathbb{R}^n$, where n is the size of the matrix), so we can think of the matrix **transforming** an input vector into its corresponding output vector geometrically. See Figure 21.4.5 for an example of this geometric transformation point of view.

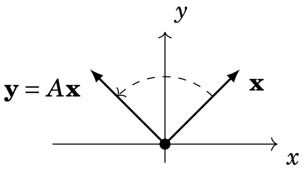
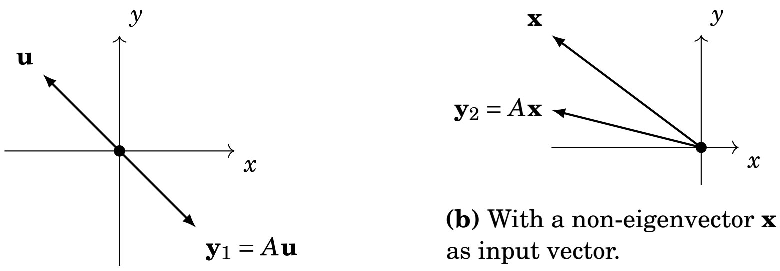


Figure 21.4.5 Example matrix function with $A = \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$ applied to input vector $\mathbf{x} = \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$ to produce output vector $\mathbf{y} = A\mathbf{x}$.

When the input vector $\mathbf{x}$ is an *eigenvector* of the transformation matrix A, then the output vector $A\mathbf{x}$ is a scalar multiple of $\mathbf{x}$ (where the scale factor is the corresponding eigenvalue). See Figure 21.4.6 for a geometric example of this view of eigenvectors.



(a) With eigenvector **u** (again from Discovery 21.1) as input vector.

(b) With a non-eigenvector **x** as input vector.

Figure 21.4.6 Two input-output examples using the same transformation matrix $A = \left[\begin{smallmatrix} 7 & 8 \\ -4 & -5 \end{smallmatrix}\right]$ (from Discovery 21.1).

Geometrically, one vector is a scalar multiple of another if and only if the two vectors are *parallel*. So we can say that ***a vector is an eigenvector of a matrix precisely when it is transformed to a parallel vector when multiplied by the matrix***.

21.5 Examples

Here we will compute eigenvalues and a basis for each corresponding eigenspace for the matrices in Discovery 21.3.

Example 21.5.1 A 2×2 **example.** From Discovery a.

First, we form the matrix

$$\lambda I - A = \begin{bmatrix} \lambda - 7 & -8 \\ 4 & \lambda + 5 \end{bmatrix}.$$

Then we compute its determinant, to obtain the characteristic polynomial of A:

$$\begin{aligned} c_A(\lambda) &= \det(\lambda I - A) \\ &= (\lambda - 7)(\lambda + 5) + 32 \\ &= \lambda^2 - 2\lambda - 3 \\ &= (\lambda + 1)(\lambda - 3). \end{aligned}$$

The eigenvalues are the roots of the characteristic polynomial, so we have two eigenvalues $\lambda_1 = -1$ and $\lambda_2 = 3$.

The eigenspace $E_{\lambda_1}(A)$ is the same as the null space of the matrix $\lambda_1 I - A$, so we determine a basis for the eigenspace by row reducing:

$$(-1)I - A = \begin{bmatrix} -8 & -8 \\ 4 & 4 \end{bmatrix} \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}.$$

This system requires one parameter to solve, as x_2 is free. Setting $x_2 = t$, the general solution in parametric form is

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -t \\ t \end{bmatrix} = t \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Associated to the single parameter we get a single basis vector, so that

$$\dim\bigl(E_{\lambda_1}(A)\bigr) = 1.$$

In particular, we have

$$E_{\lambda_1}(A) = \operatorname{Span}\left\{ \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right\}.$$

Now move on to the next eigenvalue. Again, we determine a basis for $E_{\lambda_2}(A)$ by row reducing $\lambda_2 I - A$:

$$3I - A = \begin{bmatrix} -4 & -8 \\ 4 & 8 \end{bmatrix} \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}.$$

Again, x_2 is free. One parameter means one basis vector, so again

$$\dim\bigl(E_{\lambda_2}(A)\bigr) = 1.$$

The first row of the reduced matrix says $x_1 = -2x_2$, so we have

$$E_{\lambda_2}(A) = \operatorname{Span}\left\{ \begin{bmatrix} -2 \\ 1 \end{bmatrix} \right\}.$$

□

Example 21.5.2 A 3×3 **example.** From Discovery b.

Start with

$$\lambda I - A = \begin{bmatrix} \lambda-2 & 4 & -4 \\ 0 & \lambda+6 & -8 \\ 0 & 6 & \lambda-8 \end{bmatrix},$$

and compute the characteristic polynomial,

$$\begin{aligned} c_A(\lambda) &= \det(\lambda I - A) \\ &= (\lambda-2)\big[(\lambda+6)(\lambda-8)+48\big] \\ &= (\lambda-2)(\lambda^2-2\lambda) \\ &= \lambda(\lambda-2)^2. \end{aligned}$$

The eigenvalues are $\lambda_1 = 0$ and $\lambda_2 = 2$.

The eigenspace $E_{\lambda_1}(A)$ is the null space of $0I - A = -A$, so row reduce:

$$0I - A = \begin{bmatrix} -2 & 4 & -4 \\ 0 & 6 & -8 \\ 0 & 6 & -8 \end{bmatrix} \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \begin{bmatrix} 1 & 0 & -2/3 \\ 0 & 1 & -4/3 \\ 0 & 0 & 0 \end{bmatrix}.$$

Notice that the null space of $0I - A = -A$ is the same as the null space of A, since our first step in row reducing $-A$ could be to multiply each row by -1. Since this homogeneous system has nontrivial solutions, A must be singular.

The homogeneous system $(\lambda_1 I - A)\mathbf{x} = \mathbf{0}$ requires one parameter, so

$$\dim\big(E_{\lambda_1}(A)\big) = 1.$$

The variable x_3 is free, and the nonzero rows of the reduced matrix tell us $x_1 = (2/3)x_3$ and $x_2 = (4/3)x_3$. Setting $x_3 = t$, our general solution in parametric form is

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} (2/3)t \\ (4/3)t \\ t \end{bmatrix} = t \begin{bmatrix} 2/3 \\ 4/3 \\ 1 \end{bmatrix}.$$

However, to avoid fractions in our basis vector, we may wish to pull out an additional scalar:

$$\mathbf{x} = \frac{t}{3} \begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix},$$

giving us

$$E_{\lambda_1}(A) = \operatorname{Span}\left\{ \begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix} \right\}.$$

Now row reduce $\lambda_2 I - A$:

$$2I - A = \begin{bmatrix} 0 & 4 & -4 \\ 0 & 8 & -8 \\ 0 & 6 & -6 \end{bmatrix} \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \begin{bmatrix} 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

This time we have two free variables, so $\dim\big(E_{\lambda_2}(A)\big) = 2$. Setting $x_1 = s$ and $x_3 = t$, the general solution in parametric form is

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} s \\ t \\ t \end{bmatrix} = s \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix},$$

giving us

$$E_{\lambda_2}(A) = \operatorname{Span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \right\}.$$

□

Example 21.5.3 A diagonal example. From Discovery c.

This time our matrix is diagonal, so its eigenvalues are precisely the diagonal entries, $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = 3$.

See. Subsection 21.4.2.

As usual, analyze each eigenvalue in turn.

For $\lambda = 1$:

$$1I - A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix} \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\implies \quad E_{\lambda_1}(A) = \operatorname{Span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right\}.$$

For $\lambda = 2$:

$$2I - A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\implies \quad E_{\lambda_2}(A) = \operatorname{Span}\left\{ \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}.$$

For $\lambda = 3$:

$$3I - A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\implies \quad E_{\lambda_3}(A) = \operatorname{Span}\left\{ \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

The fact that the eigenvectors of our diagonal matrix are standard basis vectors shouldn't be too surprising, since a matrix times a standard basis vector is equal to the corresponding column of the matrix, and the columns of a diagonal matrix are scalar multiples of the standard basis vectors. □

Example 21.5.4 An upper triangular example. From Discovery d.

Our final example matrix is upper triangular, so again its eigenvalues are precisely the diagonal entries, $\lambda_1 = 2$ and $\lambda_2 = -1$.

See. Subsection 21.4.2.

Note that we don't count the repeated diagonal entry 2 as two separate eigenvalues — that eigenvalue is just repeated as a root of the characteristic polynomial. (But this repetition will become important in the next chapter.)

Once again we determine eigenspaces by row reducing, one at a time.

For $\lambda_1 = 2$:

$$2I - A = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 3 \end{bmatrix} \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\implies \qquad E_{\lambda_1}(A) = \operatorname{Span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right\}.$$

For $\lambda_2 = -1$:

$$(-1)I - A = \begin{bmatrix} -3 & -1 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\implies \qquad E_{\lambda_2}(A) = \operatorname{Span}\left\{ \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

□

Example 21.5.5 Using row operations to help. Don't forget that we can use row operations to help compute determinants!

See. Subsection 9.3.1 for an example of using row operations to compute a determinant.

Let's do a 4×4 example to demonstrate. Consider

$$A = \begin{bmatrix} 5 & -4 & -27 & 46 \\ 2 & -1 & -12 & 20 \\ 2 & -2 & -8 & 14 \\ 1 & -1 & -3 & 5 \end{bmatrix}.$$

To obtain the characteristic polynomial, we want to compute the determinant of

$$\lambda I - A = \begin{bmatrix} \lambda - 5 & 4 & 27 & -46 \\ -2 & \lambda + 1 & 12 & -20 \\ -2 & 2 & \lambda + 8 & -14 \\ -1 & 1 & 3 & \lambda - 5 \end{bmatrix}.$$

Let's row reduce a bit first:

$$\begin{bmatrix} \lambda - 5 & 4 & 27 & -46 \\ -2 & \lambda + 1 & 12 & -20 \\ -2 & 2 & \lambda + 8 & -14 \\ -1 & 1 & 3 & \lambda - 5 \end{bmatrix} \quad R_1 \leftrightarrow -R_4$$

$$\longrightarrow \begin{bmatrix} 1 & -1 & -3 & 5 - \lambda \\ -2 & \lambda + 1 & 12 & -20 \\ -2 & 2 & \lambda + 8 & -14 \\ \lambda - 5 & 4 & 27 & -46 \end{bmatrix} \quad \begin{array}{l} \\ R_2 + 2R_1 \\ R_3 + 2R_1 \\ R_4 - (\lambda - 5)R_1 \end{array}$$

$$\longrightarrow \begin{bmatrix} 1 & -1 & -3 & 5-\lambda \\ 0 & \lambda-1 & 6 & -2(\lambda+5) \\ 0 & 0 & \lambda+2 & -2(\lambda+2) \\ 0 & \lambda-1 & 3(\lambda+4) & \lambda^2-10\lambda-21 \end{bmatrix}.$$

In our first step above, we performed two operations: swapping rows and multiplying a row by -1. Both of these operations change the determinant by a factor of -1, so the two effects cancel out. Our other operations in the second step above do not affect the determinant, so the determinant of this third matrix above will be equal to the characteristic polynomial of A.

Now, we cannot divide a row by zero. So we should not divide either the second or fourth rows by $\lambda-1$ in an attempt to obtain the next leading one, because we would inadvertently be dividing by zero in the case $\lambda = 1$. However, we can still simplify one step further, even without a leading one:

$$\begin{bmatrix} 1 & -1 & -3 & 5-\lambda \\ 0 & \lambda-1 & 6 & -2(\lambda+5) \\ 0 & 0 & \lambda+2 & -2(\lambda+2) \\ 0 & \lambda-1 & 3(\lambda+4) & \lambda^2-10\lambda-21 \end{bmatrix} \begin{matrix} \\ \\ \\ R_4-R_2 \end{matrix}$$

$$\longrightarrow \begin{bmatrix} 1 & -1 & -3 & 5-\lambda \\ 0 & \lambda-1 & 6 & -2(\lambda+5) \\ 0 & 0 & \lambda+2 & -2(\lambda+2) \\ 0 & 0 & 3(\lambda+2) & \lambda^2-8\lambda-11 \end{bmatrix}. \tag{$*$}$$

This last matrix is not quite upper triangular, but it's close enough that we can proceed by cofactors from here.

$$\begin{aligned} c_A(\lambda) &= \begin{vmatrix} 1 & -1 & -3 & 5-\lambda \\ 0 & \lambda-1 & 6 & -2(\lambda+5) \\ 0 & 0 & \lambda+2 & -2(\lambda+2) \\ 0 & 0 & 3(\lambda+2) & \lambda^2-8\lambda-11 \end{vmatrix} \\ &= 1\cdot \begin{vmatrix} \lambda-1 & 6 & -2(\lambda+5) \\ 0 & \lambda+2 & -2(\lambda+2) \\ 0 & 3(\lambda+2) & \lambda^2-8\lambda-11 \end{vmatrix} \\ &= (\lambda-1)\cdot \begin{vmatrix} \lambda+2 & -2(\lambda+2) \\ 3(\lambda+2) & \lambda^2-8\lambda-11 \end{vmatrix} \\ &= (\lambda-1)\big((\lambda+2)(\lambda^2-8\lambda-11)+6(\lambda+2)^2\big) \\ &= (\lambda-1)(\lambda+2)\big((\lambda^2-8\lambda-11)+6(\lambda+2)\big) \\ &= (\lambda-1)(\lambda+2)(\lambda^2-2\lambda+1) \\ &= (\lambda-1)(\lambda+2)(\lambda-1)^2 \\ &= (\lambda-1)^3(\lambda+2). \end{aligned}$$

We now see that the eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = -2$.

To determine bases for eigenspaces, we usually reduce the matrix $\lambda I - A$ with the various eigenvalues substituted in for λ. But we have already partially reduced $\lambda I - A$ with λ left variable to help us determine the eigenvalues. So we can begin from $(*)$ for both eigenvalues.

For $\lambda_1 = 1$:

$$\left[\begin{array}{rrrr} 1 & -1 & -3 & 4 \\ 0 & 0 & 6 & -12 \\ 0 & 0 & 3 & -6 \\ 0 & 0 & 9 & -18 \end{array}\right] \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \left[\begin{array}{rrrr} 1 & -1 & 0 & -2 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

$$\implies \qquad E_{\lambda_1}(A) = \operatorname{Span}\left\{ \begin{bmatrix} 2 \\ 0 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \right\}.$$

For $\lambda_2 = -2$, again starting from $(*)$:

$$\left[\begin{array}{rrrr} 1 & -1 & -3 & 7 \\ 0 & -3 & 6 & -6 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 9 \end{array}\right] \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \left[\begin{array}{rrrr} 1 & 0 & -5 & 0 \\ 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

$$\implies \qquad E_{\lambda_2}(A) = \operatorname{Span}\left\{ \begin{bmatrix} 5 \\ 2 \\ 1 \\ 0 \end{bmatrix} \right\}.$$

□

21.6 Theory

In this section.

- Subsection 21.6.1 *Basic facts*
- Subsection 21.6.2 *Eigenvalues and invertibility*

21.6.1 Basic facts

First we collect some of our observations about eigenvalues and eigenvectors from Section 21.4. We omit their proofs, as we have already discussed the ideas behind them in that section.

Proposition 21.6.1 Eigenvalues of special forms. *If square matrix A is diagonal or triangular, then the eigenvalues of A are precisely its diagonal entries.*

Proposition 21.6.2 Eigenspaces. *For an $n \times n$ matrix A, the collection of all eigenvectors that correspond to a specific eigenvalue λ, along with the zero vector, forms a subspace of $\mathbb{R}^n$.*

21.6.2 Eigenvalues and invertibility

Our observation in Subsection 21.4.5 about the possibility of eigenvalue $\lambda = 0$ allows us to add another to our list of properties that are equivalent to invertibility that we began in Theorem 6.5.2, and then continued in Theorem 10.5.3 and Theorem 20.5.5.

Theorem 21.6.3 Characterizations of invertibility. *For a square matrix A, the following are equivalent.*

1. *Matrix A is invertible.*
2. *Every linear system that has A as a coefficient matrix has one unique solution.*
3. *The homogeneous system $A\mathbf{x} = \mathbf{0}$ has only the trivial solution.*
4. *There is some linear system that has A as a coefficient matrix and has one unique solution.*
5. *The rank of A is equal to the size of A.*
6. *The RREF of A is the identity.*
7. *Matrix A can be expressed as a product of some number of elementary matrices.*
8. *The determinant of A is nonzero.*
9. *The columns of A are linearly independent.*
10. *The columns of A form a basis for $\mathbb{R}^n$, where n is the size of A.*
11. *The rows of A are linearly independent.*
12. *The rows of A form a basis for $\mathbb{R}^n$, where n is the size of A.*
13. *The scalar $\lambda = 0$ is not an eigenvalue for A.*

In particular, a square matrix is invertible if and only if $\lambda = 0$ is not *an eigenvalue for A.*

CHAPTER 22

Diagonalization

22.1 Discovery guide

A diagonal matrix is one of the simplest kinds of matrix. In this discovery guide, we will attempt to make any matrix **similar** to a diagonal one.

Recall. When $A\mathbf{x} = \lambda\mathbf{x}$, column vector $\mathbf{x}$ is called an **eigenvector** of A and scalar λ is called the corresponding **eigenvalue** of A.

Discovery 22.1 Suppose 3×3 matrices A, P, D are related by $P^{-1}AP = D$. (Remember, *order matters in matrix multiplication*, so in general $P^{-1}AP \neq A$.)

As an example, consider

$$D = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

We will leave A and P unspecified for now, but think of P as a collection of column vectors:

$$P = \begin{bmatrix} | & | & | \\ \mathbf{p}_1 & \mathbf{p}_2 & \mathbf{p}_3 \\ | & | & | \end{bmatrix}.$$

Multiplying both sides of $P^{-1}AP = D$ on the left by P, we could instead write $AP = PD$.

(a) Do you remember how we defined matrix multiplication, one column at a time?

$$AP = \begin{bmatrix} | & | & | \\ \boxed{} & \boxed{} & \boxed{} \\ | & | & | \end{bmatrix}$$

Hint. See (∗∗∗) in Subsection 4.3.7.

(b) Do you remember how multiplication on the right by a diagonal matrix affects a matrix of columns?

$$PD = \begin{bmatrix} | & | & | \\ \boxed{} & \boxed{} & \boxed{} \\ | & | & | \end{bmatrix}$$

Hint. See Remark 7.4.4.

(c) Compare your patterns for products AP and PD from Task a and Task b. For $AP = PD$ to true, each column of P must be an ________.

Hint. Reread the introduction to this discovery guide above.

(d) In Task c, we have identified a condition for $AP = PD$ to be true. But to get from $AP = PD$ back to the original equation $P^{-1}AP = D$, we also need P to be invertible, so we need the columns of P to be ________.

Hint. Theorem 20.5.5.

Discovery 22.2 Let's try out what we learned in Discovery 22.1 for matrix

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

(a) Compute the eigenvalues of A by solving the characteristic equation $\det(\lambda I - A) = 0$.

(b) For each eigenvalue of A, determine a basis for the corresponding eigenspace. That is, determine a basis for the null space of $\lambda I - A$ by row reducing.

> **Remember.** Don't row reduce with variable λ in there, substitute an actual eigenvalue for λ before row reducing. Repeat for each eigenvalue, starting back at $\lambda I - A$ and row reducing anew.

(c) Try to create a matrix P that satisfies both of the conditions from Task c and Task d of Discovery 22.1.

(d) If you succeeded in meeting both conditions in the previous step, then $P^{-1}AP$ will be a diagonal matrix. Is it possible to know what diagonal matrix $P^{-1}AP$ will be without actually computing P^{-1} and multiplying out $P^{-1}AP$?

Hint. Look back at how the diagonal entries of matrix D fit in the pattern between AP and PD that you identified in Discovery 22.1.c.

Discovery 22.3 Summarize the patterns you've determined in the first two activities of this discovery guide by completing the following statements in the case that $D = P^{-1}AP$ is diagonal.

(a) The diagonal entries of D are precisely the ________ of A.

(b) The number of times a value is repeated down the diagonal of D corresponds to ________.

(c) The order of the entries down the diagonal of D corresponds to the

________ in P.

Discovery 22.4 Repeat the procedure of Discovery 22.2 for

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

> **Careful.** Make sure the columns of P satisfy *both* necessary conditions from Task c and Task d of Discovery 22.1.

Discovery 22.5 Compare the results of Discovery 22.2 and Discovery 22.4 by filling in the chart in Figure 22.1.1 below. We will call the number of times an eigenvalue is "repeated" as a root of the characteristic polynomial its **algebraic multiplicity** , and we will call the dimension of the eigenspace corresponding to an eigenvalue its **geometric multiplicity** .

After completing the chart, discuss: What can you conclude about algebraic and geometric multiplicities of eigenvalues with respect to attempting to find a suitable P to make $P^{-1}AP$ diagonal?

	Discovery 22.2	Discovery 22.4
eigenvalues		
algebraic multiplicities		
geometric multiplicities		
suitable P exists?		

Figure 22.1.1 Comparison of examples in this discovery guide.

When we attempt to form a **transition matrix** P to make $P^{-1}AP$ diagonal, we need its columns to satisfy both conditions identified in Task c and Task d of Discovery 22.1.

In particular, consider the second of these two conditions. When you determine a basis for a particular eigenspace, these vectors are automatically linearly independent from each other (since they form a basis for a subspace of $\mathbb{R}^n$). However, unless A has only a single eigenvalue, *you will need to include eigenvectors from* different *eigenvalues together in filling out the columns of P*. How can we be sure that the collection of *all* columns of P will satisfy the condition identified in Discovery 22.1.d?

The next discovery activity will help you with this potential problem.

Discovery 22.6 Suppose $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is a linearly independent set of eigenvectors of A, corresponding to eigenvalue λ_1, and suppose $\mathbf{w}$ is an eigenvector of A corresponding to a different eigenvalue λ_2. (So $\lambda_2 \neq \lambda_1$.)

(a) Set up the vector equation to begin the test for independence for the set $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{w}\}$. Call this equation (1).

(b) Multiply both sides of equation (1) by A, then use the definition of eigenvalue/eigenvector to "simplify." Call the result equation (2).

(c) Multiply equation (1) by λ_1 — call this equation (3).

(d) Subtract equation (3) from equation (2). What can you conclude?

(e) Use your conclusion from Task d to simplify equation (1). Then finish the test for independence.

22.2 Terminology and notation

similar matrices
a pair of square matrices A and B for which there exists an invertible matrix P satisfying $B = P^{-1}AP$

transition matrix
an invertible matrix P that realizes a similarity relationship $B = P^{-1}AP$ for similar matrices A and B

diagonalizable
a square matrix that is similar to a diagonal matrix

The next two definitions apply to an eigenvalue of a square matrix.

algebraic multiplicity
the number of times the eigenvalue is repeated as a root of the characteristic polynomial of the matrix

geometric multiplicity
the dimension of the corresponding eigenspace

22.3 Motivation

Similar matrices are truly that — *similar*. While their entries contain different data, everything else about them is essentially the same. Similar matrices have the same rank, nullity, determinant, characteristic polynomial, and eigenvalues. For each shared eigenvalue, similar matrices have the same algebraic and geometric multiplicities. And via a known transition matrix to *transition* spaces from one to the other, similar matrices have "similar" column spaces, null spaces, and eigenspaces. When matrices are similar, one can essentially be replaced by the other in computations, and the transition matrix can be used to transition important vectors in those computations between the two matrices.

The simplest matrices with which to do computations are scalar matrices — matrices that are equal to kI for some scalar k. But no matrix is similar to a scalar matrix, other than the scalar matrix itself, because for $A = kI$ every possible transition matrix P would yield

$$B = P^{-1}AP = P^{-1}(kI)P = kP^{-1}P = kI = A.$$

See. Chapter 7 for a refresher on scalar matrices and their properties.

So in this chapter we consider the next simplest type of matrix with which to do computations — diagonal matrices.

Question 22.3.1 When is a matrix similar to a diagonal matrix, and how do we determine a suitable transition matrix? □

We tackle this question by concentrating on the transition matrix P. If $P^{-1}AP$ is diagonal, what relationships between P, A, and the diagonal matrix $D = P^{-1}AP$ can we discover to help us understand this situation? We have already answered these questions in Discovery guide 22.1. In the next section we summarize our findings.

22.4 Concepts

In this section.

22.4.1 The transition matrix and the diagonal form

The columns of the transition matrix. In Discovery 22.1, we transformed the equation $P^{-1}AP = D$ into the equivalent equation $AP = PD$. Thinking of P as being made up of column vectors, multiplying P on the left by A multiplies each column of P by A, and multiplying P on the right by D multiplies each column of P by the corresponding diagonal entry. So if we view P and D as having forms

$$P = \begin{bmatrix} \vert & \vert & & \vert \\ \mathbf{p}_1 & \mathbf{p}_2 & \cdots & \mathbf{p}_n \\ \vert & \vert & & \vert \end{bmatrix}, \qquad D = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix},$$

then we can view AP and PD as having forms

$$AP = \begin{bmatrix} \vert & \vert & & \vert \\ A\mathbf{p}_1 & A\mathbf{p}_2 & \cdots & A\mathbf{p}_n \\ \vert & \vert & & \vert \end{bmatrix}, \qquad PD = \begin{bmatrix} \vert & \vert & & \vert \\ \lambda_1\mathbf{p}_1 & \lambda_2\mathbf{p}_2 & \cdots & \lambda_n\mathbf{p}_n \\ \vert & \vert & & \vert \end{bmatrix}.$$

See. multiplication pattern (∗∗∗) in Subsection 4.3.7 to remind yourself of the columnwise definition of matrix multiplication, and Remark 7.4.4 for the pattern of multiplying by a diagonal matrix on the right.

The only way these two matrices can be equal is if they have equal columns, so that

$$A\mathbf{p}_1 = \lambda\mathbf{p}_1, \qquad A\mathbf{p}_2 = \lambda\mathbf{p}_2, \qquad \ldots, \qquad A\mathbf{p}_n = \lambda\mathbf{p}_n.$$

These column vector equalities exhibit the eigenvector-eigenvalue pattern. That is, the only way to make $P^{-1}AP$ diagonal is to ***use eigenvectors of** A **as the columns of the transition matrix*** P.

Moreover, P needs to be invertible, so ***the columns of** P **need to be linearly independent*** (Theorem 20.5.5).

The diagonal form matrix $P^{-1}AP$. In Discovery 22.3, we analyzed the pattern of the diagonal matrix $D = P^{-1}AP$. If λ_j is its j^{th} diagonal entry, the condition $A\mathbf{p}_j = \lambda_j\mathbf{p}_j$ from our analysis above says that λ_j is an eigenvalue for A, and the j^{th} column of P is a corresponding eigenvector. So

- D will have the eigenvalues of A for its diagonal entries,
- the number of times an eigenvalue of A is repeated as a diagonal entry in D will correspond to the number of linearly independent eigenvectors for that eigenvalue that were used in the columns of P, and
- the order of the entries down the diagonal of D corresponds to the order of eigenvectors in the columns of P.

22.4.2 Diagonalizable matrices

Is every $n \times n$ matrix similar to a diagonal one? In Discovery 22.4, we discovered that the answer is ***no***. For some matrices, it will not be possible to collect together enough *linearly independent* eigenvectors to fill all n columns of the transition matrix P. The largest number of linearly independent eigenvectors we can obtain for a particular eigenvalue is the dimension of the corresponding eigenspace. In Discovery 22.6, we discovered that eigenvectors from different eigenspaces of the same matrix are automatically linearly independent. So the limiting factor is the dimension of each eigenspace, and whether these dimensions add up to n, the required number of linearly independent columns in P.

Also see. Proposition 22.6.6 in Subsection 22.6.3.

An eigenvalue of an $n \times n$ matrix A has two important numbers attached to it — its **algebraic multiplicity** and its **geometric multiplicity**.

See. Section 22.2 to remind yourself of the definitions of these terms.

If the roots of the characteristic polynomial are all real numbers, then the characteristic polynomial will factor completely as

$$c_A(\lambda) = (\lambda - \lambda_1)^{m_1}(\lambda - \lambda_2)^{m_2} \cdots (\lambda - \lambda_\ell)^{m_\ell},$$

where the λ_j are the distinct eigenvalues of A and the m_j are the corresponding algebraic multiplicities. Since $c_A(\lambda)$ is always a degree n polynomial, the algebraic multiplicities will add up to n. To obtain enough linearly independent eigenvectors for A to fill the columns of P, we also need the geometric multiplicities to add up to n. We will learn in Subsection 22.6.3 that somehow, the **algebraic multiplicity** of each eigenvalue is the "best-case scenario" — ***the geometric multiplicity for an eigenvalue can be no greater than its algebraic multiplicity***. Thus, if any eigenvalue for A is "defective" in the sense that its geometric multiplicity is *strictly less* than its algebraic multiplicity, we will not obtain enough linearly independent eigenvectors for that eigenvalue to fill up its "portion" of the required eigenvectors. To summarize, ***a square matrix is diagonalizable precisely when* each *of its eigenvalues has geometric multiplicity equal to its algebraic multiplicity***.

See. Corollary 22.6.10 in Subsection 22.6.4.

22.4.3 Diagonalization procedure

Procedure 22.4.1 To diagonalize an $n \times n$ matrix A, if possible.

1. *Compute the characteristic polynomial $c_A(\lambda)$ of A by computing* $\det(\lambda I - A)$, *then determine the eigenvalues of A by solving the characteristic equation $c_A(\lambda) = 0$. Make note of the algebraic multiplicity of each eigenvalue.*

2. *For each eigenvalue λ_j of A, determine a basis for the correponding eigenspace $E_{\lambda_j}(A)$ by solving the homogeneous linear system $(\lambda_j I - A)\mathbf{x} = \mathbf{0}$. Make note of the geometric multiplicity of each eigenvalue.*

3. *If any eigenvalue has geometric multiplicity* strictly less *than its algebraic multiplicity, then A is* not *diagonalizable. On the other hand, if each eigenvalue has equal geometric and algebraic multiplicities, then you can obtain n linearly independent eigenvectors to make up the columns of P by taking together all the eigenspace basis vectors you found in the previous step.*

If the matrix P has successfully been constructed, then $D = P^{-1}AP$ will be in diagonal form, with eigenvalues of A in the diagonal entries of D, in order *corresponding to the order of placement of eigenvectors in the columns of P.*

22.5 Examples

In this section.

- Subsection 22.5.1 *Carrying out the diagonalization procedure*
- Subsection 22.5.2 *Determining diagonalizability from multiplicities*
- Subsection 22.5.3 *A different kind of example*

22.5.1 Carrying out the diagonalization procedure

Let's start with some examples from Discovery guide 22.1.

Example 22.5.1 A diagonalizable matrix. From Discovery 22.2. We want to compute a basis for each eigenspace, using the same method as in the examples of Section 21.5. First, we form the matrix

$$\lambda I - A = \begin{bmatrix} \lambda+1 & -9 & 0 \\ 0 & \lambda-2 & 0 \\ 0 & 3 & \lambda+1 \end{bmatrix},$$

and compute its determinant,

$$c_A(\lambda) = \det(\lambda I - A) = (\lambda+1)\bigl[(\lambda-2)(\lambda+1) - 0\cdot 3\bigr] = (\lambda+1)^2(\lambda-2),$$

to obtain the characteristic polynomial of A. The eigenvalues are $\lambda_1 = -1$ and $\lambda_2 = 2$. The first eigenvalue is repeated as a root of the characteristic polynomial, while the second eigenvalue is not, so we have algebraic multiplicities $m_1 = 2$ for λ_1 and $m_2 = 1$ for λ_2. Therefore, we will need two linearly independent eigenvectors from $E_{\lambda_1}(A)$ and one more from $E_{\lambda_2}(A)$.

We determine eigenspace bases by row reducing, assigning parameters, and extracting basis vectors. Here are the results for this matrix:

$$(-1)I - A = \begin{bmatrix} 0 & -9 & 0 \\ 0 & -3 & 0 \\ 0 & 3 & 0 \end{bmatrix} \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\implies \quad E_{\lambda_1}(A) = \operatorname{Span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\},$$

$$2I - A = \begin{bmatrix} 3 & -9 & 0 \\ 0 & 0 & 0 \\ 0 & 3 & 3 \end{bmatrix} \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \begin{bmatrix} 1 & 0 & 3 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\implies \quad E_{\lambda_2}(A) = \operatorname{Span}\left\{ \begin{bmatrix} -3 \\ -1 \\ 1 \end{bmatrix} \right\}.$$

Notice that $\dim\bigl(E_{\lambda_1}(A)\bigr) = 2$, so geometric multiplicity equals algebraic multiplicity for λ_1. Also, $\dim\bigl(E_{\lambda_2}(A)\bigr) = 1$, so again geometric multiplicity equals algebraic multiplicity for λ_2.

Let's pause to consider the result for eigenvalue λ_2. We should have expected the result for the geometric multiplicity of eigenvalue λ_2 from the relationship between algebraic and geometric multiplicites stated in Subsection 22.4.2. If we believe that a geometric multiplicity can never be greater than the corresponding algebraic multiplicity, then eigenspace $E_{\lambda_2}(A)$ in this example could never have dimension greater than 1. On the other hand, an eigenspace should never have dimension 0 because the definition of eigenvalue requires the existence of nonzero eigenvectors. So this forces the dimension of $E_{\lambda_2}(A)$ to be 1, without actually checking.

Returning to our procedure, we can see by inspection that the eigenspace basis vector for λ_2 is linearly independent from the ones for λ_1, so when we form the transition matrix

$$P = \begin{bmatrix} 1 & 0 & -3 \\ 0 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix},$$

it will be invertible because its columns are linearly independent. And we can determine the diagonal form matrix $P^{-1}AP$ *without* calculating P^{-1}, because its diagonal entries should be precisely the eigenvalues of A, with the same multiplicities and order as the corresponding columns of P. In this case,

$$P^{-1}AP = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Finally, note that we could have analyzed the eigenvalues in the opposite order, in which case we would have formed transition matrix

$$Q = \begin{bmatrix} -3 & 1 & 0 \\ -1 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix},$$

obtaining diagonal form matrix

$$Q^{-1}AQ = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

□

Example 22.5.2 A non-diagonalizable matrix. From Discovery 22.4. This matrix is upper triangular, so we can see directly that the eigenvalues are $\lambda_1 = -1$ with algebraic multiplicity 2, and $\lambda_2 = 2$ with algebraic multiplicity 1. Analyze the eigenspaces:

$$(-1)I - A = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -3 \end{bmatrix} \xrightarrow[\text{reduce}]{\text{row}} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\implies \qquad E_{\lambda_1}(A) = \operatorname{Span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right\},$$

$$2I - A = \begin{bmatrix} 3 & -1 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow[\text{reduce}]{\text{row}} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\implies \qquad E_{\lambda_2}(A) = \operatorname{Span}\left\{ \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

We could have stopped after our analysis of λ_1, since its geometric multiplicity is only 1, whereas we needed it to be equal to the algebraic multiplicity 2. Since we cannot obtain enough linearly independent eigenvectors from these two eigenspaces to fill out a 3 × 3 transition matrix P, matrix A is *not* diagonalizable. □

Remark 22.5.3 The matrices in the two examples above had the same eigenvalues with the same algebraic multiplicities, but one matrix was diagonalizable and the other was not. The difference was in the geometric multiplicities of the eigenvalues, which plays a crucial role in determining whether a matrix is diagonalizable.

22.5.2 Determining diagonalizability from multiplicities

Here is an example where we only concern ourselves with the question of whether a matrix is **diagonalizable**, without attempting to build a transition matrix P.

Is

$$A = \begin{bmatrix} -1 & 0 & -12 & 0 \\ 0 & 1 & -8 & 0 \\ 0 & 0 & 5 & 0 \\ 4 & 0 & 4 & 3 \end{bmatrix}$$

diagonalizable? Compute the characteristic polynomial:

$$\begin{aligned} \det(\lambda I - A) &= \begin{vmatrix} \lambda+1 & 0 & 12 & 0 \\ 0 & \lambda-1 & 8 & 0 \\ 0 & 0 & \lambda-5 & 0 \\ -4 & 0 & -4 & \lambda-3 \end{vmatrix} \\ &= (\lambda-5)(\lambda+1)(\lambda-1)(\lambda-3). \end{aligned}$$

So the eigenvalues are $\lambda = -1, 1, 3, 5$, each with algebraic multiplicity 1. But an eigenspace must contain nonzero eigenvectors, so eigenvalues always have geometric multiplicity *at least* 1. Since we will be able to obtain an eigenvector from each of the four eigenvalues, we'll be able to fill the four columns of the transition matrix P with linearly independent eigenvectors. Therefore, A is diagonalizable.

Remark 22.5.4 The analysis used in the above example only works for eigenvalues of algebraic multiplicity 1. If an eigenvalue has algebraic multiplicity greater than 1, then we still must row reduce $\lambda I - A$ to determine the geometric multiplicity of the eigenvalue. However, if all we are concerned with is the question of **diagonalizability**, then we don't need to carry out the full procedure — we can stop row reducing as soon as we can see how many parameters will be required, since this tells us the dimension of the eigenspace.

22.5.3 A different kind of example

Is

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

diagonalizable? Compute the characteristic polynomial:

$$\det(\lambda I - A) = \begin{vmatrix} \lambda & 1 \\ -1 & \lambda \end{vmatrix} = \lambda^2 + 1.$$

But $\lambda^2 + 1 = 0$ does not have real solutions, so A does not have eigenvalues, and cannot be diagonalizable.

A look ahead. In future studies in linear algebra you may study matrix forms in more detail, in which case you will likely work with **complex** vector spaces, where scalars are allowed to be *both* real and imaginary numbers. In that context, the matrix in the last example above *does* have eigenvalues, and in fact can be diagonalized.

22.6 Theory

In this section.

22.6.1 Similar matrices

First, we'll record just a few of the facts about general similar matrices from Section 22.3.

Proposition 22.6.1 Properties of similar matrices.

1. *Similar matrices have the same determinant.*
2. *Similar matrices have the same characteristic polynomial.*
3. *Similar matrices have the same eigenvalues, with the same algebraic multiplicities.*

Proof of Statement 1. Suppose square matrices A and B are similar, and P is a transition matrix that realizes the similarity, so that $B = P^{-1}AP$.

We know from Proposition 10.5.6 that the determinant of a product is the product of the determinants. And we also know from Proposition 10.5.8 that the determinant of an inverse is the inverse of the determinant. So we can compute $\det B$ as

$$\begin{aligned} \det B &= \det(P^{-1}AP) \\ &= \left(\det P^{-1}\right)(\det A)(\det P) \\ &= (\det P)^{-1}(\det A)(\det P) \\ &= \frac{(\det A)\cancel{(\det P)}}{\cancel{\det P}} \\ &= \det A. \end{aligned}$$

Thus, the similar matrices A and B have the same determinant.

Warning 22.6.2 Careful. In this proof, it would have been *incorrect* to cancel the P^{-1} with the P immediately, because *order of matrix multiplication matters*! It was only after we split the determinant into a *product* of determinants that we could cancel $\det\left(P^{-1}\right)$ with $\det P$ because all three of the determinants are *numbers*, and order of *number* multiplication does not matter.

■

Proof of Statement 2. Suppose square matrices A and B are similar, and P is a transition matrix that realizes the similarity, so that $B = P^{-1}AP$.

The characteristic polynomials of these two matrices are computed as

$$c_A(\lambda) = \det(\lambda I - A), \qquad\qquad c_B(\lambda) = \det(\lambda I - B).$$

Using our assumption $B = P^{-1}AP$, along with $I = P^{-1}IP$, we can express the matrix involved in the characteristic polynomial for B as

$$\lambda I - B = \lambda P^{-1}IP - P^{-1}AP = P^{-1}(\lambda I - A)P,$$

where in the last step we have factored the common P^{-1} and P factors out of the difference (making sure to factor each to the correct side, because *order of matrix multiplication matters*). We have now shown that matrices $\lambda I - A$ and $\lambda I - B$ are also similar, via the same transition matrix P, and so by Statement 1 they have the same determinant. That is,

$$c_B(\lambda) = \det(\lambda I - B) = \det(\lambda I - A) = c_A(\lambda),$$

and thus the similar matrices A and B have the same characteristic polynomial. ■

Proof of Statement 3. Statement 3 follows immediately from Statement 2, as the eigenvalues of a matrix are precisely the roots of the characteristic polynomial of the matrix, and the algebraic multiplicity of an eigenvalue is the number of times that value is repeated as a root of the characteristic polynomial. ■

22.6.2 Diagonalizable matrices

We start with the justification that a transition matrix made up of linearly independent eigenvectors will diagonalize a matrix.

Theorem 22.6.3 Characterization of diagonalizability. *An $n \times n$ matrix A is diagonalizable if and only if there exists a set of n linearly independent vectors in $\mathbb{R}^n$, each of which is an eigenvector of A. If P is an $n \times n$ matrix whose columns are linearly independent eigenvectors of A, then P diagonalizes A.*

Proof. This fact follows from our analysis of the transition matrix P and the diagonal form matrix $P^{-1}AP$ in Subsection 22.4.1. ■

We will refine this theorem using our more sophisticated notions of **algebraic** and **geometric multiplicity** in the next subsection. But first, here is a surprising result that demonstrates how central eigenvalues are in matrix theory.

Proposition 22.6.4 Determinant versus eigenvalues. *If a square matrix is diagonlizable, then its determinant is equal to the product of its eigenvalues (including multiplicities).*

Proof. Suppose A is a diagonalizable matrix. Then it is similar to some diagonal matrix D. The eigenvalues of a diagonal matrix are precisely the diagonal entries, and the algebraic multiplicity of each of these eigenvalues is the number of times that eigenvalue is repeated down the diagonal. So if $\lambda_1, \lambda_2, \ldots, \lambda_\ell$ are all of the *distinct* eigenvalues of D (i.e. there are no repeats in this list of eigenvalues), and $m_1, m_2, \ldots, m_\ell$ are the corresponding algebraic multiplicities of these eigenvalues (i.e. each m_j is equal to the number of times λ_j appears on the main diagonal of

D), then

$$\det D = \lambda_1^{m_1} \lambda_2^{m_2} \cdots \lambda_\ell^{m_\ell},$$

because the determinant of a diagonal matrix is just the product of its diagonal entries (Statement 1 of Proposition 8.5.2). But from Statement 1 of Proposition 22.6.1 we know that the similar matrices A and D have the same determinant, and have all the same eigenvalues with the same corresponding algebraic multiplicities. Thus, the expression

$$\det A = \det D = \lambda_1^{m_1} \lambda_2^{m_2} \cdots \lambda_\ell^{m_\ell}$$

can be viewed as an expression for $\det A$ as a product of the eigenvalues of A, including multiplicities. ■

Remark 22.6.5 The above fact is actually true about *all* square matrices, if you allow complex eigenvalues. In a second linear algebra course, you may learn that diagonalizable matrices are a special case of a more general theory, in which *every* matrix can be **triangularized**. That is, every square matrix is similar to a special form of triangular matrix (either upper or lower), though for many matrices both the transition matrix and the triangular form matrix might need to contain complex numbers in its entries. In this more general theory, it is again the case that the diagonal entries of the triangular form matrix will be precisely the eigenvalues of the original matrix, with each eigenvalue repeated down the diagonal according to its algebraic multiplicity, so the proof provided for the fact above can be adapted to work in this slightly more general setting.

22.6.3 The geometry of eigenvectors

We require that the columns of a transition matrix P be linearly independent, so that P is invertible. Basis vectors for a particular eigenspace are linearly independent by definition of **basis**. But when we lump basis vectors from *different* eigenspaces together, will they all remain linearly independent together? The next fact answers this question with a more general version of what we explored in Discovery 22.6.

Proposition 22.6.6 Eigenvectors from different eigenvalues are independent. *Suppose A is an $n \times n$ matrix, and S is a linearly independent set of vectors in $\mathbb{R}^n$, each of which is an eigenvector for A. Further suppose that $\mathbf{v}$ is another eigenvector for A that is linearly independent from those vectors in S that are from the same eigenspace as $\mathbf{v}$. Then the enlarged collection S' of eigenvectors consisting of all vectors in S along with $\mathbf{v}$ is also linearly independent.*

Proof. Let's write $S = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_\ell, \mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m\}$, where the $\mathbf{v}_j$ are those eigenvectors in S that are in the same eigenspace as $\mathbf{v}$, and the $\mathbf{w}_j$ are those that are not. Write λ for the eigenvalue of A corresponding to $\mathbf{v}$ (hence also to each $\mathbf{v}_j$), and write λ_j for the eigenvalue corresponding to $\mathbf{w}_j$. We have assumed that the full set S is linearly independent, and therefore so are the subsets $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_\ell\}$ and $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m\}$ (Statement 2 of Statement 17.5.3). In addition, we have assumed that the set $\{\mathbf{v}, \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_\ell\}$ remains linearly independent.

The strategy in this proof is essentially the same as explored in Discovery 22.6. To prove independence, we must prove that the assumption

$$\begin{aligned} k\mathbf{v} + a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_\ell\mathbf{v}_\ell \\ + b_1\mathbf{w}_1 + b_2\mathbf{w}_2 + \cdots + b_m\mathbf{w}_m = \mathbf{0} \end{aligned} \tag{$*$}$$

leads to the conclusion that each of the scalars $k, a_1, a_2, \dots, a_\ell, b_1, b_2, \dots, b_m$ is 0.

Since each of the vectors in the combination above is an eigenvector for A, if we multiply both sides of equation ($*$) by the matrix A, we may substitute

$A\mathbf{v} = \lambda\mathbf{v}$, $A\mathbf{v}_j = \lambda\mathbf{v}_j$, and $A\mathbf{w}_j = \lambda_j\mathbf{w}_j$. Making these substitutions, we obtain

$$\begin{aligned} &k\lambda\mathbf{v}+a_1\lambda\mathbf{v}_1 + a_2\lambda\mathbf{v}_2 + \cdots + a_\ell\lambda\mathbf{v}_\ell \\ &\qquad + b_1\lambda_1\mathbf{w}_1 + b_2\lambda_2\mathbf{w}_2 + \cdots + b_m\lambda_m\mathbf{w}_m = \mathbf{0}. \end{aligned}$$

Compare this "A times $(*)$" equation with the result of multiplying $(*)$ through by the scalar λ:

$$\begin{aligned} &k\lambda\mathbf{v}+a_1\lambda\mathbf{v}_1 + a_2\lambda\mathbf{v}_2 + \cdots + a_\ell\lambda\mathbf{v}_\ell \\ &\qquad + b_1\lambda\mathbf{w}_1 + b_2\lambda\mathbf{w}_2 + \cdots + b_m\lambda\mathbf{w}_m = \mathbf{0}. \end{aligned}$$

Notice that the $\mathbf{v}$ and $\mathbf{v}_j$ terms of both the "A times $(*)$" equation and the "λ times $(*)$" equation are identical, so if we subtract these equations and collect like $\mathbf{w}_j$-terms, we obtain

$$b_1(\lambda_1 - \lambda)\mathbf{w}_1 + b_2(\lambda_2 - \lambda)\mathbf{w}_2 + \cdots + b_m(\lambda_m - \lambda)\mathbf{w}_m = \mathbf{0}.$$

Since the collection of $\mathbf{w}_j$ vectors are linearly independent, the scalar coefficient expressions in this new linear combination must all be zero. That is, each scalar expression

$$b_j(\lambda_j - \lambda)$$

must be zero. However, none of the $\mathbf{w}_j$ is from the same eigenspace as $\mathbf{v}$, so each $\lambda_j - \lambda$ is *non*zero, which forces each of the b_j to be zero.

Substituting this new information into equation $(*)$, we have

$$k\mathbf{v} + a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_\ell\mathbf{v}_\ell = \mathbf{0}.$$

But the collection $\{\mathbf{v}, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_\ell\}$ is assumed independent, so each of the scalars in the remaining combination on the left above is also zero.

We have now successfully shown that the only way equation $(*)$ can be true is if each of the scalars involved is 0, as required. ■

The proposition above is somewhat similar in effect to Proposition 17.5.6, in that it lets us build up a linearly independent set of eigenvectors one-by-one. But the above fact is a little stronger, in that when we look to add a new eigenvector to our collection, we only need to worry about it being linearly independent from the eigenvectors we already have *from that eigenspace*. This leads to the following corollary.

Corollary 22.6.7 Eigenspaces are independent. *Given a collection of bases for the different eigenspaces of a matrix, the collection of all these eigenspace basis vectors together will still be linearly independent.*

Proof. Let A be a square matrix, and write $\lambda_1, \lambda_2, \ldots, \lambda_\ell$ for its eigenvalues. Suppose we have a basis $\mathcal{B}_1$ for eigenspace $E_{\lambda_1}(A)$, and a basis $\mathcal{B}_2$ for eigenspace $E_{\lambda_2}(A)$, and so on. Begin with $\mathcal{B}_1$, which is linearly independent because it is a basis for a subspace. Enlarge $\mathcal{B}_1$ with vectors from $\mathcal{B}_2$, one at a time. At each step we may apply Proposition 22.6.6, because each new vector from $\mathcal{B}_2$ is both

- from a different eigenspace than the vectors in $\mathcal{B}_1$, and
- linearly independent from the previous vectors from $\mathcal{B}_2$ already included in the new enlarged collection.

Proposition 22.6.6 tells us that at each step of enlarging our collection by one, the new, larger collection will remain linearly independent. Once we run out of vectors in $\mathcal{B}_2$, we begin enlarging our collection with vectors from $\mathcal{B}_3$, one at a time. Again, Proposition 22.6.6 applies at each enlargement step, so that each collection of eigenvectors along the way remains linearly independent. Carry this process through to the end, until finally all vectors from $\mathcal{B}_\ell$ are also included, and Proposition 22.6.6 will still apply at the last step to tell us that the complete set of basis eigenvectors is linearly independent. ■

In the next subsection, we will use this corollary to refine our initial characterization of diagonalizability stated in Theorem 22.6.3. In the meantime, we will formally state the relationship between geometric and algebraic multiplicities that we discussed in Subsection 22.4.2.

Theorem 22.6.8 Geometric versus algebraic multiplicity. *The geometric multiplicity of an eigenvalue is always less than or equal to its algebraic multiplicity.*

Proof. We will not include the proof of this statement here — you may encounter it in further study of matrix forms, perhaps in a second course in linear algebra. ■

Remark 22.6.9 As we've noted already, the geometric multiplicity of an eigenvalue is always *at least* one, since otherwise it wouldn't have any corresponding nonzero eigenvectors!

22.6.4 More about diagonalizable matrices

Corollary 22.6.7 tells us that when collecting eigenvectors to make up the transition matrix P, we only have to worry about linear independence *inside* eigenspaces; linear independence *between* eigenspaces is automatic. But linear independence inside an eigenspace $E_{\lambda_j}(A)$ is taken care of for us when we row reduce $\lambda_j I - A$. So our initial characterization of diagonalization in Theorem 22.6.3 can be refined so that we don't actually have to worry about linear independence of eigenvectors at all — we just have to worry about having *enough* eigenspace basis vectors. It turns out that the algebraic multiplicity of each eigenvector is exactly the necessary number of basis vectors for the corresponding eigenspace, and the next statements record this thinking.

Corollary 22.6.10 More characterizations of diagonalizability.

1. *A matrix with real eigenvalues is diagonalizable if and only if each eigenvalue has geometric multiplicity* equal *to its algebraic multiplicity.*
2. *An $n \times n$ matrix that has n different real eigenvalues must be diagonalizable.*

Note. *We present these statements as a corollary, as they follow from Theorem 22.6.8.*

Proof of Statement 1. We need n linearly independent eigenvectors to make up the columns of the $n \times n$ transition matrix P. The maximum number of linearly independent eigenvectors we can get from a single eigenspace is $\dim(E_\lambda(A))$, the geometric multiplicity of the eigenvalue λ. So the maximum number of linearly independent eigenvectors we can get in total is the sum of the geometric multiplicities of the eigenvalues. But the characteristic polynomial $c_A(\lambda)$ has degree n, and n is the sum of the *algebraic* multiplicities of the eigenvalues, because if A has all real eigenvalues, then $c_A(\lambda)$ factors as

$$c_A(\lambda) = (\lambda - \lambda_1)^{m_1}(\lambda - \lambda_2)^{m_2} \cdots (\lambda - \lambda_\ell)^{m_\ell}.$$

So if even *one* eigenvalue is deficient in the sense that its geometric multiplicity is strictly less than its algebraic multiplicity, we won't obtain enough linearly independent eigenvectors from that eigenspace to contribute to the n linearly eigenvectors we need in total.

On the other hand, if each eigenvalue has geometric multiplicity equal to its algebraic multiplicity, then forming eigenspace bases and collecting them all together will provide us with exactly n eigenvectors, and Proposition 22.6.6 tells us that these n eigenvectors will be linearly independent. ■

Proof of Statement 2. In the case that a square matrix has n *different* real eigenvalues, then each of these eigenvalues must have algebraic multiplicity 1, since otherwise these n algebraic multiplicities would add up to *more* than n, the degree of the characteristic polynomial. So each geometric multiplicity is no greater than 1. But also, as in noted in Remark 22.6.9, each geometric multiplicity must be *at least* 1. Thus, each geometric multiplicity for this matrix is *exactly* 1, and so is equal to the corresponding algebraic multiplicity.

The result now follows from the first statement of this corollary. ■

Appendices

APPENDIX A

GNU Free Documentation License

Version 1.3, 3 November 2008

0. PREAMBLE. The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS. This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The “Invariant Sections” are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The “Cover Texts” are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A “Transparent” copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not “Transparent” is called “Opaque”.

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The “Title Page” means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, “Title Page” means the text near the most prominent appearance of the work’s title, preceding the beginning of the body of the text.

The “publisher” means any person or entity that distributes copies of the Document to the public.

A section “Entitled XYZ” means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as “Acknowledgements”, “Dedications”, “Endorsements”, or “History”.) To “Preserve the Title” of such a section when you modify the Document means that it remains a section “Entitled XYZ” according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING. You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY. If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS. You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties — for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS. You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled “History” in the various original documents, forming one section Entitled “History”; likewise combine any sections Entitled “Acknowledgements”, and any sections Entitled “Dedications”. You must delete all sections Entitled “Endorsements”.

6. COLLECTIONS OF DOCUMENTS. You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS. A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an “aggregate” if the copyright resulting from the compilation is not used to limit the legal rights of the compilation’s users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document’s Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION. Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION. You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

10. FUTURE REVISIONS OF THIS LICENSE. The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See `http://www.gnu.org/copyleft/`.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

11. RELICENSING. “Massive Multiauthor Collaboration Site” (or “MMC Site”) means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A “Massive Multiauthor Collaboration” (or “MMC”) contained in the site means any set of copyrightable works thus published on the MMC site.

“CC-BY-SA” means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copyleft versions of that license published by that same organization.

“Incorporate” means to publish or republish a Document, in whole or in part, as part of another Document.

An MMC is “eligible for relicensing” if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subsequently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.

ADDENDUM: How to use this License for your documents. To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

```
Copyright (C)  YEAR  YOUR NAME.
Permission is granted to copy, distribute and/or modify this document
under the terms of the GNU Free Documentation License, Version 1.3
or any later version published by the Free Software Foundation;
with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.
A copy of the license is included in the section entitled "GNU
Free Documentation License".
```

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the “with... Texts.” line with this:

```
with the Invariant Sections being LIST THEIR TITLES, with the
Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.
```

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.

Bibliography

[1] Anton, H. *Elementary Linear Algebra*. 11th ed. Wiley, New Jersey, 2013.

[2] Dummit, D. S., Foote, R. M. *Abstract Algebra*. 2nd ed. Wiley, New Jersey, 1999.

[3] Hoffman, K., Kunze, R. *Linear Algebra*. 2nd ed. Prentice Hall, New Jersey, 1971.

Index

Colophon

This book was authored in PreTeXt.