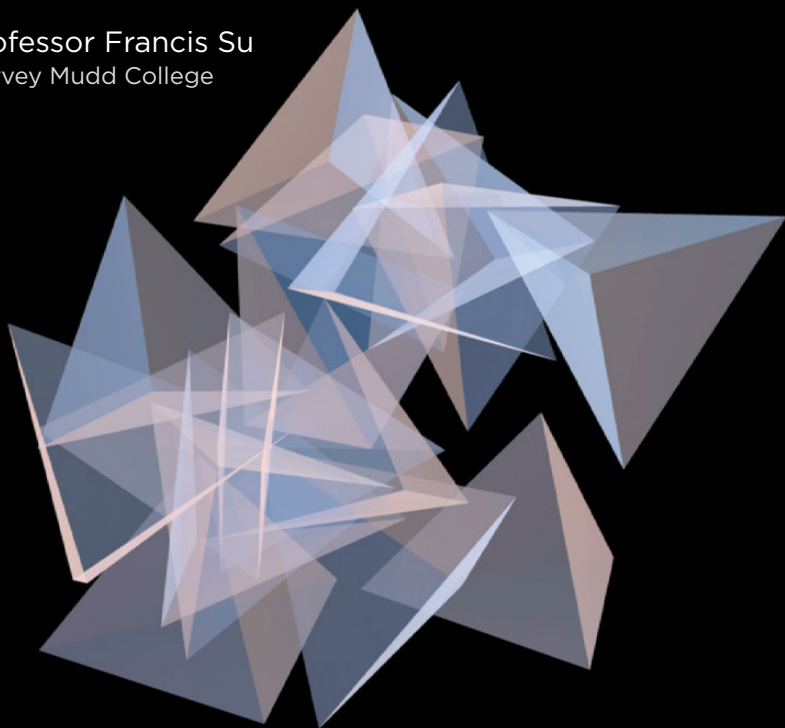


Mastering Linear Algebra

An Introduction with Applications

Course Guidebook

Professor Francis Su
Harvey Mudd College



Published by
THE GREAT COURSES

Corporate Headquarters

4840 Westfields Boulevard | Suite 500 | Chantilly, Virginia | 20151-2299

[PHONE] 1.800.832.2412 | [FAX] 703.378.3819 | [WEB] www.thegreatcourses.com

Copyright © The Teaching Company, 2019

Printed in the United States of America

This book is in copyright. All rights reserved. Without limiting the rights under copyright reserved above, no part of this publication may be reproduced, stored in or introduced into a retrieval system, or transmitted, in any form, or by any means (electronic, mechanical, photocopying, recording, or otherwise), without the prior written permission of The Teaching Company.

Francis Su, PhD

Benediktsson-Karwa
Professor of Mathematics
Harvey Mudd College



Francis Su is the Benediktsson-Karwa Professor of Mathematics at Harvey Mudd College. He earned his PhD from Harvard University, and he has held visiting professorships at Cornell University and the Mathematical Sciences Research Institute in Berkeley, California. In 2015 and 2016, he served as president of the Mathematical Association of America (MAA).

Professor Su's research focuses on geometric and topological combinatorics and their applications to the social sciences. He uses ideas from pure mathematical areas such as topology—the study of stretching things—to applied questions involving how people make decisions. His work on the rental harmony problem, the question in mathematical economics of how to divide rent fairly among roommates, was featured in *The New York Times*, and he has published numerous scientific papers. He is also the author of *Mathematics for Human Flourishing*, which will be published in late fall 2019.

Professor Su has a passion for teaching and popularizing mathematics. His speeches and writing have earned acclaim for describing the humanity of mathematics and for calling people to greater awareness of issues that contribute to inequitable mathematics education. *Wired* magazine called him “the mathematician who will make you fall in love with numbers.”

Professor Su has been nationally recognized for his teaching and mathematical exposition. From the MAA, he has received the Deborah and Franklin Tepper Haimo Award and the Henry L. Alder Award for exemplary teaching as well as the Paul R. Halmos-Lester R. Ford Award and the Merten M. Hasse Prize for distinguished writing. Three of his articles have appeared in the Princeton University Press annual anthology *The Best Writing on Mathematics*. He is the author of the popular Math Fun Facts website, has a widely used YouTube course on real analysis, and is the creator of MathFeed, the math news app. ■

TABLE OF CONTENTS

Introduction

Professor Biography	i
Course Scope	1

Guides

1	Linear Algebra: Powerful Transformations	5
2	Vectors: Describing Space and Motion	22
3	Linear Geometry: Dots and Crosses	36
4	Matrix Operations	50
5	Linear Transformations	59
6	Systems of Linear Equations	73
	QUIZ FOR LECTURES 1–6	88
7	Reduced Row Echelon Form	90
8	Span and Linear Dependence	100
9	Subspaces: Special Subsets to Look For	113
10	Bases: Basic Building Blocks	122
11	Invertible Matrices: Undoing What You Did	135
12	The Invertible Matrix Theorem	145
	QUIZ FOR LECTURES 7–12	152

13	Determinants: Numbers That Say a Lot	154
14	Eigenstuff: Revealing Hidden Structure	169
15	Eigenvectors and Eigenvalues: Geometry	179
16	Diagonalizability	191
17	Population Dynamics: Foxes and Rabbits	203
18	Differential Equations: New Applications	216
	QUIZ FOR LECTURES 13–18	224
19	Orthogonality: Squaring Things Up	226
20	Markov Chains: Hopping Around	237
21	Multivariable Calculus: Derivative Matrix	246
22	Multilinear Regression: Least Squares	255
23	Singular Value Decomposition: So Cool	270
24	General Vector Spaces: More to Explore	280
	QUIZ FOR LECTURES 19–24	294

Supplementary Material

Solutions	296
Bibliography	305

NAVIGATION TIP

To go back to the page you came from, press **Alt + ←** on a PC or **⌘ + ←** on a Mac. On a tablet, use the bookmarks panel.

Mastering Linear Algebra

An Introduction with Applications

Linear algebra is both powerful and beautiful, with classical roots and modern applications. Many structures in the world are linear or approximately linear, so linear algebra is a necessary tool for every scientist. The 24 lectures in this course form an accessible introduction to the elegant ideas of linear algebra and their applications in economics, physics, biology, computer science, and engineering, in addition to mathematics and statistics. Beyond these areas, linear algebra is of growing importance in nonquantitative fields that are taking advantage of modern techniques in data science. Moreover, the underlying geometric ideas are beautiful, and they yield insights beyond algebraic understanding. This course will prepare you to move skillfully between the abstract and concrete, between geometry and algebra, between visualization and computation.

The course begins with an overview lecture (lecture 1) that hits 4 themes that appear throughout the course: that linear algebra is a fundamental idea in mathematics that you'll find everywhere, that linear things are used to approximate nonlinear things, that linear algebra reveals hidden structure, and that the power of linear algebra comes from its interplay between geometry and algebra.

Lectures 2 through 17 discuss core topics in linear algebra, beginning in lectures 2 through 5 with the basic algebraic objects—vectors and matrices—and their geometric intuition. Lecture 5 explains the intuition behind linear transformations; this is earlier than in most treatments of linear algebra, which often favor studying linear equations first.

Lectures 6 and 7 discuss how to solve a system of linear equations and introduce the idea of simplifying a system to reduced row echelon form, which is particularly useful in unlocking the connection between various other concepts associated with matrices.

Lectures 8 through 10 develop the idea of a subspace by first explaining the concept of the span of a set of vectors—all the points that can be reached by using those vectors—and then defining linear independence of a set of vectors, which help you decide when you have a set of vectors that is minimally efficient in reaching all the points in its span. The span of the rows and columns of a matrix are special subspaces associated to a matrix, and the set of all vectors that a matrix sends to zero is another subspace, called the null-space.

Lectures 11 and 12 discuss a central concept—invertibility—and all the different ways of understanding this concept. Lecture 13 describes the determinant, a single number associated to a matrix, that helps you understand invertibility as well as how the linear transformation associated to a matrix scales volumes. Then, lectures 14 through 17 develop intuition for eigenvectors and eigenvalues, some of the most important ideas in linear algebra, in the context of an extended application to population biology: modeling predator-prey relationships.

Lectures 18 through 23 showcase many extended applications of linear algebra. Lecture 18 discusses how linear algebra helps you solve systems of differential equations and how eigenvectors and eigenvalues show up in their solutions. Only very little calculus is assumed in that lecture. Lecture 19 develops the ideas of orthogonality and the QR -factorization of a matrix. Lecture 20 discusses Markov chains, which are useful for modeling many systems of interest in the real world that evolve according to some probabilities. Lecture 21—which assumes you know some single-variable calculus—shows how linear algebra provides a window to understanding calculus in many variables. In particular, linear functions help approximate the nonlinear functions encountered in multivariable

calculus. Lecture 22 explains how linear algebra is important in statistics by shedding light on what regression is and how it's done. Lecture 23 builds up the ideas behind the singular value decomposition, a powerful way to factor matrices, and discusses an application to recommender systems like ones that recommend movies for you to watch.

Lecture 24 concludes the course with a preview of how the powerful ideas of linear algebra for n -dimensional vectors apply to more general vector spaces, where vectors could be things like functions and could be infinite-dimensional. These concepts reveal that even simple ideas in linear algebra are actually profound ideas that show up in unexpected contexts once you see the hidden structure underneath. ■

NOTE

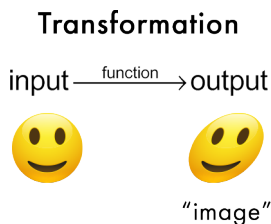
Each lecture has a few linear algebra problems associated with it that you can do to practice the concepts from the lecture, but these will not be sufficient to support your learning of linear algebra. You'll want to follow a text and work on exercises to be sure you understand the concepts. This course will follow David Lay, Steven Lay, and Judi McDonald's *Linear Algebra and Its Applications* and David Poole's *Linear Algebra: A Modern Introduction* (both listed in the Bibliography), but almost any text in linear algebra will do if you look at the sections covering the topics indicated.

LINEAR ALGEBRA: POWERFUL TRANSFORMATIONS

Linear algebra is a foundational subject in mathematics that is both powerful and beautiful, with classical roots and modern applications. The goal of this lecture is to give you a sense of what linear algebra is, why it is important, and how it can help solve some interesting problems. Specifically, the lecture focuses on what a linear transformation is and how it arises in many contexts where you wouldn't expect it.

Transformations

Transformation is really just another word for *function*, which is a rule for taking an input (such as a number, or a set of numbers) and assigning an output (another number, or set of numbers). In linear algebra, functions are called transformations because we think of them as changing one picture into another. The second picture is called the image of the first.



A transformation assigns every point or object in the old picture to an “image” in the new picture where it has moved. This kind of transformation might be important in computer graphics if you want to change perspective.

But there are many other kinds of transformations. Suppose you like to snack on nuts and chocolate, but you also care about eating healthy.

Another Transformation

$$\left(\begin{array}{cc} x & y \\ \text{chocolates, nuts} \end{array} \right) \longrightarrow \left(\begin{array}{cc} m & n \\ \text{carbs, fat} \\ \text{“image”} \end{array} \right).$$

Suppose you do some measurements on a collection of x chocolates and y nuts and find that they have m grams of carbohydrates and n grams of fat. That is a transformation of 2 numbers (x and y) into their “image”—2 numbers (m and n).

We are now ready to define the term *linear algebra*:

Linear algebra is the study of certain kinds of spaces, called vector spaces, and of special kinds of transformations of vector spaces, called linear transformations.

Our view of the night sky is the perfect backdrop for understanding linearity.

We imagine our view of the night sky as being represented on a 2-dimensional plane.

Suppose you find the planet Saturn and mark its position on nights 0 and 1.

On night 2, where should you look?

Without any knowledge of the motion of planets, there is really only one reasonable guess.



You naturally assume that Saturn's motion from night 0 to night 1 will continue in the same direction and at the same speed as it moves on subsequent nights. This motion is linear.

So, on night 3, you would expect to find Saturn here:



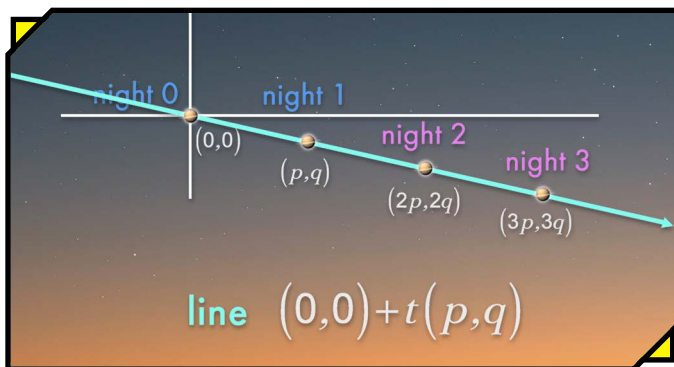
One feature of this motion is that no matter which night you start observing, the planet will traverse the same distance along that line from night to night. With linear motion, the distance and direction traveled only depends on the time elapsed, not on the starting position.



That's what's happening geometrically. What's going on algebraically?

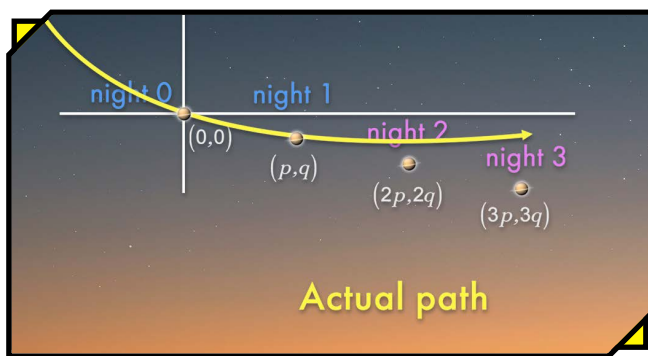
Imagine that our view of the sky has coordinates, and suppose you find Saturn at $(0, 0)$ on night 0 and at (p, q) on night 1.

Then, on night 2, you expect to find Saturn near $(p, q) + (p, q)$, which is the point $(2p, 2q)$.



More generally, at time t , Saturn should be at $(0, 0) + t(p, q)$. As time varies, this expression traces out a line with constant speed. This is a function that takes a time t to its image, which is a position in the sky.

Of course, planets do not move in straight lines, neither in the sky nor in real life.



However, if the planet's motion is relatively smooth, the formula would be actually quite good for small timescales. That's because when a path isn't linear, it is often approximately linear.

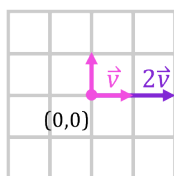
This idea is at the heart of calculus. When you encounter a crazy function, calculus suggests you try to approximate it by something you understand better. The graph of a function that is not linear can, if it's simple enough, be approximated as a line.

That's what's going on in this example. Because you only had 2 data points, you assume that Saturn moves linearly as a function of one variable, which is time.

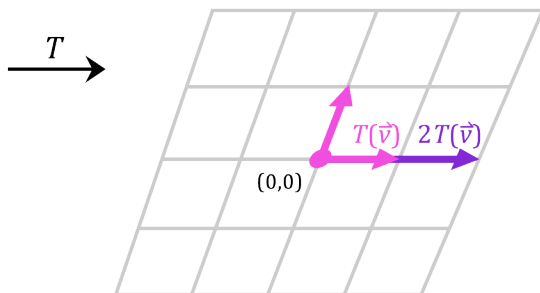
But this idea of linearity also makes sense for functions of several variables. For example, you might have a function T that takes a position in the plane to another position in the plane.

Suppose you wanted to take a square grid, in the left picture, and convert it to the picture on the right, which looks like the same square grid viewed from a different angle and distance. If you were doing computer graphics, you might want to have the ability to change perspective like this.

VIEW 1



VIEW 2



The function T that performs this conversion moves points in the plane around; the diagram suggests how the points get transformed and how lines and arrows on the diagram get transformed as well. This change of perspective is a linear transformation, and it demonstrates 2 good linearity properties.

Vectors are often denoted by boldface letters (\mathbf{r}) or by an arrow diacritic (\vec{r}). The styles have the same meaning, and both are used in this book.

Look at the pink arrow in view 1 indicated by the letter \mathbf{v} , which is based at the point $(0, 0)$. The purple arrow, twice its size and pointing in the same direction, is labeled $2\mathbf{v}$. The transformation T has the property that the image of the pink and purple arrows is related in the same way in view 2 (twice the size and pointing in the same direction).

You can think of the pink arrow as a small change in input. So, this property shows that if you double the input vector (pink arrow \mathbf{v} changing to purple arrow $2\mathbf{v}$), the output should double as well—e.g., double its magnitude but retain its direction (the pink arrow $T(\mathbf{v})$ changing to purple arrow $2T(\mathbf{v})$).

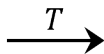
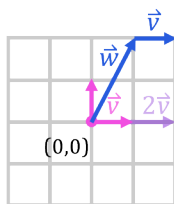
So, here's the **first linearity property**: For any real number k ,

$$T(k\mathbf{v}) = kT(\mathbf{v}).$$

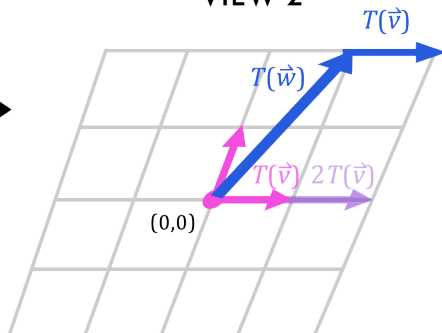
In other words, it doesn't matter whether you scale first and then apply T or apply T first and then scale.

A second property of this transformation is if you're anywhere else besides the origin and modify your input by \mathbf{v} , then the output changes by $T(\mathbf{v})$ no matter where you were to begin with. Compare the blue arrow marked \mathbf{v} , starting at \mathbf{w} , with the pink arrow at the origin marked \mathbf{v} . They represent the same change in input. Under the transformation, the pink arrow gets transformed to $T(\mathbf{v})$ at the origin and the blue arrow gets transformed to $T(\mathbf{v})$ based at $T(\mathbf{w})$, but the 2 arrows marked $T(\mathbf{v})$ are the same and represent the same change in output.

VIEW 1



VIEW 2



This transformation has the same property as the Saturn example did: Equal changes in input (in that case, it was time) lead to equal changes in output.

This highlights the **second linearity property**: For any points v and w ,

$$T(\mathbf{w} + \mathbf{v}) = T(\mathbf{w}) + T(\mathbf{v}).$$

In other words, the transformed sum is the sum of the transformed points—or, another way, it does matter whether you add first and apply T , or apply T first and then add.

So, change of perspective in computer graphics is also a linear transformation.

Returning to the nutrition example, suppose that x is the number of chocolates you have and y is the number of nuts.

$$\begin{array}{ccc} \begin{pmatrix} x & y \end{pmatrix} & \xrightarrow{A} & \begin{pmatrix} ax + by & cx + dy \end{pmatrix} \\ \begin{array}{c} | \quad | \\ \text{chocolates} \quad \text{nuts} \end{array} & & \begin{array}{cc} \underbrace{\hspace{1.5cm}} & \underbrace{\hspace{1.5cm}} \\ \text{carbs} & \text{fat} \end{array} \\ & & \text{grams} \end{array}$$

Then, if a single chocolate has a grams of carbs and a single nut has b grams of carbs, then the total carbs in your snack is $(ax + by)$ grams. Similarly, if c and d record the number of grams of fat in a chocolate and in a nut, respectively, then you'd have $(cx + dy)$ grams of fat.

So, if A is the transformation that takes the number of chocolates and nuts to the numbers of grams of carbs and fat, then it is represented by

$$A(x, y) = (ax + by, cx + dy).$$

You can check that if you vary x by 1 unit, the output changes by (a, c) grams, no matter what x and y are. If you vary y by 1 unit, you get similar behavior. So, this transformation is linear, too.

What you see here is an example of a system of equations:

$$\begin{aligned}ax + by &= \text{carb grams} \\ cx + dy &= \text{fat grams}.\end{aligned}$$

Systems of equations (often with many more equations and variables) arise in so many applications that one of the most important questions you can ask mathematically is how to find solutions to such equations. And linear algebra helps you do that. One of the ways to do that is to represent the coefficients of a system in a matrix, which is an array of numbers.

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

Matrices have their own algebra—operations that you can do with them. And because it turns out that every linear transformation can be associated to a matrix, learning how matrices behave is a big part of doing linear algebra.

At the same time, systems of equations can be understood geometrically, too. Each linear equation has a solution set that is a linear object, such as a line or plane or hyperplane. So, to solve a system of equations simultaneously can be understood as intersecting a bunch of hyperplanes and asking what (if anything) is in the intersection.

There are many ways in which linear transformations arise, and this is the first of several important themes in this course—reasons you'll want to learn linear algebra.

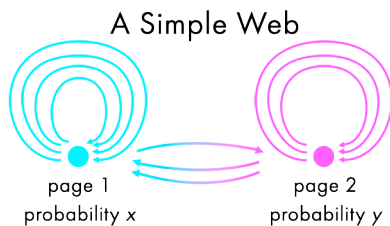
THEME 1

Linearity is a fundamental idea in mathematics and in the world—you will encounter it everywhere.

Because linear transformations describe the geometry of a change of perspective, if you know how to represent them by a matrix, you can compute them fairly easily. When producing computer graphics, rotating an object, or scaling up an object, or flattening a 3-D object to a 2-D object are all linear transformations. But there are many other linear relationships in life that may not be obvious at first glance.

For example, suppose you want to measure the importance of a webpage. One way to do that is to imagine a person surfing the web. He or she starts at one webpage and picks one link on the page at random and hops to that page. He or she repeats that process with the new webpage. Then, after many hops, you might expect that he or she is more likely to be at a popular page than at a less popular one, because a popular page will have many paths to get there. This gives you a way of ranking pages.

Imagine a very simple web with just 2 pages: page 1 and page 2. Suppose there are 4 links from page 1 to itself (which could happen if the page linked to another part of the same page), and suppose there is a single link from page 1 to page 2. And let's say that page 2 has 3 self-links and 2 links to page 1.



If the probability of being at page 1 is x_0 , and the probability of being at page 2 is y_0 , then after one hop, what are the new probabilities of being at those pages?

You can get to page 1 either from page 1 or page 2. So, the probability of being at page 1 after one hop is the probability of going from page 1 to page 1, which is $\frac{4}{5}$; times the probability of starting at page 1, called x ; plus the probability of starting at page 2 and going to page 1, which is $\frac{2}{5}$; times the probability of starting at page 2, called y .

$$\begin{aligned} P(\text{end } 1) &= P(1 \rightarrow 1)P(\text{start } 1) + P(2 \rightarrow 1)P(\text{start } 2) \\ &= \frac{4}{5} x + \frac{2}{5} y. \end{aligned}$$

The probability of being at page 2 after one hop can be computed in a very similar way. By inspecting the link diagram, you get

$$\begin{aligned} P(\text{end } 2) &= P(1 \rightarrow 2)P(\text{start } 1) + P(2 \rightarrow 2)P(\text{start } 2) \\ &= \frac{1}{5} x + \frac{3}{5} y. \end{aligned}$$

So, the 2 probabilities of ending at page 1 and page 2 are given by linear equations and can be represented by this linear transformation:

$$T(x, y) = \left(\frac{4}{5}x + \frac{2}{5}y, \frac{1}{5}x + \frac{3}{5}y \right).$$

To get the probability vector of being at page 1 and 2 at time $(n + 1)$, represented by (x_{n+1}, y_{n+1}) , you can just apply T to the probability vector at time n , which is represented by (x_n, y_n) .

$$(x_{n+1}, y_{n+1}) = T(x_n, y_n).$$

So, if you had a way of computing repeated linear transformations quickly, you could then perform this linear transformation T 100 times to get the probability of being at pages 1 and 2 after 100 steps. Larger probabilities then correspond to popular pages, and you would have your version of page rank.

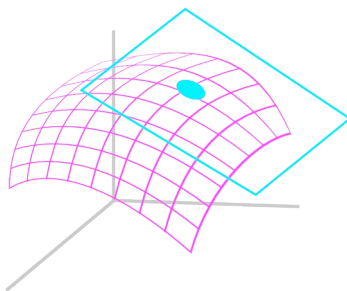
THEME 2

To understand nonlinear things, we approximate them by linear things.

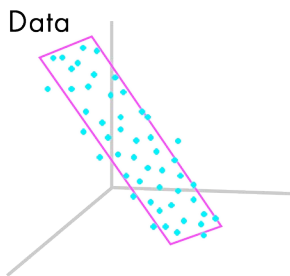
Things in the world that are not linear are often approximately linear.

In multivariable calculus, when we encounter graphs of nonlinear functions, we approximate them locally by tangent planes (if they are differentiable).

Tangent Plane



In statistics, when we encounter a cloud of data points, we often assume linear relationships and use regression to try to fit the best line or hyperplane. Or maybe we try to reduce the dimensions of the data to something manageable by finding special directions in which there is a lot of variation.



Returning again to the nutrition example, suppose you were looking at your snack of chocolates and nuts as a producer, wondering about the cost and time of producing chocolates and nuts.

Cost and time are not linear functions of quantity. If they were linear, then the cost to produce one more nut would be the same no matter how many nuts you had produced already. But usually, there are efficiencies that you gain from mass production, so the cost to produce an additional nut or chocolate is cheaper after you've produced a lot of them, and the additional time it takes is much smaller, too.

So, whatever transformation T that takes quantities to (cost, time) is nonlinear. However, like many things in this world, it is locally linear. That means that if you know how much an additional nut adds to time and cost and how much an additional chocolate adds in time and cost, then that is enough to approximate how much additional time and cost it takes to produce any small numbers of nuts and chocolates.

THEME 3

Linear algebra reveals hidden structures that are beautiful and useful.

You've seen that if you know quantities (x, y) of snacks, you can figure out (fat, carbs). But what if you know (fat, carbs) and you want to solve for quantities x and y ? This is the problem of solving a system of linear equations. It might be easy with functions of 2 variables using substitution, but if you have a large number of variables, this strategy won't work. You'll have to be more systematic.

Linear algebra can be useful here, because much of what linear algebra does is reveal hidden structure and then use those structures to give us insight into what is really going on. Such insights often allow us to solve problems in a simpler way.

For example, you may not be able to see if a system of equations has a solution just from inspection, but if you see the system of equations as arising from a linear transformation of some variables, you may be able to put those equations in a simpler form.

And matrix algebra, which underlies the algebra of linear transformations, seems strange at first and has a lot of hidden structure that turns out to be really useful. For example, let's return to the simple 2-page web. Recall that this linear transformation T converted the n -step probabilities of being at pages 1 or 2 to the $(n + 1)$ -step probabilities of being at pages 1 or 2.

$$T(x, y) = \left(\frac{4}{5}x + \frac{2}{5}y, \frac{1}{5}x + \frac{3}{5}y \right).$$

So, if you start at page 1, to find the page probabilities after 100 hops, you just apply the linear transformation T 100 times to the initial starting configuration $(1, 0)$.

This calculation can be done very efficiently, because the transformation T has special vectors, called eigenvectors, for which applying T is very easy.

$$T(2,1) = (2,1).$$

Look at this vector: $(2, 1)$. Notice when you apply T to $(2, 1)$, you just get back $(2, 1)$! In other words, T doesn't do anything to $(2, 1)$. So, if you apply T to $(2, 1)$ ninety-nine more times, it's still unchanged!

$$T^{100}(2,1) = (2,1).$$

Look at this vector: $(1, -1)$. If you compute A times $(1, -1)$, you get $(\frac{2}{5}, -\frac{2}{5})$.

The key insight here is that $(\frac{2}{5}, -\frac{2}{5})$ is actually a multiple of $(1, -1)$.

$$T(1,-1) = \left(\frac{2}{5}, -\frac{2}{5}\right) = \frac{2}{5}(1,-1).$$

So, $T(1, -1)$ is just $\frac{2}{5}$ times $(1, -1)$. And if you multiply by A ninety-nine more times, it's just $(\frac{2}{5})^{100}$ times $(1, -1)$!

$$T^{100}(1,-1) = \left(\frac{2}{5}\right)^{100} (1,-1).$$

If you start in page 1, the beginning vector of page probabilities is $(1, 0)$, because there's no chance of being in page 2. This vector is not a special vector, and the matrix A does not act on $(1, 0)$ in a very nice way.

But notice that $(1, 0)$ can be written as a combination of the special eigenvectors; in fact, it is $\frac{1}{3}$ of both eigenvectors summed.

$$(1, 0) = \frac{1}{3}(2, 1) + \frac{1}{3}(1, -1).$$

Then, something amazing happens: The first and second linearity properties ensure that you can apply T first to $(2, 1)$ and to $(1, -1)$ before you scale or sum them! But those actions are very easy. T does nothing to $(2, 1)$ and multiplies $(1, -1)$ by $\frac{2}{5}$.

$$T(1, 0) = \frac{1}{3}(2, 1) + \frac{1}{3} \frac{2}{5}(1, -1).$$

Each additional time you apply T , it just multiplies the second eigenvector by $\frac{2}{5}$. This means that $T^{100}(1, 0)$ is just, by linearity properties, the sum

$$T^{100}(1, 0) = \frac{1}{3}(2, 1) + \frac{1}{3} \left(\frac{2}{5} \right)^{100} (1, -1).$$

This is easy to compute. Notice how small the second term is. The first term is $(\frac{2}{3}, \frac{1}{3})$, and it dominates! So, after many hops, the likelihood of being in pages 1 and 2 is $(\frac{2}{3}, \frac{1}{3})$. This says that page 1 is more popular than page 2.

THEME 4

Linear algebra's power often comes from the interplay between geometry and algebra.

With every new concept in this course, you'll learn how to think about it algebraically (in terms of algebraic expressions) and geometrically (in terms of things you can visualize).

For example, the set of solutions to a linear equation in 3 variables, such as $x + 2y + 3z = 10$, represents a plane in 3-dimensional space.

If you know this, then you know that if you have 2 such linear equations involving x , y , and z , the set of solutions that satisfies both equations will be the intersection of 2 planes, which in most cases is just a line. If the 2 planes happen to be the same plane, then the set of solutions may also be a whole plane. But if the 2 planes happen to be parallel and disjoint, then the pair of equations may have no simultaneous solution.

Notice that this insight came very easily from the geometric view of these algebraic equations but may not have been obvious just by looking at the equations. This kind of interplay between algebra and geometry is indeed very powerful.

READINGS

Chartier, *When Life Is Linear*, chap. 1.

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*. Read the Introductory Example of every chapter.

Poole, *Linear Algebra*. Skim the book and read various sections—Introductions, Applications, and Vignettes—that look interesting. For now, read for the big ideas without worrying about the mathematical details.

VECTORS: DESCRIBING SPACE AND MOTION

One way to think of mathematics is that it is a study of objects with structure and of functions that can preserve that structure. In linear algebra, the basic objects of study are vectors, and mathematicians like to think of them as having a certain kind of structure—namely, you can perform certain vector operations, such as addition and scalar multiplication. The operations give a structure to the set of vectors by telling you how to do things with them. The goal of this lecture is to understand vectors and their structure.

Vectors

There are several different ways to define vectors.

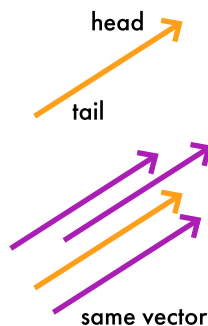
The geometric way to define a vector is as an object living in some space. We need to be careful what space we are talking about. We live in a universe with 3 spatial dimensions that are evident to us, and a vector could live in this space. But we may also be interested in limiting ourselves to a 2-dimensional space, such as points in a plane, like a sheet of paper. Or we may be interested in an n -dimensional space, which can arise in ways that don't represent physical space but are still useful to think about geometrically.

When you fix the space you're talking about, a **vector** is a line segment in space with a magnitude and a direction.

Geometric Definition of a Vector: a line segment with a magnitude and a direction.

If you've learned any physics, you might think of vectors as arrows in space that specify a magnitude and—if the length isn't zero—a direction. This is a geometric definition of a vector.

One end is called the head, and the other is the tail; the arrow points in the direction moving from tail to head. Any 2 arrows with the same magnitude and direction are considered the same vector, no matter where the arrow is placed.



This notion of a vector as an arrow or directed line segment shows why vectors are often used to represent things like motion, forces, and velocities in physics. Forces and velocities have an intrinsic magnitude and direction.

Algebraic Definition of a Vector: an ordered collection of numbers.

Another way that some people think of vectors is as an ordered collection of numbers, such as $(3, 1, 4)$ or $(-1, 2.5)$ or $(2, 3)$. This is an algebraic definition of a vector. The numbers are called **components** or **coordinates**, and for the most part, they will be considered real numbers.

The set of all vectors with n components is called \mathbb{R}^n .

The \mathbb{R} stands for the real numbers, and the superscript represents how many components the vectors have. Remember, this is just notation for a set of vectors. We are not raising numbers to a power; we are just looking at a set of vectors and giving it a name.

So, \mathbb{R}^2 is the set of all ordered pairs of real numbers, such as $(2, 3)$.

Depending on context, the components could be written as a **row vector**, in which the numbers are arranged in a row, or as a **column vector**, in which the numbers are arranged in a column.

$$\vec{v} = \begin{bmatrix} 2 & 3 \end{bmatrix} \text{ row vector}$$

or

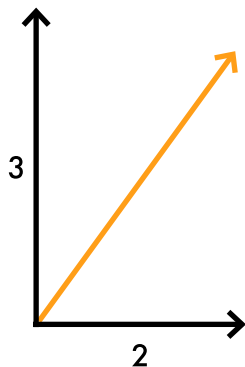
$$\begin{bmatrix} 2 \\ 3 \end{bmatrix} \text{ column vector}$$

components

This view of a vector as an ordered collection of numbers is often used when working with data.

There is a natural correspondence between the algebraic definition of a vector as an ordered collection of n real numbers and the geometric definition of a vector as an arrow in n -dimensional space.

To see this, we must first fix some coordinate directions. In the plane, we usually choose the x and y axes, drawn with the positive y -axis 90° counterclockwise to the positive x -axis. Then, a vector in \mathbb{R}^2 specified by an ordered pair of numbers like $(2, 3)$ corresponds to an arrow in the plane that moves 2 units in the x direction and 3 units in the y direction.

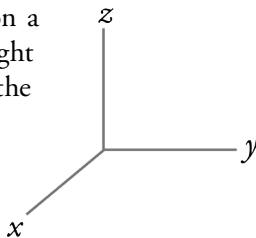


In 3-dimensional space, we usually choose the x , y , and z axes as coordinate directions.

To avoid confusion, there is a standard way to orient the x , y , and z axes: The positive directions along these axes obey the right-hand rule.

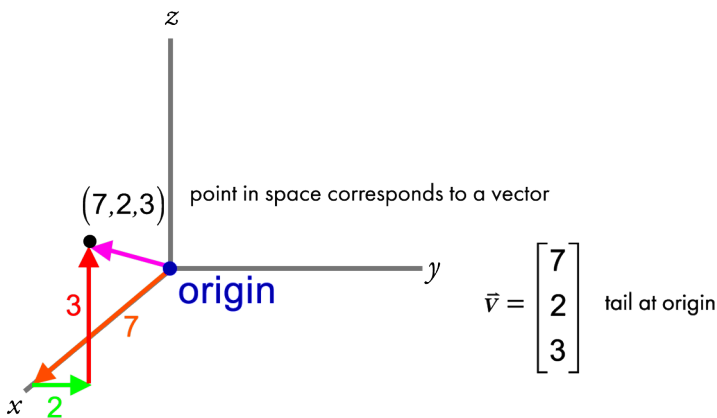
If you curl the fingers of your right hand in the direction moving from the positive x -axis to the positive y -axis, then your thumb will point in the direction of the positive z -axis.

You are probably used to drawing the xy -plane on a sheet of paper with the positive x -axis pointing right and the positive y -axis pointing up. In this case, the positive z -axis will point out of the page.



In \mathbb{R}^3 , a triple of numbers like $(7, 2, 3)$ represents an arrow in 3-dimensional space that moves 7 units in the x direction, 2 units in the y direction, and 3 units in the z direction. Similarly, every vector in \mathbb{R}^n corresponds to an arrow that moves in the coordinate directions by amounts specified by its coordinates.

This correspondence also suggests another way to view an ordered collection of numbers—not as an **arrow** in space, but as a **point** in space at the end of the arrow when you place the arrow's tail at a reference position called the **origin**. The numbers of the vector will then tell you how to get there in relation to the coordinate directions. Thus, the point called $(7, 2, 3)$ is the point that is at the head of the vector $[7 \ 2 \ 3]$ when you place its tail at the origin.



In this view, every point in 3-dimensional space can be represented by a triple of numbers, the same one that represents the vector from the origin. The origin itself is represented by the vector of all zeros, called the zero vector, because the arrow from the origin to itself involves no motion along any of the axes.

In a similar way, you can choose to think of vectors in \mathbb{R}^n as points or as arrows, and the choice you make depends on context.

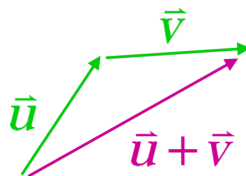
The geometric description of a vector is the picture of a vector you might have in your head, while the algebraic description is one that a computer might most easily work with.

Linear Combinations

Two of the operations you can do with vectors are addition and scalar multiplication. And each of these operations will be presented geometrically (how you might visualize them) and algebraically (how you might compute them).

The first basic operation on vectors is addition.

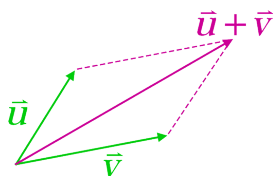
To add 2 vectors \mathbf{u} and \mathbf{v} geometrically, place the tail of \mathbf{v} at the head of \mathbf{u} and look at the arrow formed by moving from the tail of \mathbf{u} to the head of \mathbf{v} . This is $\mathbf{u} + \mathbf{v}$. The order of \mathbf{u} and \mathbf{v} doesn't matter; if you add \mathbf{v} and \mathbf{u} , you get the same result.



So, vector addition is **commutative**, which means the order of addition doesn't matter.

Because vectors have both direction and magnitude, they are often denoted by boldface letters (\mathbf{r}) or by an arrow diacritic (\vec{r}) to distinguish them from scalars. The styles have the same meaning, and both are used in this book.

Another way to view addition is to form the parallelogram with sides \mathbf{u} and \mathbf{v} with tails placed at the origin. Then, an arrow from the origin along the diagonal is the sum of \mathbf{u} and \mathbf{v} .



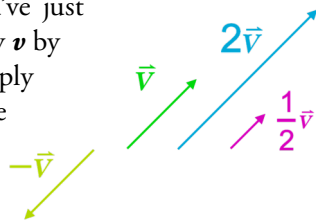
Algebraically, if you are given 2 vectors as ordered collections of numbers in \mathbb{R}^n , the sum is just given by adding the numbers coordinate-wise. For example, if the vector \mathbf{u} is $(1, 2)$ and the vector \mathbf{v} is $(3, -7)$, then the addition rule says $(1, 2) + (3, -7) = (4, -5)$, because you add coordinates 1 plus 3 to get 4 and 2 minus 7 to get -5.

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 3 \\ -7 \end{bmatrix} = \begin{bmatrix} 4 \\ -5 \end{bmatrix}.$$

You could check by drawing this out that the addition matches what is going on in the geometric picture: The new vector $(4, -5)$ says how far you have to walk in the x and y directions after walking according to the instructions given by the vectors \mathbf{u} and \mathbf{v} .

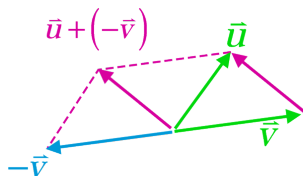
Another operation on vectors is scalar multiplication.

A **scalar** is just a real number, and it gets its name when it behaves as a scaling factor for a vector by an operation called scalar multiplication. Geometrically, to multiply a vector \mathbf{v} by 2, you just double its length while keeping its direction the same—you've just scaled the vector by a factor of 2. To multiply \mathbf{v} by $\frac{1}{2}$, you halve its length. You can also multiply by a negative number, which scales the magnitude but reverses its direction. So, $-\mathbf{v}$ has the same length as \mathbf{v} but points in the opposite direction.



To multiply a vector by a scalar c algebraically, you take every coordinate and multiply it by c . So, 2 times $(3, 1, 4) = (6, 2, 8)$. Soon, when the length of a vector is defined in \mathbb{R}^n , you will be able to see that this operation matches the geometric understanding that it scales the magnitude by 2 but does not change its direction.

Subtraction of vectors can be defined by addition and scalar multiplication. So, if you want \mathbf{u} minus \mathbf{v} , you can add \mathbf{u} and $-\mathbf{v}$.



Because $(\mathbf{u} - \mathbf{v}) + \mathbf{v}$ should be \mathbf{u} , you can also see pictorially that if you place the tails of \mathbf{u} and \mathbf{v} together, then the arrow formed by moving from the head of \mathbf{v} to the head of \mathbf{u} is also \mathbf{u} minus \mathbf{v} .

Scalar multiplication has the **distributive property** over addition:

$$c(\mathbf{u} + \mathbf{v}) = c\mathbf{u} + c\mathbf{v}.$$

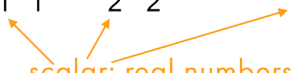
So, it doesn't matter if you add and then scale by c or scale by c first and then add.

Scalar multiplication also has an **associative property**:

If c and d are scalars and \mathbf{u} is a vector,
then $c(d\mathbf{u})$ is the same as $(cd)\mathbf{u}$.

What happens when these 2 operations on vectors—addition and scalar multiplication—are combined? If you have vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$, then a **linear combination** of \mathbf{v}_1 through \mathbf{v}_k is any vector of the form $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \dots + c_k\mathbf{v}_k$, where c_1 through c_k are scalars.

$$c_1\vec{v}_1 + c_2\vec{v}_2 + \dots + c_k\vec{v}_k$$


 scalar: real numbers

So, for example:


$$\begin{array}{ccc}
 7 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} & + 2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} & + 3 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 7 \\ 2 \\ 3 \end{bmatrix} \\
 \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\
 \hat{i} & \hat{j} & \hat{k}
 \end{array}$$

This shows that $(7, 2, 3)$ is a linear combination of the vectors $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$ —which are called the **standard basis vectors** in \mathbb{R}^3 and are often written $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$.

Linear combinations of the standard basis vectors will produce any vector in \mathbb{R}^3 . (This is one of the reasons each of these vectors is called a basis.) But if you had just 2 vectors, you wouldn't be able to obtain any

possible vector in \mathbb{R}^3 . For example, linear combinations of \mathbf{e}_1 and \mathbf{e}_2 can only produce vectors whose third coordinate is 0.

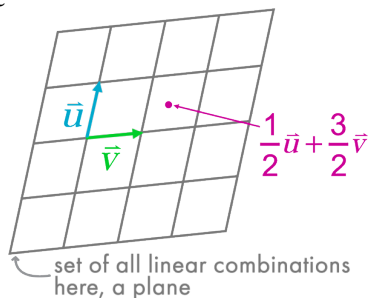
Often, it is not so obvious whether a given vector is a linear combination of a set of vectors. For example, suppose you're piloting a spaceship that sits at the origin and you have 2 thrusters that can propel you in 2 directions: $(5, 1, 1)$ and $(3, 0, -1)$. Can you reach the point $(7, 2, 3)$? In other words, does a linear combination of $(5, 1, 1)$ and $(3, 0, -1)$ exist that will produce $(7, 2, 3)$?



$$\begin{bmatrix} 5 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 3 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 7 \\ 2 \\ 3 \end{bmatrix}$$

Let's think about this abstractly and call these 2 vectors \mathbf{u} and \mathbf{v} . A linear combination of vectors \mathbf{u} and \mathbf{v} is any vector that can be obtained from doing addition and scalar multiplication on the vectors \mathbf{u} and \mathbf{v} . So, it must be of this form: a scalar times \mathbf{u} plus another scalar times \mathbf{v} . For example, $-2\mathbf{u} + 3\mathbf{v}$ is a linear combination of \mathbf{u} and \mathbf{v} , and so is $\frac{1}{2}\mathbf{u} + \frac{3}{2}\mathbf{v}$. The zero vector is also a trivial linear combination of \mathbf{u} and \mathbf{v} —namely, the scalar 0 times \mathbf{u} + 0 times \mathbf{v} .

If you look at the set of all possible linear combinations of \mathbf{u} and \mathbf{v} , you can see that it will form a plane as long as \mathbf{u} and \mathbf{v} do not point in the same direction. You may also be able to see that by fixing either the coefficient of \mathbf{u} or \mathbf{v} and letting the other coefficient vary, you will trace out natural coordinate lines along this plane. In the figure, you can see where $\frac{1}{2}\mathbf{u} + \frac{3}{2}\mathbf{v}$ is located.



This plane cannot fill out all of \mathbb{R}^3 ; some points are clearly not linear combinations of \mathbf{u} and \mathbf{v} .

The set of all linear combinations of a collection of vectors is called the **span** of the vectors. So, the span of \mathbf{u} and \mathbf{v} here will be a plane. In the degenerate case where \mathbf{u} and \mathbf{v} point in the same direction, the span of \mathbf{u} and \mathbf{v} will be just a line.

Note that the problem of determining whether a linear combination will produce a given vector is the same as solving a system of linear equations for the coefficients of the linear combination. In the previous example, if the coefficient of $(5, 1, 1)$ is x and the coefficient of $(3, 0, -1)$ is y , then you want to determine if there exist x and y such that

$$x(5, 1, 1) + y(3, 0, -1) = (7, 2, 3).$$

Looking at the components of this vector equation, with x for the first unknown coefficient and y for the second unknown coefficient, this will be true if the following system of equations can be solved simultaneously:

$$\begin{aligned} 5x + 3y &= 7 \\ x &= 2 \\ x - y &= 3. \end{aligned}$$

Abstract Vector Spaces

In addition to thinking of vectors as arrows and as ordered collections of real numbers, which we can also think of as points in \mathbb{R}^n , a third way to think of vectors is in a more abstract way. Even though this course mainly discusses vectors in \mathbb{R}^n , vectors can be seen as more general objects that apply to a wide range of things that may not at first look like an arrow or an ordered collection of real numbers.

In mathematics, we often take concrete examples and look at what properties made those objects interesting. Then, we begin to define our objects by their properties, and when we do that, we realize that our methods for working with those objects were far more general and apply to other things.

We often do 2 basic things with vectors: We add them, and we scale them by making them bigger or smaller by some factor. We form linear combinations of those vectors.

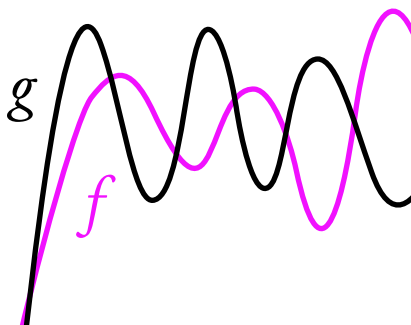
Abstract Definition of a Vector: an object that can be added and scaled (in certain ways).

So, a third way to define a vector is by an abstract definition, in which a vector is defined by its properties. In this definition, a **vector space** is, loosely, any collection of things that can be added and scaled together with some additional properties, called **axioms** (see [PAGE 34](#)), which ensure that addition and scalar multiplication play nicely by themselves and with each other.

The advantage of this definition is that we can now call other things vectors that weren't before.

For example, consider the collection of all continuous functions on the real line. If you are an engineer, such functions might arise as waveforms if you are doing signal processing. When continuous functions are added, you get another continuous function. You can scale a continuous function by multiplying by a scalar, which changes its amplitude.

The space of continuous functions satisfies all the axioms of a vector space, so it is another example of a vector space.



AXIOMS

These axioms are the properties we end up using when we prove anything about \mathbb{R}^n .

A vector space is a set V (objects are “vectors,” e.g., \mathbf{u} , \mathbf{v}) with 2 operations:

addition (write: $\mathbf{u} + \mathbf{v}$)

scalar multiplication (write: $c\mathbf{u}$, for scalar c)

such that for all \mathbf{u} , \mathbf{v} , \mathbf{w} in V and scalars c and d :

- 1 $\mathbf{u} + \mathbf{v}$ is in V (V closed under addition)
 - 2 $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ (addition is commutative)
 - 3 $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$ (addition is associative)
 - 4 There's a zero vector $\mathbf{0}$ such that $\mathbf{u} + \mathbf{0} = \mathbf{u}$
 - 5 Every vector \mathbf{u} has an additive inverse $-\mathbf{u}$ such that $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$
 - 6 $c\mathbf{u}$ is in V (V closed under scalar multiplication)
 - 7 $c(\mathbf{u} + \mathbf{v}) = c\mathbf{u} + c\mathbf{v}$
 - 8 $(c + d)\mathbf{u} = c\mathbf{u} + d\mathbf{u}$
 - 9 $c(d\mathbf{u}) = (cd)\mathbf{u}$
 - 10 $1\mathbf{u} = \mathbf{u}$
- } distributive properties

READINGS

Chartier, *When Life Is Linear*, chap. 4.

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, section 1.3.

Poole, *Linear Algebra*, section 1.1.

LINEAR GEOMETRY: DOTS AND CROSSES

Quantum mechanics says some strange things about the nature of the universe. For example, it says that quantum objects like electrons can exist in a superposition of states. Deep ideas from linear algebra underlie this strangeness. For example, these states actually live in some vector space, and superposition is basically a linear combination of states. That is like saying that a quantum object exists physically in many states at the same time—a particle can be in 2 places at once! Even though quantum states live in an infinite dimensional space that isn't \mathbb{R}^n , some of the results that can be proven for \mathbb{R}^n have analogues for quantum states. So the basic ideas of linear algebra are actually quite profound.

The Dot Product

Euclidean space \mathbb{R}^n has 2 operations—addition and scalar multiplication—and these allow you to take linear combinations of vectors.

Another kind of product structure that \mathbb{R}^n has, called a dot product, plays an important role in determining when 2 vectors are perpendicular.

Linear algebra has a theme of hidden structure. The dot product is an example of a structure that exists that is not obvious at first, but when you realize you have it, it can tell you all sorts of things!

The dot product can be described in 2 ways: one algebraic and one geometric.

Suppose you have \mathbf{u} and \mathbf{v} , 2 vectors in \mathbb{R}^n , with components that are real numbers u_1 through u_n and v_1 through v_n . Then, the algebraic definition of the dot product of \mathbf{u} and \mathbf{v} , called $\mathbf{u} \bullet \mathbf{v}$, is the sum of the pairwise products of the coordinates:

$$\vec{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}, \quad \vec{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \text{ in } \mathbb{R}^n$$

$$\vec{u} \bullet \vec{v} = u_1 v_1 + u_2 v_2 + \dots + u_n v_n.$$

There is a shorthand notation for a sum like this that uses the Greek letter sigma (Σ), which stands for sum.

$$\sum_{i=1}^n u_i v_i.$$

The i represents the index, and the $i=1$ and n tell the reader what the starting and ending indices of the sum are.

For example, if $\mathbf{u} = (3, 1, 4)$ and $\mathbf{v} = (1, -5, 0)$, the dot product turns out to be -2 .

$$\vec{u} = \begin{bmatrix} 3 \\ 1 \\ 4 \end{bmatrix}, \quad \vec{v} = \begin{bmatrix} 1 \\ -5 \\ 0 \end{bmatrix}$$

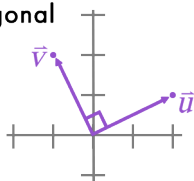
$$\begin{aligned}\vec{u} \bullet \vec{v} &= 3 \bullet 1 + 1(-5) + 4 \bullet 0 \\ &= -2.\end{aligned}$$

Notice that the dot product takes in 2 vectors and spits out a number. What does this number mean?

The dot product measures how aligned 2 vectors are. If the dot product is positive, the vectors are pointing generally in the same direction; if the dot product is negative, they are pointing generally in opposite directions; and if the dot product is zero, they are perpendicular.

Here's an example in \mathbb{R}^2 that you can visualize. If \mathbf{u} is the vector $(2, 1)$ and \mathbf{v} is the vector $(-1, 2)$, then you can see by drawing the vectors on the plane that they are perpendicular—or, in linear algebra terms, orthogonal. And if you compute the dot product of these 2 orthogonal vectors, you see that, indeed, $2 \times -1 + 1 \times 2 = 0$.

orthogonal



$$\vec{u} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \vec{v} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

$$\vec{u} \bullet \vec{v} = 2(-1) + 1 \bullet 2 = 0.$$

You can see why the dot product has this curious property by deriving a geometric definition of the dot product, which comes later in this lecture.

Properties of the Dot Product

Some properties of the dot product are evident from its algebraic definition.

$$\vec{u} \bullet \vec{v} = \vec{v} \bullet \vec{u}$$

$$\begin{aligned}\vec{u} \bullet (\vec{v} + \vec{w}) &= \vec{u} \bullet \vec{v} + \vec{u} \bullet \vec{w} \\ (c\vec{u}) \bullet \vec{v} &= c(\vec{u} \bullet \vec{v}).\end{aligned}$$

The dot product is commutative, meaning that the order of \mathbf{u} and \mathbf{v} doesn't matter. You can see this by looking at the definition.

$$\vec{v} \bullet \vec{u} = v_1 u_1 + v_2 u_2 + \dots + v_n u_n.$$

Notice that if you switch the roles of \mathbf{u} and \mathbf{v} , you get exactly the same formula, because the u_i and v_i trade places as well, but the pairwise products stay the same, because multiplication of real numbers is commutative.

The dot product is distributive: $\mathbf{u} \bullet (\mathbf{v} + \mathbf{w})$ expressed in terms of components of \mathbf{u} and \mathbf{v} is the same expression you get from $(\mathbf{u} \bullet \mathbf{v}) + (\mathbf{u} \bullet \mathbf{w})$.

If you scale one of the vectors, the dot product scales the same way: $(c\mathbf{u}) \bullet \mathbf{v}$ is the same as $c(\mathbf{u} \bullet \mathbf{v})$.

Note that if you dot a vector with itself, you get a sum of squares, which is never negative and is only zero for the zero vector.

$$\vec{u} \bullet \vec{u} = u_1^2 + \dots + u_n^2 \geq 0.$$

The magnitude, or length, of a vector is defined by taking the square root of $\mathbf{u} \cdot \mathbf{u}$. The notation for magnitude is double bars around the vector.

$$\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}.$$

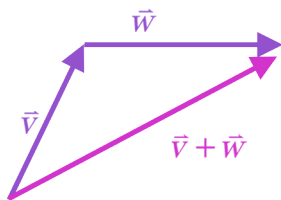
What is the length of a scalar multiple of a vector? In general, the length of c times vector \mathbf{v} is the absolute magnitude of c times the length of \mathbf{v} .

$$\|c\mathbf{v}\| = |c| \|\mathbf{v}\|.$$

What about the length of a sum? In general, it is not the sum of their lengths, because the vectors may not be pointing in the same direction. But by drawing the triangle formed by \mathbf{v} , \mathbf{w} , and $(\mathbf{v} + \mathbf{w})$, you can see that the length of $(\mathbf{v} + \mathbf{w})$ must be less than or equal to the sum of the lengths of \mathbf{v} and of \mathbf{w} .

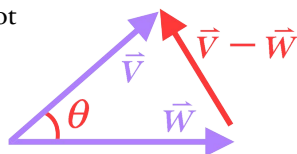
This inequality is called the triangle inequality.

$$\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|$$



A Geometric Formula for the Dot Product

Let's derive a geometric formula for the dot product. Given 2 vectors \vec{v} and \vec{w} , put their tails together and let θ be the angle between \vec{v} and \vec{w} . Notice that the third side of this triangle is $(\vec{v} - \vec{w})$.



In geometry, the law of cosines says that if a , b , and c are sides of a triangle and θ is the angle between sides a and b , then $c^2 = a^2 + b^2 - 2ab \cos \theta$.

Here, the length of $(\vec{v} - \vec{w})^2$ is the length of a^2 plus the length of b^2 minus twice the length of a times the length of b times the cosine of θ .

$$\begin{aligned} \underbrace{\|\vec{v} - \vec{w}\|^2}_{(\vec{v} - \vec{w}) \cdot (\vec{v} - \vec{w})} &= \|\vec{v}\|^2 + \|\vec{w}\|^2 - 2\|\vec{v}\| \|\vec{w}\| \cos \theta. \\ &\quad \swarrow \quad \searrow \\ &\quad \vec{v} \cdot \vec{v} \quad \vec{w} \cdot \vec{w} \\ \vec{v} \cdot \vec{v} - \vec{v} \cdot \vec{w} - \vec{w} \cdot \vec{v} + \vec{w} \cdot \vec{w} \\ &\quad \underbrace{-\vec{v} \cdot \vec{w} - \vec{w} \cdot \vec{v}}_{-2\vec{v} \cdot \vec{w}} \end{aligned}$$

But by definition, the length of a vector squared is just the vector dotted with itself. So, using the properties of dot products, you get the geometric definition of the dot product of \vec{v} and \vec{w} : It's the product of the lengths of \vec{v} and \vec{w} multiplied by the cosine of the angle between.

$$\vec{v} \cdot \vec{w} = \|\vec{v}\| \|\vec{w}\| \cos \theta.$$

This is a geometric definition because it does not involve coordinates, and it only involves things that can easily be discerned from a picture of \mathbf{v} and \mathbf{w} .

Some things can be gleaned from this interpretation of the dot product. For example, if you notice that the cosine of an angle in absolute value is always less than or equal to 1, you get the Cauchy-Schwarz inequality: The magnitude of the dot product of 2 vectors is always less than or equal to the product of their lengths.

$$|\vec{v} \bullet \vec{w}| \leq \|\vec{v}\| \|\vec{w}\| \text{ because } |\cos \theta| \leq 1.$$

And as θ varies, you see the interpretation of the dot product as a measure of alignment of \mathbf{v} and \mathbf{w} . If you keep the vector lengths fixed but change the angle θ , the dot product will change depending only on the cosine of θ . So, if θ is less than 90° , the cosine will be positive, and if it is greater than 90° , the cosine will be negative. And when θ is 90° , the most important property of the dot product applies: For nonzero vectors, the dot product is zero if and only if the 2 vectors are perpendicular.

$$\text{If } \vec{u}, \vec{v} \neq 0, \quad \vec{u} \bullet \vec{v} = 0 \quad \Leftrightarrow \quad \vec{u} \perp \vec{v}.$$

\uparrow
 if and only if

The *if and only if* here means the 2 statements are equivalent—each one implies the other.

QUANTUM MECHANICS

It turns out that something like the dot product does not exist in every vector space, but if it does, it is called an inner product, and the space is called an inner product space.

The kinds of vector spaces that are used to represent quantum states in quantum mechanics are inner product spaces. And the inner product has many analogous properties of the dot product. For example, the Cauchy-Schwarz inequality holds.

The Cauchy-Schwarz inequality turns out to be the fundamental mathematical idea behind Heisenberg's uncertainty principle in quantum mechanics, which says that the position and the momentum of a particle cannot both be specified exactly. So, this strange property of quantum particles is a direct consequence of the underlying mathematics of inner product spaces.

To learn more about quantum mechanics, watch the Great Course *Understanding the Quantum World*, taught by Professor Erica W. Carlson. You can find it online at www.thegreatcourses.com or www.thegreatcoursesplus.com.

The Cross Product

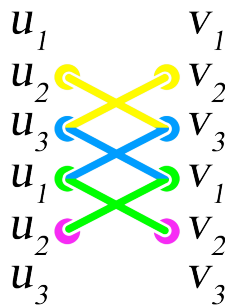
One kind of product of vectors in \mathbb{R}^n is defined only in 3 dimensions. It is called the cross product. And it can be defined algebraically and geometrically.

If you're given vectors \mathbf{u} and \mathbf{v} with components u_1 , u_2 , and u_3 and v_1 , v_2 , and v_3 , the algebraic definition is given by some crazy formulas for the components.

$$\vec{u} \times \vec{v} = \begin{bmatrix} u_2 v_3 - u_3 v_2 \\ u_3 v_1 - u_1 v_3 \\ u_1 v_2 - u_2 v_1 \end{bmatrix}.$$

There is an easy way to remember this formula: Write u_1 , u_2 , and u_3 twice in a column and write v_1 , v_2 , and v_3 twice in a column next to it.

If these are shoelace holes, form shoelace crossings in the middle 3 positions. The strings that go down to the right represent positive products, and the ones that go up to the right represent negative products. If you do this, you'll see how these crosses correspond to the terms in the cross-product expression.



What's the geometric interpretation of the cross product? Unlike the dot product, the cross product of 2 vectors is a vector, not a scalar. So, you need to know both its magnitude and direction.

$\vec{u} \times \vec{v}$ is vector with magnitude $\|\vec{u}\| \|\vec{v}\| \sin \theta$
and direction \perp to \vec{u} and \vec{v} .

The magnitude of $(\mathbf{u} \times \mathbf{v})$ is the length of \mathbf{u} times the length of \mathbf{v} times the sine of the angle θ in between them. So, the magnitude is largest when \mathbf{u} and \mathbf{v} are perpendicular, and it is zero when θ is zero or 180° —in other words, when \mathbf{u} and \mathbf{v} are multiples of one another. The magnitude is never negative, because the angle θ is always measured going from \mathbf{u} to \mathbf{v} , which is positive.

Note how the cross product differs from the dot product: The cosine in the dot product gets replaced by a sine in the cross product, and the dot product is a scalar while the cross product is a vector.

The direction of $(\mathbf{u} \times \mathbf{v})$ is always orthogonal to both vectors \mathbf{u} and \mathbf{v} . There are 2 possible directions that could be orthogonal to \mathbf{u} and \mathbf{v} , but the one that is used is the one given by the **right-hand rule**: If your first finger (your index finger) is pointing in the direction of the first vector and your second finger is pointing in the direction of the second vector, then your thumb will be pointing in the general direction of the cross product. Alternatively, if you curl the fingers of your right hand in the direction from the first vector to the second finger, your thumb will be pointing in the direction of the cross product.

Note that the order matters in both the algebraic and geometric definitions. If you look at $(\mathbf{v} \times \mathbf{u})$ instead of $(\mathbf{u} \times \mathbf{v})$, you see the signs of all the components get reversed, and your thumb will be pointing in the opposite direction. So, $(\mathbf{u} \times \mathbf{v})$ is the negative of $(\mathbf{v} \times \mathbf{u})$.

Describing Lines

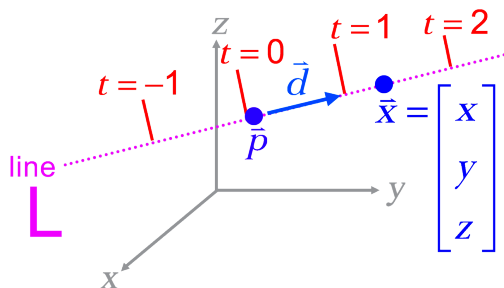
Linear combinations along with the dot product and the cross product can be used to describe lines and planes and hyperplanes.

To specify a line L in \mathbb{R}^3 , you can use a point \mathbf{p} on the line L and a direction vector \mathbf{d} that moves along L . Fixing the point \mathbf{p} and adding scalar multiples of the vector \mathbf{d} will move along this line, and every point on L arises in this way. So, every point (x, y, z) on L must be of the form $\mathbf{p} + t\mathbf{d}$, where t is a freely chosen scalar parameter. Once you choose t , the point (x, y, z) is determined.

$$L = \{ \bar{\mathbf{p}} + t\bar{\mathbf{d}} : t \in \mathbb{R} \}.$$

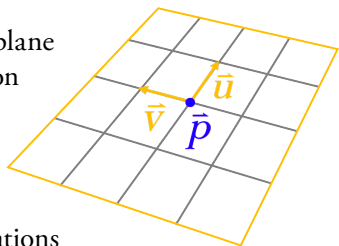
Think of the parameter t as time and the vector $\mathbf{p} + t\mathbf{d}$ as telling you the position of a bug at time t . At time $t=0$, you are at the point \mathbf{p} . If you set down tick marks at equally spaced time intervals—such as time $t=0, t=1, t=2$, etc.—you will see that the tick marks are evenly spaced along the line at $\mathbf{p}, \mathbf{p} + \mathbf{d}, \mathbf{p} + 2\mathbf{d}$, etc. If you plug in $t=-1$, you move in the other direction.

This description of a line offers a way of describing what is happening with the components (x, y, z) of a point on the line.



Describing Planes

You can also use a similar idea to describe a plane in \mathbb{R}^n . Now you need a point \mathbf{p} and 2 direction vectors, \mathbf{u} and \mathbf{v} . Then, you start at that point \mathbf{p} and add linear combinations of \mathbf{u} and \mathbf{v} to generate a plane through \mathbf{p} .



You can see in this image how linear combinations of \mathbf{u} and \mathbf{v} will generate a grid on this plane. The coefficients of \mathbf{u} and \mathbf{v} will be scalar parameters that can be chosen freely—called s and t , for example. Then, any point on this plane must be of the form $\mathbf{p} + s\mathbf{u} + t\mathbf{v}$, where s and t are real numbers.

$$\vec{x} = \vec{p} + \underbrace{s\vec{u} + t\vec{v}}_{\substack{\text{parameters} \\ \text{vector form } s, t \in \mathbb{R}}}$$

This gives you the vector form of the equation of a plane; all you need is a point and 2 directions.

If you write out the equation in components, you will get the parametric form of the equation of the plane. For example, suppose you want to describe the plane through $(1, 0, 0)$ and in the directions $(4, 5, 0)$ and $(7, 0, 9)$.

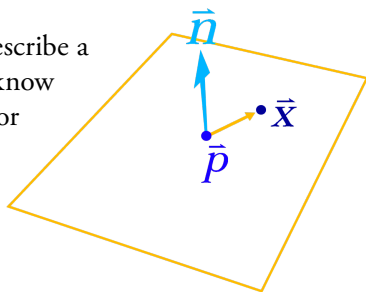
$$\vec{x} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + s \begin{bmatrix} 4 \\ 5 \\ 0 \end{bmatrix} + t \begin{bmatrix} 7 \\ 0 \\ 9 \end{bmatrix}, \quad s, t \in \mathbb{R}$$

$$\begin{aligned} x &= 1 + 4s + 7t \\ y &= 5s \\ z &= 9t \end{aligned}$$

If vector \mathbf{x} has components (x, y, z) , then looking at the components of the vector form $(1, 0, 0) + s(4, 5, 0) + t(7, 0, 9)$, you see that x must be $1 + 4s + 7t$, $y = 5s$, and $z = 9t$, where s and t can be chosen freely as any real number.

Notice that this description of a plane shows intuitively why this plane has 2 dimensions: because it has 2 degrees of freedom in the free parameters s and t . Also, a plane, like a line, may have many different parametrizations.

In \mathbb{R}^3 , you can use the dot product to describe a plane in a slightly different way. If you know a point in the plane and a direction vector perpendicular to the plane (called a normal vector), that is enough to specify the plane.



Suppose \mathbf{p} is a point in the plane and \mathbf{n} is a normal vector to the plane. If vector \mathbf{x} is any other point in the plane, then $(\mathbf{x} - \mathbf{p})$ is a vector that must be perpendicular to \mathbf{n} . So, the dot product of $(\mathbf{x} - \mathbf{p})$ with \mathbf{n} must be zero.

$$(\bar{\mathbf{x}} - \bar{\mathbf{p}}) \cdot \bar{\mathbf{n}} = 0$$

$$\boxed{\bar{\mathbf{x}} \cdot \bar{\mathbf{n}} = \bar{\mathbf{p}} \cdot \bar{\mathbf{n}}} \quad \text{normal form}$$

If you distribute $(\mathbf{x} - \mathbf{p}) \cdot \mathbf{n}$ as $(\mathbf{x} \cdot \mathbf{n}) - (\mathbf{p} \cdot \mathbf{n})$ and move $(\mathbf{p} \cdot \mathbf{n})$ to the other side of the equation, you get that $(\mathbf{x} \cdot \mathbf{n})$ must equal $(\mathbf{p} \cdot \mathbf{n})$. This relationship is sometimes called the normal form of the equation of a plane. So, any point \mathbf{x} on the plane must satisfy this relationship.

READINGS

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, section 1.1.

Poole, *Linear Algebra*, sections 1.2 and 1.3.

MATRIX OPERATIONS

This lecture introduces the matrix, which is an array of numbers. Specifically, the lecture defines matrices and their algebra and offers several applications of the idea of matrix multiplication.

What Is a Matrix?

An $m \times n$ matrix A is an array of numbers with m rows and n columns.

$A = (a_{ij})$, where a_{ij} is the entry in row i and column j .

For example, if A is the following 2×3 , then $a_{21} = 4$.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

$a_{21} = 4.$
 row ↗ ↖ column

To say that 2 matrices are equal means they have the same dimensions and their entries are identical.

To add 2 matrices, you add their corresponding entries. For example, look at the following matrices and note that, in the bottom right corner, $4 + 1 = 5$.

$$\text{If } D = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, E = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \text{ then } D + E = \begin{bmatrix} 2 & 1 \\ 3 & 5 \end{bmatrix}.$$

To multiply a matrix by a scalar, you multiply every entry by that scalar.

$$\frac{1}{2}D = \begin{bmatrix} \frac{1}{2} & 1 \\ \frac{3}{2} & 2 \end{bmatrix}.$$

Matrix Multiplication

Suppose you have some variables y_1 and y_2 written as linear combinations of 3 variables x_1 , x_2 , and x_3 .

$$\begin{aligned} y_1 &= 3x_1 + x_2 + 4x_3 \\ y_2 &= x_1 + 5x_2 + 6x_3 \end{aligned}$$

Suppose also that you have the x variables written as linear combinations of the variables w_1 and w_2 .

$$\begin{aligned} x_1 &= -w_1 \\ x_2 &= 2w_1 \\ x_3 &= w_1 + w_2 \end{aligned}$$

Simply substituting the second set of equations into the first results in expressions for the y variables as linear combinations of the w variables.

$$\begin{aligned} y_1 &= 3w_1 + 4w_2 \\ y_2 &= 15w_1 + 6w_2 \end{aligned}$$

Now suppose you write matrices that represent the coefficients in these changes of variables.

$$\begin{bmatrix} 3 & 1 & 4 \\ 1 & 5 & 6 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 2 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 4 \\ 15 & 6 \end{bmatrix}.$$

There is a relationship between the entries of the matrices.

- ◆ The dimensions of the y -in-terms-of- x matrix are 2×3 , because there are 2 y variables and 3 x variables.
- ◆ The dimensions of the x -in-terms-of- w matrix are 3×2 , because there are 3 x variables and 2 w variables.
- ◆ So, the dimensions of the y -in-terms-of- w matrix are 2×2 , reflecting that there are 2 y variables and 2 w variables.
- ◆ The fact that there is a 15 in the bottom left corner of the 2×2 matrix refers to the number of times w_1 appears in y_2 . This comes from multiplying the number of times a particular x variable appears in y_2 by the number of times w_1 appears in that particular x variable, summed over all x variables. This computation is just the dot product of the second row of the first matrix with the first column of the second matrix.

$$\begin{bmatrix} 3 & 1 & 4 \\ 1 & 5 & 6 \end{bmatrix}_{2 \times 3} \begin{bmatrix} -1 & 0 \\ 2 & 0 \\ 1 & 1 \end{bmatrix}_{3 \times 2} = \begin{bmatrix} 3 & 4 \\ 15 & 6 \end{bmatrix}_{2 \times 2}.$$

The pattern in these changes of variables suggests a natural notion of multiplication of matrices.

- ◆ If you want to multiply 2 matrices A and B , there have to be conditions on the dimensions: The number of columns of A must be the number of rows of B . Then, you can form the product AB , which can be called C .
- ◆ If A is $m \times p$ and B is $p \times n$, the matrix C will be $m \times n$.

$$c_{ij} = \sum_{k=1}^p a_{ik} b_{kj}.$$

To get the ij^{th} entry of C , take the dot product of the i^{th} row of A with the j^{th} column of B .

Note that the order of multiplication matters greatly. Even if AB is defined, the product BA may not be. And when they are both defined, it is not true in general that $AB = BA$, so **matrix multiplication is not commutative**.

Aside from representing compositions of linear functions, matrix multiplication can be used to express a system of linear equations, which you can think of as a linear combination of variables that is set equal to a constant.

For example, the following system of equations on the left is a system of linear equations because both equations are linear combinations of variables set equal to constants.

$$\left. \begin{array}{l} 3x_1 + x_2 + 4x_3 = 1 \\ x_1 + 5x_2 + 6x_3 = 5 \end{array} \right\} \text{ same } \left\{ \begin{bmatrix} 3 & 1 & 4 \\ 1 & 5 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \end{bmatrix} \right.$$

The system of equations on the left can be expressed as a matrix equation on the right. Notice that a column vector (x_1, x_2, x_3) is just a 3×1 matrix, so the usual rules for matrix multiplication applies.

What's great about a matrix equation is that you can express a large system of linear equations, such as the one shown at right, in the very compact form $A\mathbf{x} = \mathbf{b}$, where \mathbf{x} and \mathbf{b} are column vectors.

$$\begin{aligned} a_{11}x_1 + \dots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \dots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + \dots + a_{mn}x_n &= b_m \end{aligned}$$

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

It can be helpful to have this point of view: The matrix A is acting on the vector \mathbf{x} to produce the vector \mathbf{b} .

The Identity Matrix

What you may notice after doing several examples of multiplication is the following fact: If the columns of B are called \vec{b}_i , then the columns of the product AB are just the column vectors you get by doing A times \vec{b}_i in the corresponding columns.

$$\text{If } B = \begin{bmatrix} \vec{b}_1 & \cdots & \vec{b}_n \end{bmatrix} \text{ then } AB = \begin{bmatrix} A\vec{b}_1 & \cdots & A\vec{b}_n \end{bmatrix}.$$

Moreover, if you take a matrix A and multiply it by the j^{th} standard basis vector \vec{e}_j , then you just get the j^{th} column of A .

For each row of A , the 1 in the j^{th} position of \vec{e}_j just picks off a single entry in the j^{th} column of A .

$$\vec{e}_j = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \leftarrow j^{\text{th}}$$

$$A\vec{e}_j = j^{\text{th}} \text{ column of } A.$$

This means that if you fill a matrix with the standard basis vectors in order, you get a matrix that when you multiply another matrix by this one on the right, it stays unchanged!

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}.$$

Here, the highlighted matrix is called the **identity matrix**. It is always a square matrix with 1s along the diagonal and 0s everywhere else. The letter I is often used to denote it; sometimes I_n is used to remind the reader of its dimension.

Similarly, if you want to pick off a row of the matrix A , you can multiply it on the left by the row vector \mathbf{e}_i^T . That will produce the i^{th} row of the matrix A . So, multiplication by the identity matrix on the left of an appropriate dimension will leave its rows unchanged.

To summarize, if you take an $m \times n$ matrix A , you can multiply it on the left by I_m or on the right by I_n and it will stay unchanged.

$$\begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} \quad \begin{aligned} A_{m \times n} I_n &= A \\ I_m A_{m \times n} &= A. \end{aligned}$$

This identity matrix is a multiplicative identity. There is also an additive identity, which you can add to any matrix and leave it unchanged. It is just the zero matrix: the matrix of all zeros. Every matrix also has an additive inverse—that is, a matrix you can add to it to get the zero matrix. The additive inverse of A just puts a negative sign on every entry of A , and it is labeled $-A$.

$$\text{Additive identity: } A + [0] = A.$$

$$\text{Inverse additive identity: } A + (-A) = [0].$$

\uparrow
 a_{ij}

\uparrow
 $-a_{ij}$

Other Matrix Properties

First, remember that multiplication is not commutative; in general, the order of multiplication matters. So, whenever you write a matrix product, you must pay attention to the order of multiplication so that you don't accidentally switch things around.

A very important property of matrix multiplication is the way that it distributes over addition. Thus, A times $(B + C)$ is $AB + AC$. This is called the distributive property.

Distributive property: $A(B + C) = AB + AC$.

This can be shown by writing out the expression for the ij^{th} entry of both sides, using the row-column formula.

$$\begin{aligned} [A(B+C)]_{ij} &= \sum_{k=1}^p a_{ik} (b_{kj} + c_{kj}) \\ &= \sum_{k=1}^p a_{ik} b_{kj} + \sum_{k=1}^p a_{ik} c_{kj} \\ &= [AB]_{ij} + [AC]_{ij}. \end{aligned}$$

If you do this, you see that the distributive property follows from the distributive property of real numbers.

Even though matrix multiplication is not commutative, it does have an **associative property**, which may be surprising because it is not obvious at first glance.

Associative property: $A(BC) = (AB)C$.

If you look at the product $A(BC)$ and write out the expression for the ij^{th} entry, you will get a sum over some index k .

$$\begin{aligned} [A(BC)]_{ij} &= \sum_{k=1}^p a_{ik} \left(\sum_{l=1}^q b_{kl} c_{lj} \right) \\ &= \sum_{k=1}^p \sum_{l=1}^q a_{ik} b_{kl} c_{lj} \\ &= \sum_{l=1}^q \left(\sum_{k=1}^p a_{ik} b_{kl} \right) c_{lj} \\ &= [(AB)C]_{ij}. \end{aligned}$$

Now write out the expression for the kj^{th} entry of BC , and you will get another sum over an index l .

If you manipulate this by switching the order of the sums, you can see it will give the expression for the product $(AB)C$, using the row-column formula.

There is a natural operation on a matrix that switches its rows and columns called the **transpose**. To denote the transpose of A , you write A^T , and if the entries of A are a_{ij} , then the entries of A^T are a_{ji} .

Transpose: If $A = (a_{ij})$, then $A^T = (a_{ji})$.

As an example, the matrix with rows 123, 456 has as its transpose a matrix whose columns are 123, 456.

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

A matrix that stays unchanged when performing the transpose must be square and have a symmetry about the diagonal that goes down and to the right. Such a matrix is called **symmetric**.

$$\begin{bmatrix} 1 & 5 \\ 5 & 2 \end{bmatrix}$$

$$A = A^T.$$

Here are some properties of the transpose:

- a The transpose of the transpose of A is itself. $\longrightarrow (A^T)^T = A.$
- b The transpose of a sum is the sum of the transposes. $\longrightarrow (A + B)^T = A^T + B^T.$
- c The transpose of a scalar multiple of A is the scalar multiple of the transpose of A . $\longrightarrow (cA)^T = cA^T.$

Less obvious is how to take the transpose of a product. $\longrightarrow (AB)^T = B^T A^T.$
If you examine the dimensions of AB , you'll see that for the transpose, you have to reverse the order of the product for the dimensions to work out correctly.

READINGS

Chartier, *When Life Is Linear*, chaps. 2, 3, and 4.

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, sections 1.4 and 2.1.

Poole, *Linear Algebra*, sections 3.7, 8.1, and 8.2.

LINEAR TRANSFORMATIONS

In many instances, linear transformations represent many of the kinds of transformations you would use if you were coding a video game that needs to move around a virtual room and change perspective. In advanced computer graphics, you can represent rotations and translations and deal with perspective using linear transformations on \mathbb{R}^4 and represented by 4-dimensional matrices. In other words, doing these kinds of 3-dimensional transformations amounts to linear algebra in 4 dimensions. The mathematical ideas behind perspective geometry come from linear algebra, and if you understand it, you can produce realistic computer graphics.

Multivariable Functions

A function is any machine for turning an input into an output. Consider the case where the inputs and outputs are sets of numbers, which can be thought of as vectors.

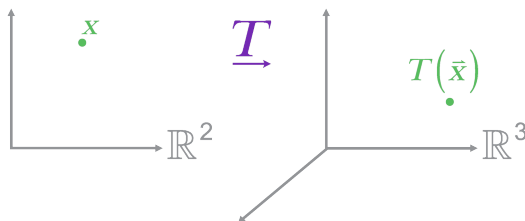
Perhaps the inputs are a vector of quantities like grams of nuts and chocolates and the outputs are numbers like calories of carbohydrates, fats, and protein. That would be a “nutrition” function from \mathbb{R}^2 to \mathbb{R}^3 .

(nut, chocolate) grams \rightarrow (carb, fat, protein) calories

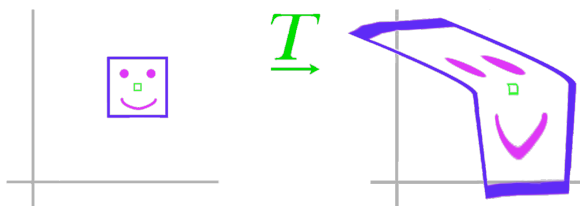
You could also have a function going from \mathbb{R}^2 to \mathbb{R}^2 ; an example would be a function that moved points around a plane. And you could have a function from \mathbb{R}^3 to \mathbb{R}^1 , such as a temperature function whose input is a position in a room and output is a temperature.

$T: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is written in this way to signify that T is a function taking an n -dimensional vector to an m -dimensional vector.

A function is also called a **mapping** or a **transformation**. A function T sends a point \mathbf{x} to the point $T(\mathbf{x})$, called the **image** of \mathbf{x} .



This terminology also applies to whole sets. The image of a smiley face is where the function T sends the entire smiley face. If the function is continuous, the image will look like a warped version of the original set.



The set of potential inputs to a function is the **domain** of the function, and the set of potential outputs is the **codomain** of the function. In this case, the domain is \mathbb{R}^n and the codomain is \mathbb{R}^m .

If T is a function, then the **range** of T is all the points of the codomain that are actually images of points in the domain. The range may or may not be all of the codomain, but if it is, the function T is said to be **onto**.

If a function T is **one-to-one**, it means you can't have 2 distinct points in the domain getting mapped by T to the same point in the codomain. In other words, if $T(x) = T(y)$ for some points x and y in the domain, then x must equal y .

Think of a function from living people to their last names. The domain is the set of all people. The codomain is the set of all possible last names. The range is the set of last names in use by living people.

This function is not onto if there are some last names that are no longer in use. This function is not one-to-one, because there are many people with the same last name.

Definition of a Linear Transformation

Not all functions are linear transformations, and there are additional properties that a linear transformation must satisfy.

Often, 2 basic things are done with vectors: adding them and scaling them by making them bigger or smaller by some factor. Linear combinations of those vectors are formed. A linear transformation is a function that plays nicely with respect to linear combinations.

A function T from \mathbb{R}^n to \mathbb{R}^m is a linear transformation if it satisfies 2 properties.

- a** $T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$ for all vectors \mathbf{u} and \mathbf{v} in \mathbb{R}^n . This means that if you add 2 vectors and then transform the result, you get the same thing as if you transform the vectors first and then add them. In other words, a linear transformation preserves the additive structure of vectors.
- b** It also preserves scalar multiplication: $T(c\mathbf{u}) = cT(\mathbf{u})$ for all scalars c and for all vectors \mathbf{u} . For example, if you transform a vector scaled by a factor of 5, you get 5 times the transformed vector.

Taken together, these 2 properties mean that linear transformations preserve the structure of linear combinations. If you take a linear combination of vectors and transform it, that will be the same result as if you transform it and then take the same linear combination of the transformed vectors.

This is where a linear transformation gets its name: It's a transformation that preserves linear combinations.

Look at this example, which applies both properties:

$$T(3\mathbf{u} - 7\mathbf{v}) = T(3\mathbf{u}) + T(-7\mathbf{v}) = 3T(\mathbf{u}) - 7T(\mathbf{v}).$$

Properties of Linear Transformations

What properties follow from the definition of a linear transformation?

The first thing to notice is that if you take the scalar to be zero, then property (b) implies that $T(\mathbf{0}) = \mathbf{0}$.

The first zero vector is the n -dimensional zero vector, and the second is the m -dimensional zero vector.

So, if a function doesn't take the zero vector to the zero vector, it can't be a linear transformation!

Another thing to notice is that a linear transformation must take lines to lines. To see this, recall that a line through a point \mathbf{p} extending along the direction vector \mathbf{d} can be expressed as $\mathbf{p} + t\mathbf{d}$, where t runs through all real numbers. If you think of t as time and $\mathbf{p} + t\mathbf{d}$ as the position of a bug, then the bug starts at some basepoint \mathbf{p} at time 0, and as time runs forward, the bug moves in the direction \mathbf{d} .

When you apply the transformation T to this line, you get $T(\mathbf{p} + t\mathbf{d}) = T(\mathbf{p}) + tT(\mathbf{d})$, where t runs through all real numbers. But this is just a point $T(\mathbf{p})$ plus t times a direction $T(\mathbf{d})$. In other words, you get a line—one that passes through $T(\mathbf{p})$ and extends along the direction $T(\mathbf{d})$. So, the transformation T takes a straight line L in \mathbb{R}^n to a straight line L' in \mathbb{R}^m .

$$\begin{aligned} T(\vec{p} + t\vec{d}) &= T(\vec{p}) + T(t\vec{d}) \\ &= T(\vec{p}) + tT(\vec{d}). \end{aligned}$$

If you take the line $\mathbf{p} + t\mathbf{d}$ and then change the basepoint \mathbf{p} to basepoint \mathbf{q} , the direction of the line is still \mathbf{d} (it doesn't change). So, the 2 lines $\mathbf{p} + t\mathbf{d}$ and $\mathbf{q} + t\mathbf{d}$ are parallel. Then, their images, $T(\mathbf{p}) + tT(\mathbf{d})$ and $T(\mathbf{q}) + tT(\mathbf{d})$, are also parallel, because they have direction vector $T(\mathbf{d})$. So, a linear transformation takes parallel lines to parallel lines!

$$\begin{aligned}\mathbf{p} + t\mathbf{d} &\rightarrow T(\mathbf{p}) + tT(\mathbf{d}) \\ \mathbf{q} + t\mathbf{d} &\rightarrow T(\mathbf{q}) + tT(\mathbf{d}).\end{aligned}$$

If you change the basepoint again, by the same amount as from \mathbf{p} to \mathbf{q} , then the image basepoint changes by the same amount as it did from $T(\mathbf{p})$ to $T(\mathbf{q})$. So, a linear transformation must take equally spaced parallel lines to equally spaced parallel lines.

This property means that linear transformations must take squares to parallelograms, because the sides must remain parallel but the angles might change. It also means that equally spaced inputs must lead to equally spaced outputs.

So, linear transformations must take a grid of squares to a grid of parallelograms, and zero must go to zero.

Linear transformations are the nicest type of multivariable function, because the outputs will depend on the inputs in a very nice way. Linearity says that if your input changes from \mathbf{v} to $\mathbf{v} + \mathbf{w}$, then the output changes from $T(\mathbf{v})$ to $T(\mathbf{v}) + T(\mathbf{w})$. In other words, if the input changes by a vector \mathbf{w} , then the output will change by a vector $T(\mathbf{w})$, no matter what \mathbf{v} is! Equal changes in input lead to equal changes in output.

Think back to the nutrition function that takes grams of nuts and chocolates to calories of various kinds.

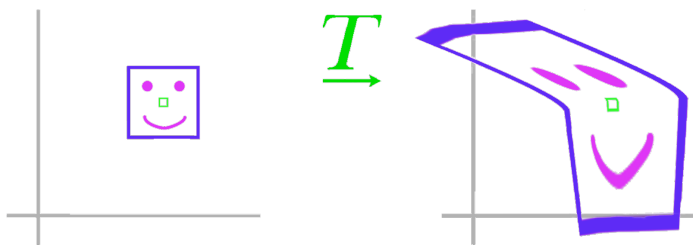
$$(\text{nut, chocolate}) \text{ grams} \rightarrow (\text{carb, fat, protein}) \text{ calories.}$$

You may not know the actual function that does this, but if you think about how such a function should behave, there are good reasons to believe it must be linear.

- ◆ Adding an additional gram of chocolate will change the calorie vector the same amount, no matter how many nuts and how much chocolate have been consumed. So, equally spaced inputs (such as adding a gram of chocolate incrementally) lead to equally spaced changes in calories, the way linear functions behave.
- ◆ The nutrition function sends the zero vector to the zero vector—another sign of linearity.

And if a multivariable function is not linear, it is often approximately linear. This is the whole message of multivariable calculus—to approximate multivariable functions by linear transformations.

For example, look again at this function from \mathbb{R}^2 to \mathbb{R}^2 . It does not send the square smiley face to a parallelogram.



However, if the function is approximately linear—or, in calculus terms, differentiable—then small-enough squares will get sent to things that are approximately parallelograms, like the nose on this face!

In multivariable calculus, you learn that the derivative of a function is a linear transformation, because that's the best linear approximation to a function at a single point!

Matrix Multiplication Is a Linear Transformation

An example of a linear transformation is multiplication by a matrix.

Suppose A is an $m \times n$ matrix, with m rows and n columns. Then, multiplication by A will take an n -dimensional vector to an m -dimensional vector. So,

$$T(\mathbf{x}) = A\mathbf{x}$$

is a function from \mathbb{R}^n to \mathbb{R}^m . Moreover, it is linear! You know from thinking about what it means to multiply matrices that for any scalar factor c and n -dimensional vector \mathbf{v} ,

$$A(c\mathbf{v}) = c(A\mathbf{v}).$$

Also, from the distributive property of matrix multiplication, you know that for n -dimensional vectors \mathbf{x} and \mathbf{y} ,

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

So, multiplication by a matrix is a linear transformation!

Amazingly, there really are no other examples of linear transformations. Matrix multiplication is the only kind of linear transformation there is for functions between finite-dimensional vector spaces, because every linear transformation from \mathbb{R}^n to \mathbb{R}^m can be represented as multiplication by some $m \times n$ matrix A .

So, understanding linear transformations is the same as understanding matrices.

This amazing fact is true because every vector can be written in terms of the standard basis vectors. Recall that the i^{th} standard basis vector \mathbf{e}_i is the vector of all 0s except for a 1 in the i^{th} position.

Any other vector can be written as a linear combination of the standard basis vectors. If you have a vector (x_1, x_2, \dots, x_n) , it equals x_1 times the first basis vector plus x_2 times the second basis vector, etc.

$$\bar{\mathbf{e}}_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i^{\text{th}}$$

$$\bar{\mathbf{x}} = x_1 \bar{\mathbf{e}}_1 + x_2 \bar{\mathbf{e}}_2 + \dots + x_n \bar{\mathbf{e}}_n.$$

If you apply a linear transformation T to this linear combination and use the linearity properties—which preserve linear combinations—you will see that

$$T(\bar{\mathbf{x}}) = x_1 T(\bar{\mathbf{e}}_1) + x_2 T(\bar{\mathbf{e}}_2) + \dots + x_n T(\bar{\mathbf{e}}_n).$$

The right side of this expression can be represented as matrix multiplication by a matrix whose columns are the transformed basis vectors. You can check this simply by thinking about what matrix multiplication means.

This matrix multiplication performs the dot product of each of the rows of the matrix with the vector (x_1, \dots, x_n) in succession. This is just taking a linear combination of the columns of this highlighted matrix.

$$\underbrace{\begin{bmatrix} | & & | \\ T(\bar{e}_1) & \dots & T(\bar{e}_n) \\ | & & | \end{bmatrix}}_{[T]} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

“The matrix representing T ” is sometimes notated by putting brackets around T : $[T]$.

A linear transformation T is determined by where T sends the standard basis vectors.

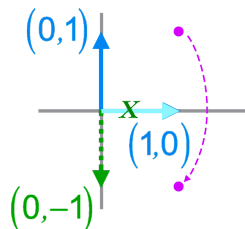
To find the matrix that represents T , all you have to do is create a matrix whose columns are the transformed basis vectors. In other words, the matrix records what it does to the standard basis vectors in its columns! Looking at the columns shows you what the matrix is doing to the standard basis vectors.

Examples of Linear Transformations

The idea that you can find a matrix that represents a linear transformation is really powerful! It means you can tell a computer how to perform a linear transformation, even if all you have in your head is a geometric picture.

You can see how powerful this idea is by doing some examples.

In the plane, a **reflection across the x -axis** is a linear transformation, because the linearity properties hold. If you scale a vector and then reflect, that is the same as reflecting then scaling. And if you add 2 vectors and then reflect, that's the same as reflecting and then adding the 2 vectors.



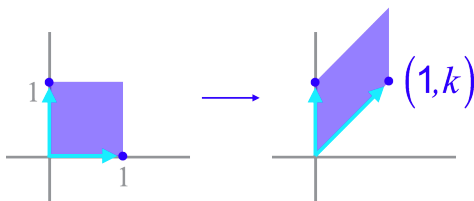
If you want to represent this reflection by matrix multiplication, you just need to figure out where $(1, 0)$ and $(0, 1)$ go. You can see that $(1, 0)$ is not changed by this reflection, so it goes to $(1, 0)$. However, $(0, 1)$ goes to $(0, -1)$. So, form a matrix with first column $(1, 0)$ and second column $(0, -1)$.

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \downarrow \downarrow \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ -1 \end{pmatrix} \text{ so } A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

This matrix A will actually perform reflection across the x -axis. You can check that $A(x, y) = (x, -y)$.

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

Another example is a **shear transformation**, which keeps one side of a square the same but pushes the other side in the direction of the first. So, $(0, 1)$ stays fixed, but $(1, 0)$ gets pushed up by $(0, k)$ to the point $(1, k)$.



Thus, its matrix has $(1, k)$ in the first column and $(0, 1)$ in the second column.

Multiplying a vector by this matrix will perform this shear.

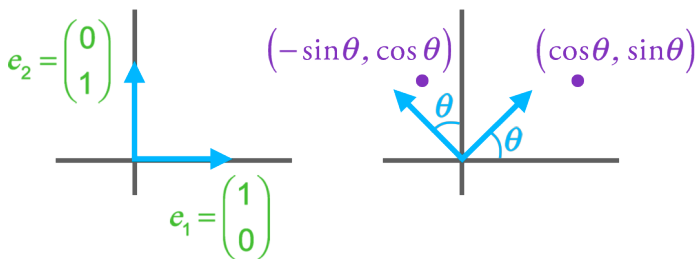
$$\begin{bmatrix} 1 & 0 \\ k & 1 \end{bmatrix}$$

where standard basis goes

A final example is a **rotation of the plane by angle θ** . This is a linear transformation. Again, you can check this by checking the linearity properties: Taking a linear combination of vectors first and then rotating is the same as rotating vectors and then taking linear combinations.

So, there must be a matrix representing rotation. Which one?

Look at where the basis vectors go and form a matrix with those vectors as columns.



Using trigonometry, e_1 , the vector $(1, 0)$, goes to the vector $(\cos \theta, \sin \theta)$, and e_2 , the vector $(0, 1)$, goes to $(-\sin \theta, \cos \theta)$.

$$[T] = \begin{bmatrix} \overset{T(e_1)}{\downarrow} \cos \theta & \overset{T(e_2)}{\downarrow} -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$$

where standard basis goes

Then, you form a matrix with the image of \mathbf{e}_1 as the first column and the image of \mathbf{e}_2 as the second column. The order matters here. And if you take a vector \mathbf{x} and multiply it by the matrix on the left, it will perform rotation by θ !

If you perform this rotation twice in a row, you get rotation by 2θ .

$$\begin{bmatrix} T_{2\theta} \end{bmatrix} \bar{\mathbf{x}} = \begin{bmatrix} T_{\theta} \end{bmatrix} \begin{bmatrix} T_{\theta} \end{bmatrix} \bar{\mathbf{x}}.$$

But this must mean that when multiplied twice in a row (in other words, squared), the matrix representing rotation by θ must be equal to the matrix representing rotation by 2θ .

The matrix representing rotation by 2θ is this matrix:

$$\begin{bmatrix} \cos 2\theta & -\sin 2\theta \\ \sin 2\theta & \cos 2\theta \end{bmatrix}.$$

But if you square the matrix representing rotation by θ , you can check that you get this matrix:

$$\begin{bmatrix} \cos^2 \theta - \sin^2 \theta & -2\sin \theta \cos \theta \\ 2\sin \theta \cos \theta & \cos^2 \theta - \sin^2 \theta \end{bmatrix}.$$

These matrices must be the same, so if you set the corresponding entries equal, you get the double-angle formulas from trigonometry!

When you perform 2 linear transformations one after the other—called a composition of functions—it is the same as multiplying 2 matrices.

READINGS

Lamb, “How to Look at Art,” <https://blogs.scientificamerican.com/roots-of-unity/how-to-look-at-art-a-mathematician-s-perspective/>.

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, sections 1.8 and 1.9.

Poole, *Linear Algebra*, section 3.6 and the Vignette that follows.

SYSTEMS OF LINEAR EQUATIONS

One of the main applications of linear algebra is solving systems of linear equations. This lecture focuses on how to think about solutions to systems of linear equations, both geometrically and algebraically.

Linear Equations

A linear equation in the variables x_1 through x_n is an equation of the form

$$a_1x_1 + a_2x_2 + \dots + a_nx_n = b,$$

where the a_i 's are real coefficients and b is a constant.

In other words, it's an equation where some linear combination of variables is set equal to a constant.

For example, the equation $3x + 2y - 7z = 5$ is a linear equation in the variables x , y , and z because you have a linear combination of x , y , and z set equal to a constant.

Linear equations have the property that each of the variables appears only to the first power and is multiplied by a constant.

Nonlinear equations, on the other hand, have a strange dependence on at least one of the variables.

For example, $3x - 2yz = 4$ is not linear, because the y and z are multiplied together. Also, $x^2 + y^2 + z^2 = 25$ is not a linear equation, because the left side depends on the squares of the variables.

Linear equations can, because of their form, always be represented by setting a dot product of some vector (a_1 through a_n) with a vector of variables (x_1 through x_n) equal to a constant.

$$(a_1 \dots a_n) \bullet (x_1 \dots x_n) = b.$$

You can also represent it as a matrix product of a single row of constants times a single column of variables.

$$\begin{bmatrix} a_1 & \dots & a_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = b.$$

The set of vectors (x, y, z) that satisfies the linear equation shown at right is a plane in \mathbb{R}^3 .

$$3x + 2y - 7z = 5.$$

Also, $3x + 2y = 5$ has a solution set that looks like a line in \mathbb{R}^2 .

This is a general feature of a linear equation in n variables: The set of all points in \mathbb{R}^n that satisfies the linear equation will be a linear object called an $(n - 1)$ -dimensional hyperplane. A **hyperplane** is a fancy name for the generalization of a plane. A point is a 0-dimensional hyperplane, a line is a 1-dimensional hyperplane, and a plane is a 2-dimensional hyperplane. The point is that linear equations have very nice solution sets that are also very linear, flat objects.

Contrast that with the nonlinear equation $x^2 + y^2 + z^2 = 25$.

This is the equation of a sphere, which means that the set of all points that satisfies this equation will form a surface in \mathbb{R}^3 that looks like a sphere. This solution set is not a hyperplane.

Systems of Linear Equations

A system of linear equations is just a collection of more than one linear equation. For example, this is a system of linear equations.

$$\begin{aligned}x_1 + x_2 + x_3 &= 4 \\2x_1 + 5x_2 - x_3 &= 11 \\-x_1 \quad \quad + 2x_3 &= -3.\end{aligned}$$

A system of linear equations naturally arises in domains such as economics, chemistry, and physics.

- ◆ In economics, a linear equation might arise from a budget constraint.
- ◆ In chemistry, a system of linear equations naturally arises when you are trying to balance a chemical equation.
- ◆ In physics, linear equations often arise from the study of electrical networks.

Linear equations arise naturally in problems where you are demanding that some linear combination of things satisfies some constraint. For example, if you want to know how many quarters and nickels you need to make \$1, you get a linear equation. If q is the number of quarters and n is the number of nickels, then by counting cents, you get this relationship:

$$q25 + n5 = 100 \text{ cents.}$$

That's a linear equation in the variables q and n .

Solving Systems of Linear Equations

Suppose you have the following system of equations.

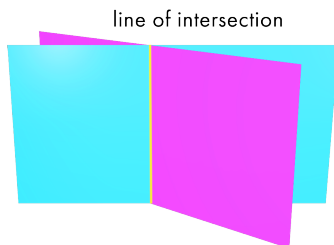
$$\begin{aligned}x_1 + x_2 + x_3 &= 4 \\2x_1 + 5x_2 - x_3 &= 11 \\-x_1 &+ 2x_3 = -3.\end{aligned}$$

To solve this system means to find all sets of numbers for (x_1, x_2, x_3) that satisfy all 3 equations simultaneously. In other words, find a point (x_1, x_2, x_3) in \mathbb{R}^3 that makes all equations true simultaneously. In other words, find a vector (x_1, x_2, x_3) that satisfies this matrix equation.

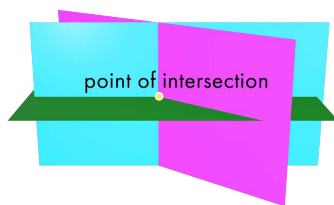
For example, $x_1 = 3$, $x_2 = 1$, and $x_3 = 0$ will work. Are there any other solutions?

Before answering this question algebraically, let's see how geometric insights can offer some perspective. The set of all solutions that satisfies each single equation is a plane. So, the set of all solutions that satisfies all equations simultaneously must be the intersection of all the planes represented by each equation.

What can the intersection of 3 planes look like? The intersection of 2 planes is usually a line.



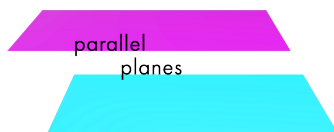
And if you intersect that with another plane, you will usually get a point. So, you expect to get a single solution vector to this set of equations.



However, there are degenerate cases. If you have all 3 planes identical, then the intersection will be that entire plane; in other words, you get a plane of solutions—infininitely many! Every point in that plane will satisfy all 3 equations.

If 2 of the planes from these 3 equations are identical, the third plane may cut this plane in a line, so you get a whole line of solutions.

But if 2 of the planes are parallel and not identical, there will not be a point that is simultaneously on both planes, so even when you intersect these planes with another plane, the set of solutions to all 3 equations will be empty.



So, without doing any algebra, you already see that a system of solutions will be either empty, a single solution, or infinitely many solutions. In fact, by reasoning about hyperplanes in a similar fashion, you will find that this is a general feature of any system of linear equations: It will either have one solution, no solution, or infinitely many solutions. And you can determine this from thinking about the geometry!

How are solutions actually found algebraically?

When you solve a system of equations like the one in this example by hand, it is tempting to just make some arbitrary choices for substitutions until you get an answer.

But how can you solve a system of equations like this one in a systematic way? The problem with doing this on a case-by-case basis is that you get no insight into what is going on. In addition, you want something that will work for a system of 3 equations or 300 equations—something you could program a computer to do.

Gaussian Elimination

One method of solving systems of linear equations is called Gaussian elimination. The idea is simple: Convert the system of equations into an **equivalent** system of equations, meaning a new system of equations that has exactly the same solutions as the original.

Although Gaussian elimination is named after the 19th-century mathematician Carl Friedrich Gauss, it dates back to the Chinese in the 3rd century BC and probably has been rediscovered many times over the centuries.

Recall a few basic facts about equations:

- ◆ You can add one equation to another equation to get a third true statement. So, if $A = B$ and $C = D$, then $A + C = B + D$.
- ◆ If you do the same things to both sides of the equation, the statement is still true. So, if $A = B$, then $AC = BC$, no matter what C is.

If you take one equation and multiply it by a nonzero constant, that will not change the set of solutions to the system.

$$\begin{array}{rcl} x_1 + x_2 + x_3 & = & 4 \\ 2x_1 + 5x_2 - x_3 & = & 11 \\ -x_1 & + & 2x_3 = -3. \end{array}$$

So, if you changed the first equation to $2x_1 + 2x_2 + 2x_3 = 8$, the set of solutions to all 3 equations would stay the same. Any solution that worked for the original system will still work for the new system; any solution that works for the new system will still work for the original.

However, if you multiplied one equation by zero, then that equation becomes $0 = 0$, which is true, but now you've lost a constraint that the variables must satisfy, and you've possibly enlarged the set of solutions. So, multiplying by zero is prohibited if you want to be sure your solution set is unchanged.

Another operation you're allowed to do is to add a multiple of one equation to another. The multiple could be positive or negative. That won't change the set of solutions, either.

So, for example, if you take the second equation and add -2 times the first equation, that will have the effect of canceling the x_1 term, and it produces a second term of $5x_2 - 2x_2$, which is $3x_2$.

Similarly, you get a third term of $-x_3 - 2x_3 = -3x_3$, and for the constant on the right side, you get $11 - 2 \times 4 = 3$.

So, the second equation becomes $3x_2 - 3x_3 = 3$.

This system has the same set of solutions as the original. The way to see that is anything that satisfied the original system now still satisfies this one. But you know you didn't enlarge the set of solutions because this step is reversible: If you want to get back the original system, all you need to do is add twice the first equation to the second (because you subtracted twice the first equation previously). So, anything that satisfies the new system still satisfies the old.

Notice how unwieldy it is to keep rewriting all these variables. You can use shorthand to represent the original linear system of equations as an **augmented matrix**.

The augmented part is the rightmost column, which is separated by a vertical line to help you remember that these numbers come from the constants in the linear equation.

$$\begin{array}{rrcr} x_1 & + & x_2 & + & x_3 & = & 4 \\ 2x_1 & + & 5x_2 & - & x_3 & = & 11 \\ -x_1 & & & + & 2x_3 & = & -3 \end{array}$$

$$\begin{array}{ccccccc} \downarrow & \downarrow & \downarrow & \downarrow & & & \\ \left[\begin{array}{ccc|c} 1 & 1 & 1 & 4 \\ 2 & 5 & -1 & 11 \\ -1 & 0 & 2 & -3 \end{array} \right] \\ \text{augmented matrix} \end{array}$$

Remember, this augmented matrix is just shorthand for a system of linear equations; it just shows you the coefficients rather than having you write out all the variables! If an equation doesn't have a variable, you can think of the variable as having the coefficient 0. For example, the third equation had no x_2 , so you have a 0 in the third row in the x_2 position.

Performing operations that don't change the solutions can be represented by row operations on the augmented matrix. For example, if you want to subtract 2 times the first row from the second row, you can represent it this way:

$$R_2 - 2R_1 \left[\begin{array}{ccc|c} 1 & 1 & 1 & 4 \\ 0 & 3 & -3 & 3 \\ -1 & 0 & 2 & -3 \end{array} \right].$$

The notation $R_2 - 2R_1$ just tells you what you did.

You're going to keep doing things that won't change the set of solutions. For example, you could next try to make the -1 in the bottom left corner into a zero. This would correspond to eliminating the x_1 term in that equation. You could do that by adding the first row to the third row. So, $-1 + 1 = 0$, $0 + 1 = 1$, $2 + 1 = 3$, and $-3 + 4 = 1$. You'll get this:

$$R_3 + R_1 \left[\begin{array}{ccc|c} 1 & 1 & 1 & 4 \\ 0 & 3 & -3 & 3 \\ 0 & 1 & 3 & 1 \end{array} \right].$$

You could just keep going, playing this game, always getting an equivalent system with the same solutions, but where are you headed?

Where you're trying to go is to an augmented matrix like the following one, which has just a single 1 in each row and each column.

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & a \\ 0 & 1 & 0 & b \\ 0 & 0 & 1 & c \end{array} \right] \text{ means } \begin{cases} x_1 = a \\ x_2 = b \\ x_3 = c \end{cases}$$

The corresponding system of equations has a very obvious solution: If the right side is a bunch of constants a , b , c , then this augmented matrix would be saying $x_1 = a$, $x_2 = b$, and $x_3 = c$. In other words, it would be telling you the solution to the system of equations. You can recognize this pattern because you will see all 0s on the left side, except for a diagonal of 1s.

You've seen this pattern before—it's called the identity matrix! And the diagonal is called the main diagonal of that square.

You may not be able to get to this pattern, but you can try.

So, in your original augmented matrix, your strategy is to try to make zeros in all these places, and you'll do it in the order shown at right.

$$\left[\begin{array}{ccc|c} 1 & \textcircled{1} & \textcircled{1} & 4 \\ \textcircled{2} & 5 & \textcircled{-1} & 11 \\ \textcircled{-1} & \textcircled{0} & 2 & -3 \end{array} \right]$$

1st 2nd 3rd

You try to zero out the entries below the main diagonal first, working column by column from left to right. Then, you work above the main diagonal to turn all those entries into zeros, and generally it will be easier at that point to work from right to left.

In this example, you zero out the first column except for the 1 at the top.

In the second column, you want to have a 1 in the main diagonal, and **you can do another thing that won't change the set of solutions: swap the rows.**

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 4 \\ 0 & 3 & -3 & 3 \\ 0 & 1 & 3 & 1 \end{array} \right]$$

Why is that allowed? It won't change the set of solutions, because it is just expressing the same equations but in a different order. And the order of the rows has nothing to do with the order of the variables corresponding to the columns; swapping rows does not change anything about which variables are related to the columns.

Swapping rows is just a move you want to use to help you get 1s on the main diagonal. Swapping rows 2 and 3 gives you the matrix shown at right.

$$\text{swap} \left[\begin{array}{ccc|c} 1 & 1 & 1 & 4 \\ 0 & 1 & 3 & 1 \\ 0 & 3 & -3 & 3 \end{array} \right]$$

Now you can get a zero on the bottom of the second column by subtracting 3 times the second row. Notice that this doesn't mess up the work you did to zero out the first column, because both row 2 and row 3 had zeros there!

So, you get the matrix shown here.

$$R_3 - 3R_2 \left[\begin{array}{ccc|c} 1 & 1 & 1 & 4 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & -12 & 0 \end{array} \right]$$

Next, you can multiply the third row by $-1/12$ to get the matrix shown at right.

$$-\frac{1}{12}R_3 \left[\begin{array}{ccc|c} 1 & 1 & 1 & 4 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 1 & 0 \end{array} \right]$$

Because you have 001 in the last row, it will be easy to zero out the entries in the third column by subtracting multiples of the last row. Doing this won't affect the other columns, because 001 has zeros in those columns.

$$R_2 - 3R_3 \left[\begin{array}{ccc|c} 1 & 1 & 1 & 4 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array} \right]$$

$$R_1 - R_3 \left[\begin{array}{ccc|c} 1 & 1 & 0 & 4 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array} \right]$$

Finally, you do a last step, as shown here.

$$R_1 - R_2 \left[\begin{array}{ccc|c} 1 & 0 & 0 & 3 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array} \right]$$

This shows you that the solution to this system, as well as to the original system of equations, is $x_1 = 3$, $x_2 = 1$, $x_3 = 0$.

So, $(3, 1, 0)$ is a solution, and, in fact, you showed that it is the only solution to the original system of equations. By plugging these into the original system (below), you can check that it works.

$$\begin{aligned} x_1 + x_2 + x_3 &= 4 \\ 2x_1 + 5x_2 - x_3 &= 11 \\ -x_1 + 2x_3 &= -3. \end{aligned}$$

You used 3 operations on the rows of the augmented matrix:

- ◆ swapping rows;
- ◆ multiplying rows by a nonzero constant; and
- ◆ adding a multiple of one row (j) to another row (i), replacing row (i).

These operations are called **elementary row operations**, and they do not change the solutions to a system of linear equations. The process of getting there is called row reduction or Gaussian elimination.

Getting Infinitely Many or No Solutions

What if you can't for some reason get the method to produce the identity matrix on the left side of the augmented matrix? Under what conditions will the method fail?

You've already seen that a system always has either no solution, one solution, or infinitely many solutions. Producing an identity matrix on the left side is equivalent to finding a unique solution to the problem. There must be other situations that will lead you to find no solution or infinitely many solutions.

Here's one bad thing that could happen when you try to do row reduction. You might end up with a row that looks something like this:

$$[0 \ 0 \ 0 \mid 8].$$

If you get a row like this, it stands for the equation

$$0x_1 + 0x_2 + 0x_3 = 8,$$

which is a nonsensical equation because zero can't equal 8.

So, this will be a system of equations that has no solution.

If you get a row of all zeros, including the augmented part, like this:

$$[0 \ 0 \ 0 \mid 0],$$

that just means you had an equation that was redundant. It was already a linear combination of the other equations, and row reduction eliminated that equation.

So, something like this

$$\left[\begin{array}{ccc|c} 1 & 0 & 3 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array} \right]$$

would represent the following system of equations:

$$x_1 + 3x_3 = 1$$

$$x_2 + x_3 = 2.$$

You can choose x_3 to be anything you want. Once you do that, x_1 and x_2 are determined. You then say you can choose x_3 “freely,” or you say x_3 is a “free variable.” If you have a free variable, then there will be infinitely many solutions!

READINGS

Chartier, *When Life Is Linear*, chap. 7.

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, sections 1.1 and 1.2.

Poole, *Linear Algebra*, sections 2.1 and 2.4.

QUIZ FOR LECTURES 1–6

- 1 What are the 4 themes that you'll encounter repeatedly throughout this course? [LECTURE 1]
- 2 In the 2-dimensional plane, is reflection about the x -axis a linear transformation? You'll want to check that it satisfies the 2 linearity properties:
 - ♦ If you scale a vector and then reflect it, do you get the same thing as if you reflect the vector and then scale it?
 - ♦ If you add 2 vectors and then reflect them, is that the same as reflecting the 2 vectors and then adding them? [LECTURE 1]
- 3 On graph paper:
 - a draw the set of all scalar multiples of the vector $\mathbf{u} = (1, -1)$ as points in \mathbb{R}^2 and indicate points that are integer multiples of the vector \mathbf{u} .
 - b draw the set of all scalar multiples of the vector $\mathbf{v} = (1, 2)$ and indicate points that are integer multiples of the vector \mathbf{v} . [LECTURE 2]
- 4 Consider the graph you drew in the previous problem. Now mark points on your graph that are linear combinations $(a\mathbf{u} + b\mathbf{v})$, where a and b are integers. From your picture, can you estimate what linear combination of \mathbf{u} and \mathbf{v} would produce the point $(4, 0)$? Does the set of all linear combinations of the vectors \mathbf{u} and \mathbf{v} cover the entire plane? [LECTURE 2]
- 5 Let $\mathbf{u} = (1, 0, 2)$ and $\mathbf{v} = (-2, 3, 1)$. What is the dot product $\mathbf{u} \cdot \mathbf{v}$ and cross product $\mathbf{u} \times \mathbf{v}$ of these 2 vectors? Based on your answers, what can you say about the angles between \mathbf{u} , \mathbf{v} , and $\mathbf{u} \times \mathbf{v}$? [LECTURE 3]
- 6 Given 3 points in \mathbb{R}^3 , describe a procedure you would use to derive an equation for the plane that passes through those 3 points. Also describe any instances where the procedure might fail, and explain why. [LECTURE 3]

- 7 Find the matrix product: [LECTURE 4]

$$\begin{bmatrix} 1 & 6 \\ 2 & 5 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

- 8 Find the matrix product of the transposes of the above matrices in the reverse order: [LECTURE 4]

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \end{bmatrix}.$$

- 9 Let $T(\mathbf{x})$ be the function that takes a point \mathbf{x} in \mathbb{R}^2 and translates it 1 unit to the right. Why is T not a linear transformation? [LECTURE 5]
- 10 Find a matrix representing R , the linear transformation of the plane that reflects the plane about the diagonal line $y = x$. [LECTURE 5]
- 11 Consider this system of equations:

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

First think about why the solution set of each equation alone is a plane in \mathbb{R}^3 . Then use row operations to show that this system has the same solution set as the following system:

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

Use this system to find a solution (x, y, z) and verify that it is a solution of the original system. [LECTURE 6]

- 12 Can you have 3 linear equations in x, y, z such that any pair of them have solutions but the 3 of them cannot be simultaneously satisfied? Think geometrically to answer this question. [LECTURE 6]

Solutions can
be found on
page 296.

REDUCED ROW ECHELON FORM

This lecture focuses on how to systematically solve any system of linear equations by using row operations to put the matrix in a special form called reduced row echelon form.

Reduced Row Echelon Form

Recall from the previous lecture that given an augmented matrix representing a system of equations, there are 3 **elementary row operations** that can be performed on it that do not change the solution set of that system of equations. You can

- ♦ swap 2 rows;
- ♦ multiply a row by a nonzero constant; and
- ♦ add a multiple of one row to another row, replacing that row.

The idea of Gaussian elimination is to use these operations to simplify the system. This could mean eliminating as many variables from the equations as possible. This corresponds to increasing the number of zeros on the left side of the augmented matrix, because they represent coefficients of variables.

Notice that the third row operation is the one that will enable you to take a particular nonzero entry of an augmented matrix and make it zero; you'll do this by adding a multiple of another row to it. The important point here is to do this in such a way that you don't unintentionally make some entries nonzero that you already zeroed earlier.

Let's try to be systematic about it. The first thing you might try is to mimic the example from the previous lecture, in which you used row operations to convert the left side of the augmented matrix to an identity matrix. This won't always be possible for many reasons; here is one example, where the left side of an augmented matrix is not square.

$$\left[\begin{array}{ccc|c} 1 & 0 & -1 & 1 \\ 0 & 1 & 2 & 2 \end{array} \right]$$

$$\left[\begin{array}{ccc|c} 1 & 0 & -1 & 1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 \end{array} \right]$$

And here is another augmented matrix, where the left side is square but no row operations will produce a 1 in the second column because all entries in the second column are 0.

So, instead of an identity matrix, you'll aim for something more general called **row echelon form** (REF).

For example, suppose you have a system of equations $A\mathbf{x} = \mathbf{b}$, where A is a 3×5 matrix. Your goal is to use elementary row operations (EROs) to reduce it to a form that has a steplike appearance.

The word *echelon* means *steplike*.

$$\left[\begin{array}{c|c} A_{3 \times 5} & \begin{bmatrix} b_1 \\ \vdots \\ b_2 \end{bmatrix} \end{array} \right] \xrightarrow{\text{EROs}} \left[\begin{array}{ccccc|c} \blacksquare & \star & \star & \star & \star & \star \\ 0 & 0 & \blacksquare & \star & \star & \star \\ 0 & 0 & 0 & \blacksquare & \star & \star \end{array} \right] \quad \begin{array}{l} \blacksquare \text{ non-0} \\ \star \text{ any} \end{array}$$

REF

Here's what makes a form an REF. On the left side of the augmented matrix, the first zero entry in any row is called a **leading entry**. To be in REF, the augmented matrix must satisfy 2 properties:

- ◆ The leading entry of any row must always be to the right of a leading entry of a previous row.
- ◆ Rows consisting only of zeros are at the bottom.

Matrices satisfying these properties will have a steplike appearance such that leading entries sit on the steps and all entries below the steps are zero. (Above the steps, apart from the leading entries, the other entries may or may not be zero.)

Look at the first column. If there is any nonzero entry in that column, use a swap (if needed) to make the first entry of the column nonzero. That is the leading entry of the first row. Then, add appropriate multiples of this row to other rows to zero out the rest of the entries under this leading entry. Now imagine you freeze the first row so that you no longer modify it. Then, consider the second row the "current" row.

Now look at the second column. If all unfrozen entries are zero, move to the next column and repeat the process. If some unfrozen entries are nonzero, swap rows (if needed) to move a nonzero entry to the current row. This becomes a leading entry of the current row, and the current row can now be used on the unfrozen rows to zero out all the unfrozen entries of the second column. These operations will not disturb the zeros that already exist in the first column, because the second row had a zero in the first column. Now freeze the current row and move on to the next row as the current row.

If you keep doing this in this fashion, you will produce a matrix in REF.

You can glean some interesting things from REF. It reveals some hidden structure of the original matrix A . For example, if the REF has a row of zeros, then one of the original rows of the matrix A must be a linear combination of the other rows. Also, the number of leading entries says something about the number of rows of A that is sufficient to span any linear combinations of the rows of A . The number of leading entries is called the **rank** of a matrix.

A matrix in REF is easy to solve by back substitution, starting with the last nonzero row. That is one reason why people often stop Gaussian elimination at the REF, because you can often figure out what you need from there.

Unfortunately, though, REF is not unique. If you tell 2 people to take a matrix and row-reduce it to put it in REF, they may come up with different forms. The set of solutions will be the same, because that's what row operations preserve, but the description of the set of solutions expressed by the REF may be very different.

This is where **reduced row echelon form** (RREF) is helpful, because it is unique for any given augmented matrix. If you give an augmented matrix to 2 different people and tell them to put it in RREF, they will give you back the same answers.

RREF takes REF a few steps further. Namely, it has 2 additional properties:

- ◆ Every leading entry is a 1.
- ◆ A column with a leading entry has zeros in all other entries of that column.

$$\left[\begin{array}{c|c} A_{3 \times 5} & \begin{bmatrix} b_1 \\ \vdots \\ b_2 \end{bmatrix} \end{array} \right] \xrightarrow{\text{EROs}} \left[\begin{array}{cccc|c|c} 1 & \star & 0 & 0 & \star & \star \\ 0 & 0 & 1 & 0 & \star & \star \\ 0 & 0 & 0 & 1 & \star & \star \end{array} \right] \quad \begin{array}{l} \blacksquare \text{ non-0} \\ \star \text{ any} \end{array}$$

RREF

This is possible because you can take REF and multiply rows by nonzero constants to turn the leading entry into a 1. Then, you can use a row with a leading 1 to zero out all the other entries in the column with that 1.

For example, consider the following system of equations.

$$\begin{aligned}x_1 + 2x_2 + 3x_3 &= 4 \\ 5x_1 + 6x_2 + 7x_3 &= 8.\end{aligned}$$

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{array} \right]$$

This has the augmented matrix shown at right.

In the first column, there is already a 1 in the first row. So, focus on changing the 5 in the bottom left corner to a 0.

From row 2, you can subtract 5 times row 1. So,

$$\begin{aligned}5 - 5 \times 1 &= 0, \\ 6 - 5 \times 2 &= -4, \\ 7 - 5 \times 3 &= -8, \\ \text{and } 8 - 5 \times 4 &= -12.\end{aligned}$$

So, you have the matrix shown here. You should recognize it as REF because the leading entry in the second row (-4) is to the right of the leading entry in the first row.

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & -4 & -8 & -12 \end{array} \right]$$

Next, you try to turn this into RREF.

You can change the -4 entry into a 1 by multiplying row 2 by the constant $(-1/4)$. If you do this, you obtain the matrix shown at right.

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & 1 & 2 & 3 \end{array} \right]$$

Finally, you need to zero out the entry in row 1, column 2, because it is in the same column as row 2's leading entry. You can do this by subtracting twice the second row from the first row.

This is now in RREF, because all leading entries are 1 and all other entries in columns with leading entries are 0.

$$\left[\begin{array}{cc|cc} 1 & 0 & -1 & -2 \\ 0 & 1 & 2 & 3 \end{array} \right]$$

Note that you had many choices you could have made to row-reduce this augmented matrix, but there is an important theorem that says no matter what choices of row operations you made, you must always get the same result for the RREF.

THEOREM

For a given matrix, the reduced row echelon form is unique.

Note that this theorem holds whether or not the matrix is augmented.

Using the RREF to Find the Set of Solutions

If you have an augmented matrix in RREF, there's an easy way to figure out the set of solutions. First of all, **check if there are any inconsistencies**—remember you can see if there are any by looking for a row of zeros on the left side of the augmented matrix equaling something nonzero on the right.

If there are none, then we identify what are called free variables. Remember that every column on the left side of an augmented matrix corresponds to a variable. Look at all the columns with leading 1s in them. The associated variables are called **leading variables**; all the other variables are **free variables**.

These variables are called free because they can be chosen freely—to be any real number. If there are no free variables, then the leading variables will be completely determined, so the solution is unique. Otherwise, the basic idea is to **use the RREF to express leading variables in terms of free variables**. Since free variables are easy to express in terms of themselves, this will give us a way to express ALL variables in terms of the free variables, which will lead to a **parametric description of the solutions using free variables as parameters**, as long as solutions exist.

To understand why, let's return to the earlier example in which the RREF for the following system was analyzed.

$$\begin{aligned}x_1 + 2x_2 + 3x_3 &= 4 \\ 5x_1 + 6x_2 + 7x_3 &= 8.\end{aligned}$$

The RREF was

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

The leading 1s in the first and second column are coefficients of x_1 and x_2 , so x_1 and x_2 are the leading variables. The remaining variable, x_3 , is thus the free variable.

You can see the importance of the free variable if you examine the system of equations that comes from the augmented matrix, which must have the same solutions as the original system.

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

The RREF guarantees that leading variables appear in exactly one equation, but the free variable x_3 can appear in many equations.

You can choose x_3 to be any number you want, and once you do, x_1 and x_2 are now determined.

So, to find the set of solutions, your goal is to express all variables in terms of the free variables, which in this case is just x_3 , and use that to determine a vector equation.

The first equation, $x_1 = -2 + x_3$, allows you to solve for x_1 in terms of x_3 .

The second equation shows you $x_2 = 3 - 2x_3$.

What about x_3 ? How do you express x_3 in terms of x_3 ?

Although there is no given equation that expresses x_3 in terms of x_3 alone, that's OK, because you don't need one. It is obvious that $x_3 = x_3$.

So, then you get a set of 3 equations expressing the set of solutions in terms of x_3 , the one free variable.

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

This can be rewritten as a vector equation if you factor out the free variable.

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -2 \\ 3 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}.$$

Because x_3 can be chosen to be any real number, it can be thought of as a parameter. So, you might recognize this vector expression as a parametric description of a line, because it is a point $(-2, 3, 0)$ plus some parameter times a vector direction $(1, -2, 1)$. So, the set of solutions to the original system is a line!

If $x_3 = 0$, you see that a solution to the original system is $(-2, 3, 0)$, and if $x_3 = 1$, you get the solution $(-1, 1, 1)$. By plugging these into the original system (below), you can check that they are both solutions.

$$\begin{aligned} x_1 + 2x_2 + 3x_3 &= 4 \\ 5x_1 + 6x_2 + 7x_3 &= 8. \end{aligned}$$

Row-Equivalent Matrices

In addition to being useful for expressing the set of solutions to a linear system, RREF can also be used to tell which matrices are related by row operations.

Let's call 2 matrices **row-equivalent** if there is a series of elementary row operations that take you from one matrix to the other. Then, the following theorem is true.

THEOREM

Matrices A and B are row-equivalent if and only if A and B have the same RREF.

Let's prove this fact. First, if A and B are row-equivalent, then there is a sequence of row operations taking A to B . But there is a series of row operations that takes you from B to B 's RREF. Putting these sequences together gives you a sequence from A to B 's RREF. But this must then be A 's RREF, because RREFs are unique. So, A and B have the same RREF.

$$A \xrightarrow{\text{EROs}} B \rightarrow \text{RREF}.$$

Next, let's suppose that A and B have the same RREF. Then, there is a sequence of row operations taking you from A to the RREF. If you could find a sequence taking the RREF to B , you would be done, because there would be a sequence of row operations going from A to B . But you know there are row operations taking B to the RREF. Because every row operation is reversible, you can reverse this sequence to obtain a sequence from the RREF to B , as desired.

This is a helpful criterion to tell when 2 matrices are row-equivalent.



READINGS

Chartier, *When Life Is Linear*, chap. 6.

Kalman, "An Underdetermined Linear System for GPS," https://www.maa.org/sites/default/files/pdf/upload_library/22/Polya/Kalman.pdf.

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, sections 1.2, 1.5, and 1.6.

Poole, *Linear Algebra*, section 2.2 and the Vignette that follows section 2.4.

Yuster, "The Reduced Row Echelon Form Is Unique," <https://www.maa.org/sites/default/files/Yuster19807.pdf>.

SPAN AND LINEAR DEPENDENCE

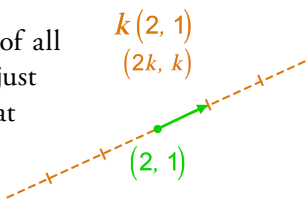
In some instances, the solution set of a system of linear equations is naturally a geometric object, such as a line or a plane, because it is a linear combination of various vectors. This lecture focuses on the geometry of linear combinations.

The Span of a Set of Vectors

Recall from lecture 2 that the **span** of a set of vectors is the set of all linear combinations of those vectors.

If you have just one vector in \mathbb{R}^n , the set of all linear combinations of one vector in \mathbb{R}^n is just multiples of that vector. This forms a set that is just a line in \mathbb{R}^n . So, for example, the vector $(2, 1)$ lives in the plane. Then, the span of the vector $(2, 1)$ is anything of

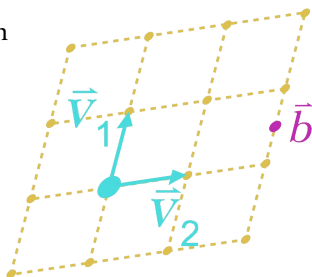
the form given by some constant k times $(2, 1)$. So, anything of the form $(2k, k)$ is in the span of $(2, 1)$. Note that k could be negative. So, this forms a line in the plane, because you can go both forward and backward in the direction $(2, 1)$.



The span of 2 vectors—for example, \mathbf{v}_1 and \mathbf{v}_2 —is the set of all linear combinations of \mathbf{v}_1 and \mathbf{v}_2 . This means it could be any multiple of \mathbf{v}_1 plus any multiple of \mathbf{v}_2 . In other words, the span of a set of vectors is the set of all points you could reach from the origin if you were limited to adding multiples of those vectors.

Consider 2 vectors \mathbf{v}_1 and \mathbf{v}_2 sitting in 3-dimensional space, as in the picture shown here, where \mathbf{v}_1 and \mathbf{v}_2 are based at a blue dot that represents the origin.

If you just draw integer multiples of \mathbf{v}_1 plus integer multiples of \mathbf{v}_2 , you will get the set of points of a grid, as in the picture. To get to the point representing $\mathbf{v}_1 + \mathbf{v}_2$, you march first in the direction \mathbf{v}_1 and then in the direction \mathbf{v}_2 . And to get to the point $\mathbf{v}_1 - \mathbf{v}_2$, you first move in direction \mathbf{v}_1 and then in the direction of $-\mathbf{v}_2$, which is the opposite of the direction \mathbf{v}_2 .



How do you get from the blue point to the point \mathbf{b} using \mathbf{v}_1 and \mathbf{v}_2 ?

You can use 2 multiples of \mathbf{v}_2 plus $\frac{1}{2}$ of \mathbf{v}_1 .

If you have 2 vectors \mathbf{v}_1 and \mathbf{v}_2 , any linear combination of \mathbf{v}_1 and \mathbf{v}_2 will produce a point on a plane containing \mathbf{v}_1 and \mathbf{v}_2 . And those are, intuitively, the only points you could reach. So, you can see that with just 2 vectors, you won't be able to reach every point in \mathbb{R}^3 . So, the span of 2 vectors is a plane.

The span of any 2 vectors is generally a plane, but if \mathbf{v}_1 and \mathbf{v}_2 are parallel (pointing in the same direction or opposite directions), their span would be a line.

And if \mathbf{v}_1 and \mathbf{v}_2 are the zero vector, then the span of \mathbf{v}_1 and \mathbf{v}_2 will be a single point—the zero vector—because that's all you can get when taking linear combinations of the zero vector!

Similarly, the span of 3 vectors can be something that is at most 3-dimensional, and the span of k vectors can be at most k -dimensional in some high-dimensional space, but it can be smaller.

Suppose you are asked to find the span of the vectors $(1, 0)$ and $(0, 1)$. The first thing you might do is think about this intuitively:

$(1, 0)$ and $(0, 1)$ are 2-dimensional vectors, so they live in \mathbb{R}^2 . And there are 2 of them, which means they span at most a plane, but maybe something smaller. If they span a plane, there is only one plane they could span—namely, all of \mathbb{R}^2 .

Can you get any vector (A, B) as a linear combination of $(1, 0)$ and $(0, 1)$?

Yes. This is just what you do with the usual coordinate system. You march over A units in the x direction and B in the y direction. This is just doing the combination $A(1, 0) + B(0, 1)$. So, the span of these 2 vectors is all \mathbb{R}^2 .

$$\text{Span}\left\{\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right\} \text{ in } \mathbb{R}^2 \text{ is all } \mathbb{R}^2.$$

On the other hand, the span of the vectors $(1, 0)$ and $(2, 0)$ is a line consisting of all points of the form $(x, 0)$ for a real number x . In this case, taking linear combinations of $(1, 0)$ and $(2, 0)$ will only produce vectors whose second coordinate is 0.

What about the span of the vectors $(1, 3)$ and $(3, 1)$?

Again, the vectors are in the plane \mathbb{R}^2 , and they should span at most a plane. You expect that as long as they don't point in the same direction, they should span the whole plane. It may not be as obvious how to construct an arbitrary vector (A, B) as a combination of $(1, 3)$ and $(3, 1)$, so let's figure out how to do that.

When Is a Vector in the Span of a Set of Vectors?

In general, when is a vector \mathbf{b} in the span of a set of vectors \mathbf{v}_1 through \mathbf{v}_k ? In other words, when is \mathbf{b} a linear combination of given vectors, and if it is, what linear combination is it?

$$\bar{\mathbf{b}} \in \text{span} \{ \bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_k \}?$$

For example, is $(1, 2, 3)$ in the span of the 2 vectors $(4, 5, 6)$ and $(7, 8, 9)$?

$$\text{Is } \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \in \text{span} \left\{ \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}, \begin{bmatrix} 7 \\ 8 \\ 9 \end{bmatrix} \right\}?$$

This is the same as finding coefficients x_1 and x_2 to make $x_1(4, 5, 6) + x_2(7, 8, 9)$ equal to $(1, 2, 3)$.

$$x_1 \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} + x_2 \begin{bmatrix} 7 \\ 8 \\ 9 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

But notice that this is just a linear system of equations! For example, the first row is $4x_1 + 7x_2 = 1$, and the other rows are similarly linear equations. So, finding coefficients x_1 and x_2 is the same as taking the augmented matrix and reducing it by row operations.

$$\left[\begin{array}{cc|c} 4 & 7 & 1 \\ 5 & 8 & 2 \\ 6 & 9 & 3 \end{array} \right] \xrightarrow{\text{EROs}} \left[\begin{array}{cc|c} 1 & 0 & 2 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{array} \right].$$

When you put it in RREF, you see that there is no inconsistency, because there is no row that is all zeros on the left while being nonzero on the augmented side. This RREF shows that a solution exists—namely, $x_1 = 2$ and $x_2 = -1$.

So, the answer to the question posed previously is yes: $(1, 2, 3)$ is in the span of the 2 vectors. And you just solved for the coefficients! So, you know that

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = 2 \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} + (-1) \begin{bmatrix} 7 \\ 8 \\ 9 \end{bmatrix}.$$

In fact, you know this is the only combination that would produce $(1, 2, 3)$ because it was the only solution to the system of linear equations you had.

Now look at the RREF of the augmented matrix.

$$\left[\begin{array}{cc|c} 1 & 0 & 2 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{array} \right]$$

The last row is a row of zeros. You could have foreseen that there would be a row of zeros on the left side because there are more rows than columns on the left side (which comes from the fact that the dimension of the vectors, 3, is greater than the number of vectors you are taking the span of, 2 in this case).

However, on the right, the last row may not have had a zero if you had row-reduced a different vector besides $(1, 2, 3)$. So, you could have had an inconsistency, which suggests that not all vectors in \mathbb{R}^3 are in the span of $(4, 5, 6)$ and $(7, 8, 9)$. This makes sense, because the span of 2 vectors is at most a plane in \mathbb{R}^3 .

THEOREM

A system of equations $A\mathbf{x} = \mathbf{b}$ has a solution if and only if \mathbf{b} is a linear combination of the columns of A . In other words, $A\mathbf{x} = \mathbf{b}$ has a solution if and only if \mathbf{b} is in the span of the columns of A .

This is an answer to the question of how to tell when a vector is in the span of a set of vectors: Just solve a certain matrix equation.

Linear Dependence of a Set of Vectors

If you have fewer vectors than dimensions, you might not span the entire space. But what if you have more vectors than dimensions?

Suppose you have 3 vectors in \mathbb{R}^2 : $\mathbf{u} = (1, 0)$, $\mathbf{v} = (0, 1)$, and $\mathbf{w} = (2, 3)$. What do they span?

$$\text{What's span} \left\{ \begin{bmatrix} \vec{u} \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} \vec{v} \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} \vec{w} \\ 2 \\ 3 \end{bmatrix} \right\} \text{ in } \mathbb{R}^2?$$

You have already seen that $(1, 0)$ and $(0, 1)$ span all \mathbb{R}^2 , so you expect that if you throw in another vector, the span should only get bigger. But it can't get bigger because you've already spanned the entire space of \mathbb{R}^2 .

The theorem suggests that finding the span of these 3 vectors is the same as seeing if a certain system of equations—the one represented by the augmented matrix at right—has a solution.

$$\left[\begin{array}{ccc|c} 1 & 0 & 2 & b_1 \\ 0 & 1 & 3 & b_2 \end{array} \right]$$

Notice how the columns on the left are vectors \mathbf{u} , \mathbf{v} , and \mathbf{w} that you were given and the right side is a generic point \mathbf{b} with coordinates (b_1, b_2) . By using generic letters for b_1 and b_2 , you are asking this question: For which b_1 and b_2 is a solution possible? In other words, which vectors \mathbf{b} are in the span of these 3 vectors?

Solving this particular augmented matrix is easy because it is already in RREF. So, it represents the system $x_1 + 2x_3 = b_1$ and $x_2 + 3x_3 = b_2$.

Recall that the variables x_1 , x_2 , and x_3 will tell you the linear combination of the columns that produce (b_1, b_2) . Also, you see from the RREF that x_1 and x_2 are leading variables (associated with the leading 1s), and x_3 is a free variable because the column for x_3 has no leading 1 in it.

So, to solve this system, you should try to express all variables in terms of free variables. When you do that, you get

$$\begin{aligned}x_1 &= b_1 - 2x_3 \\x_2 &= b_2 - 3x_3.\end{aligned}$$

And x_3 is free, so it's easily expressed in terms of x_3 by saying $x_3 = x_3$.

$$\begin{aligned}x_1 &= b_1 - 2x_3 \\x_2 &= b_2 - 3x_3 \\x_3 &= x_3\end{aligned} \quad \text{so} \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} -2 \\ -3 \\ 1 \end{bmatrix}.$$

So, you get $(x_1, x_2, x_3) = (b_1, b_2, 0) + x_3(-2, -3, 1)$.

These are the coefficient combinations that will produce the vector (b_1, b_2) from the 3 vectors \mathbf{u} , \mathbf{v} , and \mathbf{w} .

This solution should not surprise you, because if x_3 is zero, you get the result you got before: that every vector (b_1, b_2) can be produced as a linear combination of $(1, 0)$ and $(0, 1)$, with coefficients that are just the coordinates b_1 and b_2 . So, if you want to get the point $(5, 7)$, you just use 5 times $(1, 0)$ and 7 times $(0, 1)$.

$$5 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 7 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + 0 \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 5 \\ 7 \end{bmatrix}.$$

But notice from this analysis that there are multiple solutions—in fact, a line of solutions—and that each solution is given by a different choice of free variable x_3 . So, if you want to get $(5, 7)$ using 1 as the coefficient for the vector \mathbf{w} , then you'll need $x_3 = 1$.

So, then,

$$(x_1, x_2, x_3) = (5, 7, 0) + 1(-2, -3, 1) = (3, 4, 1).$$

Going back to the vectors \mathbf{u} , \mathbf{v} , and \mathbf{w} , this means

$$3\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 4\begin{bmatrix} 0 \\ 1 \end{bmatrix} + 1\begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 5 \\ 7 \end{bmatrix}.$$

Thus, the linear combination you get to produce a particular vector in this case is not unique. This is because you had the presence of a free variable—in this case, x_3 . And you had that free variable because you had more vectors than dimensions.

It's possible in a case like this that the RREF could reveal an inconsistency; for example, a set of 3 vectors might still span a line in \mathbb{R}^2 , and not every point in \mathbb{R}^2 is a linear combination of 3 given vectors. But if there is no inconsistency and you have more vectors than dimensions, then linear combinations that produce a specific vector are not unique.

In other words, when you have more vectors than dimensions, you don't need all the vectors to produce the span of those vectors. Some of the vectors are redundant for that purpose. In this example, you didn't need the vector $(2, 3)$. You could have thrown it out and the other 2 vectors would still produce all \mathbb{R}^2 .

In fact, $\mathbf{w} = (2, 3)$ is a linear combination of the other 2 vectors. It is just $2\mathbf{u} + 3\mathbf{v}$. So, anything you could form using $(2, 3)$ could be formed using $(1, 0)$ and $(0, 1)$. That's one sense in which $(2, 3)$ is redundant.

This leads to an important concept that describes when a set of vectors is redundant in that sense. It's called linear dependence.

Linear Independence of a Set of Vectors

A set of vectors is **linearly dependent** if there are coefficients (called weights) not all zero, such that a linear combination of those vectors equals the zero vector. Otherwise, the set of vectors is **linearly independent**.

$$c_1 \vec{v}_1 + c_2 \vec{v}_2 + \dots + c_k \vec{v}_k = \vec{0}.$$

Notice that this definition does not require all coefficients to be nonzero, just some of them. This situation, when some of the coefficients are nonzero, is called a **nontrivial** linear combination. (If all were zero, it would be called a **trivial** linear combination.)

In the example with \mathbf{u} , \mathbf{v} , and \mathbf{w} , you've seen that \mathbf{w} is already a linear combination of \mathbf{u} and \mathbf{v} —namely, $\mathbf{w} = 2\mathbf{u} + 3\mathbf{v}$.

$$\begin{aligned}\vec{w} &= 2\vec{u} + 3\vec{v} \\ \text{or} \\ 2\vec{u} + 3\vec{v} - \vec{w} &= \vec{0}.\end{aligned}$$

↑ ↑ ↑
a nontrivial
combination

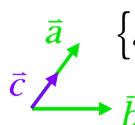
This is the same as saying that $2\mathbf{u} + 3\mathbf{v} - \mathbf{w}$ is equal to the zero vector. In other words, if one vector can be written in terms of the others, then there is a nontrivial linear combination that produces the zero vector. And if there is a nontrivial linear combination, by solving for one of the vectors in terms of the others, you'd see that one of the vectors is a linear combination of the others. You have just proved a theorem.

THEOREM

A set of vectors is linearly dependent if and only if at least one vector is a linear combination of the others.

So, the vector that is a linear combination of the others could be removed from the set of vectors and you would not change its span. You would have just removed a redundant vector.

Note that when you have a linearly dependent set, you are not saying that every vector is a linear combination of the others, just that one of the vectors is a linear combination of the others. For example, if you have vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} and vectors \mathbf{a} and \mathbf{c} are multiples of one another while \mathbf{b} points in a different direction, then you wouldn't be able to remove \mathbf{b} without changing the span, and \mathbf{b} is not a linear combination of vectors \mathbf{a} and \mathbf{c} .



$\{\vec{a}, \vec{b}, \vec{c}\}$ is linearly dependent because

$$1\vec{a} + 0\vec{b} - 2\vec{c} = \vec{0}.$$

nontrivial weights

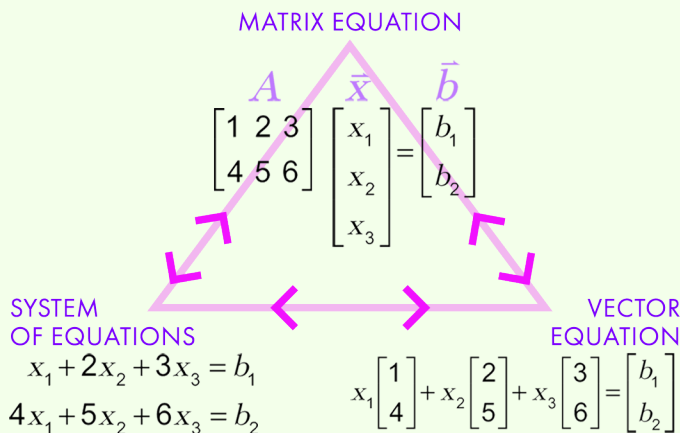
And that's because the nontrivial combination of \mathbf{a} , \mathbf{b} , and \mathbf{c} that produces the zero vector here is $(\mathbf{a} - 2\mathbf{c}) = \mathbf{0}$. The coefficient of \mathbf{b} in this nontrivial combination is zero, so this linear dependency relationship does not involve \mathbf{b} .

Notice the language that is used: Linear dependence (or independence) is a property of a set of vectors—not a property of a single vector.

A common mistake people make is to talk about a single vector as dependent or independent, but that has no meaning apart from a set you are speaking about.

THE FUNDAMENTAL CORRESPONDENCE

A matrix equation, a system of equations, and a vector equation may appear to be 3 different things, but they are really saying the same thing.



At one corner of this triangle is a **matrix equation** of the form $A\mathbf{x} = \mathbf{b}$. If you write out what it is saying, you will get the **system of linear equations** that's in the bottom left corner. But if you look at this system of equations and pull out the dependence on x_1 , x_2 , and x_3 as coefficients of vectors that are the columns of A , then you get the **vector equation** in the bottom right corner, which asks this question: What linear combination of these columns equals \mathbf{b} ?

Now let's connect the discussion of linear independence of a set of vectors to the uniqueness of a linear combination of those vectors.

If a set of vectors is linearly independent, it means that the only way to get the zero vector as a linear combination is if the coefficients are all zero. This is equivalent to saying that the zero vector is in the span of a set of vectors in exactly one way.

By the fundamental correspondence (see [PAGE 111](#)) between vector equations and linear equations, this is just a statement about a system of linear equations: If you put a set of vectors in the columns of a matrix A , then to say the columns of A are linearly independent means the only way to get the zero vector as a linear combination of the columns is if all coefficients are zero. In other words, $A\mathbf{x} = \mathbf{0}$ has only the trivial solution where all coefficients are zero.

And you can see from the RREF of A that you have linear independence of the columns if there are no free variables in the RREF.

Notice that a system of equations of the form $A\mathbf{x} = \mathbf{0}$ —called a **homogeneous system** of equations—can always be solved, because A times the zero vector gives the zero vector. So, the zero vector is always a solution, called the **trivial** solution. So, a homogeneous system of equations can never have an inconsistency.

The only question for a homogeneous system is whether the trivial solution is the unique solution. Otherwise, $A\mathbf{x} = \mathbf{0}$ has more than one solution, which happens when there are free variables in the reduced row echelon form.

READINGS

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, section 1.7.

Poole, *Linear Algebra*, section 2.3.

SUBSPACES: SPECIAL SUBSETS TO LOOK FOR

The row space, column space, and null-space are 3 special objects associated to an $m \times n$ matrix. They are subspaces of either \mathbb{R}^n or \mathbb{R}^m . These subspaces are often important objects precisely because they highlight the underlying geometry of the linear transformation associated with a matrix.

The Null-Space of a Matrix

Recall that a homogeneous system of linear equations is one of the form

$$A\mathbf{x} = \mathbf{0},$$

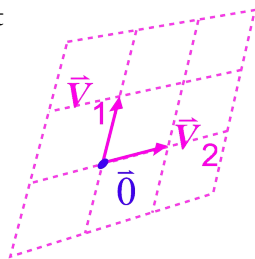
where A is a matrix, $\mathbf{0}$ is the zero vector, and \mathbf{x} is an unknown vector. Homogeneous systems always have at least one solution—namely, the trivial one, in which $\mathbf{x} = \mathbf{0}$. But there may be other solutions.

For example, consider the following simple linear equation:

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

It has $(-2, 1, 0)$ and $(3, 0, -1)$ as solutions. And if you add those 2 solutions together, you get $(1, 1, -1)$, which is another solution! This is not a coincidence.

The solution set to an equation of the form $Ax + By + Cz = D$ is a plane, and if $D = 0$, the solution set passes through the origin. So, if you add 2 vectors in this plane, it remains in the plane.



This is a general phenomenon for the solution set of any homogeneous system of linear equations. The set of all solutions to $A\mathbf{x} = \mathbf{0}$ is called the **null-space** of a matrix A , and it is denoted by $\text{Null}(A)$.

Null-space gets its name because *null* means zero, and the null-space of A is the set of all vectors that multiplication by A sends to the zero vector.

$$\text{Null}(A) = \{ \bar{\mathbf{x}} \in \mathbb{R}^n : A\bar{\mathbf{x}} = \bar{\mathbf{0}} \}.$$

The claim is that if 2 vectors \mathbf{x} and \mathbf{y} are in $\text{Null}(A)$, then $\mathbf{x} + \mathbf{y}$ is also in $\text{Null}(A)$. Let's see.

For \mathbf{x} and \mathbf{y} to be in $\text{Null}(A)$ means that $A\mathbf{x} = \mathbf{0}$ and $A\mathbf{y} = \mathbf{0}$.

Can you see why $A(\mathbf{x} + \mathbf{y}) = \mathbf{0}$?

$$\begin{cases} A\bar{\mathbf{x}} = \bar{\mathbf{0}} \\ A\bar{\mathbf{y}} = \bar{\mathbf{0}} \end{cases} \Rightarrow \begin{aligned} A\bar{\mathbf{x}} + A\bar{\mathbf{y}} &= \bar{\mathbf{0}} \\ A(\bar{\mathbf{x}} + \bar{\mathbf{y}}) &= \bar{\mathbf{0}} \end{aligned}$$

$A\mathbf{x} + A\mathbf{y}$ is equal to the zero vector because both terms are the zero vector. And because $A\mathbf{x} + A\mathbf{y} = A(\mathbf{x} + \mathbf{y})$ using the linearity of matrix multiplication, $A(\mathbf{x} + \mathbf{y}) = \mathbf{0}$, which means $\mathbf{x} + \mathbf{y}$ is in $\text{Null}(A)$. So, the set $\text{Null}(A)$ has this curious property: Adding 2 things in $\text{Null}(A)$ stays in $\text{Null}(A)$, so the set is closed under addition. In other words, you can't get anything new by adding things in $\text{Null}(A)$.

In fact, it is also closed under scalar multiplication—meaning it has the property that if \mathbf{x} is in $\text{Null}(A)$ and c is a scalar, then $c\mathbf{x}$ is in $\text{Null}(A)$, because $A(c\mathbf{x}) = cA\mathbf{x} = c\mathbf{0} = \mathbf{0}$.

Together, these 2 properties mean that linear combinations of things in the null-space remain in the null-space. Such a set is called a **subspace** if it is closed under taking linear combinations. The null-space is one example of a subspace.

One way linear systems arise naturally in economics is in setting budget constraints. For example, you may be trying to build some set of objects using some parts, and if you know the cost of those parts, then the total cost of the objects will vary linearly with the parts because each part has a particular cost. The total costs for the objects cannot exceed the budget constraints you are given.

Subspaces

The idea of a subspace is a generalization of a line or plane through the origin. Formally, a subspace can be defined as follows.

DEFINITION

A collection H of vectors in \mathbb{R}^n is called a **subspace** of \mathbb{R}^n if it satisfies 3 properties:

- 1 H contains the zero vector. (This property is here just to ensure that the subspace is nonempty.)
- 2 H is closed under addition. (If \mathbf{x} and \mathbf{y} are in H , then $\mathbf{x} + \mathbf{y}$ is in H .)
- 3 H is closed under scalar multiplication. (If \mathbf{x} is in H and c is a real number, then $c\mathbf{x}$ is in H .)

Whenever you refer to a subspace, you are always implicitly referring to a set that sits inside some bigger space. So, a given set is a subspace of something else. But the whole idea of a subspace is that when you take things in the set and perform operations like addition and scalar multiplication, you will not leave the set. You can't leave it by taking linear combinations. So, a subspace interacts with itself and is mostly oblivious to the things outside it.

Let's look at some simple examples.

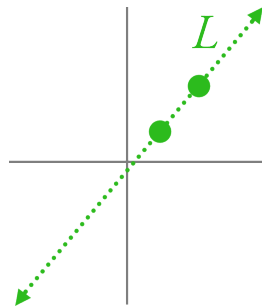
$\{\mathbf{0}\}$ in \mathbb{R}^n is subspace of \mathbb{R}^n .

The zero vector all by itself is a set that is a subspace of \mathbb{R}^n . Clearly, it contains zero, and if you add 2 things in it (both zero), then their sum is zero, so it stays in the set, and multiplying zero by a constant keeps you at zero, in the set.

\mathbb{R}^n is subspace of \mathbb{R}^n .

The entire space \mathbb{R}^n is a subspace of \mathbb{R}^n , because it contains the zero vector and is clearly closed under addition and scalar multiplication.

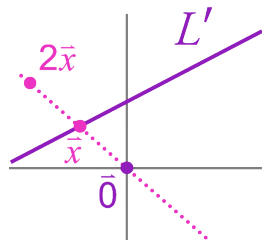
Is a line L passing through the origin in \mathbb{R}^2 a subspace? Any point on L is a multiple of some vector \mathbf{v} . So, if you take 2 points on L and add them, you are adding 2 multiples of \mathbf{v} , so you'll get another multiple of \mathbf{v} . And if you multiply a multiple of \mathbf{v} by a scalar, you'll get another multiple of \mathbf{v} . So, this line L satisfies all 3 properties of subspaces, so it is a subspace of \mathbb{R}^2 .



The same arguments show in general that a line through the origin in \mathbb{R}^n is a subspace of \mathbb{R}^n .

Can a line L' not passing through the origin be a subspace?

No, because it does not satisfy the first property: It does not contain the zero vector.



But is that the only obstacle to being a subspace? For example, suppose you include the zero vector in L' . Is that a subspace?

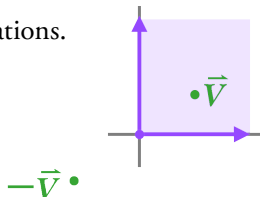
It's still not a subspace, because if you take a vector \mathbf{x} on L' and double it, you will get a vector $2\mathbf{x}$ that is not on L' . So, it is still not a subspace.

If you have 2 parallel lines in the plane, one that passes through the origin, is that a subspace?

No, because it's not closed under linear combinations.

Consider this set: the first quadrant in \mathbb{R}^2 .
Is that a subspace?

It contains zero and is closed under addition, but if you take a vector in the first quadrant and multiply it by -1 , it is no longer in the first quadrant. So, the set is not closed under scalar multiplication, so it is not a subspace, because it does not satisfy all 3 conditions.



A line through the origin is a subspace, and a plane through the origin is also a subspace. In fact, the span of a set of vectors in \mathbb{R}^n is always a subspace, because it passes through the origin, and if you add 2 vectors in the set, you generate something in the set. If you multiply a vector in the set by some scalar, the result stays in the set.

THEOREM

If H is the span of a set of vectors \mathbf{v}_1 through \mathbf{v}_k in \mathbb{R}^n , then H is a subspace of \mathbb{R}^n .

To prove this theorem, you have to check 3 conditions for a subspace that all concern membership in H . In other words, is the vector expressible as a linear combination of \mathbf{v}_1 through \mathbf{v}_k ? Let's check.

- 1 Is the zero vector in H ? Yes, because zero is the trivial linear combination of the vectors \mathbf{v}_1 through \mathbf{v}_k , where all coefficients are zero.
- 2 Suppose \mathbf{x} and \mathbf{y} are in H , meaning that they are linear combinations of \mathbf{v}_1 through \mathbf{v}_k . Then $\mathbf{x} + \mathbf{y}$ is too. To show this, write \mathbf{x} as $a_1\mathbf{v}_1 + \dots + a_k\mathbf{v}_k$ and \mathbf{y} as $b_1\mathbf{v}_1 + \dots + b_k\mathbf{v}_k$. Their sum $\mathbf{x} + \mathbf{y}$ is $(a_1 + b_1)\mathbf{v}_1 + \dots + (a_k + b_k)\mathbf{v}_k$. This is clearly a linear combination of \mathbf{v}_1 through \mathbf{v}_k , so $\mathbf{x} + \mathbf{y}$ is also in H .
- 3 If c is a scalar real number and \mathbf{x} is in H as a linear combination of \mathbf{v}_1 through \mathbf{v}_k , then $c\mathbf{x}$ is the linear combination of \mathbf{v}_1 through \mathbf{v}_k with all the coefficients multiplied by c .

So, the span of a set of vectors in \mathbb{R}^n is always a subspace of \mathbb{R}^n . And because the span of a set of vectors is always a subspace, a subspace is always the span of a set of vectors.

THEOREM

If H is a subspace of \mathbb{R}^n , then H is the span of some set of vectors.

But which set of vectors?

The entire set H would work. If you just take the span of all vectors of H , you still get H because H is closed under linear combinations. But maybe you can be more economical and choose a smaller set of vectors.

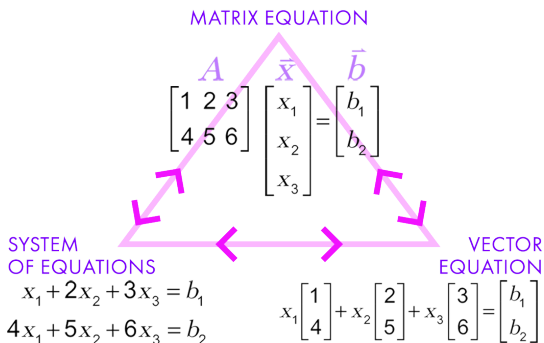
The point of this theorem is that spans of vectors are really the only kinds of things that can be subspaces. This gives us a clue, geometrically, of what subspaces must be: As sets of all linear combinations of a set of vectors, they have to be either a point (the origin), a line through the origin, a plane through the origin, or some k -dimensional flat space through the origin.

The Row Space and Column Space of a Matrix

There are 3 special subspaces that come associated with every matrix: the null-space, the row space, and the column space. These subspaces tell us something about the hidden structure of the matrix.

For example, because an $m \times n$ matrix represents a linear transformation from \mathbb{R}^n to \mathbb{R}^m , the row space and null-space are subspaces of \mathbb{R}^n and the column space is a subspace of \mathbb{R}^m , and that tells us something about the geometry of this linear transformation.

Recall the fundamental correspondence between a matrix equation, a system of equations, and a vector equation.



In the displayed example, one way to view this matrix equation is as a system of equations: 2 equations in 3 unknowns. Another way to view this matrix equation is as a linear combination of the columns—in this case, the 3 unknowns are the coefficients of the columns of the matrix A .

These 2 points of view are essentially a row view and a column view. For systems of equations, the rows are of interest because they represent equations. The column view pays attention to the column vectors: How do they combine to form the vector \mathbf{b} ? Switching back and forth between these 2 viewpoints will be continually handy.

For an $m \times n$ matrix A , the **row space** of matrix A is defined to be the span of the rows of A . It's denoted by $\text{Row}(A)$. Because it is the span of vectors in \mathbb{R}^n , it is a subspace of \mathbb{R}^n .

The **column space** of A is defined to be the span of the columns of A . It's denoted by $\text{Col}(A)$. It is the span of vectors in \mathbb{R}^m , so it is a subspace of \mathbb{R}^m .

READINGS

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, sections 2.8 and 2.9.

Poole, *Linear Algebra*, section 3.5.

BASES: BASIC BUILDING BLOCKS

A good set of vectors, called a basis, should span the space in question and be linearly independent so there isn't redundancy in how the vectors are expressed—in other words, there's only one way to write any vector as a linear combination of these vectors. It is also useful to have some of the vectors span the subspace so that coefficients of the other basis vectors are small.

Geometric Interpretation of Row, Column, and Null-Spaces

Recall that every $m \times n$ matrix A represents a linear transformation in the following way: Let the function $T(\mathbf{x})$ be defined to be $A\mathbf{x}$. Then, T is a linear transformation from \mathbb{R}^n to \mathbb{R}^m . In effect, T is the function that performs matrix multiplication by A .

$$T(\mathbf{x}) = A\mathbf{x}.$$

But $A\mathbf{x}$ can also be viewed as a linear combination of the columns of A . But this means that the **range** of T , which is the set of all possible vectors that result from multiplying A by some \mathbf{x} , is the same as the set of all linear combinations of the columns of A , which is called the column space of A .

$$\text{Range } T = \text{Col}(A).$$

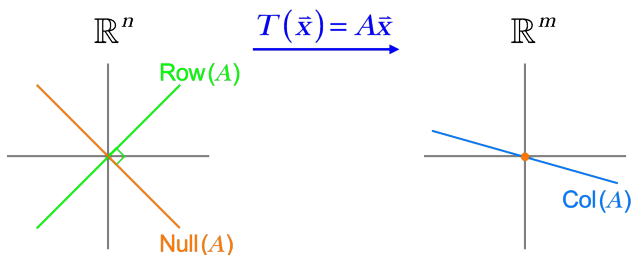
So, the column space of A has a natural geometric interpretation as the range of the linear transformation arising from multiplication by A . This column space could be all of \mathbb{R}^m , but it could also be a smaller subspace of \mathbb{R}^m .

$$\text{Kernel } T = \text{Null}(A).$$

Besides the range, another important object associated with a linear transformation $T(\mathbf{x}) = A\mathbf{x}$ is its **kernel**. The kernel of T is the set of all vectors in the domain of T that map to zero, so the kernel of T is just the null-space of the matrix A , because it is the set of all \mathbf{x} for which $A\mathbf{x}$ is zero. The null-space of A is a subspace of \mathbb{R}^n , the domain of T .

Both the kernel and the range of a linear transformation are important, because many subspaces arise as kernels or ranges of some linear transformation. For example, the solution set to a homogeneous system of equations $A\mathbf{x} = \mathbf{0}$ is the kernel of a linear transformation $T(\mathbf{x}) = A\mathbf{x}$. The span of any set of vectors is the range of the linear transformation $T(\mathbf{x}) = A\mathbf{x}$, where the vectors are the columns of the matrix A .

Let's display a picture that might suggest what is going on. A linear transformation represented by the $m \times n$ matrix A takes \mathbb{R}^n to \mathbb{R}^m .



The null-space, indicated in orange in the left picture, lives in \mathbb{R}^n and is the kernel of the linear transformation $T(\mathbf{x}) = A\mathbf{x}$. Multiplication by A takes everything in the orange line on the left and maps it to the origin, the orange dot on the right picture. Also, the entire picture on the left, representing all of \mathbb{R}^n , gets mapped to the blue line on the right, which is $\text{Col}(A)$, the range of the linear transformation T .

What about $\text{Row}(A)$? Any vector in $\text{Row}(A)$ is orthogonal to any vector of $\text{Null}(A)$. So, in the picture, the row space of A , in green on the left, is the set of vectors perpendicular to $\text{Null}(A)$ in orange.

You can see why vectors in $\text{Row}(A)$ are perpendicular to vectors in $\text{Null}(A)$ by using the observation that a $1 \times m$ row vector \mathbf{q} times an $m \times n$ matrix A is a row vector that is a linear combination of the rows of A .

If \mathbf{x} is a vector in $\text{Null}(A)$, then $(\mathbf{q}A)\mathbf{x}$ is, by associativity of matrix multiplication, the same as $\mathbf{q}(A\mathbf{x})$, but because \mathbf{x} is in $\text{Null}(A)$, you have $A\mathbf{x} = \mathbf{0}$. So, $\mathbf{q}(A\mathbf{x})$ is the 1×1 matrix consisting of the number 0. This means that $\mathbf{q}A$, which can represent any vector in $\text{Row}(A)$, must be orthogonal to any vector \mathbf{x} in $\text{Null}(A)$.

$$(\vec{q}A)\bar{x} = \vec{q}(A\bar{x}) = \vec{q}\vec{0}_{1 \times n} = 0.$$

If you had enough pictures like this, you would notice some interesting geometric relationships between these subspaces. For example, the dimension of the row space and column space are the same, even though they live in completely different spaces: The row space lives in the domain, and the column space lives in the codomain of the transformation $T(\mathbf{x}) = A\mathbf{x}$.

The Basis of a Subspace

Every subspace is the span of a set of vectors. Recall that one way to think of the span of a set of vectors in \mathbb{R}^n is to think of a spaceship at the origin and the vectors as the set of thrusters that you can fire to move in various directions. Then, the span of those vectors is all the places you can go using those thrusters. This will be a subspace of \mathbb{R}^n , but it may or may not be all of \mathbb{R}^n .

If you were designing this spaceship, you might try to be economical and use as few thrusters as possible. So, if you don't need all the vector thrusters, you might get rid of some of them. Intuitively, a basis of vectors is a minimal set of thrusters that you might need.

Formally, a **basis** B for a subspace H is a set of vectors such that 2 properties hold:

- 1 The vectors must span H .
- 2 The set B must be a linearly independent set.

The first condition ensures that your thrusters can reach all of the subspace H .

The second condition says that you have no redundancies. Recall that to be linearly independent means that there is no nontrivial linear combination of the vectors in B that will produce the zero vector. And

this is equivalent to saying that no vector in B can be written in terms of the other vectors. So, removing any thruster from your set means you won't be able to reach all of H .

Let's look at some examples. The first example is \mathbb{R}^3 , which is a subspace of itself. There is a natural basis called the **standard basis** for \mathbb{R}^3 . This basis is the set of 3 vectors that point along the coordinate axes and are unit length: $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$.

In math, the standard basis is often denoted using \mathbf{e} with a subscript—in this case, \mathbf{e}_1 , \mathbf{e}_2 , and \mathbf{e}_3 .

The physics community often uses the symbols \hat{i} , \hat{j} , and \hat{k} for these 3 vectors.

Using the standard basis, writing any other vector as a linear combination of these vectors is, in some sense, what you already do when you specify coordinates: the vector $(3, -1, 4)$ really means $3(1, 0, 0) + -1(0, 1, 0) + 4(0, 0, 1)$, or $3\hat{i} - \hat{j} + 4\hat{k}$.

From this example, it is easy to convince yourself that any vector in \mathbb{R}^3 is a linear combination of basis vectors, so they span all of \mathbb{R}^3 . And the vectors are linearly independent, because if you remove any one of them, you will not be able to get all the vectors in \mathbb{R}^3 . For example, if you remove \mathbf{e}_2 , you will not be able to get any vector whose second component is nonzero.

\mathbb{R}^3 (as subspace of itself) has basis:

$$\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

math: $\bar{\mathbf{e}}_1$, $\bar{\mathbf{e}}_2$, $\bar{\mathbf{e}}_3$

physics: \hat{i} , \hat{j} , \hat{k}

$$\text{So, } \begin{bmatrix} 3 \\ -1 \\ 4 \end{bmatrix} = 3\hat{i} - \hat{j} + 4\hat{k}.$$

But \mathbb{R}^3 has other bases. For example, consider $(1, 0, 0)$, $(1, 1, 0)$, and $(1, 1, 1)$. For simplicity, label them \mathbf{v}_1 , \mathbf{v}_2 , and \mathbf{v}_3 . You can check that these vectors form a basis for \mathbb{R}^3 by noting that the matrix A with these vectors as columns is already in row echelon form. Because there is no row of zeros, the linear system $A\mathbf{x} = \mathbf{b}$ can always be solved, which means the column vectors span \mathbb{R}^3 .

Also, the linear system $A\mathbf{x} = \mathbf{0}$ has no free variables, so it has a unique solution, which must be the trivial solution. So, the columns of A must be linearly independent.

The column vector $(3, -1, 4)$ can be written in terms of these vectors \mathbf{v}_1 , \mathbf{v}_2 , \mathbf{v}_3 by solving $A\mathbf{x} = (3, -1, 4)$. You will find $(3, -1, 4) = 4\mathbf{v}_1 - 5\mathbf{v}_2 + 4\mathbf{v}_3$.

\mathbb{R}^3 has other bases. Another:

$$\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\}$$

$$\text{So, } \begin{bmatrix} 3 \\ -1 \\ 4 \end{bmatrix} = 4\bar{\mathbf{v}}_1 - 5\bar{\mathbf{v}}_2 + 4\bar{\mathbf{v}}_3.$$

Aside from giving you a basic set that you can use to build all other vectors in your subspace, bases are important because there is only one way to build a vector as a linear combination of given basis vectors.

THEOREM

Every vector \mathbf{v} in H can be expressed as a linear combination of given basis vectors, and it can be done in only one way.

So, vectors that look different must be different! This would not necessarily be true if the basis vectors were not linearly independent.

A basis B for a subspace allows any vector in the subspace to be expressed uniquely as a linear combination of vectors in B . The coefficients of this combination are called weights or coordinates with respect to B .

So, for example, the coordinates of $(3, -1, 4)$ with respect to $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ are $(4, -5, 4)$. The coordinates are basically telling you how much to fire your rocket thrusters in particular directions. Any other basis will produce coordinates with respect to a different set of thrusters.

So, a subspace can have many bases. But it is a fact (that we won't prove here) that any 2 bases for a subspace H must have the same number of vectors. Because the number is independent of the basis chosen, that number can be called the **dimension** of the subspace H , denoted by $\dim H$. This tells you how many degrees of freedom the subspace H has.

So, if the subspace H has one degree of freedom (for example, choice of coordinate), it is called a line. If it has 2 degrees of freedom (choices for coordinates), it is called a plane—etc.

How to Find a Basis for a Column Space

If you want to find a basis for a subspace that is a span of a set of vectors, then by putting those vectors in the column of a matrix, you are changing the problem to finding a basis for the column space of a matrix.

There is an easy way to find a basis for any column space by using the reduced row echelon form of the matrix. For example, suppose matrix A is a 4×5 matrix with entries 1 through 19 and a stray 21 as the final entry.

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 \\ 11 & 12 & 13 & 14 & 15 \\ 16 & 17 & 18 & 19 & 21 \end{bmatrix}$$

Call the columns \mathbf{c}_1 through \mathbf{c}_5 . Now do row operations to get to the RREF, and after you've done that, call those columns \mathbf{c}_1' through \mathbf{c}_5' . Look at all the RREF columns with a leading 1, the so-called pivot columns—in this case, the first, second, and fifth columns (\mathbf{c}_1' , \mathbf{c}_2' , and \mathbf{c}_5').

This tells you which of the original columns will form a basis for the column space of A . In this case, \mathbf{c}_1 , \mathbf{c}_2 , and \mathbf{c}_5 are a basis; that is, the vectors $(1, 6, 11, 16)$, $(2, 7, 12, 17)$, and $(5, 10, 15, 21)$ are a basis for the span of all 5 columns. The span must be a 3-dimensional subspace of \mathbb{R}^4 .

Why does this curious method work?

The key idea is that doing row operations does not change the linear dependencies among the columns. So, if some linear combination of the \mathbf{c}_i 's equals zero, then the same linear combination of the \mathbf{c}_i' will equal zero, too!

For example, you can check that $\mathbf{c}_1 - 2\mathbf{c}_2 + \mathbf{c}_3 = \mathbf{0}$. So, $\mathbf{c}_1' - 2\mathbf{c}_2' + \mathbf{c}_3' = \mathbf{0}$ as well.

This also means that if in the RREF the first, second, and fifth columns are linearly independent, then in the original matrix, the first, second, and fifth columns are linearly independent as well. This is because whatever coefficients worked for \mathbf{c}_1' , \mathbf{c}_2' , and \mathbf{c}_5' to get the zero vector (namely, only zero coefficients), those same coefficients are the only ones that work for \mathbf{c}_1 , \mathbf{c}_2 , and \mathbf{c}_5 .

$$\begin{array}{ccccc} \mathbf{c}_1 & \mathbf{c}_2 & \mathbf{c}_3 & \mathbf{c}_4 & \mathbf{c}_5 \\ \left[\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 \\ 11 & 12 & 13 & 14 & 15 \\ 16 & 17 & 18 & 19 & 21 \end{array} \right] \end{array}$$

$$\text{Basis for } \text{Col}(A) = \{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_5\}.$$

RREF

$$\begin{array}{ccccc} \mathbf{c}_1' & \mathbf{c}_2' & \mathbf{c}_3' & \mathbf{c}_4' & \mathbf{c}_5' \\ \left[\begin{array}{ccccc} 1 & 0 & -1 & -2 & 0 \\ 0 & 1 & 2 & 3 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right] \end{array}$$

Thus, you see how to find a basis for a column space of a matrix. It is this method that produces a basis consisting of vectors from the original spanning set; it just potentially throws away some columns that weren't necessary. In this case, \mathbf{c}_3 and \mathbf{c}_4 could be thrown away.

How to Find a Basis for a Row Space

How do you find a basis for the row space of a matrix A ? You can use the RREF of the matrix A to find this as well. Just look at the nonzero rows of the RREF. That's a basis for the row space of A !

In the example, you can see that the first 3 rows of the RREF are nonzero vectors. These are a basis for the subspace spanned by the rows of A . Unlike the column space method, the row space basis produced by this method doesn't use the original vectors of the spanning set of rows. It finds some new vectors. So, in the example, a row space basis is $(1, 0, -1, -2, 0)$, $(0, 1, 2, 3, 0)$, and $(0, 0, 0, 0, 1)$. It is a 3-dimensional subspace of \mathbb{R}^5 .

This method works for a different reason than the column space method: Doing elementary row operations, by their very nature, does not change the row space of a matrix. This is because a row operation produces a new row that is a linear combination of old rows, and because elementary row operations are reversible, the reverse is true, too. In the RREF, the 3 nonzero rows clearly span the row space, and they are linearly independent because of the echelon form.

RREF

$$\begin{array}{ccccc|l} \mathbf{c}'_1 & \mathbf{c}'_2 & \mathbf{c}'_3 & \mathbf{c}'_4 & \mathbf{c}'_5 & \\ \hline 1 & 0 & -1 & -2 & 0 & \mathbf{r}'_1 \\ 0 & 1 & 2 & 3 & 0 & \mathbf{r}'_2 \\ 0 & 0 & 0 & 0 & 1 & \mathbf{r}'_3 \\ 0 & 0 & 0 & 0 & 0 & \end{array}$$

Basis for $\text{Row}(A) = \{\mathbf{r}'_1, \mathbf{r}'_2, \mathbf{r}'_3\}$.

How to Find a Basis for a Null-Space

What about finding a basis for a null-space? This means finding solutions to $A\mathbf{x} = \mathbf{0}$, and you've actually done this before. Recall that you form the augmented matrix A with a column of 0s and then row-reduce to reduced row echelon form. This enables you to figure out which are the free variables, by looking at columns without leading 1s. Then, you express all other variables in terms of the free variables, using the RREF, just as you learned to do previously.

In the example, the free variables are x_3 and x_4 , because they do not have leading 1s in them. You then get $(x_1, x_2, x_3, x_4, x_5)$ in terms of x_3 and x_4 , as x_3 times the vector $(1, -2, 1, 0, 0)$ plus x_4 times the vector $(2, -3, 0, 1, 0)$. This shows that the set of solutions is spanned by vectors $(1, -2, 1, 0, 0)$ and $(2, -3, 0, 1, 0)$.

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = x_3 \begin{bmatrix} 1 \\ -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 2 \\ -3 \\ 0 \\ 1 \\ 0 \end{bmatrix}.$$

It turns out that the vectors you obtain in this way must also be linearly independent—because if you look at the third and fourth entries of the vectors, there is no way to produce the zero vector unless the coefficients of these 2 vectors are zero.

The point is that you already know how to find null-spaces because you know how to solve linear equations using the RREF to find the null-space as the span of some vectors. The vectors you obtain in this way also turn out to be a basis.

The Rank-Nullity Theorem

Let's call the dimension of the row space of A the **row rank of A** and the dimension of the column space of A the **column rank of A** .

In the example, both these numbers turned out to be the same number. This was no accident. Both of them are equal to the size of their bases, and both bases were derived by looking at the leading 1s in the RREF. The column space basis consisted of columns of the original matrix corresponding to leading 1s in the RREF. The row space basis consisted of rows of the RREF corresponding to leading 1s. So, the dimensions of the row space and column space will always be the same, even though those objects live in different spaces (\mathbb{R}^n versus \mathbb{R}^m)!

THEOREM

The row rank and column rank of a matrix are the same.

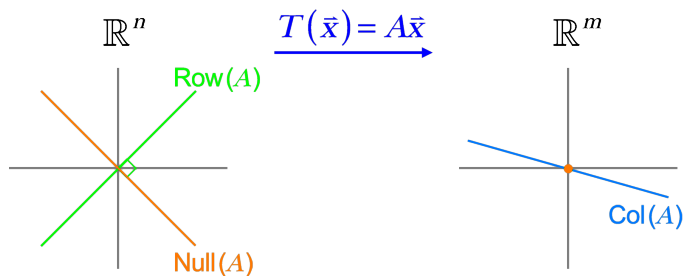
This number is called the **rank** of a matrix A , and it is the dimension of either its row space or its column space.

Let's call the dimension of the null-space of a matrix the **nullity of A** . This is equal to the number of free variables in the system of equations $A\mathbf{x} = \mathbf{0}$, which is the number of columns of the RREF that do not have a leading 1. So, this means that if you add the nullity to the number of columns with leading 1s, you get the total number of columns. For an $m \times n$ matrix, the number of columns is n .

THEOREM

The rank of A plus the nullity of A is the total number of columns of A .

You can see these relationships in the picture.



Here you see that the dimension of the row space of A plus the dimension of the null-space of A must equal the dimension of the entire domain on the left, which is n .

Also, the dimension of the row space of A is the same as the dimension of the column space of A , even though they are subspaces of different spaces.

READINGS

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, sections 4.3, 4.4, and 4.5.

McAnlis, “How JPG Works,” <https://medium.freecodecamp.org/how-jpg-works-a4dbd2316f35>.

Poole, *Linear Algebra*, section 3.5.

Lecture 11

INVERTIBLE MATRICES: UNDOING WHAT YOU DID

Solving a system of linear equations using Gaussian elimination involves row reductions to change the system to an equivalent, simpler system in reduced row echelon form (RREF). Such a system has exactly the same solutions as the original. This lecture will analyze a system of linear equations from a different point of view, by looking at the system of equations as a matrix equation: $A\mathbf{x} = \mathbf{b}$.

The Inverse of a Matrix

While Gaussian elimination remains the best way of solving a system of linear equations, looking at them as a matrix equation is useful for many theoretical reasons. It will help illuminate the general nature of the set of solutions, rather than solving a particular problem. You can also take advantage of the matrix structure of A in a way that row reductions don't.

Recall that if A is an $m \times n$ matrix, \mathbf{x} is a vector of unknown quantities x_1 through x_n , and \mathbf{b} is an m -dimensional vector, then $A\mathbf{x} = \mathbf{b}$ is a system of m linear equations in n unknowns. Notice that the left side of the previous equation is matrix multiplication, and the function

$$T(\mathbf{x}) = A\mathbf{x}$$

that performs matrix multiplication is a linear transformation that takes \mathbf{x} , a vector in \mathbb{R}^n , to $T(\mathbf{x})$, a vector in \mathbb{R}^m . It is a linear transformation because $A(\mathbf{x} + \mathbf{y}) = A\mathbf{x} + A\mathbf{y}$ and $A(c\mathbf{x}) = c(A\mathbf{x})$.

By viewing the left side of $A\mathbf{x} = \mathbf{b}$ as a linear transformation acting on the vector \mathbf{x} , you can see that solving the matrix equation $A\mathbf{x} = \mathbf{b}$ basically amounts to trying to find a vector \mathbf{x} that gets sent to the vector \mathbf{b} under multiplication by A .

For example, let's look at this system of equations:

$$2x_1 + 7x_2 = 5$$

$$x_1 + 4x_2 = 3.$$

This system can be rewritten as a matrix equation $A\mathbf{x} = \mathbf{b}$, where A is the 2×2 matrix as shown.

$$\begin{bmatrix} 2 & 7 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \end{bmatrix}$$

$$A \quad \bar{\mathbf{x}} = \bar{\mathbf{b}}.$$

The advantage of viewing it as a matrix equation is that it puts attention on the linear transformation that takes the unknown vector \mathbf{x} to the known vector \mathbf{b} . What is this transformation doing to the vector \mathbf{x} ?

Before pondering that, imagine a simpler equation that you are more familiar with, such as $5x = 3$.

What is the 5 doing to the unknown quantity x ?

It is multiplying it by 5. So, if you wanted to isolate the x to solve for it, you would need to undo what 5 is doing to x . You could think of dividing by 5 as undoing multiplication by 5, but if you were to apply this idea to matrices, you would be in trouble because you don't have a notion of division there. Is there another way to undo multiplication by 5?

You can undo multiplication by 5 by multiplying by $\frac{1}{5}$ on both sides, which gives you $x = \frac{3}{5}$.

You can try the same idea with matrix equations like $A\mathbf{x} = \mathbf{b}$. Can you multiply both sides of $A\mathbf{x} = \mathbf{b}$ by the same matrix to get something of the form $\mathbf{x} = \text{some vector}$?

In the numerical example, you could multiply $5x = 3$ by $\frac{1}{5}$ on both sides, and because $\frac{1}{5}$ times 5 equals 1 and because multiplication is associative, the left side is

$$\frac{1}{5}(5x) = \left(\frac{1}{5}5\right)x = 1x,$$

which is x .

So, the key to solving the numerical equation is to multiply both sides by the multiplicative inverse of 5—which means the number that when multiplied by 5 produces the number 1. The multiplicative inverse of 5 is $\frac{1}{5}$.

Similarly, for the matrix equation, you need a **multiplicative inverse**—something that when multiplied by A gives the equivalent of 1 for matrices, which would be a matrix for which multiplication by it leaves other matrices unchanged.

Recall that the identity matrix has that property: When you multiply the identity by any matrix, it leaves the matrix unchanged.

The identity matrix consists of 1s along the main diagonal and 0s everywhere else and is usually notated by I . You are thinking of the matrix being an $n \times n$ matrix A , so the identity matrix is the $n \times n$ identity matrix.

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

The multiplicative inverse of A is called A -inverse and is written as A^{-1} . This notation mimics how the multiplicative inverse of real numbers is sometimes written.

A^{-1} has the property that A^{-1} times A as well as A times A^{-1} equals the identity matrix I .

$$A^{-1}A = AA^{-1} = I.$$

If you had such an inverse for A , then you could take the matrix equation $A\mathbf{x} = \mathbf{b}$ and multiply both sides (on the left) by A^{-1} .

Note that when working with matrix equations in this way, you must be sure you are doing the same thing to both sides, and because order matters for matrix multiplication, you have to be sure when multiplying things on both sides that you are doing the multiplication either both on the left or both on the right.

When you multiply $A\mathbf{x} = \mathbf{b}$ by A^{-1} on the left, because matrix multiplication is associative, the A^{-1} times A becomes the identity matrix I . But I times \mathbf{x} is just \mathbf{x} . So, \mathbf{x} equals A^{-1} times \mathbf{b} , which gives the desired solution. You have thus solved for the unknown vector \mathbf{x} , as long as A^{-1} exists!

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

DEFINITION

Call a square matrix invertible (or nonsingular) if there exists a matrix A^{-1} such that A^{-1} times A and A times A^{-1} equals the identity matrix.

Let's look at an example.

Consider the 2×2 matrix with entries 2, 7, 1, 4. It has an inverse, a 2×2 matrix, with entries 4, -7, -1, 2.

$$\begin{bmatrix} 2 & 7 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 4 & -7 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 4 & -7 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 2 & 7 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

You can check that if you multiply these 2 matrices together in any order, you will get the identity matrix.

Notice if you know that A^{-1} exists, earlier arguments show that the solution to $A\mathbf{x} = \mathbf{b}$ must be $\mathbf{x} = A^{-1}\mathbf{b}$.

Let's see what this means for the system of linear equations you started with: the matrix equation $A\mathbf{x} = \mathbf{b}$, where the 2×2 matrix A has entries 2, 7, 1, 4 and the vector \mathbf{b} is (5, 3).

$$\begin{bmatrix} 2 & 7 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \end{bmatrix} \quad A \quad \bar{\mathbf{x}} = \bar{\mathbf{b}}.$$

The solution is obtained by multiplying both sides (on the left) by A^{-1} , which yields

$$\bar{x} = \begin{bmatrix} 4 & -7 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 5 \\ 3 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

You've solved for \bar{x} ! Now check that $x_1 = -1$ and $x_2 = 1$ solves the original system:

$$\begin{aligned} 2x_1 + 7x_2 &= 5 \\ x_1 + 4x_2 &= 3. \end{aligned}$$

So, if a matrix has an inverse, you can solve it! But unfortunately, not every matrix is invertible. For example, the 2×2 matrix $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ has no inverse.

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

You will not find a 2×2 matrix that you can multiply by this matrix and get the identity matrix.

One way to see this is to realize if it had an inverse, then by the previous reasoning, the system of equations

$$\begin{aligned} x_1 + x_2 &= a \\ x_1 + x_2 &= b \end{aligned}$$

would have to have a solution, no matter what a and b are. But if a and b are different numbers, then this system cannot have a solution, because the quantity $x_1 + x_2$ cannot add to 2 different numbers.

From this example, you learn 2 things.

- 1 There are some matrices that are not invertible. So, whenever you see a statement involving A^{-1} , you must be careful to remember that the statement only holds when A has an inverse.

- 2 The invertibility of a matrix A implies that the matrix equation $A\mathbf{x} = \mathbf{b}$ always has a solution for any vector \mathbf{b} , and it is unique (there is no other solution). This is because \mathbf{x} must be $A^{-1}\mathbf{b}$. In particular, if \mathbf{b} is the zero vector, the equation $A\mathbf{x} = \mathbf{0}$ has only the solution $\mathbf{x} = A^{-1}\mathbf{0}$, which is $\mathbf{0}$. This second observation can be stated in a theorem.

THEOREM

If an $n \times n$ matrix A is invertible, then the matrix equation $A\mathbf{x} = \mathbf{b}$ has a unique solution for any \mathbf{b} in \mathbb{R}^n . In particular, $A\mathbf{x} = \mathbf{0}$ has only the trivial solution $\mathbf{x} = \mathbf{0}$.

Finding the Inverse of a 2×2 Matrix

Recall that the $n \times n$ identity matrix is a matrix with 1s on the main diagonal and 0s everywhere else. To find an inverse for an $n \times n$ matrix A , you seek another matrix (called A^{-1}) that when multiplied on the left or right by A produces the identity matrix.

The simplest case is finding inverses for 1×1 matrices. Matrix multiplication for 1×1 matrices corresponds to multiplication of real numbers, and to find an inverse is just reciprocation, because the 1×1 identity matrix is just the matrix with single entry [1], and the product of a number and its reciprocal is 1.

$$\begin{bmatrix} a \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{a} \end{bmatrix}.$$

So, the inverse of the matrix with single entry $[a]$ is the matrix with single entry $\frac{1}{a}$ (the reciprocal of a) as long as a is nonzero.

If a is 8, then the inverse of $[a]$ is $[\frac{1}{8}]$, and you can check that $8 \times \frac{1}{8} = 1$.

If a is 0, the matrix is not invertible, because no number when multiplied by 0 will produce a 1.

For 2×2 matrices, the identity is the 2×2 matrix shown at right.

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

You can check that the inverse of a 2×2 matrix is given by the following formula.

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

the determinant

Notice that the 2×2 inverse formula is not valid in the case $ad - bc = 0$ because you can't take the reciprocal of zero. And there is no formula that will give an inverse in that case. When $ad - bc = 0$, the matrix has no inverse.

The quantity $(ad - bc)$ is called the **determinant** of the 2×2 matrix with entries a, b, c, d , and it plays a special role in determining whether a matrix is invertible.

A matrix inverse, if it exists, must be unique! This is true in general for any $n \times n$ matrix.

THEOREM

If A is invertible, then there is only one matrix A^{-1} with the property that A^{-1} times A equals A times A^{-1} , which equals the identity matrix I .

Is the $n \times n$ identity matrix unique? How do you know that the matrix shown here is the only matrix that behaves like an identity matrix for 3×3 matrices?

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

THEOREM

If for all $n \times n$ matrices A you have A times K equals K times A equals A , then K must be the $n \times n$ identity matrix.

Properties of Inverses

If you take the inverse of an inverse of a matrix A , you should get back the original matrix.

THEOREM

If A is invertible, then A^{-1} is invertible, and A^{-1-1} equals A .

How does the operation of taking inverses mesh with other matrix operations, such as addition or multiplication?

For multiplication, you might think that the inverse of the product is the product of the inverses, but you have to be careful.

First, remember that not every 2 matrices can be multiplied (the dimensions may not work out), but because invertible matrices are square, the product makes sense for matrices of the same dimensions. Second, the order of matrix multiplication matters.

THEOREM

If 2 $n \times n$ matrices A and B are invertible, then AB is invertible and its inverse is B^{-1} times A^{-1} .

Note that the order of B^{-1} and A^{-1} is backward from what you might first expect.

You have to be even more careful in the case of addition. The statement that the inverse of a sum is the sum of the inverses is not true.

$$(A + B)^{-1} \neq A^{-1} + B^{-1}.$$

It's not even true for 1×1 matrices, which are just real numbers. The reciprocal of a sum is not the sum of the reciprocals. For example,

$$2 + 4 = 6, \text{ but } \frac{1}{2} + \frac{1}{4} \neq \frac{1}{6}.$$

READINGS

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, section 2.2.

Poole, *Linear Algebra*, section 3.3.

THE INVERTIBLE MATRIX THEOREM

The previous lecture introduced the concept of the inverse of a matrix. You learned that for a system of solutions $A\mathbf{x} = \mathbf{b}$, if the matrix A has an inverse, then the system can be solved by multiplying both sides by A^{-1} on the left. Then, $\mathbf{x} = A^{-1}\mathbf{b}$. In practice, solving a system in this way is not the preferred method, because you must first find the inverse of A , and doing that involves row reduction. This lecture will discuss how to do that, but it's not faster than doing Gaussian elimination on the system of equations.

The Importance of Invertible Matrices

So why is the inverse of a matrix so important?

Understanding the inverse of a matrix A offers insights into the nature of the linear transformation represented by matrix multiplication by A (in which the vector \mathbf{x} maps to $A\mathbf{x}$). It also offers insights into the nature of solutions to a linear system $A\mathbf{x} = \mathbf{b}$.

For example, a linear system like $A\mathbf{x} = \mathbf{b}$ either has no solutions, one solution, or infinitely many solutions. However, if A has an inverse, then $A\mathbf{x} = \mathbf{b}$ must have a single unique solution because, in that case, $\mathbf{x} = A^{-1}\mathbf{b}$, and A^{-1} is unique and \mathbf{b} is specified in the problem.

Also, if $A\mathbf{x} = \mathbf{b}$ has a single unique solution for every \mathbf{b} , then A must be an invertible matrix. So, this property of $A\mathbf{x} = \mathbf{b}$ having a unique solution for every \mathbf{b} is actually equivalent to A being invertible. In fact, there are many more examples of criteria that are equivalent to A being invertible.

Finding the Inverse of an $n \times n$ Matrix

How do you find an arbitrary $n \times n$ matrix inverse?

You've already seen a formula for 1×1 and 2×2 matrices. But instead of having a formula for any $n \times n$ matrix, because no one really remembers the formula for matrices larger than 2×2 , there is a procedure for finding any $n \times n$ matrix. The procedure rests on the interesting fact that any elementary row operation on a matrix can be represented as left multiplication by some matrix, called an **elementary matrix**.

Let's say you have a 2×2 matrix A and you want to swap rows, but all you have is a machine that can multiply the matrix A by another matrix. Then, if you left-multiply A by the elementary 2×2 matrix E with entries 0, 1, 1, 0, you will see that this does indeed swap the 2 rows of A !

For example, carry out this multiplication and notice that the rows of the original matrix get swapped.

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 4 \\ 1 & 2 \end{bmatrix}.$$

Similarly, the row operation of adding twice the second row to the first row also has a matrix that represents it.

How do you find the elementary matrix that accomplishes a given row operation by left multiplication? You just perform that row operation on the identity matrix and you will have the matrix E .

THEOREM

Elementary matrices are invertible.

This theorem follows from the fact that row operations can be undone by another row operation. So, if you have an elementary matrix E corresponding to a row operation R and the row operation R' is the one that undoes R , then look at the elementary matrix E' corresponding to R' . The product of E' and E must be the identity matrix, because doing both row operations R and R' leaves the original matrix unchanged. So, E' is the inverse of E , and therefore E is invertible.

Here's how to find the inverse of an $n \times n$ matrix A . First, recall a theorem from the previous lecture: If A is invertible, then $A\mathbf{x} = \mathbf{0}$ has only the trivial solution. This means that there are no free variables among the coordinates of the unknown vector \mathbf{x} . So, when you try to put A in reduced row echelon form, because A is square, you must get the identity matrix! Thus, an invertible matrix must be row-reducible to the identity.

If the matrix A can be row-reduced to the identity, then there is a sequence of row operations that will get you there. This is the same as successively multiplying the matrix A on the left by some elementary matrices E_1, E_2, \dots, E_k to get the identity matrix. Thus, the product E_k through E_1 inverts A . Because inverses are unique if they exist, this product must be A^{-1} .

$$\begin{array}{l} A \rightarrow I_n \\ \text{must be } A^{-1} \\ \hline E_k \cdots E_2 E_1 A = I_n \\ [A \mid I_n] \rightarrow [I_n \mid A^{-1}]. \end{array}$$

This discussion points to a way to find the inverse of an $n \times n$ matrix A . You just form an augmented matrix by putting the matrix A on the left side and the $n \times n$ identity matrix on the right side of the augmented matrix. Then, you perform elementary row operations on the augmented matrix. This is, in effect, performing multiplication of elementary matrices on the augmented matrix. The goal is to put the left side in reduced row echelon form—and if A is invertible, then this will be the identity matrix. If you can do that, then the right side will be the product of those elementary matrices, because they must be the product of the elementary matrices with the identity. So, the right side has to be the inverse of A . This is how to find the inverse of a matrix.

Let's do an example. Suppose you want to use this method to invert the matrix A with entries 2, 7, 1, 4 (**figure a**).

$$\text{a} \begin{bmatrix} 2 & 7 \\ 1 & 4 \end{bmatrix}$$

(Yes, you already have a formula for 2×2 matrices, but let's see how this method works.)

$$\text{b} \left[\begin{array}{cc|cc} 2 & 7 & 1 & 0 \\ 1 & 4 & 0 & 1 \end{array} \right]$$

First, you augment the matrix with the identity on the right side (**figure b**).

$$\text{c} \left[\begin{array}{cc|cc} 1 & 4 & 0 & 1 \\ 2 & 7 & 1 & 0 \end{array} \right]$$

Next, try to row-reduce the left side to make the identity matrix. If you can do this, A will be invertible. Swapping rows, you get the matrix shown in **figure c**.

$$\text{d} \left[\begin{array}{cc|cc} 1 & 4 & 0 & 1 \\ 0 & -1 & 1 & -2 \end{array} \right]$$

Then, subtracting twice the first row from the second, you get the matrix shown in **figure d**.

$$\text{e} \left[\begin{array}{cc|cc} 1 & 4 & 0 & 1 \\ 0 & 1 & -1 & 2 \end{array} \right]$$

Then, multiply the second row by -1 (**figure e**).

$$\text{f} \left[\begin{array}{cc|cc} 1 & 0 & 4 & -7 \\ 0 & 1 & -1 & 2 \end{array} \right]$$

Then, subtract 4 times the second row from the first (**figure f**).

You've gotten the left side to the identity. Therefore, the right side must be A^{-1} . And indeed, the 2×2 formula for inverses gives exactly this matrix.

Inverting a 3×3 or an $n \times n$ with this method may be tedious, but the point is that if you have already programmed a computer to do Gaussian elimination, then computing inverses is easy.

What if you can't row-reduce the matrix to the identity matrix? Does that necessarily mean that the matrix is not invertible? The answer is yes.

Criteria for Telling If a Matrix Is Invertible

If a matrix is invertible, it has several consequences:

- ♦ $A\mathbf{x} = \mathbf{b}$ must have a unique solution for any \mathbf{b} .
- ♦ $A\mathbf{x} = \mathbf{0}$ must have only the trivial solution.
- ♦ The reduced row echelon form of A must be the identity.

Each of these criteria actually implies that the matrix A must be invertible. In other words, an invertible matrix is characterized by these criteria. In fact, there are a whole host of criteria that are equivalent to a matrix being invertible that tie together many of the concepts you've been exposed to so far!

THE FUNDAMENTAL THEOREM OF INVERTIBLE MATRICES

If A is an $n \times n$ square matrix, then the following are equivalent—meaning that any of them must imply any of the others.

- 1 A is invertible.
- 2 $A\mathbf{x} = \mathbf{b}$ has at least one solution for each \mathbf{b} .
- 3 $A\mathbf{x} = \mathbf{0}$ has only the trivial solution $\mathbf{x} = \mathbf{0}$.
- 4 $\text{RREF}(A)$ is the identity matrix.
- 5 A is the product of elementary matrices.
- 6 The columns of A span \mathbb{R}^n .
- 7 The columns of A are linearly independent.
- 8 The columns of A form a basis for \mathbb{R}^n .

- 9 The rank of A is n .
- 10 The nullity of A is 0.
- 11 A^T is invertible.
- 12 The rows of A span \mathbb{R}^n .
- 13 The rows of A are linearly independent.
- 14 The rows of A form a basis for \mathbb{R}^n .

The invertible matrix theorem shows that the concept of invertibility is an important one, because there are so many different ways to characterize it. For example, if you look at this matrix, call it A , you might ask if it is invertible.

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

There are many ways to decide. Perhaps if you visualized the column vectors in 3 dimensions, you could see that they span \mathbb{R}^3 . Then, you would know that the matrix is invertible. Immediately, you would also know many other things: that the row vectors must span \mathbb{R}^3 and be linearly independent, that the RREF of this matrix is the identity, that $A\mathbf{x} = \mathbf{0}$ only has one solution, that $A\mathbf{x} = \mathbf{b}$ always has at least one solution, etc.

READINGS

Anderson and Feil, “Turning Lights Out with Linear Algebra.”

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, sections 2.3 and 2.7.

Poole, *Linear Algebra*. Read the Fundamental Theorem of Invertible Matrices in section 3.5.

Stock, “Merlin’s Magic Square Revisited.”

QUIZ FOR LECTURES 7–12

- 1 Consider a system of equations $A\mathbf{x} = \mathbf{b}$, where \mathbf{b} is the zero vector. Can this system of equations ever be inconsistent? [LECTURE 7]
- 2 Consider a 4×3 matrix A . What must $\text{RREF}(A)$ look like for $A\mathbf{x} = \mathbf{b}$ to have a unique solution? [LECTURE 7]
- 3 Determine whether the vectors $(1, 2, 3)$, $(1, 1, 2)$, and $(1, 3, 4)$ are linearly independent. [LECTURE 8]
- 4 Suppose 3 vectors in \mathbb{R}^3 are not linearly independent. Can the span of those vectors ever be all of \mathbb{R}^3 ? [LECTURE 8]
- 5 Let S be a set of all points in \mathbb{R}^3 whose second coordinate is zero. Is S a subspace of \mathbb{R}^3 ? Appeal to the 3 properties of a subspace. [LECTURE 9]
- 6 Consider a basket of some number of croissants and some number of donuts. The ingredient demands of (flour, eggs, sugar) for this basket is a vector in \mathbb{R}^3 . Because the numbers of croissants and donuts vary (over both positive and negative numbers), explain why the set D of possible ingredient demands is a subspace of \mathbb{R}^3 by appealing to the properties of a subspace. [LECTURE 9]
- 7 Let A be the matrix
$$\begin{bmatrix} 1 & 1 & 2 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 3 & 4 & 5 \end{bmatrix}.$$

The RREF of this matrix is
$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Use the RREF to assist you in finding a basis for the row space and a basis for the column space of this matrix. [LECTURE 10]

- 8 Let A be the same as in the previous problem. Use the RREF to assist you in finding a basis for the null-space of this matrix. [LECTURE 10]
- 9 Find the inverse of the 2×2 matrix $\begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix}$ if it exists. [LECTURE 11]
- 10 Show that the 2×2 matrix

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

has no inverse in a different way than in the lecture by showing that if you multiply it on the right by a generic 2×2 matrix, then the result must be a matrix with identical rows, and therefore the result cannot be an identity matrix. [LECTURE 11]

- 11 Let E be the elementary matrix $\begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix}$.
- a Verify that E times a 2×2 matrix A performs the elementary row operation that adds 3 times the first row to the second row of A .
 - b Find the inverse of E and verify that $E^{-1}A$ is the matrix you get after subtracting 3 times the first row from the second row of A . [LECTURE 12]
- 12 If a matrix B has the property that $B\mathbf{x} = \mathbf{0}$ for a nonzero vector \mathbf{x} , then explain why B is not invertible. [LECTURE 12]

Solutions can
be found on
page 298.

Lecture 13

DETERMINANTS: NUMBERS THAT SAY A LOT

The previous lecture showed the importance of invertibility and offered at least a dozen different criteria that help you determine if a matrix is invertible. Many of them were conceptual criteria, such as that the rows are linearly independent, which in practice are not obvious from just looking at a matrix. The idea of a determinant is to find a single number that can be computed from the entries of a matrix that will tell you if a matrix is invertible. The formulas for inverses of 1×1 and 2×2 matrices give you a clue for how to find such a criterion.

The 1×1 and 2×2 Determinants

For 1×1 matrices, there is only a single entry, call it a , and the inverse is just the reciprocal, $1/a$.

So, the matrix has an inverse if and only if the number a is not zero.

$$\begin{bmatrix} a \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{a} \end{bmatrix}.$$

Let's define the determinant of the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ to be a .

For 2×2 matrices, recall that the inverse of the matrix a, b, c, d is

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

the determinant

The only instance where this formula fails to make sense is when $ad - bc = 0$. And when that happens, the ratio of c to a equals the ratio of d to b , assuming a and b aren't 0.

$$ad - bc = 0 \Rightarrow \frac{c}{a} = \frac{d}{b}.$$

In that case, the second row of the matrix a, b, c, d will be a multiple of the first row.

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

So, the rows of the matrix are not linearly independent, and by the invertible matrix theorem, the matrix cannot be invertible. Other cases, when a is 0 or b is 0, can be handled similarly.

So, the number $ad - bc$ is nonzero if and only if the 2×2 matrix is invertible.

Let's define the determinant of the 2×2 matrix at right to be $ad - bc$.

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$$

The determinant of the matrix A is usually written as $\det(A)$, but sometimes you'll see the determinant notated by writing the matrix entries with brackets replaced by absolute value signs.

$$\left| \begin{array}{cc} a & b \\ c & d \end{array} \right| = ad - bc$$

But don't let the absolute value notation fool you; the determinant of a matrix can be a negative number.

For example, let's compute the determinant of the following matrix.

$$\det \begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix} = (1)(4) - (-2)(-3) = 4 - 6 = -2.$$

Because -2 is nonzero, the matrix is invertible.

The 3×3 Determinant

Recall that a matrix A is invertible if and only if its determinant is not zero.

To find such a number, recall that if a matrix is invertible, it will row-reduce to the identity. So, if it's not invertible, you should run into problems.

Suppose you have a matrix with entries abc, def, ghi and you try to row-reduce it.

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

You won't deal with all the cases; you will just see what happens in a particular case.

If you make a few assumptions—that a isn't zero and that $ae - db$ is also not zero—then the row reduction looks something like this:

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \rightarrow \begin{bmatrix} a & b & c \\ 0 & e - \frac{d}{a}b & f - \frac{d}{a}c \\ 0 & h - \frac{g}{a}b & i - \frac{g}{a}c \end{bmatrix} \rightarrow \begin{bmatrix} a & b & c \\ 0 & \frac{1}{a}(ae - db) & \frac{1}{a}(af - dc) \\ 0 & 0 & \frac{1}{ae - db}\Delta \end{bmatrix}.$$

This means that if the number Δ , in the bottom right corner, is zero, then you have a row of zeros, and the matrix will not row-reduce to the identity, so the matrix is not invertible. And if Δ is nonzero, it will row-reduce to the identity!

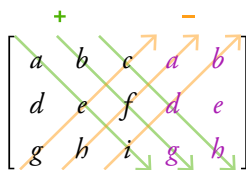
So, let Δ be the quantity called the determinant of a 3×3 matrix.

Taking a look at Δ , you'll find

$$\Delta = (aei - ahf) - (bdi - bgf) + (cdh - cge).$$

This is a crazy formula. Instead of memorizing it, consider the following device for remembering it.

Write the columns of the matrix as shown here. Draw 3 diagonals that go down and to the right and 3 diagonals that go up and to the right. If you multiply the entries along the diagonals and use a plus sign if it goes down and to the right and a minus sign if it goes up and to the right, you will get



$$\Delta = (aei - ahf) - (bdi - bgf) + (cdh - cge).$$

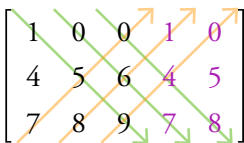
For example, if you look at the first diagonal down and to the right, you see a , e , and i . These correspond to the aei term in the formula for the determinant.

For example, suppose you want the determinant shown at right.

$$\begin{bmatrix} 1 & 0 & 0 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Then you form the diagram shown below.

$$45 + 0 + 0 - 0 - 48 - 0 = -3$$



Compute the products along the diagonals, and then add the diagonal products down and to the right and subtract the diagonal products up and to the right. You get -3 .

Notice that there were lots of zeros, because there were lots of zeros in the first row. So, sometimes computing the determinant can be really easy.

This way of remembering how to find a 3×3 determinant does not generalize to computing determinants of larger matrices. But there is another way to remember the 3×3 formula that does generalize.

Look at the last computation. There were lots of zeros, and the only terms that remained were associated to the single 1 in the first row, and they reduced to $5 \times 9 - 8 \times 6$. This looks just like a 2×2 determinant, the one you get when you ignore the first row and first column!

The formula for a 3×3 determinant builds on this idea. The expression for Δ ,

$$\Delta = (aei - ahf) - (bdi - bgf) + (cdh - cge),$$

can be rewritten like this:

$$\Delta = +a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix}.$$

Notice that the 3×3 determinant depends on computing 2×2 determinants! In that sense, this definition is a recursive definition. And the 2×2 determinants have coefficients that come from the first row. The terms are alternating in sign, starting with plus. And the 2×2 matrices come from taking the coefficient, looking at its position in the 3×3 matrix, and then crossing out the row and column that it is in.

If you try this with the example, you'll get this computation:

$$\begin{vmatrix} 1 & 0 & 0 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{vmatrix} = 1 \begin{vmatrix} 5 & 6 \\ 8 & 9 \end{vmatrix} - 0 \begin{vmatrix} 4 & 6 \\ 7 & 9 \end{vmatrix} + 0 \begin{vmatrix} 4 & 5 \\ 7 & 8 \end{vmatrix} = -3.$$

Note the terms alternate in sign. The coefficients 1, 0, and 0 come from the first row of the matrix. And if you want the submatrix associated with the second term, you cross out the first row and the second column of the matrix. This leaves the 2×2 matrix 4, 6, 7, 9.

$$\begin{vmatrix} 1 & 0 & 0 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{vmatrix}$$

And when you do this computation for the determinant, you get -3 , as before.

Note that the formula expands around terms in the first row, but you can actually compute the determinant by expanding around any row—you just have to take care to ensure that you have the correct signs on the terms.

For example, if you want to compute the 3×3 determinant by expanding around the second row, the signs still alternate, but now you start with a minus sign on the first term. So, in this example, if you chose the second row—4, 5, 6—then you could expand like this:

$$\begin{vmatrix} 1 & 0 & 0 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{vmatrix} = -4 \underbrace{\begin{vmatrix} 0 & 0 \\ 8 & 9 \end{vmatrix}}_0 + 5 \underbrace{\begin{vmatrix} 1 & 0 \\ 7 & 9 \end{vmatrix}}_9 - 6 \underbrace{\begin{vmatrix} 1 & 0 \\ 7 & 8 \end{vmatrix}}_8 = -3.$$

The 4, 5, and 6 form the coefficients, but now there is a minus sign on the 4, a plus sign on the 5, and a minus sign on the 6. The submatrices are derived the same way as before—by deleting the corresponding row and column of the row entries. So, for example, the submatrix associated with 5 is what you get by deleting row 2 and column 2 from the matrix, because that's where the entry 5 is: in the second row and second column. So, you get the submatrix 1, 0, 7, 9. And the computation, perhaps surprisingly, turns out the same: -3 !

If you want to expand around the third row, the alternating signs would start with a plus. There is a chessboard pattern that helps you remember which signs go where. If the coefficient comes from row i , column j , the term should have sign $(-1)^{i+j}$.

You can see from this figure that the second row starts with a minus.

$$\begin{bmatrix} + & - & + & \cdots \\ - & + & - & \cdots \\ + & - & + & \\ \vdots & \vdots & & \ddots \end{bmatrix}$$

So, this number, the determinant, can be obtained in a recursively defined fashion, by expanding around any row.

Even more amazingly, you can get the determinant by doing a similar computation, by expanding around any column, too!

For example, if you choose the third column, 0, 6, 9, then you get 3 terms here:

$$\begin{vmatrix} 1 & 0 & 0 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{vmatrix} = +0 \begin{vmatrix} 4 & 5 \\ 7 & 8 \end{vmatrix} - 6 \underbrace{\begin{vmatrix} 1 & 0 \\ 7 & 8 \end{vmatrix}}_8 + 9 \underbrace{\begin{vmatrix} 1 & 0 \\ 4 & 5 \end{vmatrix}}_5 = -3.$$

The alternation pattern starts with a plus, because it starts in row 1, column 3, and the sign there is $(-1)^{1+3}$, which is 1. And the submatrices are found, as before, by crossing out rows and columns of the corresponding entries. And, again, you get -3 .

The $n \times n$ Determinant

Let's try to write down a general formula for the determinant of an $n \times n$ matrix.

If you want to expand around row i , then you are looking at the entries a_{i1} through a_{in} . Those are the entries of row i . For each of these entries a_{ij} , there is a term in the sum that looks like this:

$$\det A = a_{i1}C_{i1} + a_{i2}C_{i2} + \dots + a_{in}C_{in},$$

where the C_{ij} is called the ij -**cofactor**.

$$C_{ij} = (-1)^{i+j} \det A_{ij}.$$

And the ij -cofactor contains the same chessboard sign pattern as before, such that the term with coefficient a_{ij} has sign $(-1)^{i+j}$.

$$\begin{bmatrix} + & - & + & \cdots \\ - & + & - & \cdots \\ + & - & + & \\ \vdots & \vdots & & \ddots \end{bmatrix}$$

Notice that for a 4×4 matrix, the first row would end with a minus and the second row would also begin with a minus, and that is not a problem. The key thing is that vertically or horizontally neighboring entries differ by a sign.

The determinant formula contains the determinant of this submatrix, the one you get by crossing out row i and column j , as shown at right.

$$A_{ij} = \underbrace{\begin{bmatrix} \text{delete} \\ \text{Col } j \\ \text{---} a_{ij} \text{---} \\ \text{delete} \\ \text{Row } i \end{bmatrix}}_A$$

The determinant formula that expands around columns is similar and looks like this:

$$\det A = a_{1j}C_{1j} + a_{2j}C_{2j} \dots + a_{nj}C_{nj}.$$

For each column, you have a different way to compute the determinant. This definition of the determinant, where you pick a row or column to expand around, is called **expansion by cofactors**. And you get the same answer no matter which row or column you use!

Calculating Determinants Quickly

The general determinant has been defined in a complicated recursive formula. In practice, computing the determinant this way is rather slow, because of the recursion.

The key to finding a faster way is to appeal to elementary row operations—some of which may change the determinant, but you can just keep track of what changes occur. If you row-reduce a matrix (which is square) to row echelon form, then you will be able to easily compute the determinant, because it will be the product of the diagonal entries (by looking at recursive expansion by cofactors around the first column of each determinant needed).

Here are 3 ways in which doing row operations may change the determinant.

- 1 If you swap 2 rows of a matrix, all that happens is that the determinant changes sign. You can see this most easily in the case when you swap 2 adjacent rows i and $i + 1$ and then compute the determinant by expanding around row i before the swap and row $i + 1$ after the swap. The terms will be the same, except for a sign change in each one. When you swap nonadjacent rows, the determinant flips sign also.

- 2 The second row operation is multiplying a single row by a factor k . If you compute the determinant by expanding around that row, you see that multiplication by k just multiplies each term by k . So, the determinant changes by the factor k as well.
- 3 The third row operation is adding a multiple of one row to another. This actually does not change the determinant at all!

So, you know exactly how row operations change the determinant. In fact, to get to row echelon form, you only ever need operations 1 and 3, which at most flip the sign! This is a much faster method than using the recursive formula.

Notice that if the determinant of a matrix is nonzero, row operations will not make it zero, and if the determinant is zero, row operations won't make it nonzero. This is a key to understanding why a matrix A is invertible if and only if the determinant of A is nonzero.

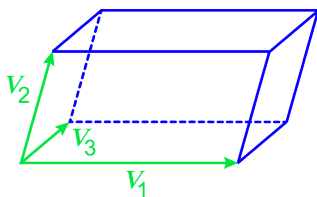
The invertible matrix theorem says that a matrix is invertible if and only if it can be row-reduced to the identity matrix. But because the identity matrix has a determinant of 1, row-reducibility to the identity means that the original matrix must have had a nonzero determinant.

THEOREM

A matrix is invertible if and only if its determinant is nonzero.

The Geometric Meaning of the $n \times n$ Determinant

If you have n vectors in \mathbb{R}^n , those vectors span a parallelepiped—an object formed by pairs of parallel sides—in a natural way: Just form all possible sums of those vectors, including the zero vector, and you will have the corners of a parallelepiped. So, a 2-dimensional parallelepiped is a parallelogram. A 3-dimensional parallelepiped is shown at right.



$$\det \begin{bmatrix} | & | & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \\ | & | & | \end{bmatrix}$$

If you have a square matrix with columns \mathbf{v}_1 , \mathbf{v}_2 , \mathbf{v}_3 , then the determinant tells you the signed volume of a parallelepiped spanned by those 3 vectors.

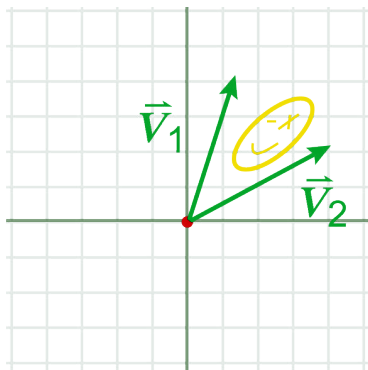
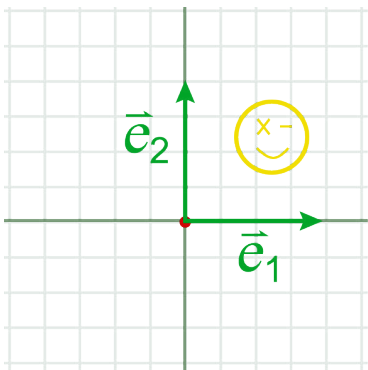
That means the number you get is the volume of the parallelepiped, but its sign may be positive or negative. Which sign it is depends on the order of the vectors.

- ◆ For a parallelogram, if \mathbf{v}_2 is counterclockwise from \mathbf{v}_1 , the sign of the volume is positive; otherwise, it is negative.
- ◆ In 3 dimensions, you use the right-hand rule on \mathbf{v}_1 and \mathbf{v}_2 and compare that to the direction of \mathbf{v}_3 : If they are in the same general direction (positive dot product), then the sign of the volume is positive; otherwise, it's negative.
- ◆ In higher dimensions, the sign will be positive if the vectors \mathbf{v}_1 through \mathbf{v}_n are oriented in the same way as the standard basis vectors and negative if the orientation is mirror-reversed.

Because the determinant of A^T is the same as the determinant of A , the determinant is also the signed volume of the parallelepiped formed by the rows of A .

It may be helpful to think about another interpretation of the determinant. Look at the linear transformation that takes the standard basis vectors \mathbf{e}_1 through \mathbf{e}_n to the column vectors \mathbf{v}_1 through \mathbf{v}_n in the order given. This is given by $T(\mathbf{x}) = A\mathbf{x}$, where A has columns \mathbf{v}_1 through \mathbf{v}_n .

Then, if you take a smiley face and watch where it goes under the transformation, it may get deformed and stretched, and it may also get mirror-reversed.



The determinant tells you the factor by which the volume of that smiley face got stretched. And the sign tells you whether the smiley face got mirror-reversed (if so, the sign is negative).

Consequences

Geometric interpretations of the determinant are useful. For example, to find the surface area of a strange shape, you can cut up the shape by coordinates and get small shapes that have known formulas for surface area, and you can use the determinant to help compute those areas.

Another property of determinants is that they behave well under multiplication. It may be a surprise that the determinant of a product is the product of the determinants!

For example, if you have 2 matrices, A and B , and their product, AB , when you compute their determinants, you get 3, -2 , and their product has a determinant of -6 . It's rather surprising that these things should be related at all.

But if you think of determinants as scaling factors, this property becomes obvious, because the product AB corresponds to the linear transformation $T(\mathbf{x}) = AB\mathbf{x}$, which first performs B , scaling the face by a factor of -2 , and then A , which scales the face by a factor of 3. So the total factor in scaling is clearly 3 times -2 , which is -6 .

$$\det(AB) = \det(A)\det(B)$$

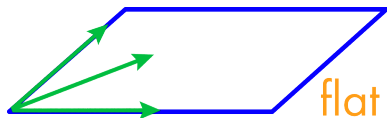
Also, determinants are not additive. It is not the case that the determinant of a sum is the sum of the determinants.

$$\begin{array}{ccccc} A & & B & & AB \\ \left[\begin{array}{cc} 1 & -1 \\ 1 & 2 \end{array} \right] & \bullet & \left[\begin{array}{cc} -2 & 1 \\ 2 & 0 \end{array} \right] & = & \left[\begin{array}{cc} -4 & 1 \\ 2 & 1 \end{array} \right] \\ 3 & & -2 & & -6 \end{array}$$

$$\det(A + B) \neq \det(A) + \det(B).$$

Another thing you might be able to see from the geometric interpretation is that a matrix A is invertible if and only if its scaling factor of $T(\mathbf{x}) = A\mathbf{x}$ is nonzero. If it is zero, then linear transformation is basically collapsing a region with nonzero volume to zero volume. As such, there will be points that get mapped to the same point, so it is not possible to find an inverse transformation.

Moreover, you can see some of the equivalences in the invertible matrix theorem. For example, if the columns of a matrix are not linearly independent, then one vector is a linear combination of the others, so the corresponding parallelepiped will be flat and have zero volume! So, it is clear that the determinant is zero.



READINGS

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, sections 3.1, 3.2, and 3.3.

Poole, *Linear Algebra*, section 4.2.

EIGENSTUFF: REVEALING HIDDEN STRUCTURE

One of the big themes of this course is how linear algebra enables you to see hidden structure. And the next several lectures will unpack one of the big ideas in linear algebra: the notion of eigenvectors and eigenvalues. Unpacking the concepts of eigenvectors and eigenvalues will illuminate how important they are and how useful they are to real applications.

Population Dynamics Application

All around you, there are examples of systems: groups of things that interact in interesting ways. For example, the solar system is a collection of planets related by the laws of gravity. Often in a system, there are things that are in motion, and such things are referred to as dynamical systems. The motion described doesn't have to be physical; it could also be relational, as in numbers of various things, such as amounts of a chemical in a chemical reaction.

Let's consider an example of a dynamical system involving the populations of foxes and rabbits. Let F_n denote the population of foxes at time n , which is measured in number of breeding cycles. Let R_n denote the population of rabbits at time n .

When people study population dynamics, they are often interested in knowing the long-term behavior of the populations. In other words, what happens to F_n and R_n as n goes to infinity?

In order to answer this question, you need to develop a model that describes how the populations of foxes and rabbits relate to each other.

Mathematical modeling is a process that uses mathematics to represent and analyze a situation or problem you want to understand. You may begin with a very simple model, but often there's a lot you can learn even from a simple model. That's because you make choices about what features belong in your model and incorporate only the most important ones. Then, your results show how those features influence the answer. So, the situation informs the model, and the model in turn informs what you can learn about the situation. And that, in turn, can inform a revision to the model if you wish to study further.

In the example of foxes and rabbits, a very basic model might begin by assuming that the relationship between the populations of foxes and rabbits from time n to $n + 1$ changes linearly. So, (F_{n+1}, R_{n+1}) depends in a linear way on both F_n and R_n . You might also assume that foxes eat rabbits, so if there are no rabbits, the total number of foxes decreases from breeding cycle n to breeding cycle $n + 1$, say, by 40% each cycle.

Similarly, if there are no foxes, you can expect the total number of rabbits to grow—say by 20% each breeding cycle. Then, your model might look like this:

$$\begin{aligned}F_{n+1} &= 0.6F_n + 0.5R_n \\ R_{n+1} &= -pF_n + 1.2R_n.\end{aligned}$$

Notice in the first equation that the coefficient of F_n is 0.6, which means that in the absence of rabbits, when $R_n = 0$, 40% of foxes are dying in each cycle so that F_{n+1} , the population of foxes at time $n + 1$, is 60% of F_n , the population at time n . The coefficient 0.5 for R_n shows how rabbits contribute to the fox population: Every 2 rabbits contributes an additional fox to the population in the next cycle (or, in other words, prevents one from dying).

In the second equation, if $F_n = 0$, the number of rabbits at time $n + 1$ is 1.2 times the number of rabbits at time n . That's the 20% growth in each breeding cycle. And the coefficient of F_n here is $-p$; the fact that it is negative means that foxes contribute negatively to the rabbit population. In the second equation, p has been left as an unspecified parameter, called the **predation parameter** because this is the term that shows how quickly foxes eat rabbits. The larger this term, the more rabbits are eaten by foxes.

$$\begin{aligned}F_{n+1} &= 0.6F_n + 0.5R_n \\ R_{n+1} &= -pF_n + 1.2R_n.\end{aligned}$$

This is a simple linear model. It is surely not an exact description of what is going on, in many ways. First, the coefficients have been made up for this toy model, but if you were determined, you could try to use data to estimate the coefficients. Also, rabbit and fox populations are always whole numbers, but that's not true of this model. And even if

you ignore that, the actual real-life fox/rabbit population dynamic is almost surely not linear. However, it is probably approximately linear, so if you explore this model, you will likely get some insight that will be valuable nonetheless.

Let's rewrite the linear model as a matrix equation, in which the population vector is \mathbf{x}_n and has components F_n and R_n . The model tells you that the vector \mathbf{x}_{n+1} is some matrix A times the vector \mathbf{x}_n .

$$\begin{bmatrix} F_{n+1} \\ R_{n+1} \end{bmatrix} = \begin{bmatrix} 0.6 & 0.5 \\ -p & 1.2 \end{bmatrix} \begin{bmatrix} F_n \\ R_n \end{bmatrix}$$

$$\bar{\mathbf{x}}_{n+1} = A \bar{\mathbf{x}}_n.$$

Given the initial populations at time 0, what is the long-term behavior of the populations of foxes and rabbits at time n as n goes to infinity? In particular, do the populations grow forever or die out? And what will happen as the predation parameter varies?

Notice that \mathbf{x}_{n+1} is written recursively in terms of \mathbf{x}_n . So if $\mathbf{x}_1 = A\mathbf{x}_0$, then $\mathbf{x}_2 = A\mathbf{x}_1$, but that is just $AA\mathbf{x}_0$, or $A^2\mathbf{x}_0$. So, if you repeatedly use the recursive formula on \mathbf{x}_n , you can inductively see that \mathbf{x}_n will be $A^n\mathbf{x}_0$.

But how can A^n be computed? Can it be done quickly?

$$\bar{\mathbf{x}}_0 = \begin{bmatrix} F_0 \\ R_0 \end{bmatrix}$$

$$\bar{\mathbf{x}}_1 = A\bar{\mathbf{x}}_0$$

$$\bar{\mathbf{x}}_2 = A\bar{\mathbf{x}}_1 = AA\bar{\mathbf{x}}_0 = A^2\bar{\mathbf{x}}_0$$

$$\vdots$$

$$\bar{\mathbf{x}}_n = A^n\bar{\mathbf{x}}_0.$$

You'll return to this example once you've developed the necessary machinery. But notice now that answering the question about fox and rabbit populations has led, in a very simple model, to the problem of how to compute powers of a matrix. You will soon see how eigenvectors and eigenvalues can help.

Understanding Matrix Powers

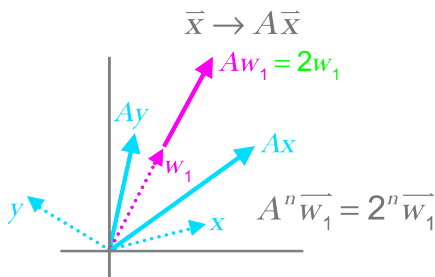
Taking a matrix A to a very large power could involve a lot of calculations. But there may be some situations where the calculations simplify greatly.

The best way to see this is to think about the associated function that takes a vector \mathbf{x} to $A\mathbf{x}$. Recall that this is a linear transformation that sends zero to zero but generally just moves vectors around. So, \mathbf{x} and \mathbf{y} go to $A\mathbf{x}$ and $A\mathbf{y}$, and these may be pointing in different directions than \mathbf{x} and \mathbf{y} .

But what if some vector, call it \mathbf{w}_1 , had the special property that $A\mathbf{w}_1$ points in the same direction as \mathbf{w}_1 ? That would mean $A\mathbf{w}_1$ is a scalar multiple of \mathbf{w}_1 .

In this case, let's suppose that multiple is 2. If so, then such a special vector is called an **eigenvector** of A , which has the property that when you apply the linear transformation that multiplies the vector by A , then you get a scalar multiple of the eigenvector. The multiple you get is called the **eigenvalue**—in this case, 2. And for eigenvectors, the action A is very simple. For example, if A times \mathbf{w}_1 is $2\mathbf{w}_1$, then \mathbf{w}_1 is an eigenvector for A and 2 is its eigenvalue. Then, $A^n \mathbf{w}_1$ will be $2^n \mathbf{w}_1$.

Computing A^n times a vector would normally be hard but is now very easy thanks to eigenvectors, because in that case, A^n times the eigenvector can be computed as an eigenvalue to the n th power times that eigenvector.



Now you know the answer to the question posed previously: What if there were such a vector? Of course, you'll need to explore—and you will, in future lectures—when you can say that such vectors exist. This is a good situation.

Even better, what if there were a second vector \mathbf{w}_2 that behaved like an eigenvector—say $A\mathbf{w}_2 = 7\mathbf{w}_2$?

And still better, what if the eigenvectors \mathbf{w}_1 and \mathbf{w}_2 formed a basis for \mathbb{R}^2 ? That would be a dream scenario—because then for any vector \mathbf{v} , computing $A^n\mathbf{v}$ would be easy.

$$\begin{aligned}\bar{\mathbf{v}} &= a\bar{\mathbf{w}}_1 + b\bar{\mathbf{w}}_2 \\ A\bar{\mathbf{v}} &= A(a\bar{\mathbf{w}}_1 + b\bar{\mathbf{w}}_2) = a2\bar{\mathbf{w}}_1 + b7\bar{\mathbf{w}}_2 \\ A^n\bar{\mathbf{v}} &= a2^n\bar{\mathbf{w}}_1 + b7^n\bar{\mathbf{w}}_2.\end{aligned}$$

If \mathbf{w}_1 and \mathbf{w}_2 formed a basis, then given any vector \mathbf{v} , you could write it as a linear combination of \mathbf{w}_1 and \mathbf{w}_2 , say $a\mathbf{w}_1 + b\mathbf{w}_2$.

Then, $A\mathbf{v}$ would be $A(a\mathbf{w}_1 + b\mathbf{w}_2)$. But by the eigenvalue property, this is just $a2\mathbf{w}_1 + b7\mathbf{w}_2$. Notice that if you were to multiply this result by A again several more times, the eigenvector property would pop out another factor 2 and another factor 7 for each multiplication by A .

Then, $A^n\mathbf{v} = a2^n\mathbf{w}_1 + b7^n\mathbf{w}_2$. This is much easier to compute than multiplying matrices! And the application of looking at the long-term behavior of foxes and rabbits depends on this kind of calculation—a matrix to the n th power times a vector.

Eigenvectors and Eigenvalues

Let's carefully define an eigenvalue and an eigenvector.

DEFINITION

Given an $n \times n$ matrix A , a number λ is an *eigenvalue* for A if

$$A\mathbf{v} = \lambda\mathbf{v}$$

for some nonzero vector \mathbf{v} .

If such a vector exists, it is called an *eigenvector* of A corresponding to λ .

Let's call the equation $A\mathbf{v} = \lambda\mathbf{v}$ the eigenvector equation.

Note a few things. First, it is traditional to notate an eigenvector using the variable \mathbf{v} , which is just a number. It could be zero or negative or positive. However, an eigenvector is prohibited to be the zero vector, because then the eigenvector property would hold for *any* matrix and *any* eigenvalue, which would be kind of silly.

In the coming lectures, you will discover how eigenvectors and eigenvalues can help you compute matrix powers quickly, provide yet another way to test if a matrix is invertible, and help you understand the linear transformation \mathbf{x} goes to $A\mathbf{x}$.

Solving the Eigenvector Equation

How do you find some vector that satisfies some equation $A\mathbf{v} = \lambda\mathbf{v}$? It may seem that you do not have enough information, especially when you realize you also don't know what λ is. Another obstacle is that the left side of the equation is a matrix times the vector \mathbf{v} and the right side is a number times the vector \mathbf{v} , so it would seem hard to manipulate this equation to isolate and solve for λ and \mathbf{v} .

The first thing you can do is to make both sides of the eigenvector equation a matrix times \mathbf{v} by simply introducing the identity matrix on the right side. You can insert it without affecting anything, because $\lambda\mathbf{v}$ equals λ times the identity matrix I times \mathbf{v} .

By grouping λI , you see that this is just a diagonal matrix with λ s on the diagonal and zeros everywhere else. The strategy now is to move everything to the left side and factor out the vector \mathbf{v} .

Doing that, you see that $[A - \lambda I]\mathbf{v} = \mathbf{0}$. In other words, because $[A - \lambda I]$ sends \mathbf{v} to the zero vector, then the vector \mathbf{v} is in the null-space of the matrix $[A - \lambda I]$.

$$\begin{aligned}[A]\bar{\mathbf{v}} &= [\lambda I]\bar{\mathbf{v}} \quad \text{for } \lambda, \bar{\mathbf{v}} \\ [A]\bar{\mathbf{v}} - [\lambda I]\bar{\mathbf{v}} &= \bar{\mathbf{0}} \\ [A - \lambda I]\bar{\mathbf{v}} &= \bar{\mathbf{0}}.\end{aligned}$$

Because \mathbf{v} was an eigenvector and eigenvectors were assumed to be nonzero, then there is a nontrivial element in the null-space of $[A - \lambda I]$. So, the nullity of $[A - \lambda I]$ is at least 1 and cannot be 0. In other words, $[A - \lambda I]\mathbf{v} = \mathbf{0}$ has more than just a trivial solution $\mathbf{v} = \mathbf{0}$. By the invertible matrix theorem, this means that the matrix $[A - \lambda I]$ is not invertible and that the determinant of $[A - \lambda I]$ is 0. So, you can solve for λ by setting the determinant of $[A - \lambda I]$ equal to 0.

$$\det [A - \lambda I] = 0.$$

Notice that now you have isolated the unknown eigenvalue λ from the unknown vector \mathbf{v} . There is no vector \mathbf{v} in this equation!

But if you take the determinant of $[A - \lambda I]$, you will find that it is a polynomial in λ , so its roots will give you the eigenvalues of A . There will be n of them, and sometimes they may be repeated roots or complex roots.

Once you've done that, then for each λ you find, you must find nonzero vectors in the null-space of $[A - \lambda I]$. These will be eigenvectors corresponding to the eigenvalue λ .

Return to Population Dynamics Application

Let's return to foxes and rabbits. You don't have enough tools to answer the question about the long-term behavior of foxes and rabbits; those tools will be built over the next few lectures. But you can at least think through the meaning of an eigenvector and an eigenvalue.

Recall that vectors for foxes and rabbits consist of the 2 population totals. If you have a special \mathbf{v} for which $A\mathbf{v} = \lambda\mathbf{v}$, then \mathbf{v} is an eigenvector, and its components represent fox/rabbit populations, which, after each successive time step, remain in the same relative proportion to each other.

All that happens is both populations get multiplied by the same factor λ . And if λ is greater than 1, then the populations are growing; if λ is less than 1, then the populations are shrinking; and if λ equals 1, then the populations are remaining constant. Thus, eigenvalues can tell you about the contraction or expansion of the eigenvector that it is associated with.

Indeed, the eigenvectors and eigenvalues of a matrix are hidden structures that are now revealing themselves, and they can convey a lot of information about a matrix and its behavior.

READINGS

Chartier, *When Life Is Linear*, chap. 8.

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*. Read the Introductory Example and section 5.1.

Poole, *Linear Algebra*, sections 4.0 and 4.1.

Lecture 15

EIGENVECTORS AND EIGENVALUES: GEOMETRY

Every matrix, which represents a linear transformation, has associated with it eigenvectors and eigenvalues, and these represent important geometric features of the transformation. It is important to think about matrix multiplication as a linear transformation on a vector space rather than some random calculation that you have to do, because if you understand the transformation, you can comprehend all at once what the transformation does.

The German prefix *eigen-* can be translated as *characteristic*, and it suggests that eigenvectors and eigenvalues characterize the matrix in some way.

The Geometry of Eigenvectors and Eigenvalues

Recall that if you have a square $n \times n$ matrix, call it A , and you can find a nonzero vector \mathbf{v} such that $A\mathbf{v}$ is just a multiple of \mathbf{v} —say, $\lambda\mathbf{v}$ —then λ is an eigenvalue of A and \mathbf{v} is an eigenvector of A corresponding to the eigenvalue λ .

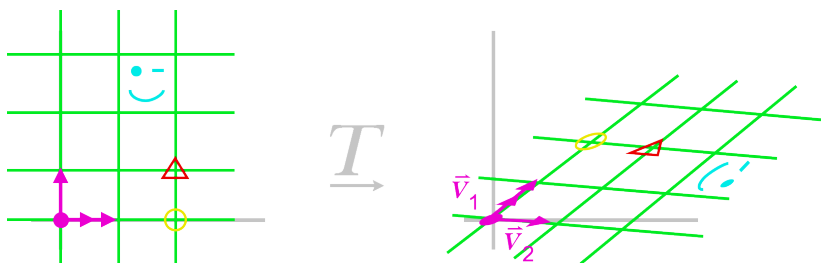
Here, the matrix A times the vector \mathbf{v} can be thought of as taking a vector and seeing how it behaves under the linear transformation $T(\mathbf{x}) = A\mathbf{x}$.

$$T(\bar{\mathbf{x}}) = A\bar{\mathbf{x}}$$

$$A = \begin{bmatrix} \bar{\mathbf{v}}_1 & \bar{\mathbf{v}}_2 \end{bmatrix}.$$

This takes the entire plane and transforms it by multiplying by A . How does it change the plane?

First, the origin goes to the origin. And the basis vectors (in red in the diagram) go to vectors that represent the columns of A .

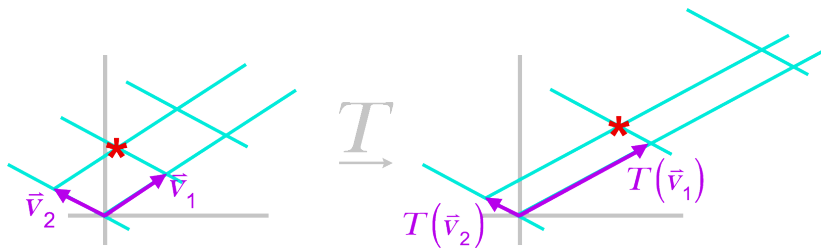


And lines get sent to lines. So, that means that a grid of squares gets deformed to a grid of parallelograms. And the winking face gets deformed to a different winking face. If the winking face is backward, it suggests that the matrix A has a negative determinant. The absolute value of the determinant is the ratio of the area of a parallelogram to the area of the original square.

Comprehending this geometrically, you can now discern where any other point should go—just figure out where it is on the grid of squares and look for the corresponding point on the grid of parallelograms.

But this grid of squares is based on the standard basis vectors, which was a rather arbitrary choice and not the most natural vectors to look at when trying to understand this transformation.

Perhaps better ones to look at would focus on the ones that don't change direction—the eigenvectors. If you could find such vectors, then the grid based on these eigenvectors wouldn't change much under the transformation; it would change scale along these directions, but the grid orientation would stay the same.



For example, vector \mathbf{v}_1 gets sent to $T(\mathbf{v}_1)$, which is a multiple of itself. So, it is an eigenvector, and the scaling factor is the eigenvalue. The eigenvector nearly doubled in size under the transformation, so the eigenvalue is about 2.

And vector \mathbf{v}_2 gets sent to $T(\mathbf{v}_2)$, pointing in the same direction, so it's an eigenvector. In this case, the eigenvalue is smaller than 1, because the vector shrank.

The nice thing here is that you can see what the transformation is doing to other vectors simply by looking at their components in the eigenvector directions. So, the red point on the left grid, which is the sum of

\mathbf{v}_1 and \mathbf{v}_2 , goes to the red point on the right grid, which is the sum $T(\mathbf{v}_1) + T(\mathbf{v}_2)$. You can easily follow a point by looking at its location on the grid and looking for the corresponding point on the image grid.

Because you are fortunate that \mathbf{v}_1 and \mathbf{v}_2 form a basis for \mathbb{R}^2 , any vector can be found on this grid; in other words, any vector can be written in terms of \mathbf{v}_1 and \mathbf{v}_2 . And once you know where the basis vectors go under a linear transformation, you know where every other vector goes just by looking at the image grid.

So, knowing the eigenvectors and eigenvalues can give you a more natural structure by which to describe the action of a linear transformation.

Verifying That a Vector Is an Eigenvector

Is the vector $(2, 5)$ an eigenvector of the 2×2 matrix A , with entries 1, 2, 5, 4?

$$\overbrace{\begin{bmatrix} 1 & 2 \\ 5 & 4 \end{bmatrix}}^A \begin{bmatrix} 2 \\ 5 \end{bmatrix} = \begin{bmatrix} 12 \\ 30 \end{bmatrix} = 6 \begin{bmatrix} 2 \\ 5 \end{bmatrix}.$$

If you take the matrix A and multiply it by $(2, 5)$, you get $(12, 30)$. That is a multiple of $(2, 5)$; in fact, it is 6 times $(2, 5)$. This means that 6 is an eigenvalue of A and $(2, 5)$ is an eigenvector of A corresponding to the eigenvalue 6.

One way to think of this is that the matrix A represents a linear transformation, which may expand, shrink, or rotate vectors, but for at least this one vector, multiplication by A scales the vector by the factor 6.

There are many other vectors that are eigenvectors corresponding to the eigenvalue 6. For example, any nonzero multiple of $(2, 5)$ is also an eigenvector with eigenvalue 6 because A times $k(2, 5)$ is just k times A times $(2, 5)$, which is k times $6(2, 5)$, which is 6 times $k(2, 5)$.

Moreover, in this example, the set of all eigenvectors corresponding to 6 (together with the zero vector) forms a subspace of \mathbb{R}^2 called the **eigenspace** of A corresponding to 6. You can check that the eigenspace satisfies the properties of subspaces. For example, if 2 eigenvectors \mathbf{v} and \mathbf{w} , with eigenvalue 6, are added together, their sum also has the eigenvalue property. You can see this by checking $A(\mathbf{v} + \mathbf{w}) = A\mathbf{v} + A\mathbf{w}$, which is $(6\mathbf{v} + 6\mathbf{w})$, which is $6(\mathbf{v} + \mathbf{w})$.

Finding Eigenvectors and Eigenvalues

To find eigenvalues and eigenvectors, you must look to solve the eigenvector equation, $A\mathbf{v} = \lambda\mathbf{v}$, for both λ and \mathbf{v} . If you move everything to one side, you get $(A\mathbf{v} - \lambda\mathbf{v})$ equals the zero vector. It is tempting to factor out the \mathbf{v} , but you can't do that because A is a matrix and λ is a real number, so they do not operate on \mathbf{v} in the same way. The key idea here is to introduce the identity matrix I as a factor. If you do that, then λI is a square matrix, just like A is a square matrix. Then, you can factor out the \mathbf{v} .

Then, the matrix $(A - \lambda I)$ times the vector \mathbf{v} equals the zero vector. λI is just a matrix with λ s down the main diagonal and zeros everywhere else.

Remember that you want to find a nonzero vector \mathbf{v} for the eigenvector, but this equation says that \mathbf{v} is in the null-space of $(A - \lambda I)$. The only way this can happen is if the null-space of $(A - \lambda I)$ is nontrivial. By the invertible matrix theorem, that is only true when $(A - \lambda I)$ is not invertible. But that means that the determinant of $(A - \lambda I)$ is zero. You must find λ for which this determinant is zero.

You now have a way to isolate the process of finding λ from the process of finding \mathbf{v} !

The first part of the method finds eigenvalues by taking the determinant of $(A - \lambda I)$ to find an expression in λ that is a polynomial in λ . In fact, if you think about the definition of the determinant, you can see this determinant will be an n th-degree polynomial for an $n \times n$ matrix A . Because you are setting this expression to zero, the eigenvalues will be the roots of this polynomial.

You can solve it by factoring the polynomial, or by using any number of other methods. These are going to be the eigenvalues of the matrix A . By a standard fact about roots of polynomials known as the fundamental theorem of algebra, there will be at most n of them that are real or complex numbers, and sometimes they will be repeated with multiplicity.

Once you have found the eigenvalues, the next step in the method is to find the eigenvectors associated to each eigenvalue. You do this for each eigenvalue λ by finding the null-space of $(A - \lambda I)$; this is called the eigenspace associated to λ and is denoted (E_λ) . The eigenvectors are the nonzero vectors in E_λ . To see why, recall that an eigenvector \mathbf{v} had to be a nonzero vector that solved this equation:

$$(A - \lambda I)\mathbf{v} = \mathbf{0}.$$

In other words, for a given eigenvalue λ , an eigenvector \mathbf{v} is any nonzero vector in the null-space of $(A - \lambda I)$.

Suppose you want to find the eigenvectors and eigenvalues of the matrix A .

$$A = \begin{bmatrix} 1 & 2 \\ 5 & 4 \end{bmatrix}.$$

The first thing you do is subtract λI from A . This yields the matrix A with λ subtracted off the diagonal entries.

$$A - \lambda I = \begin{bmatrix} 1 & 2 \\ 5 & 4 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} 1-\lambda & 2 \\ 5 & 4-\lambda \end{bmatrix}.$$

Now if you take the determinant of this matrix, you will get $(1 - \lambda)(4 - \lambda) - 10$, which factors as $(\lambda + 1)(\lambda - 6)$. If you set this to zero, you will see that λ is either -1 or 6 . These are the 2 eigenvalues of the matrix 1, 2, 5, 4.

$$\det(A - \lambda I) = (1 - \lambda)(4 - \lambda) - 10$$

$$= \lambda^2 - 5\lambda - 6 = (\lambda + 1)(\lambda - 6)$$

$$\text{set } (\lambda + 1)(\lambda - 6) = 0$$

$$\lambda = -1 \text{ or } \lambda = 6.$$

Find the eigenspace associated to 6. This is found by finding the null-space of $(A - 6I)$. That means solving $(A - 6I)$ times a vector \mathbf{v} equals the zero vector. The eigenvectors are the nonzero vectors in the null-space of $(A - 6I)$.

$$(A - 6I)\bar{\mathbf{v}} = \bar{\mathbf{0}}.$$

You know how to find a null-space: You can just form an augmented matrix with $(A - 6I)$ on the left and the column vector $(0, 0)$ on the right and solve. $(A - 6I)$ just looks like the matrix A with 6s subtracted off the diagonal.

$$\left[\begin{array}{cc|c} 1-6 & 2 & 0 \\ 5 & 4-6 & 0 \end{array} \right] \rightarrow \left[\begin{array}{cc|c} -5 & 2 & 0 \\ 5 & -2 & 0 \end{array} \right] \rightarrow \left[\begin{array}{cc|c} -5 & 2 & 0 \\ 0 & 0 & 0 \end{array} \right].$$

When you simplify this, you see the augmented matrix with rows $(-5, 2, \text{augmented with } 0)$ and $(5, -2, \text{augmented with zero})$ and notice that the second row is a multiple of the first. So, row reduction simplifies this to just one nonzero row. If the vector \mathbf{v} consists of the variables (x, y) , then this nonzero row represents the equation $(-5x + 2y) = 0$ and y is the free variable.

Now you express x and y in terms of the free variable y . Clearly, $y = y$, and x can be put in terms of y using the equation $-5x + 2y = 0$, and you find $x = (2/5)y$. So, the vector $(\mathbf{x}, \mathbf{y}) = y(2/5, 1)$, where y is freely chosen.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{2}{5}y \\ y \end{bmatrix} = y \begin{bmatrix} \frac{2}{5} \\ 1 \end{bmatrix}.$$

Thus, any multiple of $(\frac{2}{5}, 1)$ is an eigenvector of A with eigenvalue 6. In particular, $(2, 5)$ is an eigenvector of the matrix A .

$$E_6 = \text{Span} \left\{ \begin{bmatrix} \frac{2}{5} \\ 1 \end{bmatrix} \right\}.$$

What about the eigenspace associated to the eigenvalue -1 ? In this case, the matrix $(A - \lambda I)$ becomes $(A - (-1)I)$. You want the null-space of this matrix.

$$(A - (-1)I)\vec{v} = \vec{0}.$$

Once again, you can solve this by augmenting $(A - (-1)I)$ with a column of zeros. This gives the augmented matrix with rows $(2, 2)$ and $(5, 5)$ on the left and zeros on the right. The second row turns out to be a multiple of the first, so it can be row-reduced to a row of zeros.

$$\left[\begin{array}{cc|c} 1-(-1) & 2 & 0 \\ 5 & 4-(-1) & 0 \end{array} \right] \rightarrow \left[\begin{array}{cc|c} 2 & 2 & 0 \\ 5 & 5 & 0 \end{array} \right] \rightarrow \left[\begin{array}{cc|c} 2 & 2 & 0 \\ 0 & 0 & 0 \end{array} \right].$$

This shouldn't surprise you, because if the matrix $(A - (-1)I)$ is going to have a nontrivial null-space, it better not be invertible, so its rank should be less than n , so it better have a row of zeros when it is reduced.

At this point, the matrix is in row echelon form, and you see that $2x + 2y = 0$ is a relationship between x and y . Or, in other words, $x = -y$.

When you try to write all variables in terms of the free variable y , you see that $x = -y$ and $y = y$. So then, $(x, y) = y(-1, 1)$, where y is freely chosen. The -1 eigenspace is all nonzero multiples of $(-1, 1)$.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -y \\ y \end{bmatrix} = y \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$E_{-1} = \text{span} \left\{ \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right\}.$$

So, $(-1, 1)$ is an eigenvector of A , as is any multiple. You can check your work.

$$\begin{bmatrix} 1 & 2 \\ 5 & 4 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} = (-1) \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Now, if you take A times $(-1, 1)$, you get $(1, -1)$. This is indeed -1 times the vector $(-1, 1)$.

You can also check that any multiple of $(-1, 1)$, such as $(-2, 2)$, is also an eigenvector. Just try taking A times $(-2, 2)$ and you get $-1(-2, 2)$.

Matrix Powers

In the previous lecture, you saw an example in which it was useful to raise a matrix to a very high power and apply it to a given vector—in the population dynamics of foxes and rabbits. Notice that if you wanted to take the matrix A in the example and compute its effect on the eigenvectors you found, it would be very easy.

For example, what if you want to find A^{50} times the vector $(2, 5)$? Because $(2, 5)$ is an eigenvector with eigenvalue 6, the matrix A acts just like multiplication by the number 6. So, the answer is just 6^{50} times the vector $(2, 5)$.

$$\begin{bmatrix} 1 & 2 \\ 5 & 4 \end{bmatrix}^{50} \begin{bmatrix} 2 \\ 5 \end{bmatrix} = 6^{50} \begin{bmatrix} 2 \\ 5 \end{bmatrix}.$$

What if you wanted A^{50} times the vector $(0, 7)$? Unfortunately, $(0, 7)$ is not an eigenvector of the matrix A , so at first there doesn't appear to be a shortcut for computation.

However, notice that $(0, 7)$ is a linear combination of 2 eigenvectors: $(2, 5)$ and $(-2, 2)$. In fact, it's just the sum of them! So, A^{50} can be applied very easily to both parts in the sum.

$$\begin{aligned} A^{50} \begin{bmatrix} 0 \\ 7 \end{bmatrix} &= A^{50} \left(\begin{bmatrix} 2 \\ 5 \end{bmatrix} + \begin{bmatrix} -2 \\ 2 \end{bmatrix} \right) \\ &= \dots = 6^{50} \begin{bmatrix} 2 \\ 5 \end{bmatrix} + (-1)^{50} \begin{bmatrix} -2 \\ 2 \end{bmatrix} \\ &= \begin{bmatrix} 2 \times 6^{50} - 2 \\ 5 \times 6^{50} + 2 \end{bmatrix}. \end{aligned}$$

Thus, A^{50} times $(0, 7)$ is A^{50} times $((2, 5) + (-2, 2))$, which, by linearity, is just the matrix A^{50} acting on each piece in the sum. The first part is 6^{50} times $(2, 5)$, which you saw before, and the second part is $(-1)^{50}$ times $(-2, 2)$. This shows, in fact, the sum is a vector with first component $2 \times 6^{50} - 2$, and the second component is $5 \times 6^{50} + 2$.

Notice that the same kind of computation can be carried out with any other vector besides $(0, 7)$. That's because the eigenvectors $(-2, 2)$ and $(2, 5)$ form a basis of \mathbb{R}^2 . So, any other vector can be written as a linear combination. Then, linearity can be used to perform A^{50} on both pieces separately.

You've just made your understanding of the linear transformation easier by changing the description of $(0, 7)$ in the standard basis to a description in another basis—a basis of eigenvectors. In this case, $(0, 7)$ is 1 times the eigenvector $(2, 5)$ plus 1 times the eigenvector $(-2, 2)$.

This change of basis to something more convenient basically helped you see the action of A more easily, as just stretching and/or flipping along the new basis vectors.

Once you determine how the basis vectors behave, you know where everything else goes as well.

The basis of eigenvectors is hidden structure that is now revealed, and it turns out to be a better basis than the standard basis to try to understand what is going on with the matrix A .

READINGS

Fowler, "Linear Algebra for Quantum Mechanics," <http://galileo.phys.virginia.edu/classes/751.mf1i.fall02/751LinearAlgebra.pdf>.

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, section 5.2.

Poole, *Linear Algebra*, section 4.3.

DIAGONALIZABILITY

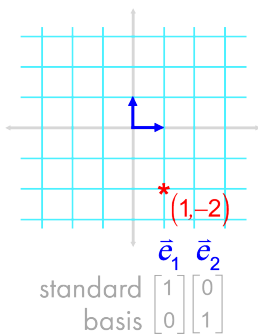
When there is a basis of eigenvectors, then it becomes relatively straightforward to compute the action of the matrix or a matrix power on an arbitrary vector. That's because when there is a basis of eigenvectors, the arbitrary vector can be written in terms of eigenvectors and the action of a matrix on an eigenvector is easy to compute. This lecture examines under what conditions you have a basis of eigenvectors, because this is a good situation to be in.

Change of Basis

Remember that a basis for \mathbb{R}^n is a set of vectors that is linearly independent and spans all of \mathbb{R}^n . And remember the key thing about a basis is that if you are given any vector in \mathbb{R}^n , you can write it as a linear combination of basis vectors in exactly one way. So, if there is only one way, then the coefficients of that linear combination are called the **coordinates** of the vector with respect to a basis.

Coordinates are just a way to tell you how much of each basis vector to use to describe a given vector as a linear combination.

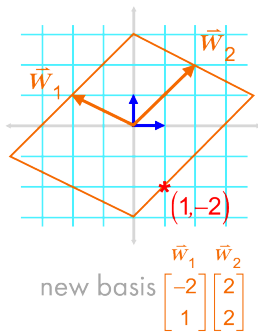
For example, the standard basis is usually used to express a vector. In the plane, the standard basis vectors are \mathbf{e}_1 and \mathbf{e}_2 . The vector \mathbf{e}_1 points right 1 unit and \mathbf{e}_2 points up 1 unit. The standard basis produces a set of horizontal and vertical grid lines. You can read the coordinates of a point from this grid.



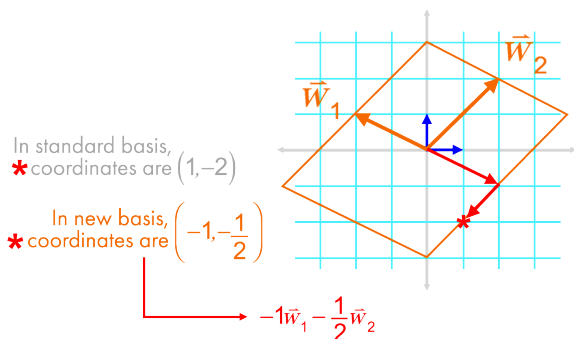
For example, the red star point is at the point $(1, -2)$ on the grid, which means you march to the right 1 unit and down 2 units. The grid is also telling you the coefficients of \mathbf{e}_1 and \mathbf{e}_2 that you need to produce a given point on the grid. So, the star point is $1\mathbf{e}_1 - 2\mathbf{e}_2$.

The standard basis is not always the best basis to use. For example, if you have a rocket ship with 2 thrusters that move in particular directions \mathbf{w}_1 and \mathbf{w}_2 , then it may be easier to express a vector in terms of \mathbf{w}_1 and \mathbf{w}_2 . How can you find the linear combination of \mathbf{w}_1 and \mathbf{w}_2 that would produce a given point, such as $(1, -2)$? In other words, how can you find the coordinates of $(1, -2)$ with respect to the new basis \mathbf{w}_1 and \mathbf{w}_2 ?

You can view this question geometrically by drawing a new grid corresponding to \mathbf{w}_1 and \mathbf{w}_2 . The grid will fill out the entire space if \mathbf{w}_1 and \mathbf{w}_2 span all of \mathbb{R}^2 . And, if you want, you can read off the coordinates of any point in the new grid, and that will be the coefficients of \mathbf{w}_1 and \mathbf{w}_2 you are looking for.



For example, look at the star point again. You saw in the standard basis that its coordinates are $(1, -2)$. In the grid from the new basis, you can get to the star point from the origin by moving in the direction $(-\vec{w}_1)$ by 1 unit and then moving in the direction $(-\vec{w}_2)$ by $\frac{1}{2}$ of a unit. This means the coordinates in the new basis are $(-1, -\frac{1}{2})$.



That's the geometric view. If you want to compute the coordinates in the new basis algebraically, then form a matrix P that contains the new basis as its columns. Then, as a linear combination of new basis vectors, the star point can be expressed as P times the new coordinates.

$$\begin{bmatrix} -2 & 2 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} -1 \\ -\frac{1}{2} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{\begin{bmatrix} 1 \\ -2 \end{bmatrix}} \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

On the other hand, if you form a matrix with the old basis as its columns, that is just the identity matrix. Then, this star point, as a linear combination of old basis vectors, must be the identity matrix times $(1, -2)$, which is just $(1, -2)$. This shows that P times the new coordinates is just the standard coordinates!

So, what you get from this discussion is the following useful way to understand the action of P , the matrix with the new basis as its columns.

If P is the matrix with the new basis as its columns, then multiplication by P (on the left) takes a vector of coordinates in the new basis to the standard coordinates. This means that multiplication by P^{-1} takes a vector of standard coordinates to coordinates in the new basis.

So, one way to view an invertible matrix like P is that it acts as a **change of basis**!

Eigenvalues and the Determinant

Recall that the process for finding eigenvalues of a square matrix A begins by subtracting from it a variable λ times the identity matrix. This is equivalent to subtracting the variable λ from each of the diagonal entries of A .

You then take the determinant of this matrix. Because of the variable λ , this expression for the determinant will be a polynomial in λ . This polynomial is called the **characteristic polynomial** of the matrix A . You can see from the recursive determinant formula that one of the terms will involve $(-\lambda)$ being multiplied n times, so the characteristic polynomial will have degree n , meaning n is the highest power of λ that appears.

The fundamental theorem of algebra says that an n th-degree polynomial always has n roots. Some of these roots may be complex, or repeated, but if you count them all, you will get n of them.

Also, these roots appear in the factorization of the polynomial. If λ is the polynomial variable and λ_1 through λ_n are the roots, then the polynomial factors as a product whose terms are $\lambda - \lambda_i$ with a $(-1)^n$ in front to account for the sign of the largest power of λ .

$$\det(A - \lambda I) = (-1)^n (\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n).$$

You can find the eigenvalues of A by setting the characteristic polynomial of A equal to zero, and this is called the **characteristic equation**.

$$\det(A - \lambda I) = 0.$$

This means that the eigenvalues are just the roots of the characteristic polynomial, so the λ_i in the factorization must be the eigenvalues of A .

On the other hand, if you look at the factored expression for the characteristic polynomial and set the variable λ equal to zero, you get the determinant of A on the left side and the product of the eigenvalues on the right side once you clear all the negative signs.

$$\det(A) = \lambda_1 \lambda_2 \cdots \lambda_n.$$

So, the **determinant of a matrix is the product of its eigenvalues!** This is an important fact that also makes intuitive sense if you recall that the eigenvalues are the expansion factors (under multiplication by A) in certain special directions, but the determinant is a kind of overall expansion factor—it's how much the volume expands under multiplication by A .

From this fact about the product of eigenvalues being the determinant, you see a new criterion for invertibility. Because a number is zero if and only if at least one of its factors is zero and a matrix having zero determinant is equivalent to the matrix being not invertible, this means that **a matrix has a zero eigenvalue if and only if it is not invertible.**

You can add this as yet another criterion to the invertible matrix theorem, which tells you when a matrix is invertible.

Algebraic Multiplicity and Geometric Multiplicity

When do you have a basis of eigenvectors? Because you want a basis in \mathbb{R}^n , you're only interested in eigenvalues that are real numbers and eigenvectors that have real entries.

The eigenvectors of a matrix play a special role as special vectors—directions that do not change under the transformation represented by that matrix. And when computing matrix powers, it is helpful to have a full basis of eigenvectors. But when is that possible?

It turns out that the answer is related to the number of repeated eigenvalues a matrix has.

Remember that the determinant of $(A - \lambda I)$ is a polynomial in λ called the characteristic polynomial of A . And eigenvalues of A are found by taking roots of the characteristic polynomial of A .

Let's define the **algebraic multiplicity** of an eigenvalue to be the number of times it appears as a root of the characteristic polynomial. So, if an eigenvalue appears twice, then its algebraic multiplicity is 2.

Note that the sum of the algebraic multiplicities has to be less than or equal to n , because the number of real eigenvalues with multiplicity is at most n but could be fewer if there are complex eigenvalues.

If the matrix A is upper triangular, then the determinant of $(A - \lambda I)$ is easy to calculate, and the characteristic polynomial has as its roots just the diagonal entries. In other words, **for upper-triangular matrices, the eigenvalues are just the diagonal entries!**

The **geometric multiplicity** of an eigenvalue λ is the dimension of E_λ , the eigenspace corresponding to λ . Then, the following fact is true: The geometric multiplicity of λ is always less than the algebraic multiplicity of λ .

Thus, the dimension of an eigenspace associated to an eigenvalue is always less than or equal to the number of times the eigenvalue is repeated.

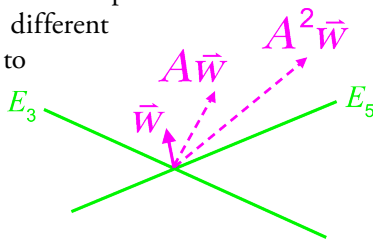
If all eigenvalues are real, then the sum of the algebraic multiplicities is n . If the geometric multiplicity of λ equals the algebraic multiplicity for every eigenvalue, and if the sum of the algebraic multiplicities is n , then the sum of the geometric multiplicities has to be n . In that case, you will have a full basis of eigenvectors—namely, just take a basis for each eigenspace.

But if the geometric multiplicity is strictly less than the algebraic multiplicity for some eigenvalue, then you are in the undesirable situation where there won't be a full basis of eigenvectors, because there won't be enough dimensions in the eigenspaces to span all of \mathbb{R}^n .

Eigenvectors from different eigenspaces must be linearly independent.

If \mathbf{v}_1 through \mathbf{v}_k are eigenvectors of some matrix corresponding to distinct eigenvalues λ_1 through λ_k , then the eigenvectors are linearly independent.

There is an easy way to make this plausible. If there were a set of eigenvectors that were not linearly independent, all from different eigenspaces, then one eigenvector, call it \mathbf{w} , would be a linear combination of the other eigenvectors. But because \mathbf{w} is a linear combination of other vectors from different eigenspaces, multiplication by A would stretch each component of \mathbf{w} from different eigenspaces by different factors. So, the vector $A\mathbf{w}$ will have to get skewed in a different direction than \mathbf{w} , so it cannot be a multiple of \mathbf{w} .



Diagonalizability

When does an $n \times n$ matrix A have a basis of eigenvectors that spans \mathbb{R}^n ?

If you have a full basis of eigenvectors (call them \mathbf{v}_1 through \mathbf{v}_n), then let P be the matrix with the eigenvectors of A as its columns.

$$P = \begin{bmatrix} | & & | \\ \bar{\mathbf{v}}_1 & \cdots & \bar{\mathbf{v}}_n \\ | & & | \end{bmatrix}$$

Let D be a diagonal matrix with eigenvalues down the diagonal and zeros everywhere else. The eigenvalues should appear on the diagonal in the same order as their corresponding eigenvectors appear in the columns of P .

$$\begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \\ & & & \lambda_n \end{bmatrix}$$

D

Then, $AP = PD$.

The reason is that this is really just expressing the eigenvector equation $A\mathbf{v} = \lambda\mathbf{v}$ for each λ simultaneously.

Then, AP is just A times the matrix with \mathbf{v}_i 's as columns. If you think about how matrix multiplication operates, that is just the matrix with $A\mathbf{v}_i$ in the i^{th} column.

$$AP = A \begin{bmatrix} | & & | \\ \bar{\mathbf{v}}_1 & \dots & \bar{\mathbf{v}}_n \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ A\bar{\mathbf{v}}_1 & \dots & A\bar{\mathbf{v}}_n \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ \lambda_1 \bar{\mathbf{v}}_1 & \dots & \lambda_n \bar{\mathbf{v}}_n \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ \bar{\mathbf{v}}_1 & \dots & \bar{\mathbf{v}}_n \\ | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_n \end{bmatrix}.$$

But then the eigenvector property shows that the i^{th} column is $\lambda_i \mathbf{v}_i$. But that can be verified to be P times the diagonal matrix D . So, indeed, $AP = PD$.

If you multiply both sides on the left by P^{-1} , you get $P^{-1}AP = D$.

Conversely, if you have this, then $AP = PD$. This means that the eigenvector equation holds for each column of A . So, the columns of P then form a full basis of eigenvectors, because P is invertible, and the entries of D are the associated eigenvalues!

THEOREM

A matrix A is diagonalizable if there exists an invertible matrix P such that $P^{-1}AP = D$, a diagonal matrix. If so, then the columns of P are the eigenvectors of A , and the diagonal matrix D contains the eigenvalues.

This theorem says that being diagonalizable is equivalent to having a basis of eigenvectors, and that's equivalent to having the geometric and algebraic multiplicities being equal and all eigenvectors being real.

One special case is that if you have a matrix with n distinct real eigenvalues (in other words, there are no repeated eigenvalues), then the matrix A is diagonalizable.

This follows by noting that the algebraic multiplicity is then less than or equal to 1, but this is always bigger than the geometric multiplicity, which is always at least 1, since eigenspaces have at least one nonzero vector in them. So, both geometric and algebraic multiplicities must be equal. So, the theorem shows that A is diagonalizable.

There's another useful way to view diagonalizability: It means that there's a change of basis that turns the action of a given matrix A into one that behaves just like a diagonal matrix with respect to the new basis. All a diagonal matrix does is scale the axes directions by various factors.

So, if you are expressing vectors with coordinates in the new basis, then the claim is that the action of D will look just like the action of first multiplying by P on the left (which changes the new basis to the standard basis), then performing A , and then changing standard basis back to new basis with P^{-1} . That's why $P^{-1}AP = D$.

Similar Matrices

There is a special name for matrices that are related by a relation like the one in the definition of diagonalizability. Two square matrices A and B are **similar** if there is an invertible matrix P such that $B = P^{-1}AP$.

If you think of the invertible matrices P and P^{-1} as changes of bases, then one way to think of similar matrices is that they represent transformations that are doing exactly the same thing, just represented in different bases.

So, then, similar matrices share some characteristics.

- ◆ Similar matrices have the same determinant.
- ◆ They have the same characteristic polynomial; therefore, they have the same eigenvalues.
- ◆ They do not have the same eigenvectors.
- ◆ They have the same rank as well as the same state of invertibility, meaning that one is invertible if and only if the other is.

Computing Matrix Powers

Matrix powers are important to compute, and diagonalizable matrices have matrix powers that are very efficient to compute!

First notice how matrix powers of similar matrices are related: If $B = P^{-1}AP$ (which means A and B are similar), then when you take the k^{th} matrix power of B , it's very easy to see how it's related to the k^{th} power of A .

$$B^k = P^{-1}APP^{-1}AP \cdots P^{-1}AP = P^{-1}A^kP.$$

If you take B and multiply it many times, the P s and P^{-1} s pair up and cancel. So, B^k is just P^{-1} times A^k times P .

For a matrix A to be diagonalizable means A is similar to a diagonal matrix D . P in this case is just the matrix whose columns are eigenvectors of A , and D consists of eigenvalues along the diagonal.

Then, the prior discussion shows that $P^{-1}A^kP = D^k$.

D^k is very easy to compute because you just raise the diagonal elements to the k^{th} power.

Then, rewrite the expression as $A^k = PD^kP^{-1}$.

READINGS

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, sections 5.3 and 5.4.

Poole, *Linear Algebra*, section 4.4.

Lecture 17

POPULATION DYNAMICS: FOXES AND RABBITS

After building up a lot of machinery to understand the eigenvectors and eigenvalues of a matrix, it's time to see some of the payoff by returning to the population model of foxes and rabbits that was introduced in lecture 14. Remember that this is just a model, which means it is not going to perfectly represent what may happen in reality. You make some simplifying assumptions and obtain some results in the hope that the results will give you some insights, even though the model isn't perfect.

Recalling the Population Dynamics Model

Let F_n represent the population of foxes at time n and R_n represent the population of rabbits at time n . Assume that the number of foxes and rabbits doesn't have to be an integer. This is unlikely to affect the character of the results, especially when the number of foxes and rabbits is large and the fractional difference between integer and noninteger values is small.

Furthermore, assume the population vector (F_{n+1}, R_{n+1}) at time $n + 1$ depends linearly on the population vector (F_n, R_n) at time n , according to the following model:

$$\begin{aligned} F_{n+1} &= 0.6F_n + 0.5 R_n \\ R_{n+1} &= -pF_n + 1.2R_n. \end{aligned}$$

Recall that the coefficient 0.6 for F_n represents the fact that foxes die without rabbits—in particular, 40% die off at each time step if there are no rabbits. The $-p$ coefficient for F_n shows that the number of rabbits is negatively influenced by the presence of foxes. In other words, foxes eat rabbits. You'll explore what happens for various values of p , called the **predation parameter**. And the 1.2 coefficient indicates that without foxes, the number of rabbits multiplies by 1.2, which means it grows by 20% each time step.

This linear system can be expressed in matrix form. If you let \mathbf{x}_n be the population vector at time n , then the time- $(n + 1)$ population vector is a matrix A times the time- n population vector. So, repeated multiplication by A gives the successive population counts as time grows.

$$\begin{bmatrix} F_{n+1} \\ R_{n+1} \end{bmatrix} = \begin{bmatrix} 0.6 & 0.5 \\ -p & 1.2 \end{bmatrix} \begin{bmatrix} F_n \\ R_n \end{bmatrix}$$

$$\vec{x}_{n+1} = A \vec{x}_n.$$

This means that if you are given the initial population vector \mathbf{x}_0 , which is (F_0, R_0) , then the time- k population vector \mathbf{x}_k is given by $A^k \mathbf{x}_0$.

What happens as the number of steps, k , goes to infinity? What is the long-term behavior of the system? Do rabbits and foxes thrive, or do they die off? And what happens if the parameter p is varied?

You'll study the cases where $p = 0.175$, 0.16 , and 0.10 , and you'll see exactly how the answers depend on the eigenvalues and eigenvectors in each instance!

To carry out the first calculation, for $p = 0.175$, you calculate the eigenvalues of A by taking A and subtracting λ off the diagonal to get $(A - \lambda I)$.

$$\begin{bmatrix} 0.6 - \lambda & 0.5 \\ -0.175 & 1.2 - \lambda \end{bmatrix} \leftarrow A - \lambda I.$$

By taking the determinant of this matrix, you'll get the characteristic polynomial:

$$\lambda^2 - 1.8\lambda + 0.8075.$$

This polynomial has 2 roots, which can be found using the quadratic formula. These are $\lambda = 0.95$ and 0.85 , and, remember, they are the eigenvalues of A . Then, you find eigenvectors associated with each eigenvalue in turn. For $\lambda = 0.95$, you plug in $\lambda = 0.95$ into the matrix $(A - \lambda I)$ to get the matrix shown at right.

$$\begin{bmatrix} -0.35 & 0.5 \\ -0.175 & 0.25 \end{bmatrix}$$

You expect this matrix not to be invertible, because the eigenvalues occur precisely where the determinant of the matrix $(A - \lambda I)$ is zero. So, it should have nontrivial null-space. And indeed it does! You could compute the null-space by solving the homogeneous system where you multiply this matrix by a vector \mathbf{x} and set it equal to zero and solve for \mathbf{x} .

But for a 2×2 , you can often see a solution by inspection. In this case, notice that one column is a multiple of the other—a sure sign the matrix is not invertible. In fact, the ratio of the first column to the second column is -7 to 10 , so if you multiply this matrix by the vector

$(10, 7)$, you get the zero vector. So, one eigenvector for the eigenvalue 0.95 is the vector $(10, 7)$.

In a similar way, you can find an eigenvector for the eigenvalue 0.85—by subtracting 0.85 off the diagonal of A and finding a vector in the null-space. If you do this, you'll find that an eigenvector is $(2, 1)$.

So, you've found the eigenvectors and eigenvalues when the predation parameter $p = 0.175$. What about predation parameters $p = 0.16$ and 0.10 ? You could do them by hand again, but you could also let software do it for you.

Here's a summary of what you find.

$p = 0.175 \Rightarrow$ e-vals of A are:

$$\lambda_1 = 0.95 \text{ w/ e-vec } \vec{v}_1 = \begin{bmatrix} 10 \\ 7 \end{bmatrix}$$

$$\lambda_2 = 0.85 \text{ w/ e-vec } \vec{v}_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$p = 0.16 \Rightarrow$ e-vals of A are:

$$\lambda_1 = 1.0 \text{ w/ e-vec } \vec{v}_1 = \begin{bmatrix} 5 \\ 4 \end{bmatrix}$$

$$\lambda_2 = 0.8 \text{ w/ e-vec } \vec{v}_2 = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

$p = 0.10 \Rightarrow$ e-vals of A are:

$$\lambda_1 = 1.1 \text{ w/ e-vec } \vec{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\lambda_2 = 0.7 \text{ w/ e-vec } \vec{v}_2 = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$$

So, the eigenvalues and eigenvectors vary as you change the predation parameter.

- ◆ When $p = 0.175$, the eigenvalues are both less than 1, and the eigenvectors are as shown.
- ◆ When you lower the predation parameter to 0.16 (so foxes eat fewer rabbits), the eigenvalues are 1 and less than 1, and the eigenvectors have changed a little, too.
- ◆ When you lower the predation parameter from 0.16 to 0.10, one of the eigenvectors now exceeds 1, and the eigenvectors have also changed a little.

These computations begin to give you clues as to what is happening. Remember that when eigenvalues are bigger than 1, that indicates an eigenvector direction that is expanding, and when the eigenvalue is less than 1, that indicates an eigenvector direction that is contracting. If the predation parameter is high, then both eigenvalues are less than 1, so this indicates that the population vector is contracting along with the special eigenvectors. This also means that the population vector is contracting everywhere else.

Let's do a computation. Suppose that the initial population $\mathbf{x}_0 = (F_0, R_0)$ can be written as some linear combination of eigenvectors: $c_1\mathbf{v}_1 + c_2\mathbf{v}_2$.

Remember that this will be possible if the matrix is diagonalizable, so there is a basis of eigenvectors. The coefficients c_1 and c_2 of the linear combination can be solved from the initial populations F_0 and R_0 , because this is just a system of equations in unknowns c_1 and c_2 .

Then, A times \mathbf{x}_0 has the effect of multiplying the \mathbf{v}_1 and \mathbf{v}_2 by their corresponding eigenvalues.

And if you keep multiplying on the left by A , the eigenvectors produce a factor of a corresponding eigenvalue each time. Then, \mathbf{x}_k , the time- k population vector, is

$$\bar{\mathbf{x}}_k = A^k \bar{\mathbf{x}}_0 = c_1 \lambda_1^k \bar{\mathbf{v}}_1 + c_2 \lambda_2^k \bar{\mathbf{v}}_2.$$

This is a very simple formula! In general, computing A^k would be much harder, especially if the matrix were large.

If A is diagonalizable, then this formula is exactly what you'd get by multiplying \mathbf{x}_0 on the left by PDP^{-1} , where P has eigenvectors of A as columns and D is a diagonal matrix with diagonal entries λ_1 and λ_2 . This is because $P^{-1}\mathbf{x}_0$ is exactly the calculation you would do to use the initial conditions to find c_1 and c_2 .

With this formula in mind, let's now consider 3 scenarios: high predation, low predation, and medium predation.

High Predation

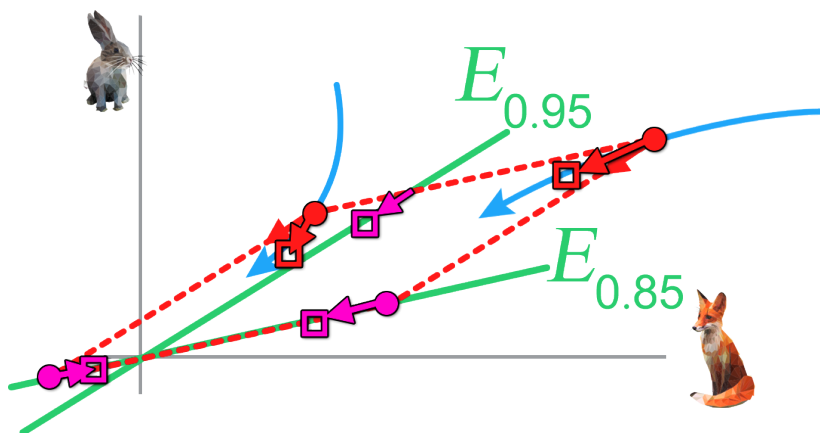
For $p = 0.175$, the above formula for the population vector becomes:

$$\begin{bmatrix} F_k \\ R_k \end{bmatrix} = c_1(0.95)^k \begin{bmatrix} 10 \\ 7 \end{bmatrix} + c_2(0.85)^k \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

Notice how the 0.95^k and 0.85^k go to 0 because both of these terms come from eigenvalues that are less than 1. This means that both terms in this expression eventually go to the zero vector.

In other words, you can interpret this as saying that as k goes to infinity, no matter what c_1 and c_2 are (remember, they come from the initial conditions), the 2 populations will approach the zero vector—meaning that both populations go extinct.

Graphically, the eigenspaces for 0.95 and 0.85 can be plotted on the fox-rabbit plane.



The 0.85 eigenspace, marked by $E_{0.85}$ in the diagram, lies along the vector direction $(2, 1)$, and any point along this eigenspace represents a population vector where the number of foxes is twice the number of rabbits. So, for a vector on this eigenspace, the population vectors will, after 1 time step, remain in that ratio of 2 to 1, but both will be multiplied by a factor of 0.85. Thus, both populations decrease by 15% in each time step.

Similarly, the 0.95 eigenspace, denoted by $E_{0.95}$ in the diagram, lies along the vector direction $(10, 7)$, so the fox-to-rabbit ratio is 10 to 7 along this eigenspace. After a single time step, the population ratio will remain the same (because it is an eigenvector), but both populations will be multiplied by a factor of 0.95. So, both populations decrease by 5% each time step.

If you start with a population vector at any other point, you can write the vector of populations as a linear combination of eigenvectors that live in the eigenspaces. The contraction will be faster along the eigenspace direction for 0.85, so the solid pink point on the 0.85 eigenspace moves toward the origin faster than the solid pink point on

the 0.95 eigenspace. The net effect is that a point that is not on either eigenspace will, as time progresses, skew toward the 0.95 eigenspace, because the component in the 0.85 eigenspace will disappear quicker. The blue curves in the diagram indicate the shape of the path that a point follows.

No matter where you start on the diagram, the population vectors eventually go to zero.

And you can see how quickly they go to zero. The eigenvalue closest to 1, which is 0.95, is the limiting factor in how slowly the populations die off, because 0.95^k goes to 0 slower than 0.85^k does.

For the predation parameter set at 0.175, you get extinction for both foxes and rabbits. It should not surprise you that if you lower the predation parameter, you may see the populations survive.

Low Predation

If you lower the predation parameter to 0.1, the eigenvalues are 1.1 and 0.7, and you calculated the eigenvectors already as (1, 1) and (5, 1), respectively.

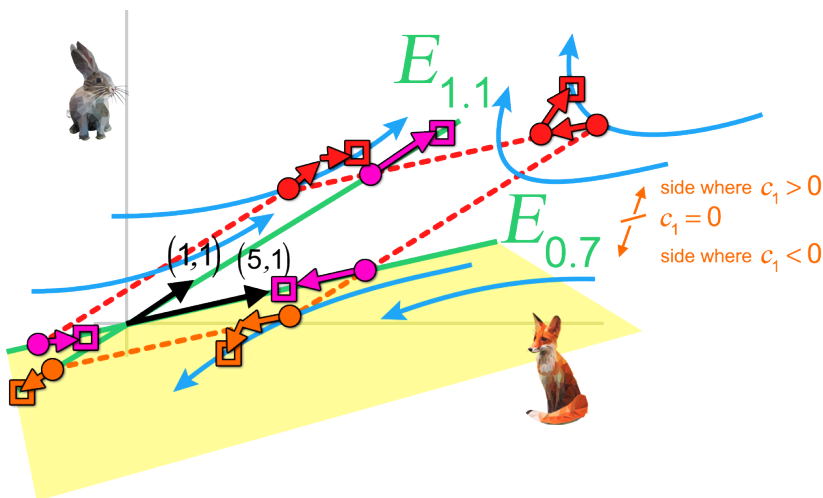
Then, using the same formula as before but with different eigenvalues and eigenvectors, you see that the time- k population vector is $c_1(1.1)^k$ times the eigenvector (1, 1) plus $c_2(0.7)^k$ times the eigenvector (5, 1).

$$\begin{bmatrix} F_k \\ R_k \end{bmatrix} = c_1(1.1)^k \begin{bmatrix} \vec{v}_1 \\ 1 \end{bmatrix} + c_2(0.7)^k \begin{bmatrix} \vec{v}_2 \\ 1 \end{bmatrix}.$$

Notice that for large k , the first term dominates—it actually blows up. The second term goes to zero as k grows. This is now saying something more interesting than when the predation parameter was large.

Remember that c_1 and c_2 are determined by the initial conditions. So, if c_2 is zero, then the initial population lies along the eigenspace for 1.1, in which the fox-to-rabbit ratio is 1 to 1 because the eigenvector is $(1, 1)$. So, if you start with a fox-to-rabbit ratio of 1 to 1, then at each time step, the population ratio stays the same, but both populations increase by 10% (as indicated by the factor 1.1). So, the fox and rabbit populations over time will grow and flourish.

But if you start with an initial condition where c_1 is 0, then the initial population lies along the 0.7 eigenspace, which means the initial ratio of foxes to rabbits is 5 to 1, because the eigenvector is $(5, 1)$. In that case, the populations after each time step will decrease by 30%, because the eigenvalue is 0.7, but the ratio will remain 5 to 1.



In this diagram, the 1.1 eigenspace is denoted by $E_{1.1}$ and the 0.7 eigenspace is denoted by $E_{0.7}$. For points in the 1.1 eigenspace, you get expansion by 10% at each time step, and for points in the 0.7 eigenspace, you get contraction by 30% at each time step.

Things get interesting when your initial populations do not lie in the eigenspaces.

As before, you'll need to take the initial vector of populations and write it as a linear combination of eigenvectors $(1, 1)$ and $(5, 1)$. Doing so produces vectors along the eigenspaces that sum to the initial vector, and the coefficients c_1 and c_2 just tell you how large these component vectors are in relation to the eigenvectors $(1, 1)$ and $(5, 1)$, shown by the black arrows along the eigenspaces.

If $c_1 > 0$, then the first term is nonzero and yields a vector of populations that is positive in each coordinate. The first will quickly become larger than the second term, if it isn't already, and when that happens, the populations will be growing by a factor of approximately 1.1 at each time step. In addition, the ratio of foxes to rabbits will be close to 1 to 1.

You can see this in the diagram, because when $c_1 = 0$, you lie on the 0.7 eigenspace, with eigenvector $(5, 1)$. So, if $c_1 > 0$, then that is all points on one side of the 0.7 eigenspace. It's the side in the direction of $(1, 1)$, so it must be the top side of the 0.7 eigenspace. This is just the set of points where the ratio of foxes to rabbits is strictly less than 5 to 1.

This kind of makes sense: If there are fewer than a certain ratio of foxes, then both populations survive and flourish!

To summarize, if $c_1 > 0$ (which corresponds to fewer than a 5-to-1 ratio of foxes to rabbits), then the populations eventually flourish in a 1-to-1 ratio of foxes to rabbits, and they grow by 10% in each time step.

In the diagram, this happens in the region between the 1.1 and 0.7 eigenspaces in the first quadrant. It also happens in the region to the left of the 1.1 eigenspace in the first quadrant. The blue curved paths show the path that the population vectors take in this region; no matter where you start, the points move toward the 1.1 eigenspace and eventually track along it.

If $c_1 < 0$, the first term will be a vector that is negative in both components—both foxes and rabbits. This corresponds to the yellow shaded region in the diagram. The first term may be outweighed by the second term if the second term is quite positive at the start, but eventually the first term will grow in absolute value and dominate, leading to one of the populations going negative. The model breaks down here because you can't have negative populations, but it suggests extinction for one of the populations.

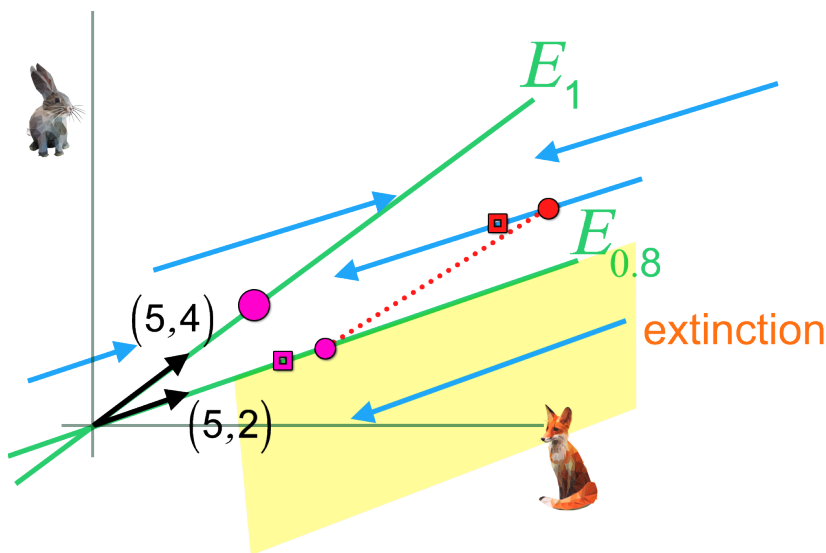
You can see which one goes extinct first by looking at the diagram. If $c_1 < 0$, this happens when you are below the 0.7 eigenspace in the diagram (shaded yellow), which is where the ratio of foxes to rabbits is greater than 5 to 1. If you start at a point in the first quadrant but below the 0.7 eigenspace, you will see that points there will move downward to the left, because they want to approach the 1.1 eigenspace and are on the negative side of the eigenspace. This leads to extinction for rabbits first, which makes sense, because in this region, there aren't enough rabbits to feed all the foxes.

So, even though the model breaks down here with negative rabbits, you learn something: The critical ratio to pay attention to is 5 foxes per rabbit. Any more than that and you have extinction of rabbits (and therefore foxes, unless they eat something else), but any less than that and both populations flourish!

Medium Predation

When the predation parameter $p = 0.16$, the eigenvalues are 1 and 0.8 with eigenvectors $(5, 4)$ and $(5, 2)$.

If you were to draw a diagram of the fox-rabbit plane, it would look a lot like the previous diagram, with eigenspaces E_1 and $E_{0.8}$ slightly shifted from the prior examples, lying along the $(5, 4)$ and $(5, 2)$ vector directions.



The key insight is that the eigenvalue 1 is not expanding nor contracting. So, if you take any initial population vector and write it in terms of components in the eigenspace directions, then any component in the $(5, 4)$ direction stays put, with no change at each time step. Any component in the $(5, 2)$ direction will contract by a factor of 0.8 at each time step. So, that component eventually vanishes and everything moves toward the eigenspace corresponding to the eigenvalue 1.

This means that if the component of the initial population vector in the $(5, 4)$ direction is positive, eventually the population vector will converge to a point on the E_1 eigenspace. Look at all the blue paths above the $E_{0.8}$ eigenspace. So, the long-term behavior is that the populations neither grow nor go extinct but approach stable populations in a 5-to-4 rabbit-to-fox ratio!

On the other hand, if the component of the initial population vector in the $(5, 4)$ direction is negative, the population vector will converge to a point on the E_1 eigenspace below the x-axis. The model breaks down there because populations can't be negative, but if you follow the blue path in the yellow shaded region below the $E_{0.8}$ eigenspace, you will find that rabbits will go extinct first. That happens in the yellow shaded region.

So, the crucial ratio here is the ratio of foxes to rabbits along the $E_{0.8}$ eigenspace, which is 5 to 2. If you have more than 5 foxes to 2 rabbits, then the rabbits go extinct; if you have less than 5 foxes to 2 rabbits, the populations converge to a 5-to-4 ratio and remain stable.

READINGS

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, section 5.6.

Poole, *Linear Algebra*, section 4.6.

Lecture 18

DIFFERENTIAL EQUATIONS:
NEW APPLICATIONS

Calculus helps explain rates of change of functions over time. If you have a function $f(t)$ that depends on time t , the rate of change of the function with respect to time is a new function that also depends on time. It is called the **derivative** of f and is usually written as $f'(t)$. When you have a physical problem that involves a rate of change of a function, it places a condition on that function's derivative. This is a **differential equation**. The solution to a differential equation is a function that satisfies the given condition.

Just for this lecture, it will be helpful if you know a little calculus. But even if you don't know any calculus, you may be able to follow the general thread of this lecture.

Solving a System of Differential Equations

If $x(t)$ is a real-valued function that satisfies $x' = ax$ for some constant a , then a solution is $x(t) = Ce^{at}$, where C is some constant.

This exponential solution is the key to solving complicated systems of differential equations of the form $\mathbf{x}' = A\mathbf{x}$. To keep things simple, let's assume that A is a 2×2 matrix and $\mathbf{x}(t)$ is the vector $(x_1(t), x_2(t))$, but everything you're going to learn will generalize readily to any square matrix, so you could study a system of n linear differential equations if you wanted to.

How can you solve the equation $\mathbf{x}' = A\mathbf{x}$? For what matrices would this system of equations be easy to solve? If you think about it, you may realize that if A is diagonal, then you can do it very easily.

In this example of a 2×2 diagonal matrix with λ_1 and λ_2 on the diagonal, the 2 equations decouple. The function x_1' will have no relationship to x_2 , and x_2' will have no relation to x_1 . And each of them is a differential equation in which the rate of change of a function is proportional to the function itself, so the solution is an exponential.

If $x_1' = \lambda_1 x_1$, then $x_1(t) = C_1 e^{\lambda_1 t}$. The exponent has the rate λ_1 in it that comes from the proportionality constant in the differential equation.

Similarly, $x_2' = \lambda_2 x_2$, then $x_2(t) = C_2 e^{\lambda_2 t}$.

$$\bar{\mathbf{x}}' = A\bar{\mathbf{x}}$$

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\begin{aligned} x_1' &= \lambda_1 x_1 \Rightarrow x_1(t) = c_1 e^{\lambda_1 t} \\ x_2' &= \lambda_2 x_2 \Rightarrow x_2(t) = c_2 e^{\lambda_2 t} \end{aligned}$$

The constants C_1 and C_2 are arbitrary; they just express the idea that many functions will work.

You can rewrite these solutions as a vector equation.

$$\bar{\mathbf{x}}(t) = c_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{\lambda_1 t} + c_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} e^{\lambda_2 t}.$$

That's a very general solution to this system of differential equations when the matrix is diagonal. Diagonalization in this case corresponds to decoupling the equations from each other.

What if the matrix is not diagonal?

Recall that diagonalization offered a good basis of eigenvectors, and this basis was a more natural way to understand the action of the linear transformation that A represents. In other words, it was a good change of basis. In this case, you'll also perform a helpful change of basis to turn the problem into the diagonal case that was just solved!

If A is diagonalizable, then you know $A = PDP^{-1}$ for some diagonal matrix D and invertible matrix P . So, $\mathbf{x}' = A\mathbf{x}$ becomes $\mathbf{x}' = PDP^{-1}\mathbf{x}$.

If you multiply both sides by P^{-1} on the left, you get $P^{-1}\mathbf{x}' = DP^{-1}\mathbf{x}$.

Now a good substitution becomes apparent. If you set the vector $\mathbf{w} = P^{-1}\mathbf{x}$, then that will be convenient. Because P is a matrix of constants, taking the derivative of \mathbf{w}' is just $P^{-1}\mathbf{x}'$. But this change of variable now turns the differential equation into one where the left side becomes \mathbf{w}' and the right

$$\bar{\mathbf{x}}' = A\bar{\mathbf{x}}$$

$$\bar{\mathbf{x}}' = PDP^{-1}\bar{\mathbf{x}}$$

$$P^{-1}\bar{\mathbf{x}}' = DP^{-1}\bar{\mathbf{x}}$$

$$\bar{\mathbf{w}}' = D\bar{\mathbf{w}}.$$

side becomes $D\mathbf{w}$. So, \mathbf{w}' is a diagonal matrix D times \mathbf{w} . You know how to solve this!

The previous method shows that \mathbf{w} equals the following expression, where λ_1 and λ_2 are eigenvalues of A and c_1 and c_2 are arbitrary constants.

$$\bar{\mathbf{w}} = c_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{\lambda_1 t} + c_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} e^{\lambda_2 t}.$$

So, you've solved for \mathbf{w} ! But you weren't interested in \mathbf{w} ; you were interested in solving for the vector \mathbf{x} .

But because $\mathbf{w} = P^{-1}\mathbf{x}$, then $\mathbf{x} = P\mathbf{w}$. Now notice that \mathbf{w} is a vector written as a linear combination of vectors, so when you multiply on the left by P , you can just multiply P times each term. Also remember that P has the eigenvectors of A in its columns. So, when you multiply P times the vector $(1, 0)$, you'll get the first column of P . When you multiply P times the vector $(0, 1)$, you'll get the second column of P . These are just the eigenvectors of A .

$$\bar{\mathbf{x}} = P\bar{\mathbf{w}} = c_1 \bar{\mathbf{v}}_1 e^{\lambda_1 t} + c_2 \bar{\mathbf{v}}_2 e^{\lambda_2 t}.$$

So, the vector \mathbf{x} is just a linear combination of eigenvectors times exponential functions related to the eigenvalues! This is a very general kind of solution when you have a diagonalizable matrix A . For larger square matrices, you would just have more terms and you'd use a basis of eigenvectors.

Complex Eigenvalues

Let's examine the behavior of matrices with complex eigenvalues. What kinds of linear transformations are they? For example, in a 2×2 matrix with no real roots, that means the matrix has no special directions in the plane that stay the same under multiplication by the matrix A . What kind of linear transformation is that?

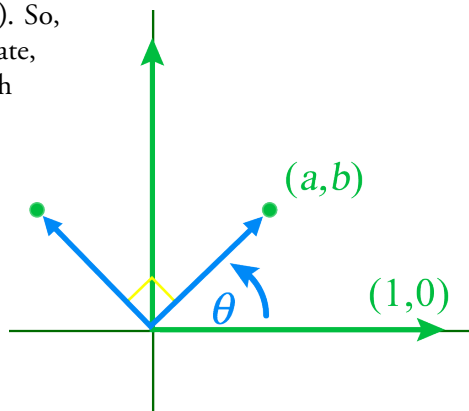
A rotation is such an example. Recall that a 2×2 rotation matrix is always of this form:

$$\begin{bmatrix} r \cos \theta & -r \sin \theta \\ r \sin \theta & r \cos \theta \end{bmatrix}.$$

If you let $a = r \cos \theta$ and $b = r \sin \theta$, then you could recognize a rotation as being of the form at right, too.

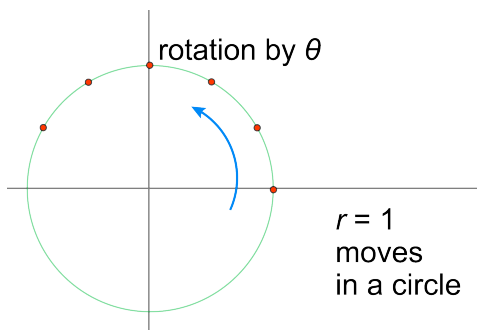
$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$$

Note that the action of this matrix is to send the standard basis vector $(1, 0)$ to (a, b) and $(0, 1)$ to $(-b, a)$. So, you can see that it does rotate, and if you calculate the length of the vector (a, b) , you get r , so it does send the unit vectors to vectors of length r . Thus, it scales by r .

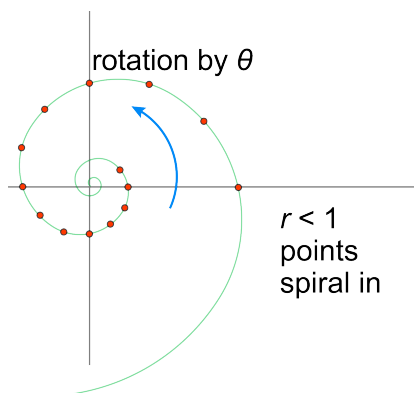


So, if you start at some point and apply the matrix over and over again, 3 things can happen, depending on whether the scaling factor r is 1, less than 1, or greater than 1.

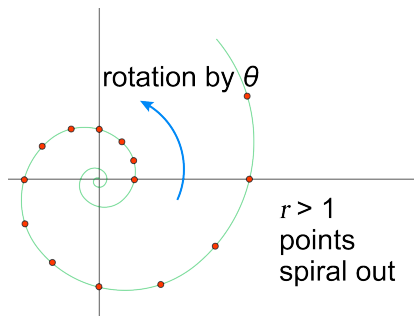
- ♦ If $r = 1$, you get a series of dots that move along a circle. Each time, the angle from the origin changes the same amount.



- ♦ If $r < 1$, you get a series of dots that spirals in toward the center. As before, the angle from the origin in each case changes the same amount.



- ♦ If $r > 1$, you get a series of dots that spirals away from the center. As before, the angle from the origin in each case changes the same amount.



Now let's determine the eigenvalues of the matrix. If you subtract λ off the diagonal and take the determinant, you'll get the characteristic polynomial:

$$(a - \lambda)^2 + b^2.$$

The roots turn out to be $(a \pm bi)$. These are the 2 eigenvalues; they are both complex. It is a fact that complex roots of any polynomial always come in pairs. You can use the same methods you've learned to find their eigenvectors, which will also be complex and come in pairs; you can learn more about this in an advanced linear algebra course.

So, you've seen that scaled rotations are matrices with complex eigenvalues. Are there any others? It turns out, in some sense, that there aren't. There's a theorem that says all other 2×2 matrices with complex eigenvalues are actually like scaled rotations if you change basis! In other words, if A is a 2×2 matrix with complex eigenvalues, then there's a basis with respect to which the action of A is just scaled rotation in that basis.

This means that 2×2 matrices with complex eigenvalues really correspond to scaled rotations in the plane, warped along certain axes. You now have a geometric picture of what complex eigenvalues mean.

READINGS

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, sections 5.5 and 5.7.

Poole, *Linear Algebra*, section 4.6.

QUIZ FOR LECTURES 13–18

- 1 Find the determinant of the matrix shown at right using any method discussed in this lecture. Are the rows linearly independent? [\[LECTURE 13\]](#)

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 5 \\ 4 & 8 & 7 \end{bmatrix}$$
- 2 In this problem, you'll see why adding a multiple of one row to another does not change the determinant of a 2×2 matrix A . [\[LECTURE 13\]](#)
 - a Verify that if the 2 rows of A are identical, the determinant of A is zero.
 - b Show that the determinant formula D , as a function of the 2 row vectors \mathbf{u} and \mathbf{w} , is linear in each argument—for example, verify that $D(\mathbf{u} + \mathbf{v}, \mathbf{w}) = D(\mathbf{u}, \mathbf{w}) + D(\mathbf{v}, \mathbf{w})$ and $D(c\mathbf{u}, \mathbf{w}) = cD(\mathbf{u}, \mathbf{w})$.
 - c Confirm that $D(\mathbf{u} + c\mathbf{w}, \mathbf{w}) = D(\mathbf{u}, \mathbf{w})$, as desired.
- 3 Let A be the diagonal matrix shown at right. The linear transformation $T(\mathbf{x}) = A\mathbf{x}$ scales the x coordinate by a factor of -2 and the y coordinate by a factor of 3 , because it sends $\mathbf{x} = (x, y)$ to $T(\mathbf{x}) = (-2x, 3y)$. Thinking about the geometric description of this transformation, locate 2 eigenvectors with different eigenvalues. [\[LECTURE 14\]](#)

$$\begin{bmatrix} -2 & 0 \\ 0 & 3 \end{bmatrix}$$
- 4 Let B be the matrix shown at right. The linear transformation $T(\mathbf{x}) = B\mathbf{x}$ takes each vector and reflects it about the line $y = x$. Thinking about the geometric description of this transformation, locate 2 eigenvectors with different eigenvalues. [\[LECTURE 14\]](#)

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$
- 5 Find all eigenvalues of $A = \begin{bmatrix} 1 & -1 \\ -6 & 0 \end{bmatrix}$, as well as the corresponding eigenspaces. [\[LECTURE 15\]](#)
- 6 Find all eigenvalues of $A = \begin{bmatrix} 3 & -1 \\ 0 & 3 \end{bmatrix}$, as well as the corresponding eigenspaces. [\[LECTURE 15\]](#)

- 7 Is $\begin{bmatrix} 1 & -1 \\ -6 & 0 \end{bmatrix}$ diagonalizable? Is $\begin{bmatrix} 3 & -1 \\ 0 & 3 \end{bmatrix}$ diagonalizable? [LECTURE 16]
- 8 If A and B are similar matrices, show that they have the same determinant. Then show that $A - \lambda I$ and $B - \lambda I$ are similar, which implies that A and B have the same characteristic polynomial. [LECTURE 16]
- 9 In the low predation example of the model explored in this lecture (when $p = 0.10$), give an intuitive explanation for why the trajectories that start close to the 0.7 eigenspace (but just above it) seem to first move left but then change direction and move right and upward. [LECTURE 17]
- 10 In each of the 3 scenarios of the model, which eigenvalue would you say was the most important eigenvalue for understanding the dynamics of the system: the smaller one or the larger one? [LECTURE 17]
- 11 Consider the system of differential equations given by

$$\begin{aligned}x' &= x - y \\ y' &= -6x\end{aligned}$$

and solve the system for $\mathbf{w}(t) = (x(t), y(t))$ using the methods of lecture 18. [LECTURE 18]

- 12 You've seen that the matrix representing a rotation has complex eigenvalues. Explain why you expect the eigenvectors to be complex as well (in other words, not all of its entries are real). [LECTURE 18]

Solutions can
be found on
page 301.

ORTHOGONALITY: SQUARING THINGS UP

Linear transformations on the plane map parallelograms to parallelograms. While they preserve parallel lines, they might not preserve angles or distances between points, so they can warp space and change volumes. But there are linear transformations that do preserve angles and volumes and distances between points. Such linear transformations are called **orthogonal**.

Orthogonal Sets

Recall that the standard basis vectors are pairwise perpendicular and are of length 1. So, if the columns of a matrix tell you where the standard basis vectors get sent under the associated linear transformation, then if the transformation preserves lengths and angles, you should expect the columns of the matrix to remain pairwise perpendicular, and you expect those vectors to be length 1 as well.

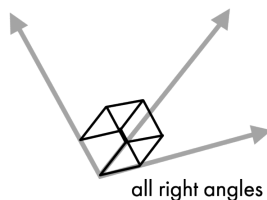
DEFINITION

A set of nonzero vectors is orthogonal if all pairs of vectors in the set are perpendicular. Or, in other words, if the vectors are \mathbf{v}_1 through \mathbf{v}_n , then for all pairs i and j where i is not j , the dot product of \mathbf{v}_i and \mathbf{v}_j equals zero. (Recall that the dot product of 2 nonzero vectors is zero if and only if the 2 vectors are perpendicular.)

Orthogonal sets are often used as reference frames to measure other things by. In particular, it's nice when a basis is orthogonal. And it's even more convenient when the vectors have length 1. In that case, the set of vectors is called **orthonormal**. An orthogonal set can be turned into an orthonormal set by simply normalizing each vector.

An orthogonal set might look something like this set of vectors in \mathbb{R}^3 . Each vector is at right angles to all the others.

And if they were all length 1, they would be orthonormal.



Because all the vectors are pairwise perpendicular, it can be shown that an orthogonal set is linearly independent.

THEOREM

An orthogonal set is linearly independent.

Given a vector in a subspace S , how can it be expressed as a linear combination in a given basis for that subspace? Normally, you'd have to solve a system of equations to find the coefficients for the combination. But if the basis is orthogonal, then finding the coefficients is easy.

If the orthogonal basis is \mathbf{v}_1 through \mathbf{v}_k and you want \mathbf{w} to equal a linear combination

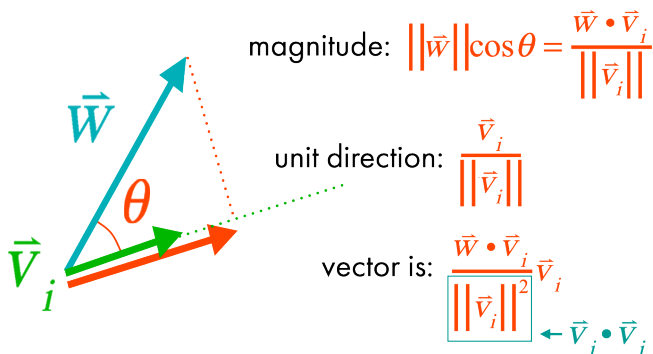
$$c_1\mathbf{v}_1 + \dots + c_k\mathbf{v}_k,$$

you can just write

$$c_i = \frac{(\mathbf{w} \cdot \mathbf{v}_i)}{(\mathbf{v}_i \cdot \mathbf{v}_i)}.$$

This is the coefficient for the component of \mathbf{w} in the \mathbf{v}_i direction.

The intuition behind this is that if you project the vector \mathbf{w} in the \mathbf{v}_i direction, you get a vector whose magnitude is the length of \mathbf{w} times the cosine of the angle in between, which by the geometric definition of the dot product is just $(\mathbf{w} \cdot \mathbf{v}_i)$ divided by the length of \mathbf{v}_i . This vector points in the direction of the unit vector \mathbf{v}_i divided by the length of \mathbf{v}_i .



So, the vector is the product of this magnitude times the unit direction. That's $(\mathbf{w} \cdot \mathbf{v}_i)$ times the vector \mathbf{v}_i , divided by the square of the length \mathbf{v}_i .

This is the vector \mathbf{v}_i times a coefficient $\frac{(\mathbf{w} \cdot \mathbf{v}_i)}{(\mathbf{v}_i \cdot \mathbf{v}_i)}$.

The representation of \mathbf{w} in terms of \mathbf{v}_i is unique, because if you take the dot product with \mathbf{v}_i , all the terms vanish by orthogonality except the i^{th} term so that $(\mathbf{w} \cdot \mathbf{v}_i)$ is indeed $c_i(\mathbf{v}_i \cdot \mathbf{v}_i)$.

Division then shows $c_i = \frac{(\mathbf{w} \cdot \mathbf{v}_i)}{(\mathbf{v}_i \cdot \mathbf{v}_i)}$.

And if the set of vectors is orthonormal, then the representation is even easier—the denominator of this fraction is just 1.

THEOREM

If \mathbf{q}_1 through \mathbf{q}_k is an orthonormal basis for some subspace S and \mathbf{w} is chosen from S , then \mathbf{w} is a linear combination of the \mathbf{q}_i 's with each coefficient $(\mathbf{w} \cdot \mathbf{q}_i)$, and this representation is unique.

This is why having an orthonormal basis is great: You can figure out the coordinates in the basis just by doing dot products!

If you put the orthonormal basis as columns of a matrix Q , then Q^T times Q must be the $n \times n$ identity matrix! This is because the ij^{th} entry of $(Q^T \text{ times } Q)$ is the i^{th} row of Q^T dotted with the j^{th} column of Q .

This dot product is 1 whenever i equals j and 0 otherwise.

Orthogonal Matrices

DEFINITION

If Q is square—meaning the number of elements of the orthonormal set is equal to n , the dimension of the vectors—then the orthonormal set spans all of \mathbb{R}^n and Q is called an **orthogonal matrix**.

(This is a strange name, because you might think such a matrix should be called orthonormal, but this is the convention that has historically been used. Just remember that an orthogonal matrix has orthonormal columns and is square.)

For an orthogonal matrix, taking the inverse is very easy, because Q^{-1} is just Q^T .

A good example of an orthogonal matrix is a rotation matrix. You can check that Q^T times Q is the identity matrix.

$$Q^{-1} = Q^T$$

$$Q = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$$

$$Q^T Q = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

But intuitively, it should make sense that the matrix is orthogonal, because the motivation for orthogonal matrices was looking for linear transformations that preserve angles and volumes and distances between points, and rotations do that!

Another example of an orthogonal matrix is the identity matrix, but because the order of the columns shouldn't matter, then any permutation of the columns of the identity matrix should be orthogonal as well.

Such a matrix is called a **permutation matrix**, and its actions on \mathbb{R}^n just swap the roles of the axes of n -dimensional space.

Properties of Orthogonal Matrices

Let Q be an $n \times n$ orthogonal matrix.

- 1 $Q\mathbf{x} \cdot Q\mathbf{y}$ is $\mathbf{x} \cdot \mathbf{y}$ for any pair of vectors \mathbf{x} and \mathbf{y} in \mathbb{R}^n . This follows because $Q\mathbf{x} \cdot Q\mathbf{y}$ is, as matrices, $(Q\mathbf{x})^T$ times $Q\mathbf{y}$, which is \mathbf{x}^T times Q^T times Q times \mathbf{y} . Because Q^T times Q is the identity, this product is $\mathbf{x} \cdot \mathbf{y}$. In other words, an orthogonal transformation preserves dot products (and therefore also angles).
- 2 The dot product preserves lengths so that the length of $Q\mathbf{x}$ equals the length of \mathbf{x} for any vector \mathbf{x} in \mathbb{R}^n . This follows from noticing that the length $Q\mathbf{x}$ is the square root of $(Q\mathbf{x} \text{ dotted with } Q\mathbf{x})$, which is the square root of $(\mathbf{x} \cdot \mathbf{x})$, using the first property. And that is just the length of \mathbf{x} .

Each of these first 2 properties is, in fact, equivalent to Q being orthogonal. In other words, both properties imply that the matrix Q is orthogonal.

- 3 The rows of Q form an orthonormal set, like the columns do. This is because the rows of Q are columns of Q^T , and Q^T inverts Q . So, Q times Q^T is the identity matrix, which is the same as saying Q^{TT} times Q^T is the identity matrix. That shows that Q^T is orthogonal.

- 4 Q^{-1} is an orthogonal matrix, because Q^{-1} is Q^T .
- 5 The determinant of an orthogonal matrix Q is ± 1 . This follows from noting that Q^T times Q equals the identity. So, the determinant of Q^T times the determinant of Q must be 1. But the determinant of Q^T is just the same as the determinant of Q . So, the determinant of Q , squared, is 1. This means that the determinant of Q must be ± 1 . If the determinant of Q is -1 , then the negative determinant is a clue that the action of Q does some kind of reflection without changing distances or angles.
- 6 If λ is an eigenvalue of Q , whether real or complex, the absolute value of λ is 1. You can see this from the eigenvalue property $Q\mathbf{v} = \lambda\mathbf{v}$ for some nonzero eigenvector \mathbf{v} . Taking lengths, you see that the length of \mathbf{v} equals the length of $Q\mathbf{v}$ (by property 2), and that is the length of $\lambda\mathbf{v}$, which is the length of λ times the length of \mathbf{v} . Taken together, this means that the length of λ must be 1. So, the eigenvalues of an orthogonal matrix must lie on the unit circle in the complex plane.
- 7 If Q_1 and Q_2 are both orthogonal matrices, then their product, Q_1Q_2 , is. That can be seen in a number of ways; you can do this as an exercise.

Taken together, these properties should give you a pretty good feel for how orthogonal matrices behave. Because they preserve dot products and lengths, they basically behave like rigid motions, possibly together with a reflection of n -dimensional space. They will be reflections if the determinant is -1 .

The Gram-Schmidt Process

Given \mathbf{v}_1 through \mathbf{v}_n as a basis for a subspace S , you will construct a new orthogonal basis by sequentially modifying each \mathbf{v}_i to form a new vector \mathbf{w}_i in such a way that the \mathbf{w}_i is length 1 and perpendicular to all the \mathbf{w} vectors before it.

The first step is to take \mathbf{w}_1 to be \mathbf{v}_1 .

$$\vec{\mathbf{w}}_1 = \vec{\mathbf{v}}_1.$$

Then, take \mathbf{w}_2 to be \mathbf{v}_2 , except that you will subtract off the component of \mathbf{v}_2 in the \mathbf{w}_1 direction.

This means subtracting off the \mathbf{w}_1 vector times the fraction $\frac{(\mathbf{v}_2 \cdot \mathbf{w}_1)}{(\mathbf{w}_1 \cdot \mathbf{w}_1)}$.

$$\vec{\mathbf{w}}_2 = \vec{\mathbf{v}}_2 - \frac{\vec{\mathbf{v}}_2 \cdot \vec{\mathbf{w}}_1}{\vec{\mathbf{w}}_1 \cdot \vec{\mathbf{w}}_1} \vec{\mathbf{w}}_1.$$

You can check that \mathbf{w}_2 is perpendicular to \mathbf{w}_1 .

You continue in this fashion with \mathbf{w}_3 , starting by modifying \mathbf{v}_3 by subtracting off the components of \mathbf{v}_3 in the \mathbf{w}_1 and \mathbf{w}_2 directions. You will get an expression like the following one.

$$\vec{\mathbf{w}}_3 = \vec{\mathbf{v}}_3 - \frac{\vec{\mathbf{v}}_3 \cdot \vec{\mathbf{w}}_1}{\vec{\mathbf{w}}_1 \cdot \vec{\mathbf{w}}_1} \vec{\mathbf{w}}_1 - \frac{\vec{\mathbf{v}}_3 \cdot \vec{\mathbf{w}}_2}{\vec{\mathbf{w}}_2 \cdot \vec{\mathbf{w}}_2} \vec{\mathbf{w}}_2.$$

You can check that \mathbf{w}_3 is, in fact, perpendicular to both \mathbf{w}_1 and \mathbf{w}_2 .

If you keep doing this process sequentially, you will get an orthogonal set. Then, you can normalize each vector \mathbf{w}_i by dividing by its length to get a unit vector \mathbf{q}_i . Then, you will have an orthonormal basis \mathbf{q}_1 through \mathbf{q}_k .

At every step, you've constructed the first k vectors of the orthonormal basis to have the same span as the first k vectors \mathbf{v}_1 through \mathbf{v}_k .

This fact helps you see the QR -factorization of a matrix.

QR-Factorization

It turns out that any $m \times n$ matrix A with linearly independent columns has a special factorization into the product of Q and R , where Q is an $m \times n$ matrix with orthonormal columns and R is an $n \times n$ invertible upper-triangular matrix.

$$A_{m \times n} = Q_{m \times n} R_{n \times n}.$$

orthonormal columns ↑ ↓ invertible upper-triangular matrix

To find Q , you can use the Gram-Schmidt process on the columns of A . Because they are linearly independent, Gram-Schmidt will work. And if you let $R = Q^T A$, then you will see that QR equals $QQ^T A$, which is just the identity times A , which is A .

The matrix R obtained in this way must be upper triangular, because if you look at how Gram-Schmidt works, then \mathbf{v}_1 can be expressed entirely in terms of \mathbf{q}_1 ; \mathbf{v}_2 can be expressed entirely in terms of \mathbf{q}_1 and \mathbf{q}_2 ; and in general, \mathbf{v}_k can be expressed entirely in terms of the vectors \mathbf{q}_1 through \mathbf{q}_k .

$$\bar{\mathbf{v}}_1 = r_{11} \bar{\mathbf{q}}_1$$

$$\bar{\mathbf{v}}_2 = r_{12} \bar{\mathbf{q}}_1 + r_{22} \bar{\mathbf{q}}_2$$

$$\bar{\mathbf{v}}_k = r_{1k} \bar{\mathbf{q}}_1 + \dots + r_{kk} \bar{\mathbf{q}}_k.$$

The diagonal elements of R must all be nonzero, because you constructed \mathbf{q}_k as a normalized version of \mathbf{w}_k , and \mathbf{w}_k was defined in terms of \mathbf{v}_k and \mathbf{w}_1 through \mathbf{w}_{k-1} .

$$\begin{matrix} \left[\bar{\mathbf{v}}_1 \dots \bar{\mathbf{v}}_k \right] \\ A \end{matrix} = \begin{matrix} \left[\bar{\mathbf{q}}_1 \dots \bar{\mathbf{q}}_k \right] \\ Q \end{matrix} \begin{matrix} \begin{bmatrix} r_{11} & \cdots & r_{1k} \\ & \ddots & \vdots \\ 0 & & r_{kk} \end{bmatrix} \\ R \end{matrix}.$$

So, the coefficient r_{kk} must be nonzero.

If you require the diagonal elements of R to be positive, then the QR -factorization is unique.

Factorizations tell you something about the structure of a matrix transformation. The matrix R here is basically playing the role of Gram-Schmidt in taking a set of linearly independent vectors and squaring it up and scaling each dimension to form vectors that are orthonormal.

Orthogonal Diagonalization

Recall that a square matrix A is diagonalizable if $A = PDP^{-1}$, where P is invertible and D is diagonal. P 's columns must then be eigenvectors of A , and the diagonal elements of D must be the associated eigenvalues in the same order. One way to think about the action of P is that it is like changing bases from the standard basis to some other basis.

But sometimes that change of basis can be done nicely—by an orthogonal matrix. Recall that for orthogonal matrices, the columns are an orthonormal set and the inverse is the transpose.

DEFINITION

A matrix A is **orthogonally diagonalizable** if in addition to being diagonalizable, P is required to be orthogonal so that $A = PDP^T$.

THE SPECTRAL THEOREM

A matrix is orthogonally diagonalizable if and only if it is real and symmetric.

A result known as the spectral theorem guarantees that orthogonally diagonalizable matrices are precisely the same as the symmetric matrices. Remember, a symmetric matrix means $A = A^T$, or the entries are symmetric when reflected around the main diagonal.

READINGS

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, sections 6.4 (and any necessary background from 6.1–6.3 to understand it), 7.1, and 7.2.

Poole, *Linear Algebra*, chap. 5, especially sections 5.0, 5.1, 5.3, and 5.4.

MARKOV CHAINS: HOPPING AROUND

Previous lectures have demonstrated how important it is to understand matrix powers. For example, in the model used in lectures 14 and 17 regarding the dynamics of foxes and rabbits, the population vector at time $(n + 1)$ was a linear transformation applied to the time- n population vector. The populations evolved by repeated application of a matrix. This lecture will show you another common scenario that involves repeated application of a matrix. In this case, the coefficients of the linear equations come from probabilities, and the matrix acts on the right rather than on the left.

Markov Chains

DEFINITION

A **Markov chain** is a process with a finite number of states. It moves from state to state. At each step, the next state only depends on the current state—not on past states. The state of the process evolves according to some probabilities of transition.

If p_{ij} is the probability of going from i to j at any step of the process, then the transition matrix A is a matrix whose entries are the transition probabilities p_{ij} .

Note that a consequence of the definition of a Markov chain is that each row of the transition matrix must sum to 1, because each row contains all the probabilities of transitioning out of some particular state in the system.

Let $x_i(n)$ be the probability of being in state i at time n . Then, let vector $\mathbf{x}(n)$ be a row vector that collects these probabilities together for time n . Then, the vector \mathbf{x} at time n equals the vector \mathbf{x} at time 0 times the matrix A^n . Note that the vector of probabilities evolves linearly according to multiplication by a matrix on the right.

Economic Mobility

Suppose you want to model economic mobility. For the sake of the model, suppose there are 3 classes of people as measured by wealth: rich, middle-class, and poor. The people in each category may move around, so you might look at the probabilities of shifting from one class to another and consider what is going to happen to the distribution of people in the long run.

If you know their current wealth, you might assume their future wealth only depends on how much money they have now, not on their status at any other moment in the past. So, this can be modeled by a Markov chain.

Let's measure time in decades and let's suppose the transition probabilities are given by the hypothetical transition matrix shown at right.

The rows represent the current state, and the columns represent the state being transitioned to.

	rich	middle class	poor
rich	0.7	0.3	0
middle class	0.02	0.78	0.2
poor	0.01	0.29	0.7

- ◆ The probabilities in the first row suggest that if you're rich, the probability of staying rich in the next decade is quite high. The probability of becoming poor is zero.
- ◆ The second row is the row where you start out in the middle class. Becoming rich is highly unlikely (2%), but the probability of staying in the middle class is very high, and there's a 20% chance of becoming poor.
- ◆ The third row shows that if you're poor, the probability of remaining poor is high, and there's a negligible chance of becoming rich in the next decade.

What happens in the long term?

If the transition matrix is A , then the matrix power A^n gives the n -step probabilities of going from one state to another. In particular, the ij^{th} entry is the probability of going from state i to state j in exactly n steps.

A^5 looks like this:

$$\begin{bmatrix} 0.44513 & 0.27261 & 0.28226 \\ 0.45906 & 0.28246 & 0.25848 \\ 0.4115 & 0.24883 & 0.33967 \end{bmatrix}.$$

A^{10} looks like this:

$$\begin{bmatrix} 0.439435 & 0.268583 & 0.291982 \\ 0.440372 & 0.269246 & 0.290382 \\ 0.437173 & 0.266984 & 0.295843 \end{bmatrix}.$$

And A^{100} looks like this:

$$\begin{bmatrix} 0.439024 & 0.268293 & 0.292683 \\ 0.439024 & 0.268293 & 0.292683 \\ 0.439024 & 0.268293 & 0.292683 \end{bmatrix}.$$

After 100 steps, it's clear that the rows are converging to the same distribution. In other words, no matter what the distribution of rich/middle class/poor looks like when you start, the distribution of people is going to look the same in the long run: 44% rich, 27% middle class, and 29% poor.

Theorems about Markov Chains

In the example, the Markov chain appears to converge to a limiting distribution, called a **steady-state vector**.

The property of a steady-state vector is that when the transition matrix is applied, it remains the same. So, if \mathbf{w} is a steady-state vector, then $\mathbf{w} = \mathbf{w}A$.

Notice that this looks like an eigenvector property, except that the vector is on the left of the matrix A and it is a row vector. In fact, this is what's called a **left eigenvector** of A .

DEFINITION

A row vector \mathbf{w} is a left eigenvector of A if $\mathbf{w}A = \lambda\mathbf{w}$.

Remember that the definition of a usual eigenvector is $A\mathbf{v} = \lambda\mathbf{v}$.

The usual eigenvector is sometimes called a **right eigenvector**, but whenever you hear the word *eigenvector* without reference to left or right, then it refers to the usual (right) eigenvector.

In a similar fashion, **left eigenspace** and **left eigenvalue** can be defined.

THEOREM

\mathbf{w} is a left eigenvector and λ is a left eigenvalue of A if and only if \mathbf{w}^T is a (right) eigenvector and λ is a right eigenvalue of A^T .

Every Markov chain has a steady-state vector.

THEOREM

A and A^T have the same eigenvalues.

What are the eigenvalues of A ?

A has eigenvalue 1 because the rows of A must sum to 1, because they are a complete set of probabilities of transitioning out of some particular state. Multiplying A by the vector of all 1s will sum these rows. But that means A times the vector of all 1s must be the vector of all 1s. So, 1 is an eigenvalue of A .

This means that A^T also has an eigenvalue of 1. So, there must be an eigenvector \mathbf{v} such that $A^T \mathbf{v} = \mathbf{v}$. You can scale \mathbf{v} so that its entries sum to 1 and it will still be an eigenvector. Then, notice that the row vector \mathbf{v}^T satisfies

$$\mathbf{v}^T A = \mathbf{v}^T.$$

This means that $\mathbf{w} = \mathbf{v}^T$ is the desired steady-state vector.

Some Markov chains have multiple steady-state vectors.

For example, if a Markov chain has a transition matrix that is the identity matrix, then every vector is a steady-state vector, because clearly $\mathbf{w} = \mathbf{w}I$ for any row vector \mathbf{w} .

With the identity as the transition matrix, this Markov chain stays put for all time: If you start in a state, you never move.

What if you require the process to move around more? Would there necessarily be a unique steady-state vector w ?

Let's call a matrix positive if all its entries are positive. Let's call a Markov chain **regular** if some power of the transition matrix is positive. This means that after some number of steps, it is possible to transition from any state to any other state with positive probability. Then, the Perron-Frobenius theorem shows that the steady-state vector is unique and every starting configuration will converge to it!

THE PERRON-FROBENIUS THEOREM

If A is the transition matrix of a regular Markov chain, then

- 1 1 is an eigenvalue of A .
- 2 The left eigenspace of A associated to the eigenvalue 1 is 1-dimensional.
- 3 There's a left eigenvector in that eigenspace with all positive coordinates.
- 4 If λ is any other eigenvalue, then the absolute value of λ is strictly less than 1 (even for complex eigenvalues).

What does this have to do with convergence of Markov chains?

Remember that the matrix A acts on left eigenspaces by multiplying on the right, just as the matrix A^T acts on its right eigenspaces by multiplication on the left.

So, the action of A on a left eigenspace scales everything in that eigenspace by a factor, just as it does for right eigenspaces. The left eigenspace associated to 1 stays fixed under action by A . Because that eigenspace is a 1-dimensional line, every eigenvector in that space is a multiple of the one with all positive coordinates. Then, there is exactly one of those multiples whose coordinates sum to 1. That must be the steady-state vector of A , and it is unique. There is no other steady-state vector!

The fact that all other eigenvalues are strictly less than 1 (even complex eigenvalues) suggests that any initial row vector will, through repeated multiplication by A , converge to the steady-state vector. This is most obvious in the case where A is diagonalizable, because there's a basis of eigenvectors and they're all contracting except the steady-state vector. So, any starting vector must contract toward the eigenspace for 1. But the theorem is true even if A is not diagonalizable.

This explains why as n goes to infinity, A^n will converge to a matrix whose rows are identical and equal to a steady-state vector \mathbf{w} —because the i^{th} row of A^n is just what you get when you multiply the standard row basis vector $(\mathbf{e}_i)^T$ by A^n . If any starting vector converges to the steady state, then we expect each row of A^n to converge to the steady-state vector as well.

There are some interesting things you can try once you know about Markov chains. For example, think about the sequence of letters you encounter as you read. That's a Markov chain on maybe 40 characters (if you allow numbers, spaces, and punctuation).

Andrey Markov himself first discussed his chains in this context—by looking at a poem by Aleksandr Pushkin and measuring the transition probabilities of Cyrillic letters.

Take a text from an author you like and train your Markov chain on that data by empirically measuring the transition probabilities that a certain letter will be followed by another letter. Then, if you run the Markov chain using those probabilities, you'll end up with some fun sentences that are nonsensical but may look very much like English.

READINGS

Note: Some texts define the transition matrix of a Markov chain to be the transpose of the way it is defined here, and if so, the theorems need to change in a corresponding way.

Chartier, *When Life Is Linear*, chaps. 10 and 11.

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, chap. 10, especially sections 10.1 and 10.2.

Poole, *Linear Algebra*, sections 2.5 and 4.6.

Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," <https://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/tutorial%20on%20hmm%20and%20applications.pdf>.

von Hilgers and Langville, "The Five Greatest Applications of Markov Chains," <http://langvillea.people.cofc.edu/MCapps7.pdf>.

Lecture 21

MULTIVARIABLE CALCULUS: DERIVATIVE MATRIX

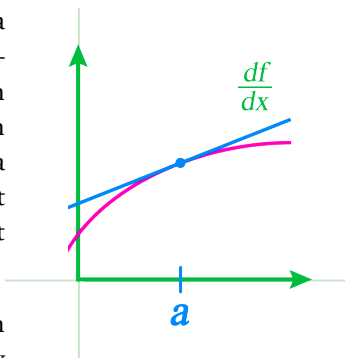
Linear algebra is an important part of multivariable calculus, also sometimes called vector calculus. Multivariable functions highlight a big theme of linear algebra: approximating nonlinear things by linear things. That idea is the foundational idea behind all the applications of the derivative matrix.

This lecture assumes you've done single-variable calculus and are familiar with those ideas but does not assume you've had multivariable calculus.

Single-Variable Calculus

Single-variable calculus studies functions of one variable; there's one input and one output.

The first big idea of calculus is the idea of an instantaneous rate of change—usually the rate of change of a function f near a point a . The derivative is often thought of as the slope of a graph near a point a . It may or may not exist, but if it does, this is called the **derivative** of f at a , written df/dx .



The derivative df/dx exists at a if the graph is **locally linear**, meaning if you magnify that region of the graph more and more, it looks more and more like a straight line. That should make sense intuitively, because then you have a well-defined notion of the slope of the graph near that point. The derivative is a limit that basically expresses this idea.

So, a function is **differentiable**—meaning it has a derivative—if the graph near the point can be approximated by a tangent line.

Another way to think of a derivative is that it is an expansion factor that results from a small change in x producing some change in f . The change in f , written Δf , is approximately the derivative times the change in x , written Δx .

$$\Delta f \approx \frac{df}{dx} \Delta x.$$

This idea supplies you with a way to approximate a function f near a point a . Namely, if x is a point you want information about, you can replace Δx with $(x - a)$ and Δf with $(f(x) - f(a))$ to get

$$f(x) - f(a) \approx \frac{df}{dx} (x - a).$$

From this, you see that $f(x)$ is approximately $f(a) + \frac{df}{dx}(x - a)$, an expression that shows you how to approximate the value of $f(x)$ knowing the value of $f(a)$ and the derivative $\frac{df}{dx}$ evaluated at a . The right side here is the equation of the tangent line to f at the point a .

$$f(x) \approx \underbrace{f(a) + \frac{df}{dx}(x - a)}_{\text{tangent line}}.$$

You've just approximated a function by something that's nice and easy to understand—a line. That's one big moral of calculus: to approximate something crazy by something simpler.

This moral appears in multivariable calculus, too; you just have to decide what you mean by a *nice* function.

Multivariable Functions

In multivariable calculus, the functions you work with can have more than one input variable and more than one output variable. There could even be different numbers of input and output variables.

$$f: \mathbb{R}^n \rightarrow \mathbb{R}^m \text{ maps } (x_1, \dots, x_n) \rightarrow (g_1, \dots, g_m).$$

$\begin{matrix} g_1(x_1, \dots, x_n) & g_m(x_1, \dots, x_n) \\ \downarrow & \downarrow \\ \uparrow & \uparrow \\ \text{component functions} \end{matrix}$

You might express such a function as a function f from \mathbb{R}^n to \mathbb{R}^m that takes n input variables (x_1 through x_n) to m output functions (g_1 through g_m). Each of these g_i 's are called **component functions**. Notice that the component functions are really functions of the input variables (see the green expressions), but to save space, they are usually omitted in writing.

Here's an example of a multivariable function: Suppose you had, for each position (x, y) on a sheet of paper, 3 quantities of interest. That could be a function like this:

$$f(x, y) = (x^2 \sqrt{y}, \sin x, 1).$$

This function takes a 2-dimensional vector to a 3-dimensional vector, so f is a function from \mathbb{R}^2 to \mathbb{R}^3 . You see that each component may depend on some or none of the input variables.

Another example of a multivariable function is one you've seen and studied in detail: a function that multiplies a vector by a matrix A .

$$\begin{array}{ccc} f(\bar{x}) & = & A\bar{x}. \\ \uparrow & \begin{array}{c} m \times n \\ \mathbb{R}^n \end{array} \rightarrow & \mathbb{R}^m \end{array}$$

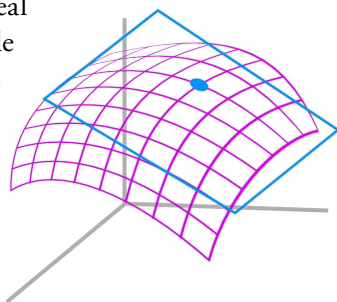
Here, the function $f(\mathbf{x})$ is $A\mathbf{x}$, where \mathbf{x} is in \mathbb{R}^n and A is an $m \times n$ matrix. Thus, $A\mathbf{x}$ is a vector in \mathbb{R}^m . This function is a linear transformation from \mathbb{R}^n to \mathbb{R}^m . As you've seen throughout this course, all linear transformations can be expressed as matrix multiplication, and you can think of linear transformations as the “nicest” examples of multivariable functions.

So, just like in calculus, when you try to approximate functions near a point by assuming they behave like lines, in multivariable calculus, you will assume that a function near a point behaves like a linear transformation!

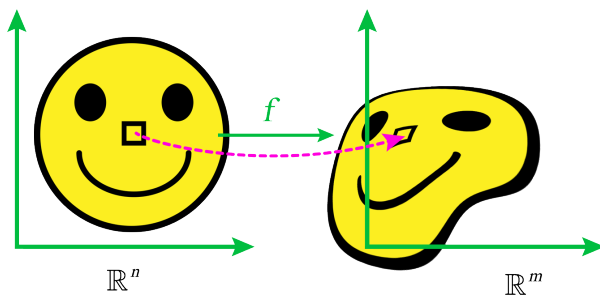
Differentiability

Just like in single-variable calculus, where a function is differentiable (meaning a derivative exists) when there is a good line that approximates it, for multivariable functions, a function will be differentiable when it has a good linear approximation.

For example, if you are looking at a function f from \mathbb{R}^2 to \mathbb{R} , then it takes in 2 real numbers x and y and spits out a single number $f(x, y)$. That means the graph of this function can be represented by a surface over the xy -plane. To say the function is differentiable at a point means that if you magnify the view of the graph near that point, it looks more and more like a plane—in other words, it is **locally planar**.



For functions from \mathbb{R}^n to \mathbb{R}^m , differentiability means the function is **locally linear**, like the following picture. Notice the function is not linear because it warps the face in a crazy fashion, but it would be locally linear if on a small scale it takes little parallelograms to little parallelograms. The square nose in this face seems to be doing this—it gets sent to a parallelogram.



The Derivative

How should the derivative of a multivariable function from \mathbb{R}^n to \mathbb{R}^m be defined? What kind of object should it be?

Let's take a cue from single-variable calculus, where a little change in f is approximately the derivative df/dx times a little change in x .

$$\Delta f \approx \frac{df}{dx} \Delta x.$$

Multivariable functions use a similar expression, except Δf is going to be a change in the output, so it's a vector in \mathbb{R}^m , and $\Delta \mathbf{x}$ will be a small change in input, so it will be a vector in \mathbb{R}^n . So, you need the derivative, whatever it is, to turn the vector $\Delta \mathbf{x}$ into the vector Δf .

The only thing that will take a vector in \mathbb{R}^n to a vector in \mathbb{R}^m in a linear way is a linear transformation. And it's represented by a matrix—in this case, an $m \times n$ matrix so that it acts on an n -dimensional vector and produces an m -dimensional vector. This matrix is playing the role that the derivative did in single-variable calculus!

So, the derivative is a matrix!

This matrix is called the **derivative matrix** of f , and it's notated Df .

$$\Delta f \approx [Df] \Delta \vec{x}.$$

\nwarrow in \mathbb{R}^m \nwarrow in \mathbb{R}^n

So, if f goes from \mathbb{R}^n to \mathbb{R}^m , the derivative matrix must be an $m \times n$ matrix, because the act of multiplying by Df needs to send \mathbb{R}^n to \mathbb{R}^m .

So, both the original function and its derivative take \mathbb{R}^n to \mathbb{R}^m .

If you replace Δf with $(f(\mathbf{x}) - f(\mathbf{a}))$ and $\Delta \mathbf{x}$ with $(\mathbf{x} - \mathbf{a})$, then you can approximate $f(\mathbf{x})$ knowing $f(\mathbf{a})$ and the derivative matrix at \mathbf{a} .

$$f(\bar{\mathbf{x}}) - f(\bar{\mathbf{a}}) \approx [Df](\bar{\mathbf{x}} - \bar{\mathbf{a}}).$$

Rewriting, you get this.

$$f(\bar{\mathbf{x}}) \approx f(\bar{\mathbf{a}}) + \underbrace{[Df](\bar{\mathbf{x}} - \bar{\mathbf{a}})}_{\text{linear adjustment}}.$$

So, you see in this formula how a small change in moving from the point \mathbf{a} to the point \mathbf{x} produces a linear adjustment to the value $f(\mathbf{a})$ —just like in single-variable calculus.

What are the entries of the derivative matrix?

This turns out to be very simple: The entry in the i^{th} row and j^{th} column is just the partial derivative of f_i with respect to x_j . So, you know what the matrix is!

$$Df = \begin{bmatrix} \vdots & & \\ \dots & \frac{\partial f_i}{\partial x_j} & \dots \\ \vdots & & \end{bmatrix} \begin{matrix} i^{\text{th}} \text{ row} \\ \\ j^{\text{th}} \text{ col} \end{matrix}$$

A **partial derivative** is just the usual derivative if you pretend all the other input variables are constant.

Chain Rule

Once you realize the derivative of a multivariable function is a matrix, then many other things begin to fall into place as analogies of the single-variable case.

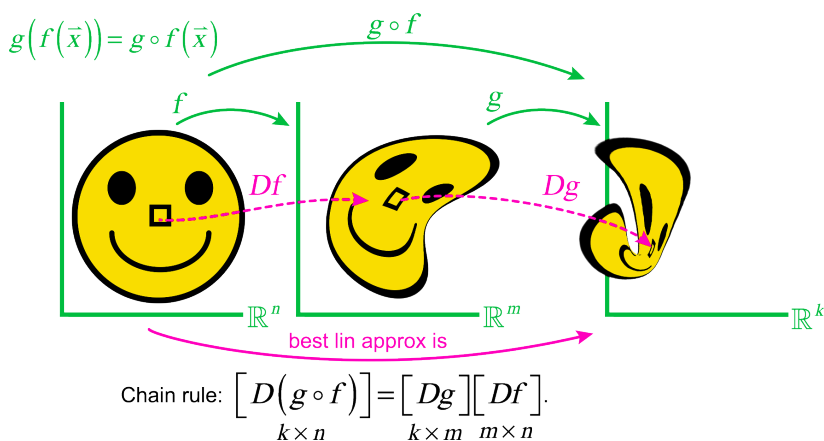
For example, the chain rule in calculus says that supposing f is a function of some variable x and x is a function of some variable t , how does f change if you change t ? This is the rate of change of a composition of functions $f(x(t))$.

It turns out you just multiply the derivatives so that $\frac{df}{dt}$ is $\frac{df}{dx}$ times $\frac{dx}{dt}$.

$$\frac{df}{dt} = \frac{df}{dx} \cdot \frac{dx}{dt}.$$

For multivariable functions, something similar happens. Suppose you have a function f going from \mathbb{R}^n to \mathbb{R}^m and a function g going from \mathbb{R}^m to \mathbb{R}^k . Then, if \mathbf{x} is a point in \mathbb{R}^n , how does $g(f(\mathbf{x}))$ depend on \mathbf{x} ?

The function $g(f(\mathbf{x}))$ is the **composition** of g and f , and it's notated $(g \circ f)$.



Just like in the single-variable case, you multiply the derivatives of g and f , which in this case are matrices. It should be plausible that you do this, because the determinants of these matrices describe the approximate scaling factor of these transformations and then the scaling factors should multiply if you first do f and then g . The dimensions of the multiplication make sense here, too, because the product of a $k \times m$ matrix and an $m \times n$ matrix is a $k \times n$ matrix.

Notice how the language of matrices makes it very easy to describe the chain rule, and conceptually you have a much clearer picture about what's going on, too!

READINGS

Colley, *Vector Calculus*.

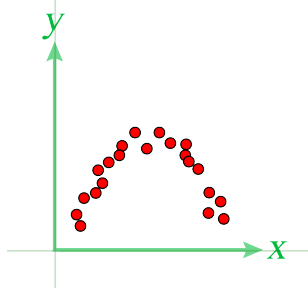
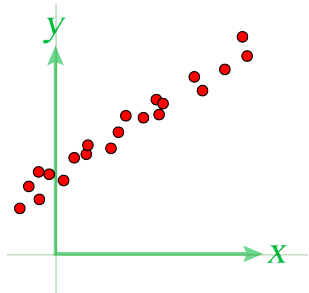
Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, section 7.3.

Poole, *Linear Algebra*, section 5.5 (especially the sections on Quadratic Forms and Graphing Quadratic Equations).

MULTILINEAR REGRESSION: LEAST SQUARES

Given a bunch of data, an important thing you might want to do is figure out how the data are related. For example, suppose you have 2 variables X and Y and you want to know if there is some relationship between X and Y . Perhaps you think Y depends on X or that knowing X helps you predict Y . For example, do SAT scores depend on family income? To try to answer a question like this, you might gather some data.

Suppose you have a collection of 100 data points for X and Y , given by (x_i, y_i) , $i = 1$ to 100. If the data looks like the graph at right, you might infer that there is some kind of approximately linear relationship between X and Y , subject to a bit of error.



But if you saw a picture like the one at left, you might be more inclined to think the relationship between Y and X is not linear—maybe it's quadratic.

Linear Regression

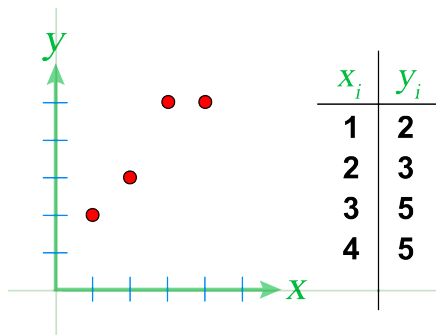
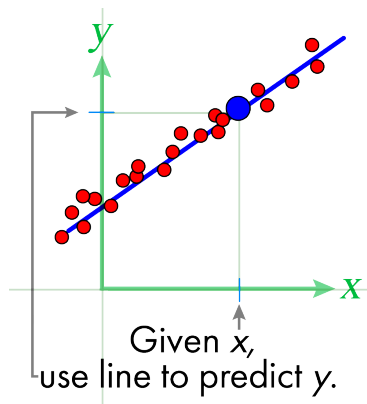
If the relationship between X and Y appears to be linear, the next thing you might do is consider what the best-fitting line is, because if you have such a line, then for any value of X , you can use the line to make a prediction about Y .

Given n data points (x_i, y_i) , i going from 1 to n , what could be meant by the *best-fitting line*?

Suppose you have 4 data points and are interested in the best-fitting line for these 4 points, sometimes called the **regression line**. The data are given in this table and plotted on the XY -plane.

Now suppose the equation for the best-fitting line is given by $Y = mX + b$. If the data fit the line perfectly, then you expect for each data point that $y_i = mx_i + b$.

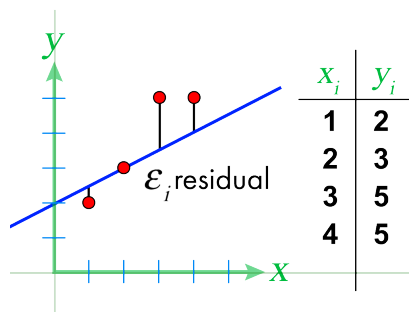
But in reality, the points don't lie perfectly on the line. Maybe they lie off the line a little. In that case, you expect $y_i = mx_i + b + \varepsilon_i$.



Notice this model assumes that the error, if any, lies only in the y direction, not in the x direction.

Here, ϵ_i represents the i^{th} **error**, also called the i^{th} **residual**. It says how far the data is from the line in the y direction.

Each data point throws off a residual, and you therefore get a vector of residuals, one for every data point.



How can the residuals be minimized? And what does that mean?

If you wrote the equations for all 4 data points simultaneously, you could put them in a matrix equation, as follows.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ x_4 & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{bmatrix}$$

\vec{Y} = X $\vec{\beta}$ + $\vec{\epsilon}$
 observation design parameter residual
 vector matrix vector vector

← minimize $\vec{\epsilon} \cdot \vec{\epsilon}$

Here, the vector \mathbf{Y} consists of all the y_i 's, and it is sometimes called the **observation vector** because it consists of the observed values of the y variable for given x values.

The matrix X is the 4×2 matrix whose rows are x_i and 1. It is called the **design matrix**.

The column vector (m, b) is called the **parameter matrix** and is written as $\boldsymbol{\beta}$, which is what you will vary to try to minimize the residuals. Notice how changing $\boldsymbol{\beta}$ will change the best-fitting line.

The design matrix may change depending on the model you choose.

And then the vector of residuals is called the **residual vector** and is written as $\boldsymbol{\varepsilon}$. This is the vector you want to minimize.

There are many ways you could do this, but the most natural thing to do is to minimize its length as a vector. This is the same as minimizing the square of the vector's length, which happens to be the dot product of the vector with itself, which in this case is the sum of the squares of all the individual residuals. It is for this reason that the best approximation is called the least squares approximation.

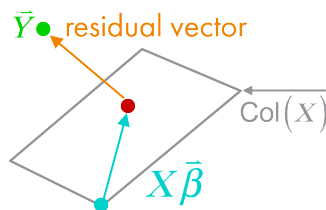
How can you minimize the length of the residual vector?

First note from the equation $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ that $\boldsymbol{\varepsilon} = \mathbf{Y} - X\boldsymbol{\beta}$, so you want to minimize $\mathbf{Y} - X\boldsymbol{\beta}$.

One way to think about this minimization is that y is a point in \mathbb{R}^4 . The residual vector is the vector between y and the point $X\boldsymbol{\beta}$. So, as $\boldsymbol{\beta}$ varies, you are trying to find $\boldsymbol{\beta}$ so that the point $X\boldsymbol{\beta}$ is closest to y . That particular $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}$, which is the estimate of the beta that accomplishes this minimization.

The point $X\hat{\beta}$, as a matrix product, lies in the column space of X , in a 2-dimensional subspace of \mathbb{R}^4 as long as there are at least 2 different x values. That's a plane in \mathbb{R}^4 . So, you are searching for a point in that plane that is closest to the point y in \mathbb{R}^4 . It will give the linear combination of the columns that you seek, which will give you the estimated parameter vector $\hat{\beta}$.

You could use multivariable calculus to find this point, but there is an easier way. At such a point, the residual vector should be perpendicular to the subspace. And because the subspace is the column space of X , this means the residual vector should be perpendicular to all the columns of X .



One way to demand that the residual vector is perpendicular to all the columns of X is to demand that it be perpendicular to all the rows of X^T . So, X^T times the residual vector $(Y - X\hat{\beta})$ should equal the zero vector.

Multiplying it out and moving one term to the other side, you get

$$X^T X \hat{\beta} = X^T Y.$$

This system of equations is called the **normal equations** for β . This always has at least one solution for β , and you can see this geometrically because there is always a closest point on a subspace to a given point. The only question is whether there might be multiple solutions.

It might depend on whether X^T times X —sometimes called the **Gram matrix**—is invertible. If it is invertible, you can solve for β as follows.

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

This is a nice theoretical solution, but in practice, it is usually easier to just solve the normal equations by Gaussian elimination rather than try to compute the inverse of X^T times X when it exists. You would then use Gaussian elimination on the augmented matrix with the Gram matrix on one side and $X^T Y$ on the other side, and this would enable you to solve for the vector $\hat{\beta}$.

Solving the normal equations can sometimes be unreliable because small errors in Gaussian elimination can lead to big errors in the solution. But if you have a QR -factorization of X , you can compute β in a more numerically stable way.

Recall from lecture 19 that an $m \times n$ matrix with linearly independent columns can be factored as Q times R , where Q is an $m \times n$ matrix with orthonormal columns and R is an invertible upper-triangular matrix.

$$A_{m \times n} = Q_{m \times n} R_{n \times n}.$$

orthonormal columns \nearrow \nwarrow invertible upper-triangular matrix

Suppose $X = QR$. Then, take the normal equations and rewrite them, substituting QR for X .

Eventually, you get $R\beta = Q^T Y$, which can be solved for β using Gaussian elimination and is less likely to exhibit the kinds of problems the normal equations did. If you want, you can multiply by R^{-1} on the left to get

$$\beta = R^{-1} Q^T Y.$$

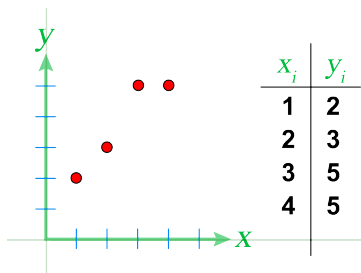
$$\begin{aligned} X^T X \bar{\beta} &= X^T \bar{Y} \\ (QR)^T QR \bar{\beta} &= (QR)^T \bar{Y} \\ R^T Q^T QR \bar{\beta} &= R^T Q^T \bar{Y} \\ R^T R \bar{\beta} &= R^T Q^T \bar{Y} \\ R \bar{\beta} &= Q^T \bar{Y} \\ \bar{\beta} &= R^{-1} Q^T \bar{Y}. \end{aligned}$$

This is a very simple formula to do least squares regression if you have a QR-factorization of X (which exists when the columns of X are linearly independent).

What does this mean for the example involving 4 points?

In this case, the observation vector \mathbf{Y} is $(2, 3, 5, 5)$.

The design matrix X is the 4×2 matrix with columns $(1, 2, 3, 4)$ and $(1, 1, 1, 1)$. You are searching for the parameter vector $\boldsymbol{\beta}$ with entries (m, b) that minimizes the length of the residual vector $\boldsymbol{\varepsilon}$.



If you compute the Gram matrix $X^T X$, you get the 2×2 matrix whose entries are 30, 10, 10, and 4. And if you compute $X^T \mathbf{Y}$, you get $(43, 15)$.

$$\begin{bmatrix} 2 \\ 3 \\ 5 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

$\bar{\mathbf{Y}} \qquad \qquad X \qquad \qquad \bar{\boldsymbol{\beta}} \qquad \qquad \bar{\boldsymbol{\varepsilon}}$

$$X^T X = \begin{bmatrix} 30 & 10 \\ 10 & 4 \end{bmatrix}$$

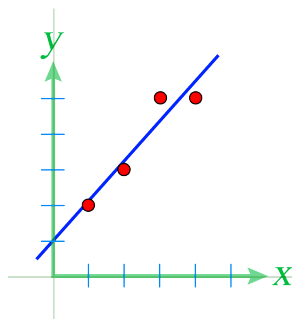
$$X^T \bar{\mathbf{Y}} = \begin{bmatrix} 43 \\ 15 \end{bmatrix}.$$

If you take the inverse of $X^T X$, you get a matrix whose entries are 0.2, -0.5, -0.5, 1.5. Multiply this by $X^T Y$, which is (43, 15), to get $\hat{\beta}$, whose entries are $m = 1.1$ and $b = 1$.

$$\hat{\beta} = \begin{bmatrix} 0.2 & -0.5 \\ -0.5 & 1.5 \end{bmatrix} \begin{bmatrix} 43 \\ 15 \end{bmatrix} = \begin{bmatrix} 1.1 \\ 1.0 \end{bmatrix}$$

$$y = 1.1x + 1.$$

So, the best-fitting line is $Y = 1.1X + 1$. If you draw this picture, you see that indeed it seems to fit the data pretty well.



What would have happened to the matrix $X^T X$ if you had more data?

$$X^T X = \begin{bmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{bmatrix} \text{ and } X^T Y = \begin{bmatrix} \sum x_i y_i \\ \sum y_i \end{bmatrix}.$$

It would still be 2×2 , but the top left entry is just the dot product of the x values with itself, so it's just the sum of the square of the x values. The bottom right corner would be the dot product of the all-1s vector with itself, which is just n , the number of data points. The other 2 entries are the dot product of the x values with the all-1s vector, so you obtain the sum of the x values.

Similarly, $X^T Y$ is the dot product of the columns of the X matrix with the Y vector. So, you get as its entries the sum of $x_i y_i$ and the sum of y_i .

Thus, when trying to find best-fitting lines, you only need to keep track of the following 5 quantities: n , $\sum x_i$, $\sum x_i^2$, $\sum y_i$, and $\sum x_i y_i$.

This is an advantage because when you enter the data one point at a time, you actually don't have to store all the data. You can just add to the running totals for each of these 5 quantities, and you have all the info you need to find the best-fitting line. So, as new data comes in, you just keep track of these running totals.

Multiple Linear Regression

You can consider least square approximations for a collection of data for multiple variables where you think one variable depends on multiple other variables in a linear way.

Suppose you have one variable, Y , the response variable, and multiple predictor variables: U , V , W . Suppose also that you make some large number of observations, n , and collect data (u_i, v_i, w_i) and (y_i) for $i = 1$ to n .

Then, if there is something close to a linear relationship between y_i and u_i , v_i , and w_i , you expect that for some (as yet) unknown coefficients b , b_u , b_v , and b_w ,

$$y_1 = b + b_u u_1 + b_v v_1 + b_w w_1 + \epsilon_1.$$

And you expect similar equations for other data points, with error terms in each.

$$\begin{aligned} y_1 &= b + b_u u_1 + b_v v_1 + b_w w_1 + \varepsilon_1. \\ &\vdots \\ y_n &= b + b_u u_n + b_v v_n + b_w w_n + \varepsilon_n. \end{aligned}$$

This can be rewritten as a matrix equation.

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & u_1 & v_1 & w_1 \\ \vdots & & & \\ 1 & u_n & v_n & w_n \end{bmatrix} \begin{bmatrix} b \\ b_u \\ b_v \\ b_w \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

$$\vec{Y} = X\vec{\beta} + \vec{\varepsilon}.$$

You can appeal to the same normal equations to solve for β and solve the problem in a similar way as before to get normal equations and then find the following estimate for the beta vector.

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Invertibility of the Gram Matrix

In the normal equations, if X is $n \times p$, note that the Gram matrix $X^T X$ is a square matrix of dimensions $p \times p$. Note that if $X^T X$ is not invertible, then β will not have a unique solution, which is a bad thing. So, it would be nice to know when the Gram matrix is invertible.

THEOREM

$X^T X$ is invertible if and only if the columns of X are linearly independent.

Here's a logically equivalent statement:

$X^T X$ is not invertible if and only if the columns of X are linearly dependent.

The columns of X come from various predictor variables in the data.

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & u_1 & v_1 & w_1 \\ \vdots & & & \\ 1 & u_n & v_n & w_n \end{bmatrix} \begin{bmatrix} b \\ b_u \\ b_v \\ b_w \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\vec{Y} = X \vec{\beta} + \vec{\varepsilon}.$$

So, for example, if it happened that the u_i column and the v_i column were identical, then you'd expect that b_u and b_v would have many possible solutions, so it makes sense that you couldn't solve for β uniquely.

How Good Is the Fit?

Are the observed values of \mathbf{Y} close to the predicted values of \mathbf{Y} ?

To find the predicted values of \mathbf{Y} means taking the estimated parameter vector $\hat{\boldsymbol{\beta}}$ and applying it to the data. You believe now that the response variable \mathbf{Y} equals \mathbf{X} times some true $\boldsymbol{\beta}$, but you only have an estimate for $\boldsymbol{\beta}$ called $\hat{\boldsymbol{\beta}}$. So, your predicted values for the data is shown by \mathbf{X} times $\hat{\boldsymbol{\beta}}$.

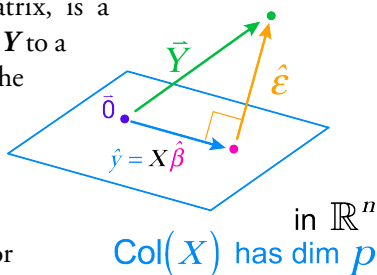
Let's call this vector $\hat{\mathbf{y}}$. If n is the number of observations, it lives in \mathbb{R}^n .

The vector $\hat{\mathbf{y}}$ tells you for each data point what you should have gotten if your estimated parameter $\hat{\boldsymbol{\beta}}$ were the true model. If you replace $\boldsymbol{\beta}$ by the solved normal equations, you find

$$\hat{\mathbf{y}} = \underbrace{\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T}_{\substack{\hat{P} \\ \text{the projection matrix}}} \bar{\mathbf{Y}}.$$

This equation is very interesting. You started off with your original data of observed predictor variables in the matrix \mathbf{X} and the observed response vector \mathbf{Y} . Based on minimizing the residuals, you got an estimate for $\boldsymbol{\beta}$, called $\hat{\boldsymbol{\beta}}$, and are now using that to return predicted values of \mathbf{Y} in $\hat{\mathbf{y}}$.

The expression \hat{P} , the projection matrix, is a matrix that projects an observed vector \mathbf{Y} to a point in the column space of \mathbf{X} , where the predicted values occur if the model is really perfectly linear. So, it takes \mathbf{Y} to $\hat{\mathbf{y}}$, just as expected. You are using your original data \mathbf{X} to turn your original response data, the vector \mathbf{Y} , into their predicted values.



The residual vector is now the difference between the observed values of \mathbf{Y} and the predicted values of $\hat{\mathbf{y}}$. This residual vector is usually called an estimated residual vector because you have only estimated parameters for the actual linear model, which is unknown. So, the residual vector becomes $\hat{\boldsymbol{\epsilon}}$.

If the data were a perfect fit, you'd expect $\hat{\boldsymbol{\epsilon}}$ to be zero. Otherwise, there might be some residuals, so $\hat{\boldsymbol{\epsilon}}$ is a nonzero vector. You are looking for an estimate of how large the individual residuals could be, and the best guess you have for that are the observed data. If you treat components of the residual vector like it were a sample, then an estimate for the standard deviation of the residual is given by using the sample standard deviation formula on the components of the residual vector.

$$s^2 = \frac{\hat{\boldsymbol{\epsilon}}^T \hat{\boldsymbol{\epsilon}}}{n - p}.$$

That's computed by taking s^2 equals the dot product of $\hat{\boldsymbol{\epsilon}}$ with itself divided by the number of degrees of freedom, which is $(n - p)$. Then, s —called the **standard error of the regression**—is the estimator for the standard deviation of the residuals.

Its units are the same units as the \mathbf{Y} variable, so it's a measure of how close you can expect the predicted values to be to the actual values. So, you expect that most of the observed values for \mathbf{Y} lie within 2 standard errors of the predicted values.

This gives you a sense of how good the fit is!

Polynomial Regression

What if the data is not linear? Suppose you want to fit data with a quadratic function or a cubic function.

Surprisingly, you can still use linear algebra! The only thing that changes is the design matrix.

Suppose you have variables X and Y , you think Y depends in a cubic way on X , and you want a best-fitting cubic. Then, you assume

$$Y = aX^3 + bX^2 + cX + d$$

and wish to solve for the coefficients a , b , c , and d .

The key here is to realize that Y depends in a linear way on X^3 , X^2 , X , and 1.

You can use multiple linear regression to do cubic regression if you take the data (x_i, y_i) and make the following substitutions.

$$\begin{aligned}u_i &= x_i \\v_i &= x_i^2 \\w_i &= x_i^3.\end{aligned}$$

Then, you are in exactly the case you were in before! You can use the normal equations to solve for the parameter vector $\hat{\beta}$, which in this case is $[d, c, b, a]$.

The order of these coefficients follows the order of the columns of the design matrix.

You can do polynomial regression of any type by using the appropriate number of variables in multiple linear regression!

READINGS

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, sections 6.5 and 6.6.

Margalit and Rabinoff, *Interactive Linear Algebra*, section 7.5 (<https://textbooks.math.gatech.edu/ila/least-squares.html>).

Poole, *Linear Algebra*, sections 7.2 and 7.3 (and any necessary background from section 7.1).

SINGULAR VALUE DECOMPOSITION: SO COOL

One of the promises, but also the perils, of the digital age we live in is that so much data can be collected about our preferences based on the choices we make online. For example, each time you buy something from Amazon or choose a movie on Netflix, you are telling these companies something about your preferences. And these companies can use this information to suggest something else that you may want to buy or another movie that you may want to watch. These are called recommender systems. This is just one of the many important—and cool—applications that the singular value decomposition has in data analysis.

The Singular Value Decomposition

Recall that if a square matrix is diagonalizable, that means it can be factored as $A = PDP^{-1}$, where P is an invertible matrix and D is a diagonal matrix. You can think of multiplication by P as performing a change of basis that takes a linear combination of the standard basis into a linear combination of the columns of P .

The singular value decomposition offers a way to compress data, such as a photo, without losing too much information.

So, if A factors as the product of P , D , and P^{-1} , then the transformation of multiplication by A can be done in 3 steps.

- 1 Multiply by P^{-1} , which takes the basis of columns of P and transforms it into the standard bases along the axes.
- 2 Multiply by D , which scales the standard basis vectors along the axes by the corresponding eigenvalues.
- 3 Multiply by P , which turns the standard basis vectors back into the basis given by the columns of P .

This sequence of 3 operations will do the same thing as taking the basis of columns of P and stretching them by factors that correspond to eigenvalues.

So, to be diagonalizable means that the transformation just looks like scaling in certain special directions. The scaling factors are the eigenvalues and the special directions are the eigenvectors, and these form the columns of P .

A special case of this is when the matrix A is symmetric. Real symmetric matrices are diagonalizable, and, in fact, the special directions can be taken to be an orthonormal set. So, then the columns of P are orthonormal, and the matrix is orthogonal. Then, P^{-1} is just P^T , revealing that any symmetric matrix A can be factored as $A = PDP^T$, where P is an orthogonal matrix and D is a diagonal matrix.

But not every matrix is symmetric, much less diagonalizable. So, it may be surprising to know that there is a standard factorization of any matrix A —even nonsquare ones—as a product of an orthogonal matrix, a diagonal matrix, and another orthogonal matrix. The catch is that the 2 orthogonal matrices are not necessarily inverses of one another.

Such a factorization is called the **singular value decomposition (SVD)**, and it says that any $m \times n$ matrix A can be factored as $U\Sigma V^T$, where Σ is an $m \times n$ diagonal matrix and U and V are orthogonal matrices of size $m \times m$ and $n \times n$, respectively. The columns of U live in \mathbb{R}^m and are called the **left singular vectors**, and the columns of V live in \mathbb{R}^n and are called the **right singular vectors**. The diagonal entries of Σ are called the **singular values** of A and are ordered by size and denoted σ_1, σ_2 , etc. The singular values are always nonnegative.

$$\begin{bmatrix} A \\ m \times n \end{bmatrix} = \underbrace{\begin{bmatrix} U \\ m \times m \end{bmatrix}}_{\substack{r \\ \text{basis Col}(A)}} \begin{bmatrix} \Sigma \\ m \times n \end{bmatrix} \begin{bmatrix} V^T \\ n \times n \end{bmatrix} \begin{matrix} \text{basis} \\ \text{Row}(A) \\ r \\ \text{basis} \\ \text{Null}(A) \end{matrix}$$

Remember that Σ has the same dimension as the original matrix A , so it's not necessarily square. It is a diagonal matrix in that the only nonzero entries occur where the row and column numbers are the same. So, a diagonal matrix could look like any of the following matrices.

$$\begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \\ 0 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \end{bmatrix} \quad \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix}$$

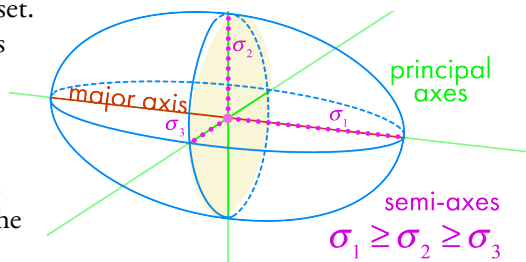
The SVD also has some other properties that are quite cool. For example, if the matrix A has rank r , then the first r columns of U will be a basis for the column space of A . And the first r columns of V (which are rows of V^T) will be a basis for the row space of A . And the last $(n - r)$ columns of V (which are the last $n - r$ rows of V^T) will be a basis for the null-space of A !

The rank of a matrix A is the dimension of the column space, the subspace that's the image of the linear transformation.

The Geometric Meaning of the SVD

The geometric meaning of the SVD can be seen by looking at how a unit sphere (a sphere of radius 1) in \mathbb{R}^n gets transformed by the matrix A . Amazingly, a unit sphere always gets sent to some kind of hollow or filled ellipsoid in \mathbb{R}^m . It will be a filled-in or collapsed ellipsoid if the matrix rank r is strictly less than the dimension m .

For any ellipsoid, the **principal axes** of the ellipsoid are a set of orthogonal axes along which the ellipsoid gets stretched. To find the principal axes, look at the largest diameter of the ellipsoid—called the **major axis**—and put that in the set. Then, look at all points in the ellipsoid perpendicular to this axis; it is another ellipsoid (of one lower dimension). Find its major axis, and put that in the set. Then, look at all vectors perpendicular to the first 2 axes and find the next-largest axis. Continue in this fashion to obtain a collection of axes that are the principal axes.



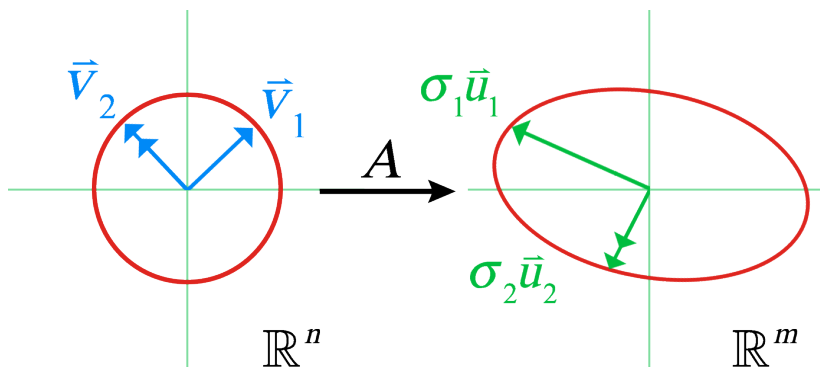
The so-called singular values of A are $\frac{1}{2}$ the lengths of the principal axes. You can also think of them as the length of the semi-axes of the ellipsoid, which are usually ordered from largest to smallest and labeled σ_1, σ_2 , etc.

The matrix Σ in the SVD is an $m \times n$ matrix that is all zeros except for r nonzero entries along the diagonal of the matrix, where r is the rank of the matrix A . You can think of it as an $m \times n$ matrix with an $r \times r$ diagonal block D and every other block filled with zeros. For this reason, Σ is called a diagonal matrix, even if it is not square. The diagonal elements of Σ are the singular values, and there will be r nonzero singular values if r is the rank of A .

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}_{m \times n}$$

The columns of the matrix U represent the directions of the principal axes in order from largest to smallest axes. And the columns of V are the unit vectors in \mathbb{R}^n that map to the directions of the principal axes of the ellipsoid.

Look at this pictorial example of a linear transformation from \mathbb{R}^2 to \mathbb{R}^2 to get a feeling for what happens in general for linear transformations from \mathbb{R}^n to \mathbb{R}^m .



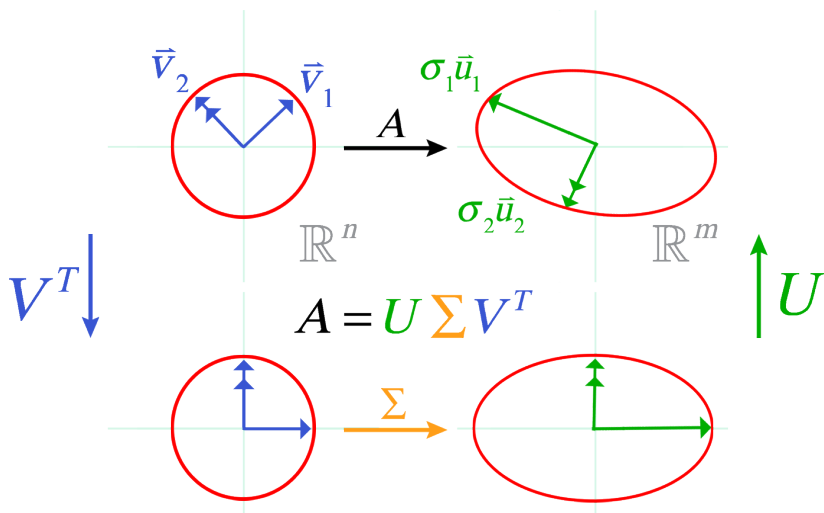
Suppose that under multiplication by A , the unit circle gets transformed to some tilted ellipse. The vectors along the principal axes of this ellipse, scaled to length 1, are the left singular vectors. Their lengths are the singular values. For example, the vector along the semimajor axis of the ellipse will be σ_1 , the largest singular value, times \mathbf{u}_1 , the first left singular vector. Similarly, the vector along the semiminor axis of this ellipse will be σ_2 times the unit vector \mathbf{u}_2 .

These vectors along the axes of the ellipse, colored in green, are images of certain unit vectors in the domain, colored in blue. Those are the corresponding right singular vectors, which form the columns of V .

It's funny that the left singular vectors appear in the right picture and the right singular vectors appear in the left picture. But that's because the vectors are named after their location in the factorization, because in the SVD, U is on the left and V is on the right.

In higher dimensions, the left singular vectors would align with the principal axes of the ellipsoid from largest to smallest. If the linear transformation is doing some collapsing or if its image doesn't span all of \mathbb{R}^m , then some of the singular values will be zero. In such cases, there may not be unique choices for the singular vectors. But the singular values are uniquely defined.

Now, let's see how the SVD factorization works. You can visualize the factorization by enlarging the previous diagram to include 2 more pictures below the first 2.



To say that the $m \times n$ matrix A factors into 3 parts means that, as a linear transformation acting on a vector in \mathbb{R}^n and producing a vector in \mathbb{R}^m , you can get exactly the same result by doing this 3-step process: first multiplying the vector by V^T on the left, then by Σ on the left, and then by U on the left.

This is indicated in the diagram by an alternate path that goes from the top left picture down to the bottom left picture, then across to the bottom right picture, and then up to the top right picture.

Look at the first map going down, which represents multiplication by V^T .

Recall that the matrix V represents the orthogonal transformation you get by rotating the standard basis vectors to the blue vectors in the first picture, which are the columns of V . Then, V^T (which is V^{-1} for orthogonal matrices) will rotate the blue vectors in the first picture into the standard basis in the bottom left picture.

Then, multiplying by the diagonal matrix Σ scales each axis direction by the corresponding singular value, so that turns the unit circle in the bottom left picture into an axis-aligned ellipse in the bottom right picture.

Next, another rotation that takes the standard basis to lie along the principal axes of the ellipse is needed, and the matrix U does that. So, that's the map going from the bottom right picture to the top right picture.

So, in the diagram, you see that performing the action of A on the top side of this diagram is the same as composing the 3 transformations along the sides and bottom.

Computing the SVD

The key to computing the SVD is to look at $A^T A$, because in the SVD, you are searching for an orthogonal, a diagonal, and an orthogonal matrix, kind of like what is done for symmetric matrices.

Though A is not symmetric, you can symmetrize it by multiplying it on the left by its transpose.

Notice that $A^T A$ is symmetric, because its transpose is just itself. And it is a real matrix. So, you can expect that it has an orthogonal diagonalization.

$$\begin{bmatrix} A^T \end{bmatrix}_{n \times m} \begin{bmatrix} A \end{bmatrix}_{m \times n} = \begin{bmatrix} A^T A \end{bmatrix}_{n \times n}.$$

So, let's assume that A has an SVD and see what that would imply about the factors.

If $A = U\Sigma V^T$, then you can compute $A^T A$, which involves first taking the transpose of $U\Sigma V^T$ and then multiplying by $U\Sigma V^T$.

You get $V\Sigma^T U^T U\Sigma V^T$.

$$\begin{aligned} A &= U \Sigma V^T \\ A^T A &= (U \Sigma V^T)^T U \Sigma V^T \\ &= V \Sigma^T \underbrace{U^T U}_I \Sigma V^T = V (\Sigma^T \Sigma) V^T. \end{aligned}$$

But U is orthogonal, so the middle terms ($U^T U$) is just the identity. So, the product becomes $V\Sigma^T \Sigma V^T$.

You've just shown that $A^T A$ is equal to $V\Sigma^T \Sigma V^T$. This is very interesting, because V^T is supposed to be orthogonal.

And $\Sigma^T \Sigma$ is a diagonal matrix, because you can check that it is square, and it is filled with dot products of the columns of Σ , and the columns of Σ were orthogonal.

So, $V(\Sigma^T \Sigma)V^T$ is really an orthogonal diagonalization of $A^T A$ if an SVD exists!

This tells you what you should expect for an SVD.

- ◆ Eigenvalues of $A^T A$ are the diagonal elements of $\Sigma^T \Sigma$, but those are the square of the singular values σ_i , so the singular values of A should be the square root of the eigenvalues of $A^T A$.
- ◆ V should be the matrix whose columns are the eigenvectors of $A^T A$.
- ◆ The image of the columns of V under the transformation A should be the columns of U scaled by the singular value so that they lie along the image ellipsoid, which means the columns of U should be $(1/\sigma_i)$ times (Av_i) .

Σ is an $m \times n$ matrix, the same dimensions as A , filled with zeros except possibly for diagonal elements, which are the singular values.

The singular values are found by taking the square root of the eigenvalues of $A^T A$. (There are n such eigenvalues, because $A^T A$ is an $n \times n$ matrix.)

The matrix V has columns \mathbf{v}_i , which will be eigenvectors of $A^T A$ normalized to make them length 1. (You know that the \mathbf{v}_i are orthogonal because they were the eigenvectors of an orthogonally diagonalizable matrix.)

The matrix U has columns \mathbf{u}_i , which are $1/\sigma_i$ times A times \mathbf{v}_i for the nonzero values of σ_i . The other columns can be completed to form an orthonormal set.

THEOREMS

- ◆ The eigenvalues of $A^T A$ must be ≥ 0 .
- ◆ $\mathbf{A}\mathbf{v}_i \cdot \mathbf{A}\mathbf{v}_i = 0$ unless $i = j$.

READINGS

Brand, “Fast Online SVD Revisions for Lightweight Recommender Systems.”

Chartier, *When Life Is Linear*, chap. 11.

Kun, “Singular Value Decomposition Part 1,” <https://jeremykun.com/2016/04/18/singular-value-decomposition-part-1-perspectives-on-linear-algebra/>.

Lay, Lay, and McDonald, *Linear Algebra and Its Applications*, sections 7.4 and 7.5.

Poole, *Linear Algebra*, section 7.4 and the Vignette on Digital Image Compression.

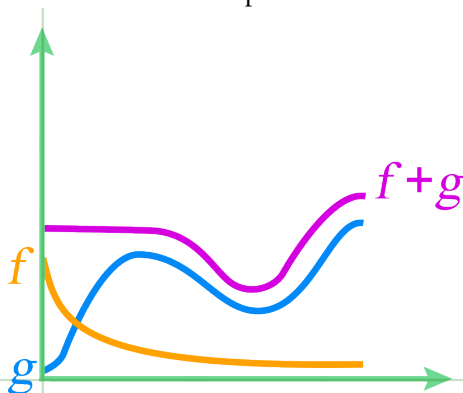
Roughgarden and Valiant, “CS168,” <http://theory.stanford.edu/~tim/s17/l/l9.pdf>.

GENERAL VECTOR SPACES: MORE TO EXPLORE

This lecture focuses on the ways that the ideas of linear algebra can apply more generally than vectors in \mathbb{R}^n . It turns out that the ideas are very powerful for many other kinds of objects, as well as the transformations between them.

Functions as Vectors

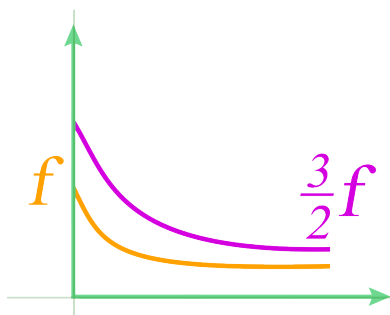
Let's explore the space of functions from \mathbb{R} to \mathbb{R} and see in what ways functions in that space behave like vectors.



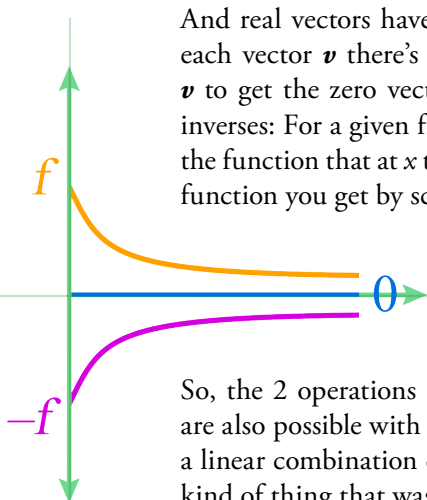
Vectors can be added. And functions can be added in a natural way: If you want the sum of 2 functions f and g , you can define a new function (call it $f+g$) at a point x by just adding the 2 functions point by point, as $f(x) + g(x)$.

Just like the addition of real numbers or real vectors, the order of addition of functions doesn't matter: $f + g$ is the same function as $g + f$. And an associative property for functions holds as well: $(f + g) + h = f + (g + h)$.

Real vectors can be scaled by a real number. Similarly, scalar multiplication of functions makes sense also. And scalar multiplication of functions obeys the laws you expect: 3 times 4 times a function f is 12 times f , and 3 times $(f + g)$ is the sum of $3f$ and $3g$.



For real vectors, there's a special vector called the **zero vector**, which, if you add to any vector, leaves that vector unchanged. A function that behaves like that is the zero function, which is identically 0 everywhere: $z(x) = 0$ for all x . Then, $f + z = f$ for any function f .



And real vectors have **additive inverses**, meaning for each vector v there's a vector $-v$ that you can add to v to get the zero vector. Functions also have additive inverses: For a given function f , you can define $-f$ to be the function that at x takes the value $-f(x)$. It is the same function you get by scalar multiplication of $f(x)$ by -1 .

So, the 2 operations that define a linear combination are also possible with functions. So, you can talk about a linear combination of functions. This is precisely the kind of thing that was important in vector spaces.

General Vector Spaces

The general definition of a vector space collects together important properties. The idea is that many of the best features of looking at vectors in \mathbb{R}^n are captured in these properties, and many of the theorems for vectors in \mathbb{R}^n will carry over for general vector spaces if they just use these properties.

A general vector space will be a set with 2 operations: addition and scalar multiplication. Things in the set will be called vectors. The scalars will be real numbers—these are the things that will be used to scale the vectors in scalar multiplication.

The operations of addition and scalar multiplication should satisfy 10 axioms. The first 5 deal with addition while the last 5 deal with scalar multiplication and how it should play nicely with addition.

AXIOMS

- 1 $\mathbf{u} + \mathbf{v}$ is in V (V closed under addition)
- 2 $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ (addition is commutative)
- 3 $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$ (addition is associative)
- 4 There's a zero vector $\mathbf{0}$ such that $\mathbf{u} + \mathbf{0} = \mathbf{u}$
- 5 Every vector \mathbf{u} has an additive inverse $-\mathbf{u}$ such that $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$
- 6 $c\mathbf{u}$ is in V (V closed under scalar multiplication)
- 7 $c(\mathbf{u} + \mathbf{v}) = c\mathbf{u} + c\mathbf{v}$
- 8 $(c + d)\mathbf{u} = c\mathbf{u} + d\mathbf{u}$ } distributive properties
- 9 $c(d\mathbf{u}) = (cd)\mathbf{u}$
- 10 $1\mathbf{u} = \mathbf{u}$

Fibonacci-Type Sequences as a Vector Space

One example of a vector space is \mathbb{R}^n . Every point is a collection of n numbers. You've been studying this vector space throughout this course, and you've seen that it satisfies all these properties.

Another example of a vector space is Fibonacci-type sequences. The Fibonacci sequence is a sequence of numbers that begins with a zero and a 1 and thereafter every number is the sum of the 2 numbers before it.

$$0, 1, 1, 2, 3, 5, 8, 13, \dots$$

There is a general formula for the n th term of the Fibonacci sequence without having to calculate all previous terms. If the Fibonacci sequence is f_n and you let the initial term be f_0 , so that $f_0 = 0$ and $f_1 = 1$, then f_n is given by the following formula, known as Binet's formula.

$$f_n = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2} \right)^n.$$

This is a remarkable formula, because the expressions on the right side involve strange combinations of the square root of 5, so it is not obvious that this will always be an integer! This formula can be found using vector space ideas.

If you keep the same rule as the Fibonacci sequence but start with different numbers, you get a Fibonacci-type sequence.

So, if you start with 1 and 0, you will get this sequence, called the **co-Fibonacci sequence**, which looks like the Fibonacci sequence with a 1 inserted in front.

$$1, 0, 1, 1, 2, 3, 5, 8, \dots$$

If you start with 2 and 1, you get a Fibonacci-type sequence called a **Lucas sequence**.

$$2, 1, 3, 4, 7, 11, 18, 29, \dots$$

The sum of 2 Fibonacci-type sequences—term by term—is still a Fibonacci-type sequence. So, if you add the Fibonacci sequence to the Lucas sequence term by term, you get

$$2, 2, 4, 6, 10, 16, 26, 42, \dots,$$

which is also a Fibonacci-type sequence.

It's also true that if you multiply every term of a Fibonacci-type sequence by the same scalar, you still get a Fibonacci-type sequence. All the other axioms of a vector space hold as well. So, the set of Fibonacci-type sequences forms a vector space!

The first thing to notice about this vector space is that a sequence is completely determined by its first 2 elements. So, every point (x, y) in \mathbb{R}^2 corresponds with a sequence that begins $x, y, x + y, \dots$.

This means that Fibonacci-type sequences have the same vector space structure as the 2-dimensional plane. So, because $(0, 1)$ and $(1, 0)$ are basis elements for \mathbb{R}^2 , you can rewrite every Fibonacci-type sequence as a linear combination of the Fibonacci sequence—which begins with 0, 1—and the co-Fibonacci sequence, which begins with 1, 0.

So, the Fibonacci sequence and the co-Fibonacci sequence form a basis for all Fibonacci-type sequences. But that's not the only basis. If you hunted for a basis where the n^{th} terms of each basis sequence were easy to calculate, then the n^{th} term of any linear combination would be easy to calculate as well.

Suppose you had a sequence $x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, \dots$.

One suggestion would be to hunt for sequences that are geometric, so every term is r times the previous term:

$$1, r, r^2, r^3, r^4, r^5, r^6, r^7, \dots$$

The n^{th} term x^n would be r^n . This would be a Fibonacci-type sequence as long as each term were the sum of the prior 2 terms. So, $r^2 = r + 1$.

This is easy to solve, and it has 2 solutions:

$$r_1 = \frac{1 + \sqrt{5}}{2}, \quad r_2 = \frac{1 - \sqrt{5}}{2}.$$

Each of these corresponds to a sequence, and these 2 geometric sequences form a basis for the set of all Fibonacci-type sequences. So, you can try to express the original Fibonacci sequence in terms of these 2 geometric sequences by seeing what condition has to hold for the first 2 terms of each sequence. So, you just need to solve how to write $(0, 1)$ in terms of $(1, r_1)$ and $(1, r_2)$. Set $(0, 1)$ equal to the following and solve.

$$(0, 1) = a(1, r_1) + b(1, r_2).$$

If you do this, you will know that the n^{th} Fibonacci term will be described by

$$f_n = ar_1^n + br_2^n.$$

If you solve for a and b , you find that

$$a = \frac{1}{\sqrt{5}} \quad b = -\frac{1}{\sqrt{5}}.$$

This yields Binet's formula!

$$f_n = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2} \right)^n.$$

Space of Functions as Vector Spaces

You've already seen that functions behave like vectors so that the space \mathcal{F} of all functions from \mathbb{R} to \mathbb{R} is a vector space. You might check that it satisfies all 10 axioms for a vector space. But there are so many kinds of functions that you might want to restrict your attention to one, and each of these, in a very natural sense, forms a subspace of this vector space of all functions from \mathbb{R} to \mathbb{R} .

Recall that a subspace of \mathbb{R}^n is a set of vectors that remains closed under addition and scalar multiplication. A subspace of functions can be described in the same way—as a subset of functions that is a vector space. But as long as the subset is closed under addition and scalar

multiplication, it will inherit all the other vector space properties from the larger vector space it is part of. So, a subspace is just a subset of a vector space that is closed under addition and scalar multiplication.

For example, there was nothing in the definition that said that functions from \mathbb{R} to \mathbb{R} have to be continuous. They don't have to be. But if you restrict your attention to the set of continuous functions, called \mathcal{C} , it is true that the sum of 2 continuous functions is continuous, and the scalar multiple of a continuous function is continuous. So, the set of continuous functions is a vector space.

$$\mathcal{C} = \{\text{continuous functions } \mathbb{R} \rightarrow \mathbb{R}\}.$$

Similarly, an even smaller subspace is the subspace \mathcal{D} of differentiable functions. Again, you can take a linear combination of differentiable functions, and it will still be differentiable. Moreover, there is a natural linear transformation on \mathcal{D} : the differential operator. With operators, you might be interested in functions that behave like eigenvectors—called eigenfunctions.

$$\mathcal{D} = \{\text{differentiable functions } \mathbb{R} \rightarrow \mathbb{R}\}.$$

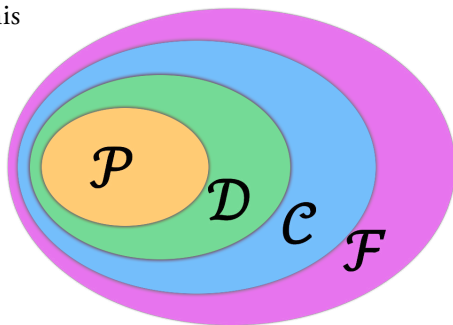
One question you might ask about any vector space is this: Does it have a basis? This means what it did before: a linearly independent set that spans the whole space. If it does, you might want to know how to express any vector in the space in terms of a convenient basis. (You've already seen this lesson with the Fibonacci-type sequences.) As it turns out, every vector space does, in fact, have a basis. And for a given vector space, every basis must have the same size, so you can talk about the dimension of the vector space.

You could look at an even smaller subspace—the set of polynomials—again noting the sum of polynomials is a polynomial and the scalar multiple of a polynomial is a polynomial. You can check that the monomials $1, x, x^2, x^3, \dots$ are linearly independent and span the set of all polynomials, so they are a basis! This means that the space of polynomials is infinite-dimensional.

$$\mathcal{P} = \{\text{polynomial functions}\}.$$

Polynomials that have degree no more than n form a subspace of this space of polynomials with dimension $(n + 1)$.

Schematically, you might represent the relationship between all these subspaces as in the drawing shown at right, which is only meant to showcase inclusion, rather than the linear aspects of these spaces.



Solutions of Differential Equations

When solving differential equations, linear algebra can be very powerful.

A differential equation is an equation that expresses a condition that the derivative of a function must satisfy. There is a special kind of differential equation called a **linear differential equation**, in which the space of solutions turns out to be a vector space!

For example, think of a spring with a mass that is oscillating, but the oscillation is damped by friction. The function $f(t)$ describes the position of the mass at time t . This is a damped harmonic oscillator, and the governing equation for this is the following second-order differential equation (meaning it involves the second derivatives of a function).

$$mf'' + cf' + kf = 0.$$

The first thing to notice is that if f and g are solutions, then $(f + g)$ also satisfies the same differential equation. Similarly, a constant times f will also satisfy the same differential equation. So, the space of solutions is a subspace of the space of differentiable functions!

The theory of existence and uniqueness of second-order differential equations says that this differential equation has a unique solution once the function value and its derivative are specified at time 0.

So, if you have a set of functions where

$$\begin{array}{lll} f(0) = 1 & g(0) = 0 & h(0) = a \\ f'(0) = 1 & g'(0) = 0 & h'(0) = b, \end{array}$$

consider the function $af + bg$.

You can check that this function has the same initial conditions as h . So, by the existence and uniqueness theorem, that means it must be h . So, every solution to this differential equation can be expressed in terms of f and g . This also means the subspace of solutions is 2-dimensional.

This doesn't tell you what the solutions are, but if the subspace is 2-dimensional, this means if you can find 2 linearly independent solutions, you will have a different basis, and all solutions will be expressible in terms of that basis!

So, the key idea is to look for a solution that will be of the form

$$f(t) = e^{\lambda t}$$

because this fits with the intuition that an oscillator that is damped heavily will have a displacement that exponentially decays.

When you differentiate this function once, you get λe^t , and when you differentiate twice, you get $\lambda^2 e^{\lambda t}$. Plugging that into the differential equation, you get the following.

$$m\lambda^2 e^{\lambda t} + c\lambda e^{\lambda t} + k e^{\lambda t} = 0$$

$$\left[m\lambda^2 + c\lambda + k \right] e^{\lambda t} = 0.$$

When factored, this shows that $m\lambda^2 + c\lambda + k$ must be zero. This is a quadratic equation with 2 roots! So, as long as the roots λ_1 and λ_2 are distinct, you will get 2 linearly independent solutions to the differential equation: $e^{\lambda_1 t}$ and $e^{\lambda_2 t}$.

Because you know that the subspace of solutions is 2-dimensional, this means that you now know all solutions to the differential equation: All solutions are linear combinations of these 2 solutions. In other words, all solutions are of the form

$$C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}.$$

The constants C_1 and C_2 can be found by using the initial conditions.

Ideas of Fourier Analysis

The idea of a good basis is a motivation for the idea of Fourier analysis. In signal processing, we are interested in taking a signal, some function of time—call it $f(t)$ —and breaking it down into its component frequencies. In effect, we are asking if an arbitrary function $f(t)$ can be expressed as a sum of sines and cosines.

In general, it turns out that any continuous function on a closed interval can be expressed as a sum of sine and cosines if you allow infinite sums! In other words, sines and cosines form a very nice kind of “basis” with which you can write other functions. (“Basis” here is not quite the same as the notion of basis that was defined for a vector space, because you are allowing infinite sums here, but in most other respects, the ideas are the same.)

Moreover, it turns out that if you are looking at spaces of continuous functions on an interval, you can define something like a dot product that makes sense, called the **inner product**. The inner product of functions f and g is the integral of $f(x)g(x)dx$, which is an integral of a product—so it looks just like the dot product, except the sum is replaced by an integral.

$$\langle f, g \rangle = \int_{x=a}^b f(x)g(x)dx.$$

This inner product has similar properties to the dot product. Once you have a dot product, then you know what it means for one function to be orthogonal to another function. And then you see that the sines and cosines form an **orthonormal basis** for this vector space.

The ideas of Fourier analysis are powerful for many reasons. For example, it allows you to break a function into basis functions that are simpler. If you take advanced courses in differential equations, you'll see that writing a function this way can help you solve certain partial differential equations, such as the heat equation, which is what Joseph Fourier is well known for.

And if you have a function expressed as an infinite sum, then if you ignore the low-amplitude terms, you will have an efficient approximation to a function. In signal processing, low-amplitude terms can arise from noise in the signal, or background noise. So, taking only the largest terms can make the sound clearer. In this case, the vectors are the functions, and an arbitrary infinite sum is being approximated by a finite linear combination of sines and cosines. Nevertheless, this powerful idea is in action once again.

The idea of linearity, and linear combinations, is such a powerful idea. You might not have expected to see it in the study of functions that themselves could be very wild and nonlinear. But the ways they are being combined are linear.

READING

Poole, *Linear Algebra*, chaps. 4 (especially the Introductory Example) and 6 (especially 6.0 and 6.1 and the Exploration after section 6.2).

This course has only scratched the surface of many great ideas of linear algebra. If you want to learn more, you can take almost any further advanced courses in mathematics, because linear algebra shows up in some of the most unexpected places.

QUIZ FOR LECTURES 19–24

- 1 Given an $n \times n$ matrix Q , show that if $(Q\mathbf{x}) \cdot (Q\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$ for all vectors \mathbf{x} and \mathbf{y} in \mathbb{R}^n (in other words, Q preserves dot products), then Q must be an orthogonal matrix. [LECTURE 19]
- 2 If Q_1 and Q_2 are both orthogonal matrices, then show that $Q_1 Q_2$ is orthogonal, too. [LECTURE 19]
- 3 If the probability of going from state i to state j in a Markov chain is the same as the probability of going in the reverse direction, what must be true about the transition matrix of the Markov chain? [LECTURE 20]
- 4 Give an example of a 2×2 transition matrix (whose rows sum to 1) that is not positive, but its square is positive. (Hint: Think about a Markov chain with 2 states where not every state is reachable in 1 step but is reachable in 2 steps.) [LECTURE 20]
- 5 What is the derivative matrix of the function T from \mathbb{R}^2 to \mathbb{R}^2 that leaves all points fixed: $T(x, y) = (x, y)$? [LECTURE 21]
- 6 Let $f(x, y) = (y, xy + y)$, $g(x, y) = (y, x + y)$, and $h(x, y) = (xy, y)$. Observe that $f(x, y) = g(h(x, y))$ and verify that their derivative matrices obey the chain rule: $[Df] = [Dg][Dh]$. [LECTURE 21]
- 7 Use software to compute the QR -factorization of X ; then, use this to calculate $\hat{\beta}$ using the method described in the lecture. Check this $\hat{\beta}$ against the one found in the lecture. [LECTURE 22]
- 8 What meaning would it have for the lecture's example of a 4-data-point linear regression if the columns of X were not linearly independent? [LECTURE 22]

- 9 Find the singular values of the 2×2 matrix A (at right) by computing the eigenvalues of the Gram matrix $A^T A$.
[LECTURE 23]

$$A = \begin{bmatrix} -4 & 6 \\ 3 & 8 \end{bmatrix}$$

- 10 Find the SVD of the 2×2 matrix A (at right). [LECTURE 23]
- 11 Check that the set of $n \times n$ matrices form a vector space under matrix addition and scalar multiplication. [LECTURE 24]
- 12 The inner product of 2 continuous functions f and g over an interval $[0, 2\pi]$ is defined as

$$\langle f, g \rangle = \int f(x)g(x)dx,$$

where the integration runs over x going from 0 to 2π . Verify that $\sin(x)$, $\sin(2x)$, and $\cos(x)$ are mutually orthogonal using this inner product.
[LECTURE 24]

Solutions can
be found on
page 303.

SOLUTIONS

Lectures 1–6

- 1 The 4 themes are as follows:
 - a Linearity is a fundamental idea in mathematics—and in the world—and it appears everywhere.
 - b Nonlinear things are approximated by linear things.
 - c Linear algebra reveals hidden structures that are beautiful and useful.
 - d Linear algebra's power comes from the interplay between geometry and algebra.
- 2 Reflection is a linear transformation because it does satisfy the 2 linear properties.
- 3 If you draw these on graph paper, you should be drawing 2 lines with a set of equally spaced tick marks on one line and another set of equally spaced tick marks on the other line.
- 4 The points that are integer linear combinations $(au + bv)$ will form a grid of points in the plane. If you estimate the linear combination that produces $(4, 0)$, you should see that it is $\frac{8}{3}u + \frac{4}{3}v$. The set of all linear combinations does cover the entire plane.
- 5 The dot product $u \cdot v = 0$. The cross product $u \times v = (-6, -5, 3)$. Because the dot product is 0, the vectors u and v are perpendicular. Also, the cross product $u \times v$ has the property that it is perpendicular to u and v . So, all 3 vectors— u , v , and $u \times v$ —are mutually orthogonal.
- 6 To find the equation of a plane through 3 points a , b , and c (as vectors in \mathbb{R}^3) first find the vectors $(b - a)$ and $(c - a)$, which must lie parallel to the plane. You can use these vectors and the point a to express the plane in parametric form.

Alternatively, you can take the cross product $(\mathbf{b} - \mathbf{a}) \times (\mathbf{c} - \mathbf{a})$, which will be normal to the plane and, together with the point \mathbf{a} , can be used to find the equation of the plane in normal form. These procedures will fail if the 3 points you start with are collinear.

7 The product is $\begin{bmatrix} 6 & 7 & 1 \\ 5 & 7 & 2 \\ 4 & 7 & 3 \end{bmatrix}$.

8 The product is $\begin{bmatrix} 6 & 5 & 4 \\ 7 & 7 & 7 \\ 1 & 2 & 3 \end{bmatrix}$.

9 There are many ways to see why the translation T is not linear. Perhaps the easiest is to note that $T(\mathbf{0})$ does not equal the zero vector $\mathbf{0}$, which must be true for any linear transformation. But you can also check that T does satisfy either of the 2 linearity properties.

10 To find the matrix $[R]$ representing to the linear transformation R , you just note where the standard basis vectors go and put those in the columns of the matrix in the same order. Because $(1, 0)$ goes to $(0, 1)$ under reflection, $(0, 1)$ goes in the first column of $[R]$. And because $(0, 1)$ goes to $(1, 0)$, the second column of $[R]$ is $(1, 0)$.

$$\text{So, } [R] = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

11 Each equation in the original system has a solution set that is a plane in \mathbb{R}^3 because it is the equation of a plane in normal form. If you modify the second equation by subtracting from it 3 times the first equation, you'll get this system, which has the same set of solutions:

$$\begin{aligned} x + y - z &= 1 \\ -y + 3z &= -1. \end{aligned}$$

Then, if you modify the first equation by adding the second equation to it, you get:

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

Now multiply the second equation by -1 to get:

$$x + 2z = 0$$

$$y - 3z = 1.$$

To obtain a particular solution, fix z to be any number you want, and use the above equations to find x and y . If you set $z = 1$ in the above equations, you find $x = -2$ and $y = 4$, so $(-2, 4, 1)$ is a solution of this system, and you can check that it is a solution of the original system, too.

- 12** Yes. Think of 3 vertical planes in \mathbb{R}^3 where any 2 of them intersect in a line, but these lines of intersection are all parallel and distinct. Then there is no point in \mathbb{R}^3 that satisfies all 3 equations simultaneously. For example, $x = 0$, $y = 0$, and $x + y = 1$ are all planes in \mathbb{R}^3 that pairwise have solutions, but the 3 equations together have no solutions.

[Click here to go back to the quiz.](#)

Lectures 7–12

- 1** No, the system must always be consistent, because $\mathbf{x} = \mathbf{0}$ will always be a solution. Moreover, when row-reducing the augmented matrix $[A|\mathbf{0}]$, the zeros on the right side will remain zero under any row operations, so you will never get an inconsistency.
- 2** Given a 4×3 matrix A , in order for $A\mathbf{x} = \mathbf{b}$ to have a unique solution, $\text{RREF}(A)$ must be a 4×3 matrix where the first 3 rows look like the 3×3 identity matrix and the fourth row is a row of zeros.

- 3 The given vectors are linearly dependent. This is because if you put those vectors as columns of a matrix A in the given order and solve $A\mathbf{x} = \mathbf{0}$, you will find $\mathbf{x} = (2, -1, -1)$. So, there is a nontrivial combination of the vectors that produces the zero vector:

$$2(1, 2, 3) - 1(1, 1, 2) - 1(1, 3, 4) = (0, 0, 0).$$

- 4 If 3 given vectors are not linearly independent, then one of them is a linear combination of the other 2; thus, that vector lies in the span of the other 2. So, the span of all 3 vectors can be at most a plane (and might be smaller), so the 3 vectors cannot span all of \mathbb{R}^3 in this instance.
- 5 The set S is a subspace because: It contains the zero vector (which has second coordinate zero); any 2 vectors with second coordinate zero, when summed, still has second coordinate zero; and if you multiply a vector with second coordinate zero, you still get second coordinate zero.
- 6 The set D of ingredient demands is a subspace because: The ingredient demands for an empty basket is the zero vector; if \mathbf{u} is a vector of ingredient demands for one basket and \mathbf{v} is a vector of ingredient demands for another basket, then $\mathbf{u} + \mathbf{v}$ will be a vector of ingredient demands for the combined baskets; and if c is a scalar, then $c\mathbf{u}$ is the ingredient demand for the basket that has c times as many things as the basket that \mathbf{u} represents.
- 7 Because the leading entries of $\text{RREF}(A)$ are in columns 1, 2, and 4, this means that columns 1, 2, and 4 of the original matrix A will form a basis for the column space of A . A basis for the row space of A is the nonzero rows of $\text{RREF}(A)$.
- 8 Solving $A\mathbf{x} = \mathbf{0}$ can be done by row-reducing the matrix formed by augmenting A with the zero vector. Then, row-reducing produces $\text{RREF}(A)$ augmented with the zero vector. Using x , y , z , and w for the variables, this augmented matrix corresponds to these equations:

$$\begin{aligned}x + z &= 0 \\y + z &= 0 \\w &= 0.\end{aligned}$$

Because z is a free variable, you can use these equations, plus the equation $z = z$, to express all variables in terms of the free variable z :

$$x = -z$$

$$y = -z$$

$$z = z$$

$$w = 0.$$

This shows that any solution (x, y, z, w) is a multiple of $(-1, -1, 1, 0)$. Thus, the vector $(-1, -1, 1, 0)$ is, by itself, a basis for the 1-dimensional null-space.

9 The inverse is $\begin{bmatrix} -3/2 & 1/2 \\ 1 & 0 \end{bmatrix}$.

10 Multiplying $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ produces the matrix $\begin{bmatrix} a+c & b+d \\ a+c & b+d \end{bmatrix}$, which cannot be the identity matrix.

11 Multiplying $\begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ produces the matrix $\begin{bmatrix} a & b \\ c+3a & d+3b \end{bmatrix}$.

The inverse of $\begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix}$ is $\begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix}$.

Multiplying $\begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ produces the matrix $\begin{bmatrix} a & b \\ c-3a & d-3b \end{bmatrix}$.

12 This fact follows from the invertible matrix theorem (also called the fundamental theorem of invertible matrices).

Click here to go back to the quiz.

Lectures 13–18

- 1 The determinant is zero. (This may be easiest to check by expanding the determinant formula around the second row.) Therefore, the matrix is not invertible, so by the invertible matrix theorem, the rows are not linearly independent.

- 2 Check:

a For $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, if $a = c$ and $b = d$, the determinant $ad - bc = 0$.

b Express \mathbf{u} , \mathbf{v} , and \mathbf{w} in coordinates and verify that the indicated equations hold.

c Use the results of parts (a) and (b) to see that

$$\begin{aligned} D(\mathbf{u} + c\mathbf{w}, \mathbf{w}) &= D(\mathbf{u}, \mathbf{w}) + D(c\mathbf{w}, \mathbf{w}) \\ &= D(\mathbf{u}, \mathbf{w}) + cD(\mathbf{w}, \mathbf{w}) \\ &= D(\mathbf{u}, \mathbf{w}) + c \cdot 0 = D(\mathbf{u}, \mathbf{w}). \end{aligned}$$

- 3 Any vector of the form $(x, 0)$ is an eigenvector of A with eigenvalue -2 because it gets sent by the action of A to $(-2x, 0)$. Any vector of the form $(0, y)$ is an eigenvector with eigenvalue 3 because it gets sent by the action of A to $(0, 3y)$.
- 4 Any vector of the form (x, x) is an eigenvector of B with eigenvalue 1 because it gets sent by the action of B to (x, x) . Any vector of the form $(x, -x)$ is an eigenvector of B with eigenvalue -1 , because it gets sent by the action of B to $(-x, x)$.
- 5 The eigenvalues of A are 3 and -2 , which are obtained as the roots of the characteristic polynomial $\lambda^2 - \lambda - 6$. The eigenspace corresponding to 3 is the line that is the span of $(-1, 2)$. The eigenspace corresponding to -2 is the line that is the span of $(1, 3)$.
- 6 The eigenvalues of A are 3 and 3 . The eigenspace corresponding to 3 is the line that is the span of the single vector $(1, 0)$.

- 7 The matrix $\begin{bmatrix} 1 & -1 \\ -6 & 0 \end{bmatrix}$ is diagonalizable because it has a basis of eigenvalues.

However, the matrix $\begin{bmatrix} 3 & -1 \\ 0 & 3 \end{bmatrix}$ is not diagonalizable because it does not have a basis of eigenvalues. Also, the algebraic multiplicity of the eigenvalue 3 is 2, but the geometric multiplicity is 1.

- 8 If A and B are similar matrices, then $A = PBP^{-1}$ for an invertible matrix P .

a Taking determinants of both sides, you see $\det(A) = \det(P)\det(B)\det(P^{-1}) = \det(B)\det(P)\det(P^{-1}) = \det(B)$.

b Noting that $A - \lambda I = PBP^{-1} - \lambda PIP^{-1} = P(B - \lambda I)P^{-1}$, you see that $\det(A - \lambda I) = \det(B - \lambda I)$; thus, the characteristic polynomials of A and B are the same.

- 9 Under low predation, trajectories that start near the 0.7 eigenspace have barely enough rabbits to keep both populations from going extinct. The fox population initially decreases because there aren't enough rabbits to sustain a large number of foxes, but there are enough rabbits to sustain a smaller population of foxes. Once the fox population reaches that point, both populations start growing again.

- 10 The largest eigenvalue (and, in general, the largest eigenvalue in absolute value) is the most important for understanding the dynamics of the system because it governs whether or not under many initial conditions both species will flourish.

- 11 This system can be viewed as $\mathbf{w}'(t) = \begin{bmatrix} 1 & -1 \\ -6 & 0 \end{bmatrix} \mathbf{w}(t)$, where $\mathbf{w}(t) = (x(t), y(t))$.

The eigenvectors and eigenvalues of the matrix have been calculated in the problems for lecture 15, and these can be used to write the general solution:

$$\mathbf{w}(t) = C_1(-1, 2)e^{3t} + C_2(1, 3)e^{-2t}.$$

- 12 If λ is complex, you expect any corresponding eigenvector \mathbf{v} to be complex as well, because if \mathbf{v} had only real entries, then $A\mathbf{v}$ would only have real entries, but $\lambda\mathbf{v}$ would have at least one complex entry, contradicting the eigenvector equation $A\mathbf{v} = \lambda\mathbf{v}$.

Click here to go back to the quiz.

Lectures 19–24

- Because Q preserves dot products, apply that fact to the standard basis vectors by setting $\mathbf{x} = \mathbf{e}_i$ and $\mathbf{y} = \mathbf{e}_j$. Then, $Q\mathbf{x}$ is just the i^{th} column of Q and $Q\mathbf{y}$ is the j^{th} column of Q . Then, $(Q\mathbf{x}) \cdot (Q\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$ implies that the pairwise dot product of the columns of Q are 0 unless $i = j$, in which case the dot product is 1. So Q must be an orthogonal matrix.
- Because Q_1 and Q_2 are both orthogonal matrices, then $Q_1^T Q_1 = I$ and $Q_2^T Q_2 = I$. But then $Q_2^T Q_1^T Q_1 Q_2 = Q_2^T Q_2 = I$. From this, you see that $(Q_1 Q_2)^T (Q_1 Q_2) = I$, which shows that $Q_1 Q_2$ is orthogonal.
- The transition matrix must be symmetric.
- The transition matrix $\begin{bmatrix} 1/2 & 1/2 \\ 1 & 0 \end{bmatrix}$ is not positive, but its square is $\begin{bmatrix} 3/4 & 1/4 \\ 1/2 & 1/2 \end{bmatrix}$, which is positive.
- The derivative of this function T is the 2×2 identity matrix because the entries arise from the partial derivatives of x and y with respect to either x or y .
- The derivative matrices are

$$[Df] = \begin{bmatrix} 0 & 1 \\ y & x+1 \end{bmatrix}, [Dg] = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \text{ and } [Dh] = \begin{bmatrix} y & x \\ 0 & 1 \end{bmatrix}.$$

You can check that they indeed satisfy $[Df] = [Dg][Dh]$.

- 7 You should get the same answer that was found in the lecture.
- 8 If the columns of X are not linearly independent, that would describe the situation where for every data point the x value is the same. Clearly, such a cloud of data points cannot have a best-fitting line.

9 The Gram matrix $A^T A = \begin{bmatrix} 25 & 0 \\ 0 & 100 \end{bmatrix}$.

Because this is upper triangular, the eigenvalues of $A^T A$ lie along the diagonal: They are 25 and 100. The singular values are square roots of these: 5 and 10.

10 If you arrange singular values from largest to smallest, then you set $\Sigma = \begin{bmatrix} 10 & 0 \\ 0 & 5 \end{bmatrix}$.

Then, to find V , you must find normalized eigenvectors of $A^T A$. These are $(1, 0)$ corresponding to the eigenvalue 25 and $(0, 1)$ corresponding to the eigenvalue 100.

Putting those in columns corresponding to 10 and 5, you get $V = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

You can compute the columns of U as A times the columns of V divided by the corresponding singular value.

Thus, $U = \begin{bmatrix} -\frac{6}{10} & \frac{4}{5} \\ \frac{8}{10} & \frac{3}{5} \end{bmatrix}$.

- 11 Check all 10 axioms for a vector space.
- 12 Check that the integrals $\int \cos(x) \sin(x) dx$, $\int \cos(x) \sin(2x) dx$, and $\int \sin(x) \sin(2x) dx$ are each 0 when integrated over $x = 0$ to 2π .

Click here to go back to the quiz.

BIBLIOGRAPHY

Anderson, Marlow, and Todd Feil. “Turning Lights Out with Linear Algebra.” *Mathematics Magazine* 71, no. 4 (October 1998): 300–303. Explains the linear algebra behind solving the game Lights Out.

Brand, Matthew. “Fast Online SVD Revisions for Lightweight Recommender Systems.” *Proceedings of the 3rd SIAM International Conference on Data Mining*, 2003. Discusses methods for completing incomplete data when using the SVD.

Chartier, Tim. *When Life Is Linear: From Computer Graphics to Bracketology*. Mathematical Association of America, 2015. This book is a very readable and friendly introduction to linear algebra in the context of motivating examples from computer graphics and ranking systems, and it is great supplemental reading.

Colley, Susan. *Vector Calculus*. 4th ed. Pearson, 2011. A very good text on multivariable calculus that will be easy to read if you know some linear algebra.

Fowler, Michael. “Linear Algebra for Quantum Mechanics.” <http://galileo.phys.virginia.edu/classes/751.mf1i.fall02/751LinearAlgebra.pdf>. This is a good primer on linear algebra written for physicists learning quantum mechanics.

Kalman, Dan. “An Underdetermined Linear System for GPS.” *The College Mathematics Journal* 33, no. 5 (November 2002): 384–390. https://www.maa.org/sites/default/files/pdf/upload_library/22/Polya/Kalman.pdf. This article contains a discussion of how GPS works.

Kun, Jeremy. “Singular Value Decomposition Part 1: Perspectives on Linear Algebra.” April 18, 2016, <https://jeremykun.com/2016/04/18/singular-value-decomposition-part-1-perspectives-on-linear-algebra/>.

A readable discussion of the interpretation of the SVD for movie recommender systems.

Lamb, Evelyn. “How to Look at Art: A Mathematician’s Perspective.” *Scientific American*, *Roots of Unity* (blog), April 28, 2016, <https://blogs.scientificamerican.com/roots-of-unity/how-to-look-at-art-a-mathematician-s-perspective/>. Contains a good discussion about one-point perspective in art based on the work of Annalisa Crannell. See also this paper: Annalisa Crannell, Marc Frantz, and Fumiko Futamura. “Dürer: Disguise, Distance, Disagreements, and Diagonals!” *Math Horizons* (November 2014).

Lay, David, Steven Lay, and Judi McDonald. *Linear Algebra and Its Applications*. 5th ed. Pearson, 2015. A standard text on linear algebra. Useful for exercises and worked examples. There are also several good sections on applications.

Margalit, Dan, and Joseph Rabinoff. *Interactive Linear Algebra*, 2018. <https://textbooks.math.gatech.edu/ila/1553/index.html>. This is an online textbook with interactive graphics but without exercises.

McAnlis, Colt. “How JPG Works.” April 26, 2016. <https://medium.freecodecamp.org/how-jpg-works-a4dbd2316f35>. A readable explanation of the JPEG file format and how compression works.

Poole, David. *Linear Algebra: A Modern Introduction*. 4th ed. Cengage Learning, 2014. A standard text on linear algebra. Useful for exercises and worked examples. See the Applications sections as well as the Vignette sections for stories about applications.

Rabiner, Lawrence. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.” *Proceedings of the IEEE* 77, no. 2 (February 1989): 257–286. <https://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/tutorial%20on%20hmm%20and%20applications.pdf>.

Gives an overview of the basic theory of hidden Markov models and describes fundamental open questions.

Roughgarden, Tim, and Gregory Valiant. “CS168: The Modern Algorithmic Toolbox Lecture #9: The Singular Value Decomposition (SVD) and Low-Rank Matrix Approximations.” Course notes, <http://theory.stanford.edu/~tim/s17/l/l9.pdf>. Discusses the matrix completion problem and low-rank approximations using SVD.

Stock, Daniel. “Merlin’s Magic Square Revisited.” *The American Mathematical Monthly* 96, no. 7 (August/September 1989): 608–610. Explains how to solve the game Merlin by inspection.

von Hilgers, Philipp, and Amy Langville, “The Five Greatest Applications of Markov Chains.” *Proceedings of the Markov Anniversary Meeting*, 2006. <http://langvillea.people.cofc.edu/MCapps7.pdf>. Contains a discussion of Markov’s own interest in sequences of letters in text.

Yuster, Thomas. “The Reduced Row Echelon Form Is Unique: A Simple Proof.” *Mathematics Magazine* 5, no. 2 (March 1984): 93–94. <https://www.maa.org/sites/default/files/Yuster19807.pdf>. This is a short demonstration explaining why the reduced row echelon form of a matrix is unique.