

Linear Algebra for Data

Michael W. Mahoney¹

May 5, 2018

¹University of California at Berkeley, Berkeley, CA. E-mail: mmahoney@stat.berkeley.edu.

Contents

I Overview	13
1 Introduction and Overview	15
1.1 How to think about data	16
1.2 One motivating example in more detail	16
1.3 Trying to quantify the inference step ... with geometry	19
1.4 Graphs and connections with matrices	21
1.5 Problems	25
1.5.1 Pencil-and-paper Problems	25
1.5.2 Implementations and Applications of the Theory	25
II Linear Algebra: Extending Basic Algebra and Geometry	27
2 Matrices, vectors, and \mathbb{R}^n	29
2.1 Overview of the chapter	29
2.2 Ways to label points, elements, vectors, matrices, etc.	30
2.2.1 A familiar way	30
2.2.2 A more powerful way	31
2.2.3 Many points in a few dimensions, or a few points in many dimensions	32
2.2.4 Many points in many dimensions	33
2.3 What is \mathbb{R}^n ?	34
2.3.1 Examples of \mathbb{R}^n	34
2.3.2 Some very basic properties of \mathbb{R}^n	35
2.3.3 Some very basic subsets of \mathbb{R}^n	37
2.4 Measuring the size of vectors in \mathbb{R}^n	38
2.4.1 Norms	38
2.4.2 Balls	43
2.5 Visualizing elements of \mathbb{R}^n	45

2.6	Two basic operations on vectors in \mathbb{R}^n	47
2.6.1	Example illustrating possible operations on matrices	47
2.6.2	Vector addition and scalar multiplication	49
2.6.3	Using norms to measure distances	50
2.6.4	Using norms to normalize	50
2.7	Looking forward: algebra and geometry	52
2.8	Problems	52
2.8.1	Pencil-and-paper Problems	52
2.8.2	Implementations and Applications of the Theory	53
3	Vector spaces, matrices, and linear functions	55
3.1	Introduction to vector spaces and subspaces	55
3.1.1	Vector space	55
3.1.2	Subspaces of a vector space	56
3.1.3	Standard basis vectors and not-standard basis vectors	57
3.1.4	Some initial examples of subspaces and not-subspaces in two dimensions	60
3.1.5	Subspaces in \mathbb{R} , \mathbb{R}^2 , \mathbb{R}^3 , and beyond	63
3.1.6	Proving something is a subspace	65
3.2	Matrices and operations on matrices, including matrix multiplication	66
3.2.1	Two operations: matrices are vectors	67
3.2.2	A third operation: matrices are more than just vectors	67
3.2.3	Examples of matrix multiplication	69
3.2.4	A complementary perspective on matrix multiplication	72
3.3	Functions, linear functions, and linear transformations	73
3.3.1	Functions and transformations	73
3.3.2	Connection between functions and matrices	75
3.3.3	Transformations and matrices as transformations	75
3.4	Special types of matrices	78
3.4.1	Transpose of a matrix	78
3.4.2	Inverse of a matrix	78
3.4.3	Symmetric, triangular, diagonal, and Identity matrices	81
3.5	Examples of matrices as transformations	83
3.6	Problems	86
3.6.1	Pencil-and-paper Problems	86
3.6.2	Implementations and Applications of the Theory	89

4 Geometry: angles, spans, bases, and projections	91
4.1 Geometry of \mathbb{R}^n : dot products, angles, and perpendicularity	91
4.1.1 Dot products	91
4.1.2 Angles	93
4.1.3 Orthogonality between two vectors in \mathbb{R}^n	95
4.2 Linear combinations, spans, and linear dependence/independence	96
4.2.1 Linear combinations	97
4.2.2 Span	98
4.2.3 Linear dependence and independence	100
4.2.4 Testing for linear dependence and independence	103
4.3 Bases, orthogonal bases, and orthonormal bases	104
4.3.1 Basis vectors	104
4.3.2 Orthogonal and orthonormal bases	107
4.3.3 Computing an orthonormal basis from an orthogonal basis	109
4.3.4 Computing an orthonormal basis from any set of columns	111
4.3.5 Orthogonality more generally	115
4.4 Projections	115
4.4.1 Projecting onto a vector, i.e., onto a one-dimensional subspace	116
4.4.2 Projecting onto higher-dimensional subspaces	118
4.5 Problems	121
4.5.1 Pencil-and-paper Problems	121
4.5.2 Implementations and Applications of the Theory	124
III Basic Probability: A Way to Understand High-Dimensional Spaces	127
5 Introduction to probability	129
5.1 Overview of the chapter	129
5.2 Simple models for understanding probability	130
5.2.1 Flipping coins, rolling dice, and throwing darts	130
5.2.2 A warm up	132
5.3 Foundations of Probability	133
5.3.1 Sample spaces and events	133
5.3.2 Some basic set theory	136
5.3.3 Basics of probability	137
5.3.4 Mass/Volume as an intuition	141

5.4	Conditional Probabilities	142
5.4.1	Conditional Probability	142
5.4.2	Independence	144
5.4.3	Bayes' Theorem	145
5.4.4	Computing more complex probabilities	149
5.5	Problems	150
5.5.1	Pencil-and-paper Problems	150
5.5.2	Implementations and Applications of the Theory	154
6	Random variables and their properties	157
6.1	Random Variables	157
6.1.1	Random variables are just functions	157
6.1.2	Two different definitions of the probability of an event	158
6.1.3	Measure concentration: a first example	162
6.1.4	Other stuff on random variables	162
6.2	Moments of random variables	164
6.2.1	Mean/expectation	165
6.2.2	Variance and other moments	166
6.2.3	Two basic properties of expectations of random variables	168
6.2.4	Conditional expectation	170
6.2.5	How good are your estimates of the mean and the variance?	172
6.3	More complex combinations: Covariances and correlations	174
6.4	Problems	179
6.4.1	Pencil-and-paper Problems	179
6.4.2	Implementations and Applications of the Theory	179
7	Quantifying variability and concentrating measure	181
7.1	Large, small, and typical variability	181
7.2	Bounding deviations from the mean (Part 1: weak bounds)	183
7.2.1	Markov's Inequality.	183
7.2.2	Chebychev's Inequality.	185
7.3	A baseline to aim for	188
7.3.1	The Gaussian/normal distribution	188
7.3.2	Types of claims one can make: limiting versus non-limiting statements	191
7.4	Bounding deviations from the mean (Part 2: strong bounds)	192
7.4.1	Chernoff bounds	193

7.4.2	Discussion of Chernoff bounds	195
7.5	Examples of random variables	196
7.6	Problems	200
7.6.1	Pencil-and-paper Problems	200
7.6.2	Implementations and Applications of the Theory	202
8	A retrospective: Probability and high-dimensional linear algebra	205
8.1	Connections between probability theory and linear algebra	205
8.1.1	A geometric approach to probability	206
8.1.2	A probabilistic approach to Euclidean geometry	206
8.2	Properties of high dimensions versus low dimensions	207
8.3	More on measure, concentration, and measure concentration	208
8.3.1	Concentration in flipping coins	208
8.3.2	Concentration in throwing darts	210
8.4	Advanced Aside: Insights from calculus	213
8.5	Problems	214
8.5.1	Pencil-and-paper Problems	214
8.5.2	Implementations and Applications of the Theory	215
IV	The Spectral Theorem: The Central Result in Linear Algebra	219
9	Eigendecompositions: Eigenvectors and Eigenvalues	221
9.1	Overview of the chapter	221
9.2	Introduction to eigenvectors and eigenvalues	221
9.3	Some simple examples of eigenvectors and eigenvalues	224
9.4	Computing eigenvectors and eigenvalues	230
9.5	Basic properties of determinants	230
9.6	Using determinants to understand eigendecompositions	233
9.7	Using determinants to compute simple eigendecompositions	234
9.8	Expressing matrices in terms of their eigendecompositions	238
9.9	A larger example	246
9.10	Problems	248
9.10.1	Implementations and Applications of the Theory	248
9.10.2	Pencil-and-paper Problems	249
10	Eigendecompositions: The Quadratic Forms Perspective	251

10.1 Quadratic forms and matrices	251
10.2 Some simple examples	257
10.3 Symmetric bi-linear functions	259
10.4 Connections with conic sections	261
10.5 Definiteness, indefiniteness, and quadratic forms as a sum/difference of squares	266
10.6 Two other topics	271
10.7 Problems	273
10.7.1 Implementations and Applications of the Theory	273
10.7.2 Pencil-and-paper Problems	273
11 The Spectral Theorem: EVD and SVD	275
11.1 The EigenValue Decomposition (EVD)	276
11.1.1 Efficiently Expressing the EVD	276
11.1.2 Finding the EVD	279
11.1.3 Computing the EVD	287
11.2 Singular Value Decomposition (SVD)	288
11.2.1 The basic SVD	288
11.2.2 Equivalent ways to view the SVD	289
11.2.3 SVD and thin SVD and rank-deficient thin SVD	289
11.3 Additional properties of the SVD	290
11.3.1 SVD and the structure of \mathbb{R}^m and \mathbb{R}^n	290
11.3.2 SVD and the norm of a matrix	291
11.3.3 SVD and inverses of non-invertible matrices: the pseudoinverse	293
11.4 Problems	295
11.4.1 Implementations and Applications of the Theory	295
11.4.2 Pencil-and-paper Problems	295
V Applications: PCA, Least-Squares, Linear Equations, PageRank, High-Dimensional Calculus, Et Cetera	297
12 Principal Components Analysis	299
12.1 The basic PCA method	300
12.2 PCA and PD/PSD matrices	302
12.3 When vanilla PCA is not particularly appropriate to use	304
12.4 Statistical interpretation of PCA	305
12.5 More on variable transformations underlying PCA	310

12.6 Problems	315
12.6.1 Implementations and Applications of the Theory	315
12.6.2 Pencil-and-paper Problems	315
13 Least-squares (LS) regression	317
13.1 Least-squares (LS) regression	317
13.1.1 Trying to model data to make predictions	317
13.1.2 The basic LS method: simple linear regression	320
13.1.3 The basic LS method: multiple linear regression	323
13.1.4 Complementary Perspectives on LS	324
13.1.5 When is LS the “right thing” to do?	326
13.2 Comparison of PCA and LS	326
13.3 Regression Diagnostics and Related Methods	327
13.4 Regularized LS Regression	327
13.5 Problems	328
13.5.1 Implementations and Applications of the Theory	328
13.5.2 Pencil-and-paper Problems	329
14 Systems of Linear Equations	331
14.1 Simple Example of System Linear Equations	331
14.2 Basics Ideas Underlying of Linear Equations	331
14.3 Some Mechanical Procedures to Solve Linear Equations	334
14.4 Direct Methods to Solve Linear Equations	334
14.5 Iterative Methods to Solve Linear Equations	334
14.6 Numerical Issues	334
14.7 Problems	335
14.7.1 Implementations and Applications of the Theory	335
14.7.2 Pencil-and-paper Problems	335
15 PageRank for Ranking, Clustering, and Classifying	337
15.1 Random Walks, Diffusions, Markov Chains, and Other Approaches to PageRank	337
15.2 Graphs and representing graphs as matrices	339
15.3 Probability Perspective	342
15.4 Linear Algebra Perspective	345
15.5 Usefulness of these results	349
15.6 Problems	351
15.6.1 Implementations and Applications of the Theory	351

15.6.2 Pencil-and-paper Problems	351
16 High-dimensional Calculus: Integration and Differentiation	353
16.1 Integration and Differentiation in One Dimension	353
16.2 An Obvious but Not Good Way to Extend to Higher Dimension	353
16.3 High-dimensional Integration with Markov Chain Monte Carlo	353
16.4 High-dimensional Differentiation with Stochastic Gradient Descent	353
16.5 Problems	354
16.5.1 Implementations and Applications of the Theory	354
16.5.2 Pencil-and-paper Problems	354
VI Additional Miscellaneous Stuff To Incorporate Somewhere	355
17 Additional Homework Questions to Incorporate Somewhere	357
17.1 From putting together s18 final and final prep	357
17.2 From putting together s18 midterm	375
17.3 XXX. MORE	380
18 Additional Material to Incorporate Somewhere	383
18.1 Notes of things to cover or fix in the future	383
18.2 Ideas to take from other books	384
18.3 Additional Miscellaneous Stuff Maybe To Put Somewhere	385
18.3.1 Misc Stuff	385
18.3.2 Some higher-level things that will take some thought	385
18.4 Additional Miscellaneous Stuff Probably To Remove	386
18.4.1 Some philosophy	386
18.5 XXX REVIEWS FROM PREVIOUS WEEKS/CHAPTERS	388
18.5.1 Review from last chapter	388
18.5.2 Review from last chapter	388
18.5.3 Review from last chapter	389
18.5.4 Review of the chapter XXX ON QUADRATIC FORMS	389
18.6 Advanced Aside: Discussion of other uses of the spectral decomposition	392
18.6.1 Describing a matrix as a linear transformation	392
18.6.2 Describing a matrix as a quadratic form	393
18.6.3 Computing higher powers of a matrix, and iterative algorithms	394
18.6.4 Generalize differential calculus to multiple variables	395

18.6.5 Matrices that are non-symmetric	396
--	-----

Part I

Overview

Chapter 1

Introduction and Overview

This class is designed to serve as an introduction to what may be termed the mathematics of data. By that, I mean that we want to identify some basic mathematical topics that are common to all areas of Data Science, and in particular understand how they are used in statistical theory, computer science theory, and Data Science practice. Some of the topics we will cover are actually quite advanced, but they are ubiquitous and extremely important, both in theory and in practice. Thus, we will take an elementary approach. In many cases, the mathematical ideas we will discuss appear “under the hood” rather than explicitly. For example, they might manifest themselves in the improved running time of a complicated algorithm, or they might manifest themselves in bounds for showing that one obtains good inference, or they might simply explain why some heuristic that is done in practice works well or doesn’t work well. Having some understanding of “why” different methods do and don’t work can be quite important, e.g., to understand tools that are used as black boxes, to diagnose problems, to develop new methods, etc.

In particular, since it is so ubiquitous, basic linear algebra will be our main focus in this class, but an important secondary and complementary focus will be on some discrete probability and optimization. Among other things, we will be interested in how these methods are used with matrices and graphs. Matrices and graphs are two very common ways to model data. There are some important differences between them, but they also have many important similarities, and they provide a nice introduction to how linear algebra and probability and optimization are used in all sorts of areas of Data Science.

Perhaps surprisingly, much of the linear algebra that is taught in traditional linear algebra classes is actually of rather limited usefulness for the practicing Data Scientist. The reason is that these more traditional classes have been designed for different reasons. For example, linear algebra is often used in engineering, physics, and related areas, and from this perspective numerical aspects of it, connections with differential equations, solving general linear equations, etc., are of particular importance. Similarly, a lot of probability theory emphasizes continuous random variables and/or different types of distributions. While important in certain ways, this approach can hide the intuitions of when these methods are and are not useful in Data Science, and this approach also hides the connections between these probability ideas and very related linear algebra ideas. Similarly with optimization. In this class, we’ll cover linear algebra from the perspective of Data Science. This will involve making connections with parts of probability and optimization that are most useful for Data Science. In particular, for probability and optimization, we’ll cover a small fraction of what you will see in more traditional versions of those classes, we’ll cover it in a different way, and we’ll also highlight connections that aren’t typically made in those classes; and for linear algebra, we’ll cover the basic topics, but in a very different order and with a quite different emphasis.

Depending on your interests, this class has one of two complementary goals. If this is your last class on theoretical aspects of data, the goal is to give you a basic understanding and intuition of why things work, so that you can use them in practical applications you will encounter in the future. If this is your first of many classes on theoretical aspects of data, the goal is to give you a basic understanding and intuition of why things work, so that in later classes you can build on them to gain a more rigorous theoretical understanding.

1.1 How to think about data

Perhaps you don't know what the subjects of linear algebra and probability and optimization involve. Even if you do, perhaps it is not obvious to you why they are so central to the foundations of data. Let's start with this. To do so, let's ask: how do we represent and think about data?

- One answer is that data is/are a bunch of stuff in a file on which we run jupyter, python, R, ipython, etc. Representing data in a computer in this way involves using lists, tables, floats, ints, etc., and other things like that that computers can understand and on which they can operate. This is a very common way to represent data, and one can get a lot of insight into data by doing relatively-simple operations on data represented in these ways. For example, one can select from a table all the rows that satisfy a given condition, and then if there is another field in the table corresponding to time one can ask how does this number of rows vary with time. Many things one wants to do fit within this framework of selecting, filtering, counting, etc.
- Often, one wants to do more complex operations. For example, one might want to represent the dependence between one variable and several other variables, or one might want to represent all the pair-wise interactions between data elements. In those cases, matrices and graphs are two very basic mathematical structures that arise naturally. Basically, they arise since—informally—they are at a “sweet spot” between what I'll call *descriptive flexibility* (meaning, informally, how well does it describe the data, or how much is lost in the mathematical abstraction step, which relates at least informally to how well statistical inference can be controlled) and what I'll call *algorithmic tractability* (meaning, essentially, how quickly can one compute things of interest).

That is, matrices and graphs can describe real data relatively well and one can compute things of interest relatively quickly. This is an important tradeoff, often formalized in more advanced classes in statistical learning theory. Moreover, more sophisticated representations of data can be built up from them. For these and other reasons, understanding basic properties of matrices and graphs, including both linear algebraic and probabilistic/optimization issues related to their use in Data Science, will be central to what we will cover.

Of course, neither of these things (characters or words in a file, or more abstract things like matrices and graphs) correspond immediately to what one might view as primary data, i.e., things like an image or the record of a purchase or the genetic properties of a person or other things that. We'll get to several examples of this connection soon.

1.2 One motivating example in more detail

As a motivating example of both more simple and more complex operations, imagine that you have the text of a book as a long string of characters. For example, in the main class (at least last year), you may have heard about a simple analysis of the *Little Women* book. Let's look in more detail about how the data could be represented, transformed, and interpreted, as this illustrates a common pipeline in data science. Here is a summary of that pipeline.

1. Represent the entire text as a long string, representing words at the sentence level, meaning that each word is a string, separated by a space or maybe other characters like a period.
2. Split the entire text into, e.g., 47 different strings based on a keyword, CHAPTER, which is used to denote the beginning of each chapter. This gives a table with 47 rows, each of which has a long string of words.
3. In each row, count the number of times each of the, e.g., 4 names of the main characters appear, doing so by performing a count or sum operation on the number of times that word appears in that line. This gives 4 ordered sequences of 47 numbers, each one representing a running sum.

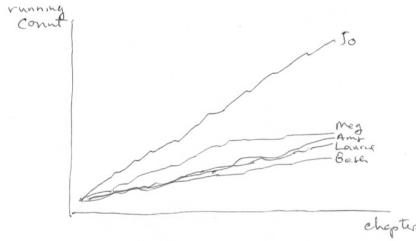


Figure 1.1: Illustration of running total of frequency of different names.

4. Plot each of those 4 ordered sequences versus time to get a plot like in Figure 1.1.
5. Look at that plot and make some hypothesis about what was going on in the book, e.g., about who fell in love and lived together and ran off to New York. This was done by looking at the “similarity” of two of the curves in the last few chapters of the book.

Observe that the data are originally stored as a long character string, i.e., not as numbers or something else; and then the data are effectively stored as what may be called a *flat table*, each row of which is a *key*, in this case a string corresponding to a number that is the chapter number, followed by some sort of *value*, in this case a long character string that corresponds to an ordered list of words.

This idea of a flat table is a very general and common way to think about data. It is a table, in that it has several rows or records and one or more columns describing properties of the thing described by that row/record; and it is called flat since there is usually no structural relationships between the rows/records.

Here is another example of a flat table, as would be illustrated in a database course.

Name	SSN	Year	Major	HW1	HW2	HW3	HW4	Test1	Test2	Grade
Alice	123	Fr	Math	95	90	70		100		
Bob	234	Fr	Comp Sci	80		65		80		
Bob	345	So	Undecided	50	60	60		70		
Charlotte	456	So	Stats	85	70	55		80		
...										
Zaccheus	789	Fr	Undecided	100	70	55		70		

Note that the elements of this table could be represented by two indices. So, if we wanted to call the table A , then we could represent an element of the table by A_{ij} . That can be helpful sometimes, but it's power is limited. In particular, there aren't many mathematical operations defined on this table A or on the rows i or columns j that describe interactions between parts of the table in a rich way. For example, one could add the numbers in the HW1 column/field, e.g., to get the average score on that homework, but it doesn't make much sense to add "Alice" and "Bob" and so on. In particular, while this table *looks* like a matrix that we will get to soon, it isn't "really" a matrix—since a matrix qua matrix isn't defined as a thing that has two subscripts, but instead by the *operations* that are allowed on it. This is true for a flat table, but there the operations were rather limited. For matrices and graphs, the operations will be much richer.

Question: For the tables above, what operations are allowed?

Answer: There are many, but they are typically combinations of a small number of primitive operations. Examples of those primitives are the following.

- **Query.** Here, e.g., we might want to ask if a word appeared in a row or it appeared more than a certain number of times.
- **Filter.** Here, e.g., we might want to select just those rows/chapters where a given word appeared or appeared more than a certain number of times.
- **Join.** Here, e.g., we might want to combine two rows into one, which might be of interest if we mistakenly split one chapter into two.
- **Count or Sum.** Here, e.g., we might want to count the number of words or the number of times a given word appears.

Observe that these are the types of operations that were used to generate the plots in Figure 1.1.

Flat tables and extensions of them are very important in data science. Here are two places in particular where they are widely-used.

- **In databases.** In databases, i.e., that area of computer science that studies how to store, manipulate, etc. data, they are very common, in particular when the data are very large. In particular, if the data are really large then they are often stored somewhere in a database that the data scientist can access with queries and related operations.
- **For simple operations.** When you are first exploring a new data set, even if it is rather small. Often you want to see what you have, do a few simple operations to determine the size and shape of the data, do some initial visualization, etc., to determine whether there is anything crazy, outliers, etc.

In spite of the advantages, there are some disadvantages to working with flat tables. In particular, there are two related things to note about these.

- **Assumptions about noise.** These operations listed above (i.e., query, filter, join, count, etc.) essentially assume that there is no “noise” in the data. By that, I don’t mean that the data are noise-free, and I don’t mean that the data scientist doesn’t keep noise in the back of his or her mind. Instead, by that, I mean that the transformations transform the data “exactly,” and so, e.g., quirks in the data, including errors/noise/etc. get propagated into the results. That fact has pros and cons:
 - *Pros.* If you have a bank account that has \$123,456.78, and I give you a check for \$100.00, then you hope to see \$123,556.78 in your account once you deposit the check. That is, you won’t be happy if the bank tells you that there is a lot of noise in data and therefore they are rounding your balance to \$100,000.00, i.e., to two significant digits. That “noise” gets propagated means that the least significant digits in your account are correct.
 - *Cons.* If you want to do some sort of operation that removes noise or tries to do inference or prediction to unseen data, then it is very difficult. Of course, when we look at plots generated by these operations, our eyes are often robust to those quirks, and we might form a hypothesis that “denoises” such noise. Importantly, though, the algorithmic tools we used to query the data didn’t do that, and it is of great interest to develop algorithmic and statistical tools that are more robust in that sense, e.g., in cases that are less easy to visualize.
- **More sophisticated statistics.** These operations work on one record, or they just join two records and sum, but they do not permit us to do more sophisticated things. For example, they don’t in general permit us to quantify what we mean by the two curves look similar in the last few chapters of the book; they don’t permit us to describe more sophisticated correlations in the data, etc. To do this, we need to model the data in a way that permits us to run meaningful operations that have an interpretation in terms of denoising or inference or prediction.

Linear algebra will permit us to do these things.

1.3 Trying to quantify the inference step ... with geometry

Continuing with our motivating example about the text of a book, let's go into some more detail on why those two curves for those two characters are close in some sense. Recall Figure 1.1. Let's ask ourselves why we think that the two curves are close? Here are several possible answers.

- They have similar magnitudes.
- They have similar changes.
- They have similar changes at the time and chapter at which we looked.

Big Question: While eyeballing is good, and in this case it worked, how would we quantify these things to do this more generally, e.g., for data that may be harder to visualize?

In light of this question, and before we move on, note that when we plotted as a function of the chapter the running total of word counts, the goal was to look at the plots and achieve some sort of insight/inference, at least in an informal sense. What extra assumptions went into the plots? By that, I mean: what else are we implicitly assuming by viewing the data in the way we did? Here are two things.

- The time/chapter on the X axis was more meaningful than viewing things in some other way, e.g., alphabetically or lexicographically. Clearly, we could have plotted the same data in these other ways, but presumably it would have been less useful.
- The running total might (or might not) be more meaningful than plotting individual frequencies, or differences between frequencies between subsequent chapters, etc. Again, we could have plotted it either way.

One might wonder: what effect do decisions like these have? In different cases, one option or another might be more appropriate, and it is worth playing with both/several. A few things to note about these two decisions:

- Time was not in the flat table, i.e., the record was just a character string that was a number corresponding to the chapter, and we separately interpreted it as a real number or integer with an ordering.
- The particular way the sum was done, i.e., starting from the first chapter, introduced the time order and running total into the plots.
- The running total corresponds to an integral/sum, and integrals/sums tend to smooth things, which might be good to do denoising away some noise. Recall Figure 1.1.

Let's get a little more precise about these things.

For each of the 4 individuals, we have an ordered set of 47 numbers. To use notation that we will describe in more detail next time, let's call this a vector

$$x = (x_1, x_2, \dots, x_{47}) \in \mathbb{R}^{47}.$$

Here, the notation \mathbb{R}^n means that we have an ordered list of n numbers that we will view as a vector. (The notation \mathbb{R} means that, while the numbers are positive integers corresponding to counts, we will think of them as real numbers; we will spend a lot more time on this issue later.)

For simplicity, let's start with $n = 2$. When the data consist of 2 real numbers, i.e., x_1, x_2 on the Euclidean plane, then we have a number of things with which you should be familiar from high school mathematics. See, e.g., Figure 1.2.

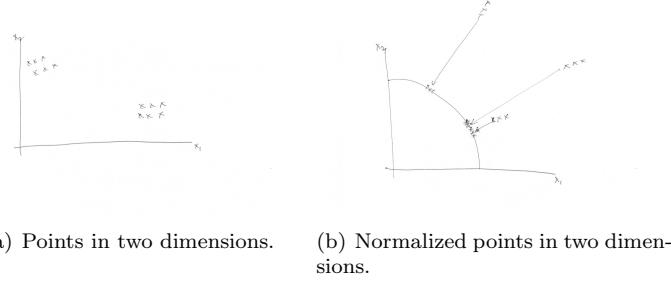


Figure 1.2: Points in two dimensions, unnormalized and normalized to the unit circle, illustrating distance and angles between points.

- **Dot product between two vectors.**

If $x = (x_1, x_2)$ and $y = (y_1, y_2)$ are points on the plane, then the dot product or inner product between those two vectors is

$$x \cdot y = x_1 y_1 + x_2 y_2 = \sum_{i=1}^2 x_i y_i.$$

- **Norm of a vector.**

If $x = (x_1, x_2)$ is a point on the plane, then one way to measure the norm or size of x is

$$\|x\|_2 = (x_1^2 + x_2^2)^{1/2} = \left(\sum_{i=1}^2 x_i^2 \right)^{1/2} = (x \cdot x)^{1/2}.$$

- **Angles between two vectors.**

If $x = (x_1, x_2)$ and $y = (y_1, y_2)$ are points on the plane, then the (cosine of the) angle between those two vectors is

$$\gamma = \cos(\theta) = \frac{x \cdot y}{\|x\|_2 \|y\|_2}.$$

Note that the dot product (as well as the norm and the angle that depend on it) is a more complicated operation than operations such as query, filter, join, and count. (Of course, it is more restricted, e.g., in that we are working with data that are real numbers and not more general things like strings, etc.) This will allow us to do more complicated types of data analysis. Moreover, it has strong connections with geometry and geometric ideas like distances, angles, etc.

Informally, and based on what we know about two-dimensional examples, we might expect that two things that are closer, i.e., that have a smaller distance between them, are more similar. See, e.g., Figure 1.2(a). On the other hand, since some things are just bigger, and since this might be “real” or it might be just a question of measuring things in different “units,” the angle between two points in a sort of distance between two points that are normalized that might be more appropriate. See, e.g., Figure 1.2(b). We will see both of these notions are useful more generally, but we will also see that there are important caveats to using these notions more generally.

As we will see, these ideas are *not* special to points on the two-dimensional plane, or even points in three-dimensional space, and indeed they generalize to high-dimensional spaces. For example, for vectors in \mathbb{R}^{47} , we will be able to *define* the following.

- **Dot product between two vectors.**

If $x = (x_1, x_2, \dots, x_{47})$ and $y = (y_1, y_2, \dots, y_{47})$ are points in \mathbb{R}^{47} , then the dot product or inner product between those two vectors is

$$x \cdot y = \sum_{i=1}^{47} x_i y_i.$$

- **Norm of a vector.**

If $x = (x_1, x_2, \dots, x_{47})$ is a point in \mathbb{R}^{47} , then one way to measure the norm or size of x is

$$\|x\|_2 = \left(\sum_{i=1}^{47} x_i^2 \right)^{1/2} = (x \cdot x)^{1/2}.$$

- **Angles between two vectors.**

If $x = (x_1, x_2, \dots, x_{47})$ and $y = (y_1, y_2, \dots, y_{47})$ are points in \mathbb{R}^{47} , then the (cosine of the) angle between those two vectors is

$$\gamma = \cos(\theta) = \frac{x \cdot y}{\|x\|_2 \|y\|_2}.$$

These ideas of norms and angles and other related notions that we will get to in the next few classes will help us to generalize some of the ideas that you are familiar with to data that are represented by high-dimensional vectors and high-dimensional vector spaces. In particular, these can be used to provide a notion of closeness between two high-dimensional vectors (we will see others that have some similarities and some differences). Importantly, this notion depends on \mathbb{R}^{47} having a Euclidean vector space structure (we'll say more later on what exactly that means) and not just a flat table structure. That is, the operations of norms and dot products and angles go well beyond flat table operations and will depend on the geometry of \mathbb{R}^{47} (or, more generally, of \mathbb{R}^n , as there is nothing special about $n = 47$). This will permit us to define certain types of transformation and operations that will permit us to use the matrix properties of, say, the 4×47 table to answer questions about correlations, denoising, inference, etc.

As we go forward, it is worth keeping in mind three canonical examples of matrices.

- **Term-document matrices.** This is what we have been talking about, where we have m things (e.g., documents, but they could be other things, as we will discuss), each of which is described n attributes (e.g., frequency of words, although it could be other things also).
- **Correlation matrices.** These are matrices constructed from the variance and covariance properties of random variables. They provide an important connection between linear algebra and probability and optimization. We will get to those notions and thus to this class of matrices later.
- **Matrices associated with a graph.** We will turn to these matrices next.

1.4 Graphs and connections with matrices

Let's switch gears and consider something that is seemingly-different but will turn out to have a lot of similarities: Graphs.

To start, by a graph, we do *not* mean a plot, e.g., of a function $y = f(x)$, on the x-y plane. See Figure 1.3 for two different meanings of the word “graph.” Instead, by graph, we will mean a structure that can be used to model data that consists of a bunch of “things” and pairwise “connections” between those things.

Somewhat more formally, a graph G is a set of things, often denoted V , called nodes, and a bunch of edges, often denoted E , where each edge in E consists of two nodes from the set V . The edges might be directed (see Figure 1.3(c), in which case it matters which node is pointing to which one) or undirected (in which case it does not, as in Figure 1.3(b)), weighted or unweighted, etc. Graphs are a very popular way to model things that interact via pair-wise interactions.

Here are several examples of data that are often modeled as graphs.

- Social networks: people and friendships.
- Protein-protein networks: proteins in an organism and whether two proteins have some sort of biochemical interaction.

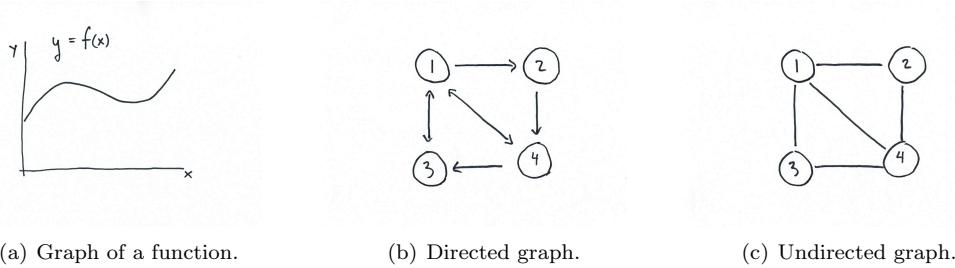


Figure 1.3: Several different notions of a graph. By graph, we will *not* be primarily interested in that of Figure 1.3(a).

- Network of closeness of characters in a book. For example, closeness could be measured by some sort of similarity measure, e.g., distance or angle between other vectors, that would be appropriate to determine whether those two characters were close in the book.
- WWW or Internet. For example, think of the Web as a directed graph, where the nodes are web pages and where the edges are directed links between different web pages.

Let's look at the last example in some more detail. Here is a question: how do we determine how "important" is a web page, either in general or for a specific topic or query. Here are possible answers.

1. Count the frequencies of certain words on a page, and the page with the largest count of certain words is the most important.
2. Count the frequency of other pages that point to a given page, and the page that has the most links to it are the most important ones.
3. Important pages are in some sense pages that are linked to by important pages.

All three of these have been used historically. One advantage of the first two metrics is that they are relatively-easy to compute, e.g., they just involve join and count operations. Both have to do with a given page, or a simple count of things that link to a given page, and they can be computed with flat table operations. Moreover, it has similarities with what we saw earlier with the text analysis of the book. For example, for the first metric, to find important pages, here is what we would do.

1. Crawl all the web pages on the web.
2. Read all the words on all of those pages.
3. When the user enters a query, find pages with the most mentions of that word or related words.

Alternatively, for the second metric, here is what we would do.

1. Crawl all the web pages on the web.
2. For each page, scan over the other pages and count the number of other pages that link to it, saving that number as another field in the record.
3. When the user enters a query, find pages with the largest value in that field.

Here are some things to note about these two metrics.

- Both involve relatively simple operations (like query, filter, join, and count) on the graph.

- Both may or may not be relevant before people start ranking pages and evaluating those rankings, but once people do that then both are relatively easy to “spam.” That is, you can easily make a web page with a given word repeated 100,000 times, but that doesn’t mean that the page is of particular interest to that word. Similarly, it is only slightly more difficult to make 100,000 web pages and have them all link to a given page, but that doesn’t make that page of particular interest to a given topic. This means that these two metrics may be less useful for the original goal of helping to determine important pages.

The third option is somewhat more subtle: it is recursive, in that it doesn’t define what is the importance of pages that link to the original page are, except recursively. As we will see, this third notion can be given a meaningful interpretation as a certain matrix computation on a graph associated with the Web/Internet.

In particular, here are two related developments in term-document and web page analysis, the first of which is a form of term-document matrix analysis, and the second of which is basically our third metric.

- Do term-document analysis on the content of the web page. Here, we might form a matrix—in the sense of linear algebra— A , where the element A_{ij} corresponds to the frequency of the i^{th} word in the j^{th} document, and we might do some linear algebra to identify ‘latent’ structure in that matrix. This involves more sophisticated computations than simply counting, and thus this might be better than our first metric.
- Do link analysis of the entire web. Here, we might do linear algebra on a matrix constructed from a graph associated with the web. The idea is that the importance of a page is not given by the content of the page, whether they are modeled well or poorly, but instead by the entire universe of other pages and how they link together. This too involves more sophisticated computations than just counting.

Both of these generalize simpler options and are much more common in practice, and both of these involve matrix and graph computations. To understand each of these, as well as the many extensions of these basic ideas, one needs to know something about linear algebra and probability and optimization.

To illustrate this latter notion, and in particular the connections between graphs and matrices, consider the directed graph shown in Figure 1.3(c). Given the graph in Figure 1.3(c), we can define the following matrix, which is known as the *adjacency matrix*, since the (i, j) element of this matrix is 1 or 0, depending on whether or not there is an edge *to* node i *from* node j in the graph (i.e., whether the two nodes, taking into account directedness, are adjacent).

$$A = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}. \quad (1.1)$$

Note that this matrix is not “symmetric,” in the sense that $A_{ij} \neq A_{ji}$ for some i and j . This is since the graph is directed. If, instead, we consider the undirected version of this graph, as shown in Figure 1.3(b), then we would have the following adjacency matrix.

$$A' = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}. \quad (1.2)$$

Let’s go back to the adjacency matrix A of the directed graph. The following matrix is known as the *diagonal degree matrix*, since it has 0 entries everywhere, except along the diagonal, where it has the total number of

edges going out of the i^{th} node.

$$D_{out} = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}. \quad (1.3)$$

There is a lot of information that can be extracted from be extracted about a graph from it's adjacency matrix. For example, one can extract the diagonal degree matrix by summing the elements along each row, but one can obtain much more. For example, the following matrix is a little more interesting (although that might not be obvious yet); it is known as the *random walk matrix*.

$$W = \begin{pmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1 & 0 & 0 \end{pmatrix}. \quad (1.4)$$

This random walk matrix can obtained by taking the adjacency matrix and dividing each (vertical) column by the sum of the entires in that column (which, note, also equals the corresponding diagonal value in the diagonal degree matrix). We will represent this later more conveniently by an operation known as matrix multiplication, but for now just follow the process just described.

This matrix is quite important, and it has many uses. Among it's uses, we will see later that it can be used to provide precise sense in which important pages are those lined to by important pages.

Before we get to that, note the following. In particular, let's switch gears again to consider something that is seemingly-different, but that also has a lot of connections with what we have been discussing. Say that we go back to simple algebra from high school, where we want to find solutions to equations, where it is common to write out, e.g., two equations in two unknowns. The web has billions of pages, so it would be hard for a person to solve, but computers can handle this fairly easily. For now, let's consider this simple example with four nodes. Then, we could write out the following set of *linear* equations.

$$\begin{aligned} y_1 &= & 1 \cdot x_3 + \frac{1}{2} \cdot x_4 \\ y_2 &= \frac{1}{3} \cdot x_1 & \\ y_3 &= \frac{1}{3} \cdot x_1 & + \frac{1}{2} \cdot x_4 \\ y_4 &= \frac{1}{3} \cdot x_1 + 1 \cdot x_2 & \end{aligned} \quad (1.5)$$

Next, note that the coefficients here are the same as those in the matrix in Equation (15.3), suggesting that we can write Equation (1.5) as $y = Wx$ or as

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} & & & \\ & & & \\ & W & & \\ & & & \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}. \quad (1.6)$$

If we define for a matrix $A \in \mathbb{R}^{m \times n}$ and a vector $b \in \mathbb{R}^n$ the product as $c = Ab$ where $c \in \mathbb{R}^m$, where $c_i = (Ab)_i = \sum_{j=1}^n A_{ij}b_j$, for each $i = 1, \dots, m$, then we have exactly that $y = Wx$. This is an example of what we will call a matrix-vector product that we will get to in more detail soon.

This is tedious but not impossible to do for \mathbb{R}^4 ; doing it for \mathbb{R}^{47} is already challenging even to think about; but there exist more than 10^9 web pages, i.e., we would be doing it in \mathbb{R}^{10^9} , and so we would like an easy way to think about and express these manipulations. For example, if we write $y = Wx$ and forget about the dimensionality.

Two things to note.

- The original vector x is a vector in \mathbb{R}^4 can be interpreted as a score or importance score for each node of the graph.
- The interpretation of $y = Wx$ is that we take one vector in \mathbb{R}^4 and we apply W to get another vector in \mathbb{R}^4 , so we have applied some sort of function that transforms one 4-vector into another 4-vector.

This holds for all vectors in \mathbb{R}^4 , and note that y is “linear” in x . Linear algebra is basically about generalizing this to much more general situations.

However, if, in addition, the following is true:

$$0 \leq x_i \leq 1 \quad \text{and} \quad \sum_{i=1}^4 x_i = 1,$$

then x is what we will call a *probability distribution* (in the sense that it generalizes the probabilities of 50% heads and 50% tails from flipping a fair coins to more complicated dependency situations). The reason is that, in this case, if x is a probability distribution, then because of the way we have normalized the columns of W , it will be the case that y is also a probability distribution. This will help us to tap into ideas from probability theory to get better algorithms and better statistical properties.

All in all, there will be two complementary interpretation to a lot of what we do.

1. Linear algebra: relate to linear functions and linear transformations.
2. Probability theory: relate to random walks and random variables.

So, in addition to having an informal interpretation in terms of voting/importance scores, this process will have a more rigorous interpretation in terms of random walks and diffusion processes that are central to probability theory. For example, if we start with our probability mass at one node and compute $y = Wx$ and then $z = W(Wx) = W^2x$, and so on, then we might hope that probability mass will spread out. This is known as a Markov chain, and it is central to probability theory. But it also captures something called an eigenvector, which is central to linear algebra; and this vector has a very natural interpretation in terms of data.

In fact, if you iterate that process many steps, which again might be hard to do by hand but which is something that we will do on the computer, then we will get a vector, call it z^* (sometimes this is called the spectral ranking vector or the Pagerank vector) that has been used to rank all sorts of things from web pages to sports teams to academic departments, and so on. By the way, what we just discussed is true if we work with a directed graph. If, on the other hand, we ignore the direction, then we get a vector that can be used for classification, clustering, and other common data science tasks.

(By now, it’s okay if you don’t follow everything—that’s the point of taking this class. But keep these topics in the back of your mind, since by the end of the term we will have revisited all of them.)

1.5 Problems

1.5.1 Pencil-and-paper Problems

XXX. DO WE WANT ANY IN THIS CHAPTER.

1.5.2 Implementations and Applications of the Theory

XXX.

Part II

Linear Algebra: Extending Basic Algebra and Geometry

Chapter 2

Matrices, vectors, and \mathbb{R}^n

2.1 Overview of the chapter

Let's start with an introduction to matrices, vectors, and \mathbb{R}^n .

Matrices and linear algebra are central to many areas of applied mathematics, and in particular they are widely-used in machine learning, data analysis, and data science. In Chapter 1, we said that a matrix is more than just a thing A_{ij} with two subscripts. Essentially, it is a thing that is characterized by a certain set operations that can be performed on it (that happens to have two subscripts, for reasons we will see soon). It is these operations that make matrices so useful. We will get into that now, starting from the beginning.

Take a step back. To set the more general context, there are four major ways to view a matrix.

- **As a set of points.** In this case, m points, each of which are in \mathbb{R}^n (i.e., in an n -dimensional Euclidean space, which we will define soon, but which is a generalization of the familiar one-dimensional line, two-dimensional plane, and three-dimensional space) can be written as an $m \times n$ matrix.

A very simple example of this is one point on the real line, i.e., on \mathbb{R} ; and slightly less simple examples of this are one point on the plane, two points on the line, two points on the plane. All of these simple examples can be written as matrices (1×1 , 2×1 , 1×2 , and 2×2 matrices), but the examples are so simple that this is not usually done.

- **As a linear transformation.** Matrix-vector products, and matrix-matrix products, e.g., the iterated random walk $y \leftarrow Ax$, $z \leftarrow Ay$, etc., that we discussed in Chapter 1, can be viewed as defining linear functions taking vectors as inputs and returning vectors as outputs. If the matrix is $m \times n$, we will see that this involves linear functions from \mathbb{R}^n to \mathbb{R}^m .

This is a generalization of the familiar $y = ax$, the equation of a line with slope a going through the origin, where y and a and x are all real numbers, i.e., elements of \mathbb{R} , and we are thinking of y as a simple function of x . In this case, given the value of a , one may want to compute y as a function of the input x .

- **As a quadratic form.** This is a generalization of $y = ax^2$, the familiar equation for a simple quadratic form involving one variable $x \in \mathbb{R}$. For example, if we have a symmetric matrix A and a vector x , then we can define the transpose of x^T and then extend the matrix-vector multiplication to write $y = x^T Ax$, which is a quadratic form in $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$.

We know that the one-dimensional quadratic form $y = ax^2$ has very different properties depending on whether $a > 0$ or $a = 0$ or $a < 0$. We will see that for higher-dimensional quadratic forms, i.e., involving vectors $x \in \mathbb{R}^n$ rather than numbers $x \in \mathbb{R}$, there is a rich array of properties that are possible.

- **In terms of linear equations.** Given a vector y and a matrix A , we might want to find a vector x such that $y = Ax$. This could be a vector where the random walk process leaves the vectors unchanged, if such a vector exists, but it could also be all sorts of other things. This has similarities with the linear transformation perspective, except when one thinks about matrices in terms of linear equation solving one typically knows y and A and one wants to find x .

When restricted to the familiar case, one wants to solve for x in the equation $y = ax$, and we know that the solution is $x = a^{-1}y = y/a$, assuming that $a \neq 0$. Again, when we generalize to higher-dimensional linear equations, a rich array of properties are possible.

While the linear transformation and linear equation perspectives both involve expressions of the form $y = Ax$, there is an important difference in terms of what we assume we know.

- For the linear transformation perspective, we know A and x and we are interested in computing y . This is sometimes known as a forward process.
- For the linear equation perspective, we know A and y and we are interested in computing x . This is sometimes known as a backward or inverse process, since at least formally/naïvely this involves the inverse of A .

Inverse processes tend to be more difficult. Most linear algebra classes start with this inverse process of solving linear equations as the main motivation. We will get to that eventually as an important application, but we won't start with it.

Viewing matrices from each of these perspectives has its advantages and disadvantages. In particular, the last perspective (i.e., in terms of linear equations) is most common and is the most important for many traditional applications of linear algebra, e.g., in engineering, physics, and scientific computation, and so it gets the greatest emphasis in many traditional linear algebra classes. This approach leads to an emphasis on things like Reduced Row Echelon Forms, QR decompositions, and so on. (If you are familiar with these things, great, that's just for context; but don't worry if you aren't, since we won't be covering them nearly as much here.) The other perspectives are more useful in data science, data analysis, and machine learning, and so in this class we will be much more interested in those perspectives. In particular, we will start by viewing matrices as consisting of a bunch of points in a high-dimensional vector space, in which case the linear transformation and quadratic form perspectives will arise naturally.

2.2 Ways to label points, elements, vectors, matrices, etc.

2.2.1 A familiar way

To start, consider three points on the plane, as illustrated in Figure 2.1. It may seem natural or more familiar to label them as (x, y) pairs, with a number indicating which point it is, i.e., as in the following:

$$\begin{aligned}(x_1, y_1) &\quad \text{is the first point} \\ (x_2, y_2) &\quad \text{is the second point} \\ (x_3, y_3) &\quad \text{is the third point.}\end{aligned}$$

That convention is fine, but we will adopt a numbering convention that will help us generalize to many dimensions, e.g., from \mathbb{R}^2 to \mathbb{R}^{47} to \mathbb{R}^n , for arbitrary n . In general, given a vector x , i.e., x is a vector and not a number, we will label the different elements of x with subscripts, as in $x = (x_1, x_2)$ for the two components of a two-dimensional vector x , and $x = (x_1, x_2, x_3)$ for the three components of a three-dimensional vector x . This will be since the generalization to an n -dimensional vector will be simply to write $x = (x_1, \dots, x_n)$, i.e., we will just add additional subscripts and work only with operations that respect that.

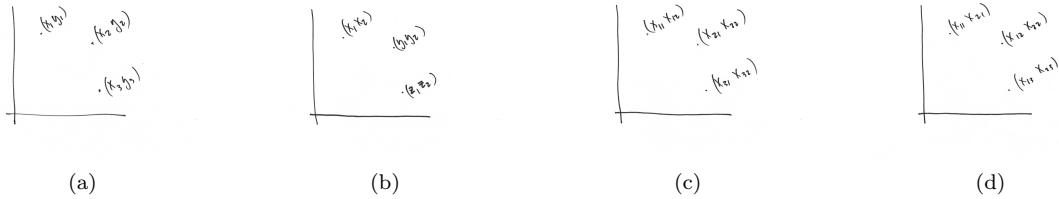


Figure 2.1: Three points on the plane, with several different notational conventions.

For now, since we have three points, let's call them x , y , and z , in which case we have the following:

$$\begin{aligned} x &= (x_1, x_2) && \text{is the first point} \\ y &= (y_1, y_2) && \text{is the second point} \\ z &= (z_1, z_2) && \text{is the third point.} \end{aligned}$$

See Figure 2.1(b), and contrast this with Figure 2.1(a). As we will see, this will be a common way we will write data and the elements of data.

2.2.2 A more powerful way

Alternatively, we could label the elements in a way that *seems* more difficult, but that makes the interpretation of matrices as consisting of high-dimensional vectors more immediate. In particular, if we label the three two-dimensional points as x_1 , x_2 , and x_3 , where x_1 , x_2 , and x_3 are vectors and not numbers, and where the subscript denotes which point we are dealing with and not which component of a given vector, i.e., subscripts represent different two-dimensional vectors, then we can let a second subscript denote the component of that vector and write it as follows:

$$\begin{aligned} x_1 &= (x_{11}, x_{12}) && \text{is the first point} \\ x_2 &= (x_{21}, x_{22}) && \text{is the second point} \\ x_3 &= (x_{31}, x_{32}) && \text{is the third point.} \end{aligned}$$

This is illustrated in Figure 2.1(c). Given this notational convention, we could write this as a matrix, with x_{ij} in the ij position, as follows:

$$A = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{pmatrix}. \quad (2.1)$$

Here, we have used the notation that element ij refers to the element that is in the i^{th} horizontal row and the j^{th} vertical column. In this case, the *rows* of A are data points. Alternatively, we could change the order of the subscripts, letting the first subscript denote the component and the second subscript denote the point, to obtain

$$\begin{aligned} x_1 &= (x_{11}, x_{21}) && \text{is the first point} \\ x_2 &= (x_{12}, x_{22}) && \text{is the second point} \\ x_3 &= (x_{13}, x_{23}) && \text{is the third point.} \end{aligned}$$

This is illustrated in Figure 2.1(d). In this case, we could also write these as a matrix, with x_{ij} in the ij position, as follows.

$$A = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{pmatrix}. \quad (2.2)$$



(a) 10 points on a two-dimensional plane.
(b) 100 points on a two-dimensional plane.

Figure 2.2: Illustration of points on the two-dimensional plane.

Here, we have still used the notation that element ij refers to the element that is in the i^{th} horizontal row and the j^{th} vertical column, but we have swapped what rows and columns mean. In this case, the *columns* of A are data points.

The point here is that we can shuffle around subscripts to write data points as either the rows or columns of a thing with two subscripts that we will identify as a matrix. In fact, depending on one's perspective, they are both, i.e., one approach or the other may be more natural, depending on one's perspective. It is helpful to be comfortable with both perspectives. The reason is that some areas assume one approach, and other areas assume the other approach. Relatedly, depending on the area, e.g., mathematics, computer science, statistics, economics, physics, biology, etc., particular letters can be numbers, vectors, or matrices, and subscripts can refer to points or elements. We will try to be clear about using this in different ways and in different cases in this class, but you should note that different areas and different books do it differently.

2.2.3 Many points in a few dimensions, or a few points in many dimensions

So far, we have considered 3 points on the plane, i.e., in a 2-dimensional Euclidean space. Clearly, we can easily extend this to 10 or 100 points on \mathbb{R}^2 . See Figure 2.2, which provides an illustration of what 10 and 100 points on the plane look like. These too can be written as a matrix, as follows:

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} & A_{14} & A_{15} & A_{16} & A_{17} & A_{18} & A_{19} & A_{1,10} \\ A_{21} & A_{22} & A_{23} & A_{24} & A_{25} & A_{26} & A_{27} & A_{28} & A_{29} & A_{2,10} \end{pmatrix}. \quad (2.3)$$

expresses 10 points on the plane. (Here, we have written the ij element of the matrix A using the more conventional A_{ij} rather than x_{ij} .) Alternatively, if rows rather than columns are data points, then this could be written as:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \\ A_{41} & A_{42} \\ A_{51} & A_{52} \\ A_{61} & A_{62} \\ A_{71} & A_{72} \\ A_{81} & A_{82} \\ A_{91} & A_{92} \\ A_{10,1} & A_{10,2} \end{pmatrix}. \quad (2.4)$$

Note that in this case of points on \mathbb{R}^2 , we can see that measuring distance or angles might be meaningful. Visualizing points in 3 dimensions is somewhat harder, but it is still possible. See, e.g., Figure 2.3, which provides a common way to visualize three-dimensions on a two-dimensional piece of paper. In that figure, two axes are shown at perpendicular angles, and the third axis is drawn at some other angle,



(a) 10 points on a three-dimensional plane.
(b) 100 points on a three-dimensional plane.

Figure 2.3: Illustration of points in a three-dimensional space.

as if you are viewing a three-dimensional object rotated slightly so you aren't looking directly down it's third axis.

Clearly, this becomes awkward to write out as the number of points gets large. Similarly, this becomes awkward to write out as the number of dimensions becomes large. Since writing data in this way is so useful, much of linear algebra is about developing techniques to make this much less awkward. (Of course, keep in mind that what is awkward for a person may not be awkward for a computer, and vice versa.)

2.2.4 Many points in many dimensions

More generally, a matrix can be written as:

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & & & \\ \vdots & & & \vdots \\ A_{m1} & \cdots & & A_{mn} \end{pmatrix}. \quad (2.5)$$

Here, A_{ij} refers to the (ij) element of A , which by convention is the element in the i^{th} horizontal row and the j^{th} vertical column. We can think of this matrix A in any of the four ways described above, in particular as m points in n -dimensional space or as n points in m -dimensional space. In this case, if we view it as m points in n -dimensional space, then the data consist of m data points, each of which is a vector in \mathbb{R}^n .

If we consider only one (vertical) column, say the first, call it x , then we have:

$$A_{:,1} = \begin{pmatrix} A_{11} \\ A_{21} \\ \vdots \\ A_{m1} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} = x. \quad (2.6)$$

Here, $x \in \mathbb{R}^m$ is an ordered list of real numbers, which we are viewing as a (vertical) column vector.

Alternatively, we could have considered only one (horizontal) row of A , say the first, to be the vectors in which we are interested, in which case

$$A_{1,:} = (A_{11} \ A_{12} \ \cdots \ A_{1n}) = (x_1 \ x_2 \ \cdots \ x_n) = x. \quad (2.7)$$

We will spend a lot of time understanding matrices of the form of Eqn. (2.5), their columns (e.g., as in Eqn (2.6)) and rows (e.g., as in Eqn. (2.7)) as vectors, as well as operations on the columns and rows. For reasons that should not be obvious now but that should become obvious over the next few chapters, matrices are really "about" their columns and rows. Thus, before we talk about matrices and operations on matrices more generally, let's spend some time on vectors.

One of the points of linear algebra and probability is that they will provide tools such that we will be able to identify a few key very simple ideas from one, two, and three dimensional spaces, i.e., from \mathbb{R} , \mathbb{R}^2 , and \mathbb{R}^3 , that we will be able to generalize to n -dimensional Euclidean space, i.e., \mathbb{R}^n . This will enable us to ask questions about \mathbb{R}^n . This in turn will enable us to do better term-document analysis, better web page ranking, etc., i.e., to make statements about data that we are modeling as vectors in \mathbb{R}^n .

2.3 What is \mathbb{R}^n ?

Advanced comment. If you are interested, pages 6 to 8 of Hubbard have more details on set theory, and pages 9 to 17 of Hubbard have more details on functions.

We will use the notation \mathbb{R}^n a lot, so let's describe it in a bit more detail. The notation \mathbb{R}^n refers to a set—special cases of which should already be familiar to you—with a particular structure. We will be a little more precise about sets qua sets when we discuss discrete probability, and we will be a little more precise about the definition of \mathbb{R}^n later, so here we will give the basic idea as well as examples of what we mean by the set \mathbb{R}^n . Basically, \mathbb{R}^n refers to a generalization of the familiar Euclidean plane, which is the special case \mathbb{R}^2 .

Recall that a point on \mathbb{R}^2 can be specified by two numbers (x, y) or (x_1, x_2) , which are the values or coordinates of the point projected onto the two perpendicular axes that are used to describe the plane. Generalizing this, elements of \mathbb{R}^n are known as vectors and can be thought of as a thing with n elements that is subscripted by one index, e.g., $x = (x_1, \dots, x_n)$, where each x_i is a real number, with properties that generalize the familiar properties of points on the Euclidean plane. Thus, \mathbb{R}^n is often thought of as a list/array of n numbers, with properties that generalize the familiar properties of points on the one-dimensional line, the two-dimensional plane, and the three-dimensional space.

2.3.1 Examples of \mathbb{R}^n

In the notation \mathbb{R}^n , the n is a positive integer. It is the dimension of \mathbb{R}^n (the dimension of the plane is two, and we will define dimension more generally below, but for now let's just think of it as the number of elements in the list that specify a vector $x \in \mathbb{R}^n$). Here is what \mathbb{R}^n is for various values of n .

- **One-dimensional Euclidean space:** \mathbb{R}^1 , a.k.a., \mathbb{R} : This is just the set of real numbers with which you are already familiar, i.e., this is the ordered real number line. Typically visualized as a line, the elements of \mathbb{R} can be added and multiplied, and both of these operations have an inverse, except for multiplication by 0 which doesn't have an inverse, etc. (By that last comment, I mean that $y = ax$ has the solution $x = a^{-1}y$ for all $a \neq 0$; and if $a = 0$, then there is in general no solution.)
- **Two-dimensional Euclidean space:** \mathbb{R}^2 : This consists of all pairs of real numbers, i.e., we say that

$$x \in \mathbb{R}^2 \text{ if } x = (x_1, x_2), \text{ where } x_1 \in \mathbb{R} \text{ and } x_2 \in \mathbb{R}.$$

This has the familiar interpretation/visualization of points on a plane, which is also easily visualized. We will see below that we will be able to define operations like addition and multiplication for points of the plane—some of these operations will be similar to addition and multiplication of points on the line, but there will be some important differences.

- **Three-dimensional Euclidean space:** \mathbb{R}^3 : This consists of all triples of real numbers, i.e., we say that

$$x \in \mathbb{R}^3 \text{ if } x = (x_1, x_2, x_3), \text{ where } x_i \in \mathbb{R}.$$

This has the familiar interpretation/visualization of points in three-dimensional space. Visualizing points in three dimensional Euclidean space is a little harder than visualizing points in two dimensions,

but since we live in a physical world that is well-approximated by a three-dimensional Euclidean space we have some intuition about \mathbb{R}^3 . For now, though, let's just view \mathbb{R}^3 as a mathematical object that may or may not come with familiar intuitions. Here, too, we define addition and multiplication of points in three dimensional space, etc.

- **Four-dimensional Euclidean space:** \mathbb{R}^4 . This consists of all quadruples of real numbers, i.e., we say that

$$x \in \mathbb{R}^4 \text{ if } x = (x_1, x_2, x_3, x_4), \text{ where } x_i \in \mathbb{R}.$$

This is harder to visualize and interpret, but we will be able to define operations on it in a way that cleanly generalizes familiar operations from \mathbb{R}^2 and \mathbb{R}^3 , and thus we will be able to talk about the properties of points in \mathbb{R}^4 .

- **n -dimensional Euclidean space:** \mathbb{R}^n . This consists of all sets of n real numbers, i.e., we say that

$$x \in \mathbb{R}^n \text{ if } x = (x_1, x_2, \dots, x_n), \text{ where } x_i \in \mathbb{R}.$$

This is even harder to visualize and interpret, especially as n gets large, but here too we will be able to define operations on it in a way that cleanly generalizes familiar operations from \mathbb{R}^2 and \mathbb{R}^3 . Importantly, as n gets larger, e.g., for 10 or 20, not to mention 10^2 or 10^9 , many properties of \mathbb{R}^n are *very* different than the corresponding properties of \mathbb{R}^2 or \mathbb{R}^3 , but a few properties do generalize cleanly. We can use these latter properties to analyze data that are modeled as points in \mathbb{R}^n and thus understand better data modeled as matrices and graphs.

A word of caution. While we can use these latter properties, we have to be careful: our intuition (which is built up from experiences with \mathbb{R} and \mathbb{R}^2 and \mathbb{R}^3) is often completely wrong about the properties of \mathbb{R}^n . We will revisit this issue.

These examples of \mathbb{R}^n will be key players for us in the linear algebra part of the class, and in the probability part of the class we will see that there are strong not-immediately-obvious connections between them and discrete probability.

2.3.2 Some very basic properties of \mathbb{R}^n

As a first step toward the goal of defining operations that cleanly generalize familiar operations from \mathbb{R}^2 and \mathbb{R}^3 to \mathbb{R}^n , let's ask what are some of the most basic intuitions that we have about points on a line or on a plane or in three dimensional space? Here are several:

- In \mathbb{R} .
 - **Size or absolute value.** Given a real number x , some are larger, and some are smaller, and this is quantified by the absolute value, which is typically denoted $|x|$. Given this absolute value, we can define a distance between two points, i.e., given $x, y \in \mathbb{R}$, we can define

$$\text{dist}(x, y) = |x - y|.$$

In many cases, one often has the intuition that points that have smaller distance between them are more alike. For example, see Figure 2.4(a), which plots a possible distribution or amount of some variable like age or wealth (actually, it is a quite unrealistic plot for these two quantities), in which case one might interpret this to consist of two groups, e.g., old and young or rich and poor.

- In \mathbb{R}^2 .
 - **Size or norm.** Given a point on the plane, let's reference it to an origin, in which case we will call it a vector. (We'll be more precise on this below, but this is analogous to choosing an origin for \mathbb{R} .) Then, we might want to measure the size of the vector. As opposed to the line, where any two numbers are comparable in terms of being larger or smaller, that is not true on the plane.

But we can associate a real number to any point on the plane which measures its size. There are actually many ways to do this, and we will discuss several in detail below, but the simplest version of this should be familiar as the Euclidean norm, which we will denote as $\|\cdot\|_2$. Here, given $x = (x_1, x_2) \in \mathbb{R}^2$, we define that

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2}.$$

For this and other norms, we can define a distance between two vectors as the norm of the difference, i.e., given $x, y \in \mathbb{R}^2$, we can define

$$\text{dist}(x, y) = \|x - y\|_2.$$

This generalizes the notion of size or absolute value to \mathbb{R} .

Note that since there are other notions of size, and since we can associate a distance with a norm as the norm of the vector difference, there are other notions of distance, even for \mathbb{R}^2 .

One reason this is of interest is that one often has the intuition that points that have smaller distance between them are more alike. See, e.g., Figure 2.4(b) and Figure 2.4(c), which shows two clusters and the distribution of distances for those points. In each case, points in one cluster are closer to other points in that cluster than they are to points in the other cluster, and this is often used in data science.

- **Angle.** In addition, for $x, y \in \mathbb{R}^2$, by using this particular norm (i.e., the Euclidean norm), we can define an angle θ between x and y as

$$\cos(\theta) = \frac{x^T y}{\|x\|_2 \|y\|_2} = \frac{x_1 y_1 + x_2 y_2}{\|x\|_2 \|y\|_2}.$$

Here too, one often has the intuition that points that have smaller angles between them are more alike. See, e.g., Figure 2.4(d), which shows the distribution of angles for those points in Figure 2.4(b). In each case, points in one cluster are closer to other points in that cluster than they are to points in the other cluster. (A caveat in the last statement is that angles “wrap around” the circle, and thus $\theta + 360^\circ = \theta$.)

- In \mathbb{R}^3 .

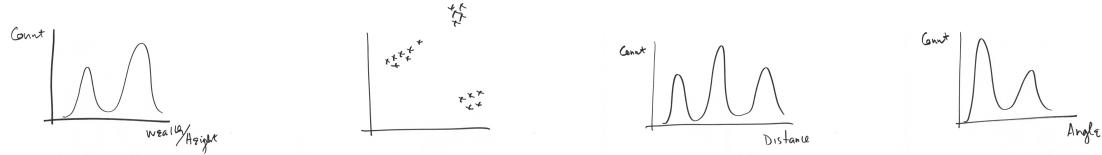
- **Norms, distances, and angles.** Here too, the notion of norms, distances, and angles arise. Other notions arise too (e.g., curls and divergences, if you have heard of those; if not, okay, since we won’t be interested in them); and, while they are important in physical applications, they generalize less well to very high dimensional spaces. They are less important in data science, and so we won’t go into them. We will see that norms and angles do generalize to \mathbb{R}^n , for $n \gg 3$. These are very useful in data science, and so we will spend a great deal of time on them.

As an example of what can be gained by this, recall Equation (2.1) and Equation (2.2). Depending on whether one encodes points as columns or rows, one can view these expressions as describing the location of 2 points in \mathbb{R}^3 or the location of 3 points in \mathbb{R}^2 . Indeed, both interpretations are valid. Similarly, although we introduced A in Equation (2.3) as representing the location of 10 points in \mathbb{R}^2 , where each column of A corresponds to one of those 10 points, we could alternately view A as representing 2 points in \mathbb{R}^{10} , where each row of A corresponds to one of those 2 points. By the way, this discussion is not pedantic. It basically means that we can view a term-document matrix in one of two ways: either as encoding a bunch of documents, each of which is described by the terms it contains, or as encoding a bunch of terms, each of which is described by the documents in which it appears. Both perspectives are useful.

Remark. A gotcha with angles, which holds for \mathbb{R}^2 and \mathbb{R}^3 , as well as \mathbb{R}^n more generally, is that they are measured in two common ways, i.e., with two different units, degrees and radians:

$$360^\circ = 2\pi \text{ radians.}$$

Thus, in particular, two lines on the plane are perpendicular if the angle between them is 90° or $\frac{\pi}{2}$ radians. For most languages, including python, there is a default as to which is used, but you can provide an argument to specify whichever you prefer.



(a) Example of the distribution of some quantity. (b) Illustration of points in 2D. (c) Distribution of distances between points in 2D. (d) Distribution of angles between points in 2D.

Figure 2.4: Illustration of distribution of norms and angles.

2.3.3 Some very basic subsets of \mathbb{R}^n

Understanding the properties of high-dimensional spaces and data modeled by high-dimensional spaces can be difficult, i.e., it can be difficult to have an intuition of what \mathbb{R}^n “looks like,” and it is often best to do this by understanding the properties of very simple subsets of \mathbb{R}^n . Often, these subsets are so simple in \mathbb{R} or \mathbb{R}^2 or \mathbb{R}^3 that they seem uninteresting and almost trivial.

For the moment, let’s restrict ourselves to \mathbb{R}^2 , i.e., the familiar plane. Here are three very simple subsets of the plane.

- **Unit ball.** This is the set of points (x_1, x_2) defined by the following equation:

$$\left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1 \right\}.$$

The generalization to \mathbb{R}^n is immediate: $\{x \in \mathbb{R}^n : \sum_{i=1}^n x_i^2 = 1\}$.

- **Positive orthant.** This is the set of points (x_1, x_2) defined by the following inequalities:

$$\left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 : x_1 \geq 0, x_2 \geq 0 \right\}.$$

The generalization to \mathbb{R}^n is immediate: $\{x \in \mathbb{R}^n : x_i \geq 0\}$.

- **Probability Simplex.** This is the set of points (x_1, x_2) defined as follows:

$$\left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 : x_1 + x_2 = 1, x_1 \geq 0, x_2 \geq 0 \right\}.$$

The generalization to \mathbb{R}^n is immediate: $\{x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1, x_i \geq 0\}$.

- **Subspace, e.g., a line through the origin.** Given $a \in \mathbb{R}$, one example of this is the set of points (x_1, x_2) defined by the following equation:

$$x_2 = ax_1.$$

A second example of this is the set of points (x_1, x_2) defined by the following equation:

$$x_2 = -\frac{1}{a}x_1.$$

Note that the second equation defines a line through the origin which is perpendicular/orthogonal to the line defined by the first equation. The generalization of a subspace to \mathbb{R}^n is more subtle. For example, it requires the ideas of linear combinations, basis, span, linear independence, etc. that form the heart of linear algebra. We will spend a lot of time on it since it is extremely important.

For the plane, all three of these classes of sets seem very simple, perhaps too simple to be interesting. We will spend time on each of these since we will see some surprising properties even for them, and these properties will be the tip of the iceberg in terms of counterintuitive properties for \mathbb{R}^n .

2.4 Measuring the size of vectors in \mathbb{R}^n

2.4.1 Norms

Definition of a norm. A norm is a function that measures the size of things. We used the Euclidean norm above (since it is probably most familiar and since it provides us the notion of an angle which provides important connections to geometry), but there are many other norms that are of interest, and we will discuss a few of the most important. Here is the definition of a norm, which is a generalization of the Euclidean norm in \mathbb{R}^2 .

Definition 1 Given a vector $x \in \mathbb{R}^n$, we say that $\rho(x) \in \mathbb{R}$ is a norm or a vector norm of x if it satisfies the following properties:

- $\rho(x) \geq 0$, and $\rho(x) = 0$ iff $x = 0$,
- $\rho(\alpha x) = |\alpha|\rho(x)$, for all numbers $\alpha \in \mathbb{R}$,
- $\rho(x + y) \leq \rho(x) + \rho(y)$.

The last condition is known as the triangle inequality.

Observe that when $n = 1$ the absolute value function is a vector norm on $\mathbb{R} = \mathbb{R}^1$.

When a function $\rho(\cdot)$ is a norm, it is common to represent it as $\|\cdot\|$. In this case, the three defining properties of a norm from Definition 72 can be expressed as follows.

$$\|x\| \geq 0, \text{ and } \|x\| = 0 \text{ iff } x = 0, \quad (2.8)$$

$$\|\alpha x\| = |\alpha|\|x\| \text{ for all real numbers } \alpha, \quad (2.9)$$

$$\|x + y\| \leq \|x\| + \|y\|. \quad (2.10)$$

Norms in \mathbb{R}^2 and \mathbb{R}^3 . Perhaps the most well-known norm is the so-called Euclidean norm, also known as the L_2 norm. We will see below that the Euclidean norm as well as other norms generalize to \mathbb{R}^n , but for the moment, let's restrict ourselves to \mathbb{R}^2 . Here it is for vectors x in \mathbb{R}^2 .

- L_2 : $\|x\|_2 = \left(\sum_{i=1}^2 x_i^2 \right)^{1/2} = (x_1^2 + x_2^2)^{1/2}$

The Euclidean norm is sometimes called the L_2 norm due to the “2” in the square (of each element) and square root (of the sum); see the Advanced Comment on L_p norms below.

The interpretation of the L_2 norm as the length or magnitude of a vector should be familiar. But what does this really mean? Said another way, what are the key/fundamental properties of size or length. Here are the key properties of the L_2 norm that justify its interpretation as the length or magnitude of a vector:

- **Positivity/non-negativity.** The length of a vector is always greater than 0, unless it is the zero vector, in which case its length is equal to 0.
- **Positive scalability.** The length of product of a vector and a scalar real number is the length of the vector multiplied by the absolute value of the scalar.
- **Triangle inequality.** The length of one side of a triangle is not larger than sum of the lengths of the other two sides of that triangle.

Each of these three properties should be easily-understood in terms of the familiar geometric properties of the two-dimensional plane. Indeed, it is these three properties that motivated the three conditions in the more general Definition 72.

Introducing a more general definition isn't so interesting if there is only one example of it. Fortunately, in this case, there are actually other norms. Two others are the L_1 norm and the L_∞ norm. Here they are for vectors x in \mathbb{R}^2 .

- L_1 : $\|x\|_1 = \sum_{i=1}^2 |x_i| = |x_1| + |x_2|$
- L_∞ : $\|x\|_\infty = \max_{i \in \{1,2\}} |x_i| = \max\{|x_1|, |x_2|\}$

These provide two different ways than the L_2 norm to measure the size of a vector.

Examples of norms of vectors in \mathbb{R}^2 . We can consider a given vector $x \in \mathbb{R}^2$, and we can measure the size of that vector with respect to different norms. Which one is appropriate depends on the application.

Example. Let $x = (1, 2)$. In this case:

- $\|x\|_1 = 1 + 2 = 3$
- $\|x\|_2 = (1 + 4)^{1/2} = \sqrt{5} \approx 2.24$
- $\|x\|_\infty = \max\{1, 2\} = 2$

This example illustrates the more general property that $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$, i.e., that the numerical value of the size of a vector depends on the norm used to measure its size. This can also be seen from Figure 2.6. (For the former inequality, observe that only one term (the largest) entering into the sum of the L_2 is used in L_∞ . For the latter inequality, observe that the grid/rectilinear/Manhattan distance between two points is never shorter than the length of the line segment between them (the Euclidean or “as the crow flies” distance, e.g., since the crow flies along the hypotenuse of the triangle.)

Fact. For all of these inequalities, there exists vectors $x \in \mathbb{R}^2$ such that the inequality holds as an equality.

Question. Can you give an example of a vector $x \in \mathbb{R}^2$ such that each of the above inequalities holds as an equality?

Example. Let $x = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix}$. In this case:

- $\|x\|_1 = |\cos(\theta)| + |\sin(\theta)|$
- $\|x\|_2 = \cos^2(\theta) + \sin^2(\theta) = 1$
- $\|x\|_\infty = \max\{|\cos(\theta)|, |\sin(\theta)|\}$

Similar statements hold for the vector $x = \begin{pmatrix} -\sin(\theta) \\ \cos(\theta) \end{pmatrix}$.

These properties are not peculiar to \mathbb{R}^2 . Consider, e.g., a vector in \mathbb{R}^3 .

Example. If we take $x = (1, 2, 3)$, then we have:

- $\|x\|_1 = 6$
- $\|x\|_2 = (1 + 4 + 9)^{1/2} = \sqrt{14} \approx 3.74$
- $\|x\|_\infty = 3$

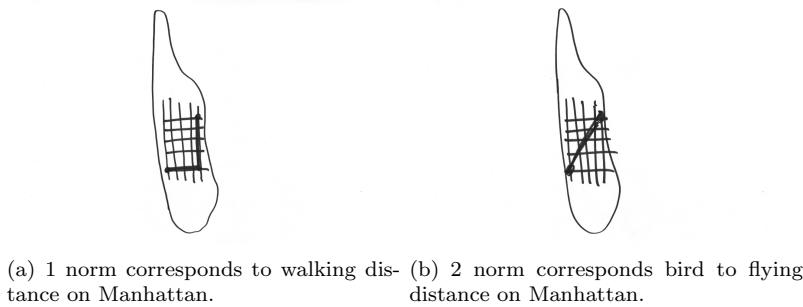


Figure 2.5: Illustration of 1-norm versus 2-norm.

Again, this illustrates that $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$, but here for a vector in \mathbb{R}^3 .

Similarly for a vector in \mathbb{R}^4 .

Example. If we take $x = (1, 2, 3, 4)$, then we have:

- $\|x\|_1 = 10$
- $\|x\|_2 = (1 + 4 + 9 + 16)^{1/2} = \sqrt{30} \approx 5.48$
- $\|x\|_\infty = 4$

Similar properties will generalize to \mathbb{R}^n .

Example. While we have defined these norms, it might be less easy to see—initially at least—why these other notions of size or norm are interesting. To see one example of this, consider Figure 2.5, which illustrates a grid-like street map of a city like Manhattan. Perhaps a bird can ignore the streets and fly directly from point A to point B, but to walk from point A to point B means that one must go along streets, horizontally and vertically, as opposed to along the hypotenuse. This is captured by the L_1 norm, and thus this notion of norm is a more meaningful notion of distance when walking around city blocks such as those shown in Figure 2.5. We will see more examples of this as well as where one can use the L_∞ norm below.

Remark. A large part of the importance of the L_2 norm is that it is associated with angles (since it is associated with dot products, i.e., it is the dot product of a vector with itself, up to normalization of the vectors), which we will get to below. This means that when we are measuring these quantities on the \mathbb{R}^2 plane, there are two complementary perspectives: algebra; and geometry. This should be familiar from the Cartesian approach from high school mathematics, and understanding these connections form most of the basis for understanding high dimensional data. This connection will generalize to high dimensional Euclidean spaces and will be very important in what we will talk about in this class.

Norms in \mathbb{R}^n . We have gone through this since a lot of linear algebra and matrices for data science is about generalizing these ideas from \mathbb{R}^2 and \mathbb{R}^3 to \mathbb{R}^n , where $n \gg 3$, and where intuitions that one has from \mathbb{R}^2 and \mathbb{R}^3 break down. We will cover this in more detail next time, but for now simply note that a point in \mathbb{R}^n can be represented as

$$\begin{pmatrix} x_1 & x_1 & \dots & x_n \end{pmatrix} \quad \text{or as} \quad \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}. \quad (2.11)$$

That is, as an ordered list of numbers. We have shown it both as a column and as a row, we will get to some differences later. This is often interpreted as a point representing position or as a vector representing displacement, e.g., from an origin. For now, simply note that the equations above were all represented as sums and thus they can be generalized to sum not from 1 to 2 or 3 but instead up to n .

The definition of norm given in Definition 72 does not make any reference to the value of n in \mathbb{R}^n , and so it holds for \mathbb{R}^n . In addition, the definitions of the three norms we have given easily generalize to \mathbb{R}^n . Here they are:

- L_1 : $\|x\|_1 = \sum_{i=1}^n |x_i| = |x_1| + |x_2| + \cdots + |x_n|$
- L_2 : $\|x\|_2 = (\sum_{i=1}^n x_i^2)^{1/2} = (x_1^2 + x_2^2 + \cdots + x_n^2)^{1/2}$
- L_∞ : $\|x\|_\infty = \max_{i \in [n]} |x_i| = \max\{|x_1|, |x_2|, \dots, |x_n|\}$

These three norms (there are also many others we won't discuss) provide three different ways to measure the "size" of a vector in \mathbb{R}^n . The fact that you may have seen one of them before does not mean that it is "better" or more "natural" than the others. Instead, you should ask what does a particular norm capture and when is it useful. All of these norms are useful in different ways, and we will see several examples of this later.

Observe that all three of these norms reduce to the same thing in $\mathbb{R} = \mathbb{R}^1$, namely the absolute value.

Example. Let $x = (1, 1, \dots, 1) \in \mathbb{R}^n$. Then,

$$\begin{aligned}\|x\|_1 &= n \\ \|x\|_2 &= \sqrt{n} \\ \|x\|_\infty &= 1.\end{aligned}$$

Example. Let $x = (1, 0, \dots, 0) \in \mathbb{R}^n$. Then,

$$\begin{aligned}\|x\|_1 &= 1 \\ \|x\|_2 &= 1 \\ \|x\|_\infty &= 1.\end{aligned}$$

These two examples are interesting since they represent two extreme cases—one in which the "mass" of the vector is spread out evenly over all the components of that vector, and the other in which the "mass" of the vector is spread out very unevenly.

Relationship between different norms. Before that, however, let's ask how do these different measures of size of a vector relate to each other. These norms are related to each other with the following inequalities. For a given vector $x \in \mathbb{R}^n$, we have that:

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2 \tag{2.12}$$

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty \tag{2.13}$$

$$\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty. \tag{2.14}$$

Proving that these inequalities are true will be done in the homework exercises.

Remark. Whenever you see an inequality such as one of these, you should ask whether it is "tight," i.e., whether it can be "saturated." By that, we mean does there exist a vector $x \in \mathbb{R}^n$ such that the inequality holds as an equality. If the answer is yes, then it is a much more informative inequality than if the answer is no.

- As an example, consider the function $f(x) = \sin(x)$, for $x \in \mathbb{R}$. For all $x \in \mathbb{R}$, it is true that

$$f(x) \leq 10^6,$$

and it is also true that

$$f(x) \leq 1.$$

- As a second example, consider the function $g(x) = x^2$, for $x \in \mathbb{R}$. For all $x \in \mathbb{R}$, it is true that

$$g(x) \geq -10^6,$$

and it is also true that

$$g(x) \geq 0.$$

While both pairs of inequalities are true, in each case the latter inequality is much more informative than the former inequality, e.g., it tells us more about the properties of the function being bounded. The basic reason for this is that there exists values of $x \in \mathbb{R}$ such that $f(x) = 1$ and there exists a value of $x \in \mathbb{R}$ such that $g(x) = 0$. Thus, each of the latter inequalities is tight and can't be "improved" by choosing a different (smaller or larger, respectively) value for the right hand side of the inequality.

Let's go back the the three sets of inequalities given in (2.12), (2.13), and (2.14). It is true that all of these inequalities are tight, i.e., for each of these six inequalities, there exists a vector $x \in \mathbb{R}^n$ such that it is an equality. Proving this will be done in the homework exercises.

You might be wondering what is the significance of this. Since for every n each of these six inequalities tight, there are vectors whose "size" is very different, depending on precisely how "size" is being measured. For \mathbb{R}^2 , the numerical values of the norms of a given vector can be different, but not extremely different. That is, they can only differ by a factor of $\sqrt{2}$ or 2, depending on which pair of norms we are considering. In particular, the person walking on Manhattan does not have to walk much more than the crow flies over Manhattan. For \mathbb{R}^n , on the other hand, a given vector can have numerical values of norms that differ by a factor of up to \sqrt{n} or n . If $n = 10^6$, that is a large difference. We will see soon why this matters.

Advanced comment. Equivalence and non-equivalence of norms. You may have heard, or you may at some point hear, some comment to the effect that "all norms are equivalent" in \mathbb{R}^n . That's true ... or false, depending on what you mean by "equivalent." That comment relates to our discussion in the following way.

- In calculus, one is interested in the convergence properties of sequences of things like numbers, vectors, functions, etc., and this convergence is defined in terms of the absolute value (in \mathbb{R}) or a norm (in \mathbb{R}^n) of something going to zero in some limit. (Recall ϵ - δ arguments when defining derivatives and integrals in terms of limits.) In this case, it is important that the convergence does not depend on the particular norm that is chosen to measure the size of the numbers/vectors/functions. One formalizes this in the following way.
 - Let's fix the value of n , or equivalently fix that we are interested in vectors in \mathbb{R}^n . (In one-dimensional/single-variable calculus, $n = 1$, and this is just \mathbb{R} .) Then we say that two vector norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are equivalent if there exists constants C_1 and C_2 that are independent of the vector x such that for any vector $x \in \mathbb{R}^n$, it holds that

$$C_1\|x\|_\alpha \leq \|x\|_\beta \leq C_2\|x\|_\alpha.$$

That is, for a fixed n , the numerical values of the norms are not too different, in the sense that they are within a factor of C_1 or C_2 of each other, and this does not depend on x , i.e., the same C_1 and C_2 hold for all vectors x . If two norms are equivalent in this sense, then a sequence of vectors that converges to a limit with respect to one norm will converge to the same limit in the other norm (although, of course, how quickly it converges might be different, depending on the norm).

- In our discussion, and in many applications of data science, we are interested in a different question. We are less interested in limits for a fixed value of n than in the properties of norms as the value of n varies. The reason is that n is the number of elements in the vector, i.e., the number of data points and/or the number of features that are used to describe each data point, and we want to know how norms and other things behave as we get more data and/or more features. In particular,

we want to know how C_1 and C_2 in the above expression vary with n . The reason is that the values of C_1 and C_2 will determine the algorithmic and statistical behavior of computations we perform (e.g., we will see that the running time of algorithms depends on the number of data points n and the statistical properties of noise depends on the number of data points n), and we will get different results if they depend on n than if they do not depend on n . What (2.12), (2.13), and (2.14) say is that they do depend on n , and thus we need to be careful about which norm we are using in data science applications.

Thus, while all norms in finite dimensional vector spaces are equivalent (in terms of their convergence properties), all norms in finite dimensional vector spaces are not equivalent (in terms of how quickly they converge and thus in terms of their algorithmic and statistical properties).

Advanced comment. We can view all these norms as members of a more general family of norms, the L_p norms, defined as

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}, \quad (2.15)$$

which, for the case of \mathbb{R}^2 reduces to:

$$\|x\|_p = (x_1^p + x_2^p)^{1/p}. \quad (2.16)$$

All three norms we have discussed are special cases of this.

- The L_1 norm is clearly the special case of Equation (2.15) with $p = 1$.
- The Euclidean norm is clearly the special case of Equation (2.15) with $p = 2$, and this is why it is often called the L_2 norm.
- The L_∞ norm has that name since if one were to compute $\|x\|_p = (x_1^p + x_2^p)^{1/p}$ for larger and larger values of p , then in the limit as $p \rightarrow \infty$ it equals the max $\max\{|x_1|, |x_2|\}$.

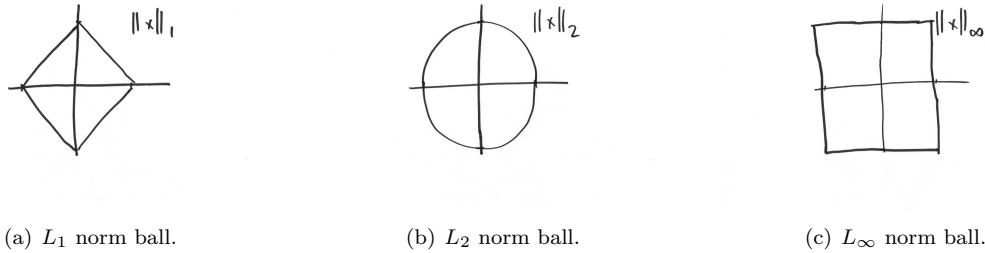
2.4.2 Balls

Given the idea of a norm of a vector, as Definition 72 provides, and which provides a way to measure of the size of a vector, it is also useful to have the idea of of a ball or sphere, which informally is the set of points—in that norm—at a given distance from (say) the origin. This is simply a generalization to an arbitrary norm of the notion of a round ball/sphere in two and three dimensions, i.e., a circle or sphere; but, as with many simple ideas, there are surprisingly rich and useful extensions of the basic idea. This will permit us to make statements about the set of points that are a given distance (without loss of generality assumed to be 1) from a given point (often but not necessarily taken to be the origin) with respect to other notions of distance.

Definition 2 Given a point $x_0 \in \mathbb{R}^n$ and a norm $\|\cdot\|$,

- the unit sphere (around x_0 with respect to $\|\cdot\|$) is the set of points $x \in \mathbb{R}^n$ such that $\|x - x_0\| = 1$;
- the unit ball (around x_0 with respect to $\|\cdot\|$) is the set of points $x \in \mathbb{R}^n$ such that $\|x - x_0\| < 1$; and
- the closed unit ball (around x_0 with respect to $\|\cdot\|$) is the set of points $x \in \mathbb{R}^n$ such that $\|x - x_0\| \leq 1$.

Sometimes, when it doesn't matter, we'll use these terms interchangeably, e.g., say unit ball when we mean the set of points that is at unit distance in that norm from the origin, or vice versa. In a few cases, we'll

Figure 2.6: The unit ball for the L_1 , L_2 , and L_∞ norms.

	L_1 Ball	L_2 Ball	L_∞ Ball
$n = 1$	$2r$	$2r$	$2r$
$n = 2$	$\frac{(2r)^2}{2!}$	πr^2	$(2r)^2$
$n = 3$	$\frac{(2r)^3}{3!}$	$\frac{4}{3}\pi r^3$	$(2r)^3$
	\vdots	\vdots	\vdots
n	$\frac{(2r)^n}{n!}$		$(2r)^n$

Table 2.1: Volumes of balls of radius r for different norms in different dimensions. (The volume of an L_2 norm ball in n -dimensions is known but is a little messy, and so we will compute/simulate it in a few weeks.)

want to distinguish between the ball, i.e., points inside the sphere, and the sphere, i.e., the points at the boundary of the ball. When it matters, we'll be specific. (The reason is that while the nomenclature is fairly standard, it isn't perfectly standardized.)

To help visualize these norms, consider Figure 2.6, which shows the “unit ball” of these three norms.

Three-dimensional space looks different than two-dimensional space (e.g., you can go up and down, in addition to left and right and back and forth), which in turn looks different than one-dimensional space (where you can only go left and right—or, of course, back and forth, or up and down, depending on what way you turn your head, but not two or three of those pairs at once). These differences are important since they inform your intuition about what is possible, and this intuition is important since it helps you to make decisions about what methods to use, when those methods might be returning a crazy answer, etc. High-dimensional vectors, i.e., vectors in \mathbb{R}^n , when n is 10^2 or 10^6 or even when it is only 10, have very different properties than vectors in \mathbb{R}^n , when $n = 1$ or 2 or 3. How can we begin to understand this?

To begin to get see how we can use this to get an understanding of the geometry of \mathbb{R}^n , let's ask: what is the difference between an L_2 ball in \mathbb{R} versus \mathbb{R}^2 versus \mathbb{R}^3 . More specifically, what about the ratio between the L_2 ball, i.e., $\{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$ and the Minimum Enclosing Box, which recall is an L_∞ ball, i.e., $\{x \in \mathbb{R}^n : \|x\|_\infty \leq 1\}$. Recall Figure 2.6, and see Table 2.4.2 for the volumes of balls of radius r for different norms in different dimensions.

- \mathbb{R} : Here the two balls are the same: $\frac{A_{ball}}{A_{box}} = \frac{2r}{2r} = 1$.
- \mathbb{R}^2 : $\frac{A_{ball}}{A_{box}} = \frac{\pi r^2}{(2r)^2} = \frac{\pi}{4} \approx 0.78$.
- \mathbb{R}^3 : $\frac{A_{ball}}{A_{box}} = \frac{\frac{4}{3}\pi r^3}{(2r)^3} = \frac{\pi}{6} \approx 0.55$.

We can ask the same question for the ratio of the L_1 ball, i.e., $\{x \in \mathbb{R}^n : \|x\|_1 \leq 1\}$, and the L_∞ ball.

- \mathbb{R} : Here the two balls are the same: $\frac{A_{L_1 ball}}{A_{L_\infty ball}} = \frac{2r}{2r} = 1$.
- \mathbb{R}^2 : $\frac{A_{L_1 ball}}{A_{L_\infty ball}} = \frac{\frac{1}{2}(2r)^2}{(2r)^2} = \frac{1}{2}$.
- \mathbb{R}^3 : $\frac{A_{L_1 ball}}{A_{L_\infty ball}} = \frac{\frac{1}{3}(2r)^3}{(2r)^3} = \frac{1}{6}$.

Note that the ratio decreases as the dimension increases. We will see that this is true more generally, and it is this phenomenon that is responsible for many of the important and counterintuitive properties of \mathbb{R}^n , for $n \gg 1$. For example, while it shouldn't be obvious, this is the phenomenon that is related to the fact that a fair coin flipped many times almost always comes up heads. It is also related to the fact that if you have a class with more than a few dozen people then it is very unlikely that there are no two people who share the same birthday. It is also related to many other related things in data science. Proving properties related to this is what you do in advanced classes, but many of the points can be illustrated computationally/simulationally, and this will be the substance of one of your homeworks.

2.5 Visualizing elements of \mathbb{R}^n

If, instead of working with a small number of points in \mathbb{R}^2 or \mathbb{R}^3 , we have 10^9 points, each of which is described by a set of 10^6 numbers, then we have a much larger matrix. Not only would it be difficult for a human to write out that many points, but it is not possible to visualize points in that many dimensions.

Question: What do 10^9 points in 10^6 “look like”? For example, do they look similar to or different than on \mathbb{R}^2 or \mathbb{R}^3 ? This is very common in data science, but it can be a very difficult question to answer.

Another Question: What does it even mean to ask that previous question? That is, how can we quantify what “look like” means? That ambiguity being noted, it can be difficult to tell what such a data set “looks like,” and indeed one of the main challenges here is going to be that high dimensional spaces are *very* different than low dimensional spaces.

While the algebraic aspect of linear algebra (to which we will get soon) is the way quantities of interest are actually computed, especially on a computer, it is the geometric aspects of linear algebra that drives much of the intuition people have about it as well as much of its usefulness. As such, it is often helpful to “visualize” \mathbb{R}^n as well as elements and subsets of \mathbb{R}^n . For $n = 1, 2, 3$, this is easy—it simply corresponds to the familiar one-dimensional line, two-dimensional plane, and three-dimensional space, with which we are familiar, as well as familiar subsets of these spaces. For $n = 4$, not to mention $n = 47$ or $n = 10^6$, this is much more difficult.

Here, we will describe one way to “visualize” higher-dimensional vectors. Since people have such strong intuition about one and two and three dimensional spaces, this method tends to be less useful than the usual way (shown in Figure 2.2 and Figure 2.3 for $n = 2$ and $n = 3$); but since people have such weak intuition about higher dimensional spaces, for $n > 3$ it can be very helpful. For lack of a better term, we will call this the *augmented visualization method*.

The basic idea of the *augmented visualization method* is to view a vector $x \in \mathbb{R}^n$ as n separate numbers $x_i \in \mathbb{R}$, which can be represented on a two-dimensional piece of paper by n points on n separate number lines, plotted vertically and parallel to each other. On the i^{th} vertical line, we plot the point x_i . See Figure 2.7

for plots of the points $(1) \in \mathbb{R}^1$, $\begin{pmatrix} 1 \\ 1/2 \end{pmatrix} \in \mathbb{R}^2$, $\begin{pmatrix} 1 \\ 1/2 \\ 1/3 \end{pmatrix} \in \mathbb{R}^3$, and $\begin{pmatrix} 1 \\ 1/2 \\ 1/3 \\ 1/4 \end{pmatrix} \in \mathbb{R}^4$.

We have not drawn the n vertical number lines, but they are the lines vertical and parallel to the Y-axis, going through $x_1 = 1, 2, \dots$

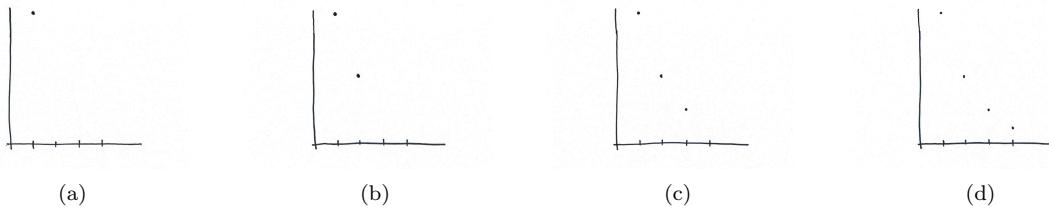


Figure 2.7: Augmented visualization of four vectors in \mathbb{R}^n , for $n = 1, 2, 3, 4$.

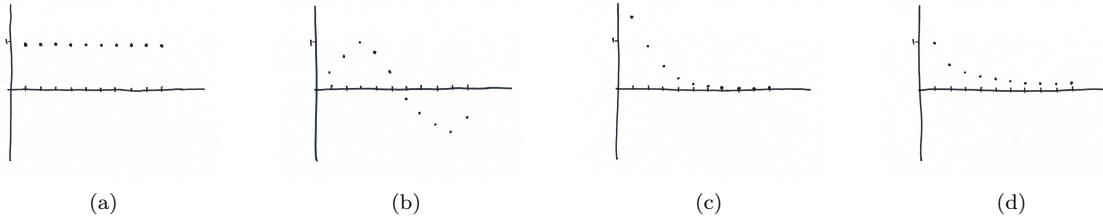


Figure 2.8: Augmented visualization of four vectors in \mathbb{R}^{10} .

Clearly, for the first three vectors, we could plot them in the more familiar way. Moreover, any vector in \mathbb{R} or \mathbb{R}^2 or \mathbb{R}^3 can be represented in this way, with one or two or three vertical number lines. It should be obvious, however, that this less familiar way could be used to plot any vector, e.g., in \mathbb{R}^{10} . That is, any configuration of points on n vertical numbers corresponds to a point $x \in \mathbb{R}^n$. See Figure 2.8 for plots of four such vectors in \mathbb{R}^{10} : an all-ones vector, a discrete sinusoid, a discrete exponential, and a vector in which subsequent elements get smaller and smaller as $x_i = 1/i$. For simplicity, each of these vectors comes from an easily-describable continuous function, but we could plot any set of numbers, e.g., an arbitrary set or a random set of numbers.

As with any representation of or way to view a mathematical object, it is worth asking what is gained and what is lost in that approach as well as what familiar things might “look like” when viewed that way. To do this, let’s consider points on a two-dimensional sphere, i.e., a circle. A circle has a geometric interpretation, but algebraically it is just the set of pairs of numbers (x_1, x_2) that satisfy the constraint $x_1^2 + x_2^2 = 1$. See Figures 2.9(a) and 2.9(b) for one point on the two-dimensional circle, visualized in the usual method and the augmented visualization method; and see Figures 2.9(c) and 2.9(d) for a different point, visualized in both ways. Clearly, e.g., for points that are more axis-aligned, one of the points in the augmented visualization method is much larger than the others, and this is very easy to see. On the other hand, it is harder to visualize the rotational properties with the augmented visualization method. Similar results hold more generally for vectors in \mathbb{R}^n . Recall Figure 2.8, and see also <https://www.youtube.com/watch?v=zWAD6dRSVYI> for a nice description, which we will revisit in later chapters.

Observe that, although we didn’t emphasize it, the illustration in Figure 1.1 was exactly of this form.

Important. Although it looks like we are plotting points on the familiar plane, we are really plotting n separate vertical number lines, and the i^{th} component of the vector is plotted on the i^{th} vertical number line. This visualization is designed to be suggestive: For example, we described the vectors we were plotting in Figures 2.7 and 2.8 in terms of an underlying real-valued function that is continuous along the x_1 -axis. For the augmented visualization of an arbitrary vector in \mathbb{R}^n , the different values along the “ x_1 axis” are unrelated, and there doesn’t need to be any connection or relationship between them. (There may be such a relationship, but there does not need to be. We could plot any vector in \mathbb{R}^n in this way.)

Advanced comment. One issue to keep in mind is that in Figure 1.1 there was a temporal structure. That is, earlier chapters in the book happen before later chapters, and this is reflected in the left-to-right

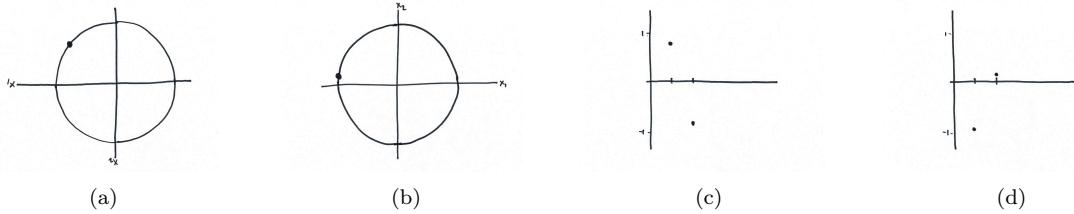


Figure 2.9: Usual visualization and augmented visualization of two points on a two-dimensional circle.

along the plot. The augmented visualization method can be applied to any vector, whether or not there is such an ordering, since the n vertical number lines don't necessarily have any "ordering" with respect to each other, and so the ordering in the method isn't necessarily meaningful. Of course, there might be a "temporal" ordering, as in the sinusoid or exponential decaying vectors, but in general there is not.

2.6 Two basic operations on vectors in \mathbb{R}^n

2.6.1 Example illustrating possible operations on matrices

Here is another example of a type of data set. Microarrays arise in genetics, and DNA SNP measurements are often performed in genetics and related areas. In these applications, one measures m organisms or tissue samples or environmental conditions in each of n time steps; or m individuals have their base pairs measured at each of n DNA sites. One can think of the data as an $m \times n$ matrix, where

$$A_{ij} = \begin{cases} \text{real number} & \text{if a microarray, depending on the measured values} \\ \{-1, 0, +1\} & \text{if a SNP, depending on SNP values} \end{cases}$$

Question. What are meaningful operations one might want to perform on a data matrix such as this?

Answer. For the "meaningful" condition, let's consider properties of the domain from which the data were generated, and then try to abstract out some operations that go beyond flat table operations that will be of interest more generally. For the microarray example, a very simple model is the following. The experimentalist is working with a bacteria sample, and he or she does some sort of heat/chemical shock, and he or she watches the response. As an idealization, the bacteria may have two types of responses, depending on the genes involved: an oscillating sinusoidal response; and an exponentially decaying response. See Figure 2.10 for the idealized sinusoids and exponentials as well as a noisy version of them, the latter of which might correspond to what is actually measured.

Note that here we are simply providing a plausibility argument, drawn from a data science problem, of meaningful operations on vectors in \mathbb{R}^n . In the next chapter, the operations will be introduced as axioms in the definition of a vector space.

Question: Why is this a reasonable toy model?

Answer: There are some phenomena that have cyclic trends, e.g., bacteria reproduce on the time scale of a day, and there are other things that involve some response/relaxation, e.g., genes that respond to the shock and then relax back to the normal state.

In this case, what we measure in different rows of the data matrix might be the following.

$$\begin{aligned} y_1(t) &= \sin(\omega_1 t) + \varepsilon(t) \\ y_2(t) &= \exp(\omega_2 t) + \varepsilon(t) \\ y_3(t) &= \sin(\omega_3 t) + \sin(\omega_4 t) + \varepsilon(t), \end{aligned}$$

where ε is something that captures the "noise" or other phenomena that aren't sinusoid or exponentials.

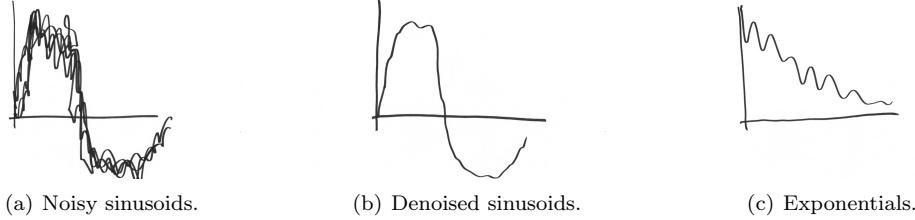


Figure 2.10: Illustration of sinusoids and exponentials.

(As an aside, what are other examples of data that might be a combination of two or more sinusoids? How about temperature, that varies daily and seasonally; or traffic patterns, that vary daily, weekly and seasonally.)

Observe that we could view the above data as a flat table and query it. For example, we can return all rows such that $y(t) > y^*$, for any t or for some value of $t > t^*$. Alternatively, we can return all rows such that some row is above a critical value. These types of queries are of interest in certain data science applications, either as the last step in the analysis or as a first clean-up step. In practice, however, in other applications, people are typically interested in performing more complex computations.

Question: What other operations might make sense here?

Answer: Here are two.

- $y(t) \rightarrow \alpha y(t)$: That is, rescale a data vector by a scalar/number α . Why? There are many possible reasons, but here is one. The measurements might be from different machines or in different units, and thus we might want to rescale them to make the measurements comparable. For example:

$$\begin{aligned} - y(t) &\rightarrow \frac{1}{\max_t y(t)} y(t) \\ - y(t) &\rightarrow \frac{1}{\text{avg}_t y(t)} y(t) \end{aligned}$$

Note that the denominators in both of those cases are norms, the L_∞ and L_2 norms, respectively. Thus, this is an example of normalization: we are normalizing or rescaling by two different notions of size.

- $y(t) \rightarrow \alpha_1 y_1(t) + \alpha_2 y_2(t)$. That is, take a linear combination of two rows. Why? There are many possible reasons, but here is one. We might have multiple measurements of what we think should be the same things, and we might want to take the average of two or more to get a single less-noisy measurement. For example, see Figure 2.10(b), which gives an illustration of how this can lead to less noise. Alternatively, these two measurements might both be related to some underlying hypothesized state that we want to identify.

An important is the following. While these two operations were introduced as plausible domain-specific operations, each of which can be interpreted as a mathematical operation, the theory of linear algebra as well as its broad usefulness boils down to understanding the mathematical consequences of these two operations. Thus, we will spend a lot of time discussing this. The consequences of these two operations will lead to many useful results, even when the operations are less obviously-useful in terms of domain considerations.

Before that, a note to emphasize that point: These two types of transformations sometimes make sense in terms of processes generating the data, and sometimes they do not. For example, what does

$$\frac{1}{2} \sin(\omega_1 t) + \frac{1}{2} \exp(\omega_2 t)$$

“mean” in terms of processes generating the data, in the simple example we gave above? It is worth keeping in mind whether or not the matrix and graph computations “make sense” in terms of the domain that generated the data. Sometimes, doing operations even if they aren’t easily interpretable can lead to useful results, and sometimes it can get you in hot water, e.g., due to invalid reification.

As we will see in more detail below, the above two operations essentially define a linear function and thus by building on them we can tap into a lot of mathematics that has important algorithmic and statistical benefits, even if it isn’t obviously meaningful in terms of the data.

The motivation is very different than microarray/SNP data or term-document data or social network data, but when viewed in terms of the abstraction of vectors and matrices, similar properties will hold. In both cases, there is a lot of domain-specific work to do to turn the raw data into a “meaningful” matrix, but the general idea of representing data in this way is common. Actually, we have already seen an example of this. In natural language processing, such as might arise in the analysis of web pages that we discussed last time with describing web pages, it is called the term-document model, where one has m documents, each of which is described by n terms. Clearly, there are many differences between microarray and DNA SNP matrices, not to mention microarray/SNP matrices and typical term-document matrices and stock price matrices. Nevertheless, from a mathematical perspective, it is worth thinking of them in terms of their similarities. In both cases, we have a bunch of things, each of which is described by a bunch of features, and we want to do operations on one or the other.

2.6.2 Vector addition and scalar multiplication

Taking a step back, the reason that we are interested in vectors is not that we can place them next to each other and play with subscripts to get something called a matrix. We are interested in vectors since they will allow to define something called a vector space, which will allow us to do more complex operations than just flat tables, e.g., to answer questions like the above, and from this matrices will follow. To do this, let’s proceed operationally.

What can we do with vectors? Basically, we are allowed to do two things:

- **Add two vectors:** If $x, y \in \mathbb{R}^n$, then

$$z = x + y = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix} \in \mathbb{R}^n. \quad (2.17)$$

See Figure 2.11(a) for an illustration.

- **Multiply a vector by a scalar:** If $x \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ then

$$y = \alpha x = \alpha \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{pmatrix} \in \mathbb{R}^n. \quad (2.18)$$

See Figure 2.11(b) for an illustration.

We often say that vector addition and scalar multiplication are defined elementwise. By this, we mean the following. For two vectors in \mathbb{R}^n , each entry of the sum vector is defined to be the sum of the elements of the two vectors; and for a vector in \mathbb{R}^n and a scalar in \mathbb{R} , each entry of the product vector is defined to be the product of the scalar and the corresponding element of the vector.

We already know this for \mathbb{R}^2 and \mathbb{R}^3 , both of which are actually simple examples of vector spaces, but these notions generalize to the more general vector space \mathbb{R}^n .

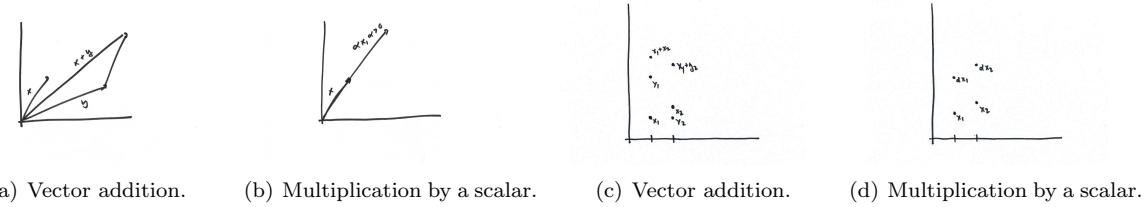


Figure 2.11: Usual visualization and augmented visualization of vector addition and multiplication by scalar.

Remark. Composability. While those two operations seem almost trivial, and while they can be implemented with basic operations from a flat table, their strength comes in their composability: both of those operations give as output a vector in the same vector space, and thus we can apply them recursively. (This is not true of queries and other flat table operations, which in general produce a new quite different flat table.) This composability then leads to a wide range of more complex operations. As a very simple example of this, given two (or more) vectors, $x, y \in \mathbb{R}^n$, we can define their mean/average: $z = \frac{1}{2}(x + y)$, i.e., the mean of two or more vectors is defined to be the vector in which each element is the mean of the elements of those vectors. (Note that this mean/average of a set of vectors is itself a vector, and that is different than the mean/average of the elements of a vector, which is a number. We will see much more of both of these later.)

2.6.3 Using norms to measure distances

Norms are useful for many things, one of which is to measure the distance between two vectors.

Example. If $x = (0, 0)$ and $y = (5, 5)$, then:

- $\|x - y\|_1 = 10$
- $\|x - y\|_2 = \sqrt{25 + 25} \approx 7.07$
- $\|x - y\|_\infty = 5$

Here, we are measuring the norm of the difference vector, and we are considering two vectors in \mathbb{R}^2 , the difference between which is also a vector in \mathbb{R}^2 . Again, $\|x - y\|_\infty \leq \|x - y\|_2 \leq \|x - y\|_1$ and also $\|x - y\|_1 \leq \sqrt{2}\|x - y\|_2 \leq 2\|x - y\|_\infty$.

Two things should be noted. First, the distance between two vectors depends on the norm you use to measure distance. Second, the relationship between those distances is the same as the relationship between the corresponding norms (that we discussed earlier). Clearly, these are true, since the distance is just a norm—of the difference vector.

2.6.4 Using norms to normalize

Another way in which norms are useful is that they can be used to normalize vectors. Normalization refers to the process of taking a vector, computing its norm (with respect to some norm), and then constructing a unit-length vector (in that norm) by multiplying the vector by the inverse of the norm. This is an example of multiplying a vector by a scalar to construct another vector. Any vector, except for a vector in which each entry equals 0 and which thus has norm equal to 0, can be normalized.

Here are several examples.

Example. Let $x = (1, 2)$. In this case:

- Since $\|x\|_1 = 3$, the normalized vector is $x' = \frac{1}{3}(1, 2) = (\frac{1}{3}, \frac{2}{3})$.

- Since $\|x\|_2 = \sqrt{5}$, the normalized vector is $x' = \frac{1}{\sqrt{5}}(1, 2) = (\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}})$.
- Since $\|x\|_\infty = 2$, the normalized vector is $x' = \frac{1}{2}(1, 2) = (\frac{1}{2}, 1)$.

Example. Let $x = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix}$. In this case:

- Since $\|x\|_1 = |\cos(\theta)| + |\sin(\theta)|$, the normalized vector is $x' = \frac{1}{N}x$, where N is the normalization factor.
- Since $\|x\|_2 = 1$, the normalized vector is $x' = x$. That is, for every value of θ , the vector x is normalized, with respect to the Euclidean norm.
- Since $\|x\|_\infty = \max\{|\cos(\theta)|, |\sin(\theta)|\}$, the normalized vector is $x' = \frac{1}{N}x$, where N is the normalization factor.

Similar statements hold for the vector $x = \begin{pmatrix} -\sin(\theta) \\ \cos(\theta) \end{pmatrix}$. While right now these are just vectors with entries, we will see below that these vectors x can be interpreted as rotated versions of $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ or $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Thus, what these results say is that the Euclidean norm does not change when one does a rotation, but that the L_1 norm and the L_∞ norm do change when one does a rotation.

Example. Let $x = (1, 2, 3)$. In this case:

- Since $\|x\|_1 = 6$, the normalized vector is $x' = \frac{1}{6}(1, 2, 3)$.
- Since $\|x\|_2 = \sqrt{14}$, the normalized vector is $x' = \frac{1}{\sqrt{14}}(1, 2, 3)$.
- Since $\|x\|_\infty = 3$ the normalized vector is $x' = \frac{1}{3}(1, 2, 3)$.

Example. Let $x = (1, 1, \dots, 1) \in \mathbb{R}^n$. In this case:

- Since $\|x\|_1 = n$, the normalized vector is $x' = \frac{1}{n}(1, 1, \dots, 1) = (\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n})$.
- Since $\|x\|_2 = \sqrt{n}$, the normalized vector is $x' = \frac{1}{\sqrt{n}}(1, 1, \dots, 1) = (\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})$.
- Since $\|x\|_\infty = 1$, the normalized vector is $x' = (1, 1, \dots, 1) = x$.

Example. Let $x = (1, 0, \dots, 0) \in \mathbb{R}^n$. In this case:

- Since $\|x\|_1 = 1$, the normalized vector is $x' = x$.
- Since $\|x\|_2 = 1$, the normalized vector is $x' = x$.
- Since $\|x\|_\infty = 1$, the normalized vector is $x' = x$.

Note that, in each case, the normalized vector points in the same direction as the original vector, but its norm depends on the norm used to normalize. The vector is unit-norm in the norm used to normalize, but its norm in some other norm is typically not unity.

2.7 Looking forward: algebra and geometry

Looking forward, there are two complementary ways to interpret such a vector x .

- Geometrically: Here, we want to generalize the ideas from Euclidean geometry that you learned in high school. To do so, we will have a dot product between two vectors, i.e., $x \cdot y = \sum_{i=1}^m x_i y_i$, and this will give a particular norm $\|x\|_2 = (x \cdot x)^{1/2}$ (the Euclidean norm), and this will give a notion of angle between two vectors as $\cos(\theta) = \frac{x \cdot y}{\|x\|_2 \|y\|_2}$. Central to this approach are the notions of parallel lines; finding the closest point on a line to a point not on the line (this is related to Euclid's postulates, and it is a special case of a projection we will get to below); conic sections such as circles, ellipses, parabolas, and hyperbolas; etc.
- Algebraically: Here, we want to generalize the ideas from basic algebra that you learned in high school. In this case, we just have vectors and there are the operations of addition of two vectors and multiplication of a vector by a scalar that are defined for them. Central to this approach is that basically all of algebra you learned and the operations you did boil down to following a few notions like commutativity, associativity, distributivity, etc. In a similar way, we will be able to define a small number of operations (that will be slightly different) that will permit us to derive a lot of other useful things.

These two approaches should be familiar from high school mathematics, and we will see that they can be applied to the study of matrices and graphs in data science more generally

2.8 Problems

2.8.1 Pencil-and-paper Problems

1. Recall the definition of the 1-norm: $\|x\|_1 = \sum_{i=1}^n |x_i|$. By showing that it satisfies each of the conditions in the definition of a norm prove this is a vector norm. First do this for \mathbb{R}^2 , and then do this for \mathbb{R}^n .
2. Recall the definition of the ∞ -norm: $\|x\|_\infty = \max_{i \in \{1, \dots, n\}} |x_i|$. By showing that it satisfies each of the conditions in the definition of a norm prove this is a vector norm. First do this for \mathbb{R}^2 , and then do this for \mathbb{R}^n .
3. Consider $x \in \mathbb{R}^n$. For each of the following, prove whether or not it is a norm.
 - (a) The number of non-zero elements of x (which is sometimes called the L_0 norm)
 - (b) $\rho(x) = \sum_{i=1}^n x_i$
 - (c) $\rho(x) = \|x\|_2^2$
4. To prove that the Euclidean norm is a norm is more difficult than we want to do now. (We will get to it soon.) For now, let's assume that it is a norm; and, for simplicity, let's work with \mathbb{R}^3 . For each of the following, either prove that it is a norm, or prove which conditions in the definition of a norm are violated.
 - (a) $\rho(x) = \alpha \|x\|_2$, for $\alpha \in \mathbb{R}$
 - (b) $\rho(x) = \sqrt{4x_1^2 + x_2^2 + x_3^2}$
 - (c) $\rho(x) = \sqrt{4x_1^2 + x_2^2 - x_3^2}$
 - (d) $\rho(x) = \sqrt{\frac{1}{2}(x_1 + x_2)^2 + \frac{1}{2}(x_1 - x_2)^2 + x_3^2}$
 - (e) $\rho(x) = \sqrt{(x_1 + x_2)^2 + (x_1 - x_2)^2 + x_3^2}$
 - (f) $\rho(x) = \sqrt{x_1^2 + x_2^2}$

$$(g) \rho(x) = |x_1| = \sqrt{x_1^2}$$

$$(h) \rho(x) = |x_1 + x_2|$$

Hint: For several of these, it will help to do some sort of change-of-variables or variable substitution and/or try to find a non-zero vector x such that $\rho(x) = 0$.

5. Let $n = 3$, and let $x \in \mathbb{R}^n$ be a vector with $x_i = i^{-1}$. By hand, compute the 1-norm, the 2-norm, and the ∞ -norm. Do the same for the all-ones vector, i.e., $x_i \in \mathbb{R}^3$, where $x_i = 1$, for $i \in \{1, \dots, n\}$.
6. Let $n = 10$, and let $x \in \mathbb{R}^n$ be a vector with $x_i = i^{-1}$. By hand, compute the 1-norm, the 2-norm, and the ∞ -norm. Do the same for the all-ones vector, i.e., $x_i \in \mathbb{R}^{10}$, where $x_i = 1$, for $i \in \{1, \dots, n\}$.
7. For each of the following vectors, compute the corresponding normalized vector. Do this for the L_1 , the L_2 , and the L_∞ norm.
 - (a) $(10, -1, 0, 0, 0)$
 - (b) $(10, -10, 10, -10, 10)$
 - (c) $(10, 10, 10, 10, 10)$
 - (d) $(1, 2, 3, 4, 5)$
 - (e) $(1.0, \frac{1.0}{2.0}, \frac{1.0}{3.0}, \frac{1.0}{4.0}, \frac{1.0}{5.0})$
 - (f) $(-2, -1, 0, 1, 2)$

What can you say about how the magnitudes of the entries of a vector x relate to the ratios $\frac{\|x\|_1}{\|x\|_2}$, $\frac{\|x\|_1}{\|x\|_\infty}$, $\frac{\|x\|_2}{\|x\|_\infty}$ of norms of that vector?

8. Let $u = \begin{pmatrix} -1 \\ 2 \end{pmatrix} \in \mathbb{R}^2$ and $v = \begin{pmatrix} -3 \\ -1 \end{pmatrix} \in \mathbb{R}^2$. Compute and draw on a two-dimensional plane the vectors $u, v, -v, -3v, u+v, u-v$. Can you find numbers $a, b \in \mathbb{R}$, not both equal to zero, such that $au + bv = 0$, where $0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \mathbb{R}^2$? If yes, then provide values for a, b ; if no, then explain why not.
9. In addition, let $w = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \in \mathbb{R}^2$. Can you find numbers $a, b, c \in \mathbb{R}$, not all equal to zero, such that $au + bv + cw = 0$, where $0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \mathbb{R}^2$. If yes, then provide values for a, b, c ; if no, then explain why not.

2.8.2 Implementations and Applications of the Theory

1. In python, using an array, write functions to compute each of the 1-norm, the 2-norm, and the ∞ -norm.
 - (a) Let $n = 10$, and let $x \in \mathbb{R}^n$ be a vector with $x_i = i^{-1}$. Use your functions to compute the 1-norm, the 2-norm, and the ∞ -norm of this vector. Confirm that you obtain the same results as when you compute these norms by hand.
 - (b) Use your functions to compute all three norms for the all-ones vector in \mathbb{R}^n , i.e., where $x_i = 1$, for $i \in \{1, \dots, n\}$, where again $n = 10$. Again, confirm that you obtain the same results as when you compute norms by hand.
 - (c) Do the same for the vector $x \in \mathbb{R}^{10}$, where each x_i is a random number from $[0, 1)$.
2. In python, for $n = 100$ and $n = 1000$ and $n = 1000000$, and for each of the three vectors ($x_i = i^{-1}$; $x_i = 1$ for all i ; and x_i is random between $[0, 1)$):
 - (a) Compute the 1-norm, the 2-norm, and the ∞ -norm.
 - (b) For each value of n , which of these three vectors is the ratio $\|x\|_1/\|x\|_\infty$ the largest? The smallest?

- (c) Explain this in terms of how uniform or nonuniform are the elements of the vector?
- (d) How does this vary with n ?
3. Let's consider different ways to compute area of a "ball" in a "box" in two-dimensions, as in Section 2.4. Let $n = 2$, and let B_p be the unit ball in the ℓ_p norm.
- (a) By hand, compute $\text{Area}(B_2)/\text{Area}(B_\infty)$.
 - (b) In python, write a script to estimate $\text{Area}(B_2)/\text{Area}(B_\infty)$.
 (Hint: first, choose a random point from within B_∞ (call this a trial); second, check whether this point lies within B_2 (call this a success); third, return the ratio $\rho = N_{\text{success}}/N_{\text{trial}}$ of the number of successes to the number of trials.)
 - (c) Do this for $N_{\text{trial}} = 1, 10, 10^2, 10^3, 10^4, 10^5, 10^6$ trials, i.e., throws of the dart. Return the list of numbers; and make a plot of ρ versus $\log_{10}(N_{\text{trial}})$.
 - (d) How does this compare with $\pi/4$? Equivalently, how does the plot of 4ρ versus $\log_{10}(N_{\text{trial}})$ compare with π .
4. Do the same things in three-dimensions, i.e., for $n = 3$ (except, here, compare with $\pi/6$ and how the plot of 6ρ versus $\log_{10}(N_{\text{trial}})$ compares with π). Based on your results from one, two, three dimensions, how do you expect the ratio of volumes to behave as the dimension increases?
5. Consider the function $f(\theta) = \sin(\theta)$ over the interval $\theta \in [0, 6\pi]$ (when measured in radians, so three full cycles, i.e., $\theta \in [0, 1080]$ when measured in degrees).
- (a) In python, using an array, discretize the interval $[0, 6\pi]$ into (say) 40 elements, call them $\{\theta_i\}_{i=1}^{40}$, and define the vector $x \in \mathbb{R}^{40}$ to have elements $x_i = f(\theta_i)$, for $i \in \{1, \dots, 40\}$. Plot this vector, i.e., plot (i, x_i) , for $i \in \{1, \dots, 40\}$.
 - (b) Let $\varepsilon \in \mathbb{R}^{40}$ be the vector in which each element ε_i is a random number between $[-0.1, 0.1]$. On the same figure as the previous plot, plot this vector, i.e., plot (i, ε_i) , for $i \in \{1, \dots, 40\}$.
 - (c) Generate ten vectors of the form $y = x + \varepsilon$, where the randomness in ε is different each time, and plot all of them on a single plot.
 - (d) On the same figure as the previous plot, plot the mean/average of these ten vectors. What effect does taking the mean have on the noisiness of these vectors?
6. In addition to $f(\theta) = \sin(\theta)$ over the interval $\theta \in [0, 6\pi]$, consider $g(\theta) = \exp(-\kappa\theta)$, where $\kappa = 2\pi$, and $h(\theta) = 1$. In each case, discretize the interval into (say) 40 elements, and let $x, y, z \in \mathbb{R}^{40}$ be a discrete sinusoidal vector, a discrete decaying exponential vector, and a discrete all-ones vector.
- (a) Plot $\alpha x + (1 - \alpha)y$, for $\alpha \in \{0, 0.2, 0.4, 0.6, 0.8, 1.00\}$ (use a different color for each value of α). What do you notice?
 - (b) Plot $\alpha y + (1 - \alpha)z$, for $\alpha \in \{0, 0.2, 0.4, 0.6, 0.8, 1.00\}$ (use a different color for each value of α). What do you notice?

Chapter 3

Vector spaces, matrices, and linear functions

3.1 Introduction to vector spaces and subspaces

We have been talking about vectors as things that are indexed by integers $\{1, \dots, n\}$. Let's be somewhat more formal. More formally, vectors are abstract things that are elements of a vector space, where a vector space is an abstract thing that satisfies certain rules. These rules are essentially the vector addition and scalar multiplication that we discussed at the end of Chapter 2. The operations of adding two vectors and multiplying a vector by a scalar are simple, but they are very powerful. Indeed, these two operations form the foundation for all of linear algebra. One reason they are so powerful is that the output of these two operations is itself a vector, upon which those operations can be performed again. This leads to the notion of a vector space.

3.1.1 Vector space

We start with the definition of a vector space. One could define vector spaces more generally, but the following will be sufficient for our purposes. We'll actually give two definitions.

Definition 3 A nonempty set V is a vector space iff

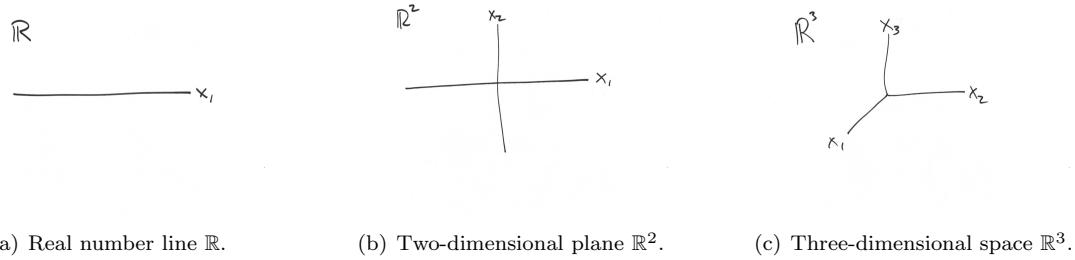
$$x, y \in V \text{ and } a \in \mathbb{R} \text{ implies that } x + y \in V \text{ and } ax \in V. \quad (3.1)$$

Definition 4 A nonempty set V is a vector space iff

$$x, y \in V \text{ and } a, b \in \mathbb{R} \text{ implies that } ax + by \in V. \quad (3.2)$$

It is not hard to show that Definition 3 and Definition 4 provide equivalent definitions of a vector space, in the sense that if a set V satisfies Condition 3.1 of Definition 3 then V satisfies Condition 3.2 of Definition 4, and vice versa.

Condition 3.1, equivalently Condition 3.2, is sometimes referred to by saying that *a vector space V is “closed” under the operations of vector addition and under multiplication by scalars*. This is an important concept (in linear algebra and in mathematics more generally). Let's parse that phrase.

Figure 3.1: Illustration of \mathbb{R} , \mathbb{R}^2 , and \mathbb{R}^3 .

- **Vector addition.** This is an addition operation on two vectors, the output of which is another vector. From Eqn. (2.17), for two vectors in \mathbb{R}^n , this is defined elementwise.
- **Scalar multiplication.** This is a multiplication operation between a vector and a scalar number, the output of which is another vector. From Eqn. (2.18), for a vector in \mathbb{R}^n and a scalar in \mathbb{R} , this is defined elementwise.
- **Closed.** The “closed” property means that if we do addition of two vectors or the scalar multiplication of a vector and a scalar, then we end up with another vector in the same vector space.

It's not hard to show that if we consider the set of vectors $V = \mathbb{R}^n$ and scalars \mathbb{R} , then V is closed under vector addition and scalar multiplication, i.e., that $V = \mathbb{R}^n$ satisfies the definition of a vector space.

Note that the operations of vector addition and scalar multiplication are very different than operations (such as query, filter, join, and count) that are common on flat tables, where the output is typically a very different flat table, and thus can't be combined in a meaningful way with the original tables.

Advanced comment. While we will be most interested in the vector space \mathbb{R}^n , there are many other examples of vector spaces. For example, the set of continuous functions, the set of square integrable functions, and even the set of $m \times n$ matrices.

3.1.2 Subspaces of a vector space

We will be most interested in the vector space \mathbb{R}^n , and we will be interested in it since elements of it, i.e., vectors $x \in \mathbb{R}^n$, can be used to model data.

The next question is: what does a vector space such as \mathbb{R}^n “look like”? While this question is not particularly precise, we feel like we know the answer for $n = 1$ and $n = 2$ and $n = 3$. In these cases, roughly, it looks like the familiar real number line, the two-dimensional plane, and three-dimensional space, respectively. See Figure 3.1 for an illustration. If $n > 3$, this question can be more difficult to answer. The typical approach is to consider very simple subsets of \mathbb{R}^n , so simple in fact that they are often trivial for \mathbb{R} or \mathbb{R}^2 or \mathbb{R}^3 . Here are several examples that we mentioned previously.

- Unit ball.
- Positive orthant.
- Positive simplex.
- Subspace, e.g., a line through the origin.

Unit balls and positive orthants and positive simplices are relatively straightforward things, although there can be some subtle/interesting issues, such as our discussion in Chapter 2 regarding how the unit ball depends on the choice of norm. A more subtle but extremely import concept is that of a subspace.

A line through the origin of a two-dimensional plane is the simplest non-trivial example of a subspace, but subspaces are very important more generally. We will go into a lot more detail about this, since it forms the basis for a lot of linear algebra, and we will give more examples in Chapter 3.1.4. Let's start with the abstract definition.

Here is the definition of a subspace of the vector space \mathbb{R}^n .

Definition 5 *A nonempty subset $V \subset \mathbb{R}^n$ is a subspace if it is closed under addition and closed under multiplication by scalars, i.e., V is a subspace iff $x, y \in V$ and $a \in \mathbb{R}$ implies that $x + y \in V$ and $ax \in V$.*

Here, we have arbitrarily chosen to use Condition 3.1 rather than Condition 3.2. As with Definitions 3 and 4, however, here in Definition 5, we could equivalently have written this condition as: $x, y \in V$ and $a, b \in \mathbb{R}$ implies that $ax + by \in V$.

Remark. The two operations under which a subspace is closed are exactly the same two operations that define a vector space. In particular, this means that a subspace of a vector space is itself a vector space in its own right. Thus, anything we say about vector spaces qua vector spaces also holds true for all the subspaces of that vector space. This holds not only for the easy-to-visualize line through the origin in \mathbb{R}^2 , but also for more general harder-to-visualize subspaces of \mathbb{R}^n . This will turn out to be extremely useful, since understanding the properties of the subspaces of a high-dimensional vector space is a powerful way to understand that vector space.

Remark on the remark. A line through the origin is a one-dimensional subspace of the two-dimensional plane \mathbb{R}^2 . To describe a point on that line, one needs two coordinates, x_1 and x_2 , but they are constrained by an equation of the form $x_2 = ax_1$, and so only one of them is “free” or “independent.” This suggests that simply equating the number of elements in a vector with the dimension, as we did naïvely in Chapter 2, is not completely correct in general. We will be more precise and general about this later, but informally the solution is the following. If we fix a , then we need one number to determine the position of a point on the line in \mathbb{R}^2 . For example, once a is fixed, i.e., once the line is specified, then the point $x = \begin{pmatrix} \alpha \\ a\alpha \end{pmatrix} \in \mathbb{R}^2$ is completely determined by the one number $\alpha \in \mathbb{R}$, and vice versa. Thus, we will say that a line is a one-dimensional subspace of \mathbb{R}^2 .

3.1.3 Standard basis vectors and not-standard basis vectors

Definition. Let's start with the following definition.

Definition 6 *The standard basis vectors in \mathbb{R}^n , denoted e_i , have n entries, with a 1 in the i^{th} position and a 0 in all the other positions.*

For example, in \mathbb{R}^2 , the standard basis vectors are

$$e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix};$$

and in \mathbb{R}^3 , the standard basis vectors are

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{and} \quad e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Note that the dimensional dependence of e_i is typically implicit, and thus the notation e_i is overloaded. Usually, this is not a problem, since the dimension with which one is working is known.

Usefulness in general. The standard basis vectors are very important since they can be used—along with the operations of vector addition and scalar multiplication—to “describe” or “write” or “express” any vector. For example, consider the vector $x \in \mathbb{R}^2$ with coordinates

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

In this expression $x \in \mathbb{R}^2$, and $x_1, x_2 \in \mathbb{R}$. Then, it should be clear that

$$x = x_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = x_1 e_1 + x_2 e_2.$$

Here is an example where we need to be careful with notation: $x_i \in \mathbb{R}$, while $e_i \in \mathbb{R}^2$.

Similarly, consider the vector $x \in \mathbb{R}^3$ with coordinates

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

It should be clear that

$$x = x_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = x_1 e_1 + x_2 e_2 + x_3 e_3.$$

More generally, for a vector

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n,$$

we can express x in terms of the n standard basis vectors $\{e_i\}_{i=1}^n$ as

$$x = x_1 e_1 + \cdots + x_n e_n = \sum_{i=1}^n x_i e_i.$$

Question. What is important here?

Answer. The important thing here is that any vector in \mathbb{R}^n can, using the operations of vector addition and scalar multiplication, be expressed in terms of the standard vectors in a unique way. That is why it is called a basis. Other sets of vectors for which this is true are also interesting, and some of those other sets of vectors will be even better for use in data science.

Question. What is less important here?

Answer. A *much* less important thing is that the standard basis vectors look “nice” to humans. By “nice,” we mean the following. Since standard basis vectors have a single non-zero entry which equals 1, and the rest of the entries are 0, they are simple for humans to look at and to think about. What is simple for humans, however, is not necessarily simple for computers, and vice versa. A lot of linear algebra will involve generalizing these ideas to vectors that are “nice” more generally, e.g., in terms of the data, for the computer to work on, etc.

Usefulness in data science. In data science, the standard basis vectors define the way the data are given to you, e.g., the way they are measured or collected in the application area that generated the data, but they are not always the most useful way to describe, understand, and use the data. To see a simple example of this, consider Figure 3.2, which illustrates several different examples of sets of data points on \mathbb{R}^2 . Figures 3.2(a), 3.2(b), and 3.2(c) illustrate a scattering of data points, with somewhat different properties.



(a) Data points on a two-dimensional plane, scattered in a round manner.
(b) Data points on a two-dimensional plane, scattered in an elongated manner.
(c) Data points on a two-dimensional plane, scattered in an elongated manner, in a different elongated manner.
(d) Same data points on a two-dimensional plane, scattered in an elongated manner, but with rotated axes.

Figure 3.2: Examples of data points on a two-dimensional plane and how to describe them.

In Figure 3.2(a), different data points have different values for the two variables/axes, x_1 and x_2 , but both axes seem important to capture properties of the data. We'll discuss later more precise ways to capture this informal claim, but at this point this should be contrasted with the data in Figures 3.2(b) and 3.2(c), both of which are much more "elongated" in two slightly different ways.

In Figure 3.2(b), the elongation is along the standard axes. This could be due to measuring different features in different units, or it could be due to one feature being more "important" in some sense. In either case, two things should be noted. First, the elongation conforms well with the variables used to measure the data. Second, one might hope or expect that one could ignore the second variable and use just the first variable and do nearly as well in downstream data science applications. That is, the numerical value of x_2 is much smaller than the numerical value of x_1 , and this *may* mean that most of the information of interest is captured by x_1 , while x_2 might be less important or simply random noise. In this case, we might hope or expect get very similar results by considering only (x_1) , rather than (x_1, x_2) , for each data point.

In Figure 3.2(c), on the other hand, the data are elongated along some other direction. Again, there could be different reasons or interpretations for this, but two things should be noted. First, as in Figure 3.2(b), there is one direction on the plane that seems more "important" than the other perpendicular direction on the plane. Second, this direction is not the same as the actual variables used to measure the data, i.e., it is not axis-aligned and does not correspond to either of the standard axes, but instead it is some sort of linear combination (a rotation here) of the standard axes. In this case, we have also plotted the line $x_2 = ax_1$, which points in the direction of the elongation.

Let's go in to the example in Figure 3.2(c) in more detail. Recall that any data point on the plane as well as any point on the line can be described by a combination of the standard basis vectors e_1 and e_2 . This requires giving a coefficient with which to multiply e_1 and a coefficient with which to multiply e_2 . For the points on the line $x_2 = ax_1$, those two numbers are not independent—specifying one nails down the other.

If we rotate by the angle θ between e_1 and $\begin{pmatrix} 1 \\ a \end{pmatrix}$, then we obtain a new vector that points in the direction $e'_1 = \begin{pmatrix} 1 \\ a \end{pmatrix}$. If we rotate by the same angle θ the vector e_2 , then we obtain a new vector that points in the direction $e'_2 = \begin{pmatrix} a \\ -1 \end{pmatrix}$. Note that just as $e_1 \cdot e_2 = 0$, since we have rotated both vectors by the same angle, so too $e'_1 \cdot e'_2 = 0$, i.e., e'_1 and e'_2 are also perpendicular. This is illustrated in Figure 3.2(d).

With respect to these new basis vectors: any point on the line can be described by one number (the magnitude along e'_1); and, in the same way as the example in Figure 3.2(b) suggests that we might be able to consider only one coordinate axis, here we might be able to consider only one of the new coordinate axes (e'_1) and ignore the other (e'_2) and still be able to do something useful with the data. (To do this will require being more precise about the notion of eigenvectors and eigenvalues, a topic to which we will return later in the term.)

Summarizing this discussion, there are several points here.

- The vector $\begin{pmatrix} 1 \\ a \end{pmatrix}$ seems more natural to describe the data in Figure 3.2(c).
- Using this more natural description it takes 1 number rather than 2 numbers.
- The line through the origin defined by $\begin{pmatrix} 1 \\ a \end{pmatrix}$ —as well as the line through the origin defined by the $\begin{pmatrix} a \\ -1 \end{pmatrix}$ perpendicular to it—is a one-dimensional subspace of \mathbb{R}^2 .

When we go to \mathbb{R}^n , for $n \geq 3$, there is a much richer variety of subspaces, and so subspaces can be used to find less obvious structure.

Not-standard basis vectors. In the same way that any point on the plane can be expressed in terms of the standard basis vectors, so too any point on the plane can be expressed (via the two basic operations of linear algebra) in terms of the two vectors $\begin{pmatrix} 1 \\ a \end{pmatrix}$ and $\begin{pmatrix} a \\ -1 \end{pmatrix}$. (This should be clear informally, and we will be more precise about this below. Getting the algebraic details right takes some effort, to which we will get soon, and the computations are more tedious than with the standard basis vectors, but this is the sort of thing at which computers excel.) As such, they are a second example of a set of basis vectors for \mathbb{R}^2 . These two vectors are also perpendicular, and they have many other “nice” properties. For example, the first seems to be very natural to describe the data in Figure 3.2(c). We will see that, in many cases, basis vectors like these, where one or more seem to be a natural way to express the data, are even more appropriate than the standard basis vectors for use in data science.

3.1.4 Some initial examples of subspaces and not-subspaces in two dimensions

Let’s discuss in more detail the operations of vector addition and scalar multiplication, what it means to be closed under these operations, and what are the implications of being closed. To do so, let’s provide examples of things that are and are not subspaces of \mathbb{R}^2 .

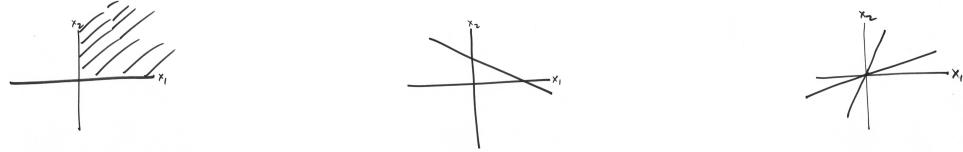
Examples of not-subspaces. Here are examples of things that are *not* subspaces of \mathbb{R}^2 .

- **Example.** $S^1 \subset \mathbb{R}^2$, i.e., the L_2 ball, is *not* a subspace.
 - It is not closed under addition: for example, $\begin{pmatrix} 1 \\ 0 \end{pmatrix} \in S^1$ and $\begin{pmatrix} 0 \\ 1 \end{pmatrix} \in S^1$, but their sum $\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \notin S^1$.
 - Also, it is not closed under scalar multiplication: for example, $2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix} \notin S^1$.

The same is true for the L_2 sphere.

- **Example.** The L_1 ball/sphere is *not* a subspace.
- **Example.** The L_∞ ball/sphere is *not* a subspace.
- **Example.** The positive orthant is *not* a subspace.
 - It is closed under addition: if $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ and $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ are in the positive orthant, then they have nonnegative entries, and so

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \end{pmatrix}$$



(a) Positive orthant. (b) Line not through the origin. (c) Two lines through the origin.

Figure 3.3: Illustration of subsets of \mathbb{R}^2 that are *not* a subspace.

has nonnegative entries and so is in the positive orthant.

- But it is not closed under scalar multiplication, in particular when the scalar is negative: for example:

$$-2 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -2x_1 \\ -2x_2 \end{pmatrix}$$

has negative (or nonpositive) entries, if x_i are positive (or nonnegative).

The positive orthant is, however, closed under multiplication of nonnegative scalars, a fact that is sometimes of interest.

- **Example.** The positive simplex is *not* a subspace.

- It is not closed under the addition of two vectors, and it is nor closed under scalar multiplication.
- It is, however, closed under the special class of vector additions of the form $z = ax + by$, where x, y, z are vectors, and a, b are numbers such that $a, b \geq 0$ and $a + b = 1$. We will see later why this is of interest.

- **Example.** The line L given by the equation $x_1 + x_2 = 1$, which can also be written as $x_2 = 1 - x_1$ or $x_2 = -x_1 + 1$, is *not* a subspace.

- For example, $\begin{pmatrix} 1 \\ 0 \end{pmatrix} \in L$ and $\begin{pmatrix} 0 \\ 1 \end{pmatrix} \in L$, but $\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \notin L$.

- **Example.** The set of points Ω that is the union of two lines through the origin is *not* a subspace. That is, let

$$\begin{aligned} \Omega_1 &= \{(x_1, x_2) : x_1 = ax_2\} \\ \Omega_2 &= \{(y_1, y_2) : y_1 = ay_2\} \\ \Omega &= \Omega_1 \cup \Omega_2. \end{aligned}$$

- The set Ω is closed under multiplication by scalar, since multiplying a point by a scalar generates another point on that same line.
- The set Ω is not closed under addition of two vectors, since if you add two vectors from two different lines lead to a vector that is not on either of those lines (except in the degenerate case when the two lines are the same).

Thus, the set Ω is not a subspace.

See Figure 3.3 for an illustration of several of these examples of things that are *not* a subspace of \mathbb{R}^2 .

Examples of subspaces. Here are examples of things that are subspaces of \mathbb{R}^2 .

- **Example.** Lines through the origin. To see this, let's start with a question. Question: why is or is not the line $x_2 = ax_1 + b$ (perhaps more familiar as $y = ax + b$) a subspace? To answer this, let $x \in \mathbb{R}^2$ and $y \in \mathbb{R}^2$ be two points on this line, i.e., let

$$\begin{aligned} x_2 &= ax_1 + b \quad \text{and} \\ y_2 &= ay_1 + b \end{aligned}$$

both hold. Then, consider the point

$$z = x + y = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \end{pmatrix}.$$

By adding the above two equations, we have that

$$x_2 + y_2 = a(x_1 + y_1) + 2b.$$

This point is *not* on the same line—unless $b = 0$, in which case this point is on the same line.

- If $b \neq 0$, then the line $x_2 = ax_1 + b$ is *not* a subspace.
- If $b = 0$, then we have that $x_2 = ax_1$, and so the line the line $x_2 = ax_1 + b = ax_1$ is a subspace.

Compare and contrast Figure 3.3(b) and 3.4(b) for an illustration of this.

If that seemed backward, let's prove this directly. Consider the line $x_2 = ax_1$ (where, recall, this is $y = ax$ in the high school notation, but we use x_1 and x_2 since the ideas will generalize to \mathbb{R}^n). If the points $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ and $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ are both on this line, then the point

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \end{pmatrix}$$

is also on the line, since $x_2 + y_2 = a(x_1 + y_1)$, i.e., vector addition is satisfied. In addition, the point

$$\alpha \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \alpha x_1 \\ \alpha x_2 \end{pmatrix}$$

is also on the line, since $\alpha x + 2 = a\alpha x_1$, i.e., multiplication by a scalar is satisfied. (By the way, that is a “proof” of the claim that the set of points that satisfy $x_2 = ax_1$ is a subspace.)

Strictly speaking, we haven't proved that *any* line through the origin is a subspace, since the vertical line through the origin can't be written in the form $x_2 = ax_1$, for $a \in \mathbb{R}$. It can, however, be written in the form

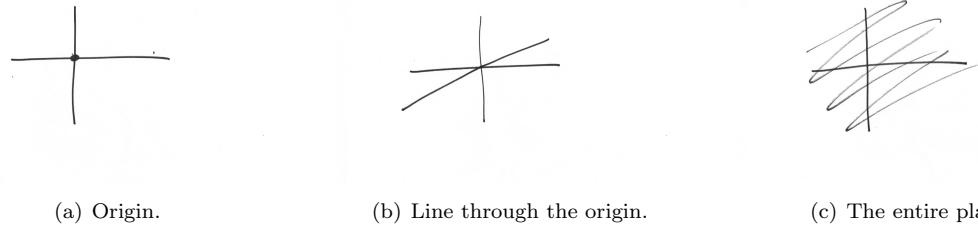
$$\Omega_2 = \{x \in \mathbb{R}^2 : x = \alpha e_2\},$$

for $\alpha \in \mathbb{R}$, where, recall, $e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ is the second standard basis vector. In the same way, the horizontal line through the origin can be written as

$$\Omega_1 = \{x \in \mathbb{R}^2 : x = \alpha e_1\},$$

where e_1 is the first standard basis vector. Clearly, if $x \in \Omega_1$ and $y \in \Omega_2$, then x and y are perpendicular. This completes the proof that *any* line through the origin on \mathbb{R}^2 is a subspace.

The caveats in the previous example about lines expressed in particular parametric forms are awkward, and they make it difficult to think about things more generally. We can deal with them in a one-off manner on \mathbb{R}^2 , but we would like to avoid them when dealing with more general vector spaces, e.g., \mathbb{R}^n , where we have less intuition about “corner cases.” Before we do that, here are two other examples of subspaces of \mathbb{R}^2 . They seem trivial, but understanding why they are subspaces is important as we generalize.

Figure 3.4: Illustration of subsets of \mathbb{R}^2 that are a subspace.

- **Example.** The origin. We can see that the origin $0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ is a “trivial” subspace of \mathbb{R}^2 since: if we multiply it by any scalar or add it to itself, then the output is still the 0 vector.
- **Example.** All of \mathbb{R}^2 . We can see that all of \mathbb{R}^2 is a “trivial” subspace of \mathbb{R}^2 since: \mathbb{R}^2 is a vector space; and \mathbb{R}^2 is a subset of itself. (If that latter observation isn’t clear, we’ll discuss it in more detail when we discuss basic set in the probability chapters.)

See Figure 3.4 for an illustration of several of these examples of things that are a subspace of \mathbb{R}^2 . (We will see later that these are all the subspaces of \mathbb{R}^2 , there are no others.)

3.1.5 Subspaces in \mathbb{R} , \mathbb{R}^2 , \mathbb{R}^3 , and beyond

Before discussing more general situations and higher dimensions, let’s discuss in a bit more detail some (hopefully) intuitive examples.

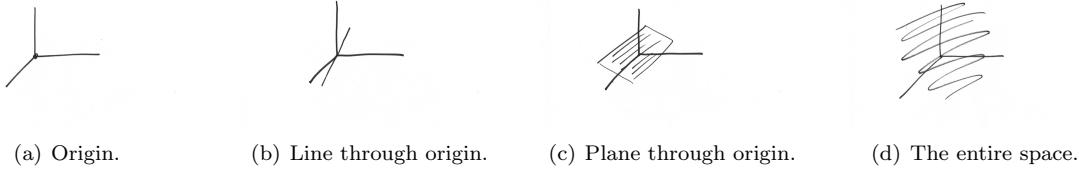
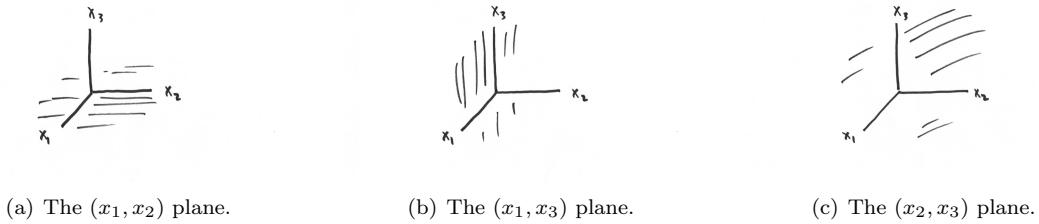
To start, before discussing even \mathbb{R}^3 , let’s consider an even lower dimensional space, i.e., $\mathbb{R}^1 = \mathbb{R}$. There are two types of subspaces for \mathbb{R} :

- All of \mathbb{R} .
- The origin $0 \in \mathbb{R}$.

The one-dimensional \mathbb{R} is itself a vector space; and it has a one-dimensional subspace (itself) and a zero-dimensional subspace (the origin). Both of these are “trivial,” in the sense that there is not too much interesting going on, and so one typically does not spend much time discussing the subspace aspects of \mathbb{R} , but it’s good to understand such “boundary cases” in the definitions of vector spaces and subspaces.

The examples of the previous section suggest (correctly, as we discussed above) that there are three kinds of subspaces for \mathbb{R}^2 :

- A plane through the origin, i.e., the entire two-dimensional plane, is a subspace of dimension 2, and it takes 2 numbers to specify a point on the plane.
- A line through the origin is a subspace of dimension 1, and it takes 1 number to specify a point on a line. (It may take more than that to specify the line, but once the line is specified, there is only one “free” variable, the other is specified by the equation defining the line.)
- The origin itself $\{0\}$ is a subspace of dimension 0, and it takes 0 numbers to specify the location of a point in it (since specifying it specifies everything there is to specified).

Figure 3.5: Illustration of subsets of \mathbb{R}^3 that are a subspace.Figure 3.6: Illustration of two-dimensional axis-aligned subspaces of \mathbb{R}^3 .

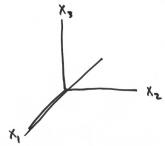
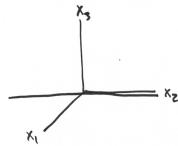
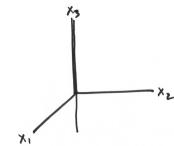
Before going forward, e.g., to describe \mathbb{R}^n , think about how these ideas generalize and apply to \mathbb{R}^3 . We will see that there are four kinds of subspaces for \mathbb{R}^3 :

- A 3-space through the origin (i.e., the entire three-dimensional space) is a three-dimensional subspace of \mathbb{R}^3 , and one needs 3 numbers to specify the location of a point in all of \mathbb{R}^3 .
- A plane through the origin is a two-dimensional subspace of \mathbb{R}^3 , and one needs 2 numbers to specify the location of a point on a plane through the origin of \mathbb{R}^3 .
- A line through the origin is a one-dimensional subspace of \mathbb{R}^3 , and one needs 1 number to specify the location of a point on a line through the origin of \mathbb{R}^3 .
- The origin itself is a zero-dimensional subspace of \mathbb{R}^3 , and it takes 0 numbers to specify the location of a point in it.

See Figure 3.5 for an illustration of several of these examples of things that are a subspace of \mathbb{R}^3 .

By the way, these claims should be “obvious” in certain special cases. For example, consider two-dimensional planes that are axis-aligned, i.e., defined by $x_3 = 0$ or $x_2 = 0$ or $x_1 = 0$, in which case it reduces to the usual (x_1, x_2) or (x_1, x_3) (x_2, x_3) plane, respectively. See Figure 3.6 for illustrations of two-dimensional axis-aligned subspaces of \mathbb{R}^3 . As an example that may be easier to visualize, think of the floor and two perpendicular walls of a typical room as defining these planes, e.g., the floor or walls of the classroom. Similarly, consider one-dimensional lines that are axis-aligned, e.g., the intersection of two walls or between a wall and the floor of the classroom, as that too may be easier to visualize as a one-dimensional subspace of \mathbb{R}^3 . See Figure 3.7 for illustrations of one-dimensional axis-aligned subspaces of \mathbb{R}^3 .

These examples are “obviously” lower-dimensional, e.g., since if one performs computations then one just has to “carry along and ignore” all the zeros. The important point is that a two-dimensional subspace of \mathbb{R}^3 could just as well be at some other orientation, and a one-dimensional subspace of \mathbb{R}^2 or \mathbb{R}^3 could also just as well be at some other orientation. These other subspaces that are not axis-aligned may be harder to visualize, but they are no less interesting mathematically, and they are often much more important for data science applications. *The reason they are often much more important for data science applications is that*

(a) The x_1 line.(b) The x_2 line.(c) The x_3 line.Figure 3.7: Illustration of one-dimensional axis-aligned subspaces of \mathbb{R}^3 .

the canonical axes are often not the most useful to describe the data. They might be easier for a human to think about, but they are not much easier for the computer to deal with. Thus, we want to generalize these ideas to arbitrarily-oriented subspaces that are no harder for a computer to deal with and that are much more useful to describe the data.

This is the intuition you should have about going to higher dimensional spaces like \mathbb{R}^n , since it will form the basis for understanding matrix operations in higher dimensional spaces. Being somewhat more precise about these claims will allow us to generalize all these ideas to \mathbb{R}^n , which we are not able to visualize, but this will require some mathematical machinery, e.g., the ideas of linear combinations, linear dependence/independence, span, basis vectors, etc., and we will get to this soon. It's important to keep in mind, however, while we discuss the mathematical machinery, that these are the intuitions that we are trying to generalize to \mathbb{R}^n .

3.1.6 Proving something is a subspace

Here is an example of how one tries to prove things more generally about subspaces.

Often we would like to prove that certain things we construct or define are or are not subspaces. (Note that we have already done this: we defined the positive orthant, and we proved that it is not a subspace; and we defined lines through the origin on the plane, and we proved that they are subspaces.) Moreover, we would like to do this assuming as little as possible. As a simple example of this, consider the following.

Claim 1 Given a vector $x \in \mathbb{R}^2$, such that $x \neq 0$, let x^\perp be the set of vectors that are perpendicular to x . Then, x^\perp is a subspace.

Proof: To see this, let $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, in which case an element $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in x^\perp$ satisfies the condition

$$y_1 x_1 + y_2 x_2 = 0.$$

Observe that this is the equation of a line through the origin—it is simply the line perpendicular to the vector x . See Figure 3.4(b). Thus, by the previous example, this is a subspace. \diamond

We know from the previous discussion that the line that contains the vector x is a subspace. Thus, Claim 1 says that the set of points perpendicular to the one-dimensional subspace defined by $x \neq 0$ is a one-dimensional subspace. (It actually doesn't say that it is a one-dimensional subspace, but that is revealed in the proof.)

Question. Why do we have the requirement that $x \neq 0$?

Answer. Because we were not thinking as generally as we could. It turns out it isn't necessary. We can show the following.

Claim 2 Given a vector $x \in \mathbb{R}^2$, let x^\perp be the set of vectors that are perpendicular to x . Then, x^\perp is a subspace.

Proof: Either $x = 0$ or $x \neq 0$. If $x \neq 0$, then the result follows by Claim 1. If $x = 0$, then the set of vectors that are perpendicular to x is all of \mathbb{R}^2 , which we know is a subspace. \diamond

In Claim 2, the set x^\perp is either a line through the origin or the entire plane, depending on whether or not x is the all-zeros vector.

Question. Why do we only consider the set of points that are zero-dimensional or one-dimensional subspaces to be perpendicular with respect to?

Answer. Because we were not thinking as generally as we could. It turns out it isn't necessary. We can show the following.

Claim 3 Let V be a subspace of \mathbb{R}^2 , and let V^\perp be the set of vectors that are perpendicular to V . Then, V^\perp is a subspace.

Proof: Either V is the origin, or V is a line through the origin, or V is all of \mathbb{R}^2 . The first two cases reduce to Claim 2. For the third case, observe that the set of vectors that are perpendicular to all the vectors on \mathbb{R}^2 consist only of the origin $0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, which we know is a subspace. \diamond

Question. Why do we only consider the set of points that are a subspace V of \mathbb{R}^2 rather than some more general vector space like \mathbb{R}^n ?

Answer. Because we were not thinking as generally as we could. It turns out it isn't necessary. As we will show later, a similar statement holds for any subspace of \mathbb{R}^n . (It actually holds even more generally, although we won't discuss that).

All of this discussion may seem pedantic or like over-kill, since we already know that the origin and lines through the origin and all of \mathbb{R}^2 are subspaces of \mathbb{R}^2 . What is interesting here is the form of Claim 3. That is, look at what Claim 3 claims, not how we proved it. Claim 3 basically says the following:

- if we have a vector space and *any* subspace of that vector space, then if we consider the set of vectors that satisfy some rule (here, those that are perpendicular to the hypothesized subspace), then we have another subspace.

That is quite different than our previous claims. In particular, we need to know very little about V except that it is a subspace. We don't need to know its dimension or a parametric form for it or that it is easy to compute or that it is useful or anything else like that. When we can prove statements of this form, they are very powerful, since they hold very generally, even if we know very little. A lot of thinking about matrices and vectors and high-dimensional spaces takes this form.

3.2 Matrices and operations on matrices, including matrix multiplication

Sometimes, matrices are simply “defined” as an array of numbers, without talking about operations. (While not incorrect, and while sometimes useful for performing certain matrix computations, this approach highlights the structural form and downplays the operations that are of interest to us.) Here is such a definition.

Definition 7 An $m \times n$ matrix is a rectangular array of entries, m high and n wide, i.e., with m horizontal rows and n vertical columns.

In the same way that we say that a real number is an element of \mathbb{R} and that a vector with real-valued entries lies in \mathbb{R}^n , we will say that an $m \times n$ matrix with real-valued entries lies in $\mathbb{R}^{m \times n}$.

Of greatest interest is when the elements of that array are real numbers, i.e., in \mathbb{R} , but they could be other things, e.g., Boolean values, integers, complex numbers, polynomials, other matrices, etc.

Note that, from this perspective, vectors and numbers are simple matrices.

- A vector $x \in \mathbb{R}^m$, viewed as a column vector, is an $m \times 1$ matrix.
- Alternatively, a vector $x \in \mathbb{R}^n$, viewed as a row vector, is a $1 \times n$ matrix.
- A number $x \in \mathbb{R}$ is a 1×1 matrix.

3.2.1 Two operations: matrices are vectors

When we go to operations, as we have said, matrices are of interest not just since they are an array of numbers, but really due to the operations defined on them. Here are two operations of interest.

- **Addition of two matrices.** This operation is defined for two matrices A and B of the same size, e.g., two $m \times n$ matrices, in which case the sum matrix $C = A + B$ is a matrix of the same size that has entries

$$C_{ij} = B_{ij} + A_{ij},$$

for $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$. For example,

$$\begin{pmatrix} 1 & 0 \\ 2 & -1 \\ 4 & 2 \end{pmatrix} + \begin{pmatrix} 0 & -3 \\ 1 & -2 \\ 3 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -3 \\ 3 & -3 \\ 7 & 3 \end{pmatrix}.$$

- **Multiplication of a matrix by a scalar.** Given any matrix A and any real number α , this operation gives a matrix of the same size with each entry multiplied by the scalar, i.e., the matrix $B = \alpha A$ has elements

$$B_{ij} = \alpha A_{ij},$$

for $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$. For example,

$$2 \begin{pmatrix} 1 & 3 \\ -2 & 4 \end{pmatrix} = \begin{pmatrix} 2 & 6 \\ -4 & 8 \end{pmatrix}.$$

Given these two results, it is not too hard to show that the set of $m \times n$ real-valued matrices, sometimes denoted $\mathbb{R}^{m \times n}$ is itself a vector space, i.e., that it satisfies the (addition and multiplication by scalar) conditions in the definition of a vector space. (By the way, this is an example of why it is best to think of vectors and matrices in terms of the operations they support rather than as something with a certain number of subscripts.)

Problem. Show that the set of $m \times n$ real-valued matrices, with addition and scalar multiplication defined above, form a vector space.

3.2.2 A third operation: matrices are more than just vectors

Question. How is the array of mn numbers, call them x_i , for $i \in \{1, \dots, mn\}$, different when we view them as a vector $x \in \mathbb{R}^{mn}$ than when we view them as a matrix, call it X , with entries X_{ij} , for $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$? We can define norms and dot products and angles between vectors, and (since the set of $m \times n$ matrices is itself a vector space) it turns out that we can define the same operations for the matrix. We will see some examples of this below, and this is often quite useful. So, what is the “real” difference?

Answer. Recall that for vectors, there were basically two operations you could do: you could add two vectors, and you could multiply a vector by a scalar. You can do those for matrices, but you can also do a third thing: you can multiply two matrices to get a third matrix. There are several ways to *define* the multiplication of two matrices, one of which is extremely important. Let's go into more detail on that.

To start, note that we have actually seen a few examples of this before.

- **Random walk transition matrix.** Recall the random walk transition matrix: given $x \in \mathbb{R}^n$, compute $y = Ax$, then we can compute $z = Ay$, and so on. From this it follows that $z = A(Ax)$. We will see that we can define the multiplication of two matrices, and that multiplication will be associative, so that $z = A(Ax) = (AA)x = A^2x$. This has the interpretation of applying the random walk transition matrix twice.
- **The dot product.** We defined this as a way to “multiply” two vectors to get a number, and from that we could talk about norms, angles, etc. Alternatively, note that vectors themselves can be viewed as matrices with 1 row/column, and numbers can be viewed as 1×1 matrices, so this is an operation that takes as input two matrices of a special form and returns a matrix of a special form.

As a reminder, recall some of the properties of addition and multiplication on real numbers: in particular, they are both commutative and associative, and they satisfy the distributive property, and there is an identity for each and inverse, with the exception of 0 for multiplication. Keep these in mind, since some of these properties will generalize, but some won’t.

Onto matrix multiplication. There are multiple ways one can define the product of two matrices. Here we present two: a perhaps more intuitive but less useful notion; and an initially less intuitive but much more important notion. Both are presented for context, but we are interested in the more useful notion.

- **Hadamard product of matrices.** (*Initially more intuitive, but less useful.*)
This is a matrix product that is defined for any two matrices A and B that are the same size, i.e., that have the same dimensions, e.g., two $m \times n$ matrices. In this case, the Hadamard matrix product $C = A \circ B$ is a matrix of the same size that has entries

$$C_{ij} = B_{ij} * A_{ij},$$

for $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$. While simple to define, and while of some use in certain situations, this is not a particularly useful or important notion of matrix multiplication.

- **Matrix multiplication.** (*Initially less intuitive, but much more useful.*)
This is in a sense the most important matrix property qua matrix property. A *much* more useful notion of the product of two matrices is given by an initially-confusing definition, to which we now turn. Given an $m \times n$ matrix A and an $n \times p$ matrix B , the product $C = AB$ is an $m \times p$ matrix with elements

$$C_{ij} = \sum_{k=1}^n A_{ik}B_{kj}.$$

This definition is given as Definition 8 below, and we now turn to describing it in more detail.

To understand the more useful notion of multiplication of general matrices, let's return to the dot product. Before we talked about it as an operation on vectors that gave norms, angles, etc., and now we'll describe the same thing as a simple example example of matrix multiplication and as a way to introduce general matrix multiplication.

Recall the dot product:

$$a \cdot b = \sum_{i=1}^n a_i b_i, \quad \text{where } a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \quad \text{and } b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix},$$

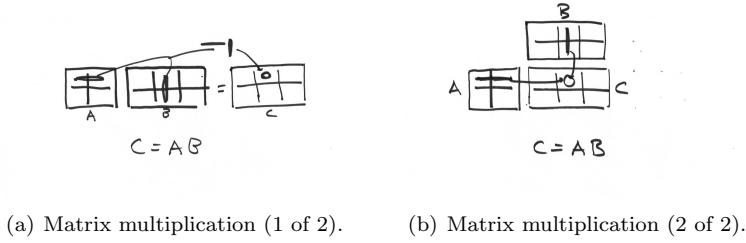


Figure 3.8: Illustration of matrix multiplication.

which basically takes two vectors, multiplies them element-wise, and sums them up. Let's write the same thing in a different way:

$$c = \text{dot}(a, b) = a \cdot b = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix}^T \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = \sum_{i=1}^n a_i b_i \in \mathbb{R},$$

where we have defined the $1 \times n$ matrix a^T to be a row matrix with the same entires as the $n \times 1$ column matrix a . This is a simple example of matrix multiplication: given as input two matrices, a $1 \times n$ matrix a^T and a $n \times 1$ matrix b , it returns as output a number $c \in \mathbb{R}$, i.e., a 1×1 matrix, i.e., $\mathbb{R}^{1 \times 1}$, and it does so by taking the element-wise product and “summing up” the n terms in the inner dimension.

More generally, we can define the product of two matrices as follows.

Definition 8 Given an $m \times n$ matrix A and an $n \times p$ matrix B , the product $C = AB$ is an $m \times p$ matrix with elements

$$C_{ij} = \sum_{k=1}^n A_{ik} B_{kj}.$$

That is, for every one of the mp elements of C , the element C_{ij} is computed as the dot product of the i^{th} row of A and the j^{th} column of B . If we denote the i^{th} row of A by $A_{i:} \in \mathbb{R}^n$ and the j^{th} column of B by $B_{:j} \in \mathbb{R}^n$, then

$$C_{ij} = \text{dot}(A_{i:}, B_{:j}) = \sum_{k=1}^n A_{ik} B_{kj}.$$

See Figure 3.8 and the equation below for illustrations of this.

For this definition of matrix multiplication, the “inner dimension” must be the same, but the “outer dimensions” can be different (from each other as well as from the inner dimension). This is very important, as otherwise the product of two matrices isn't even defined. In fact, this is often not even stated, and you might see a product written as AB . This assumes that the inner dimension is the same, since otherwise the product isn't even defined.

3.2.3 Examples of matrix multiplication

Here are some examples of matrix multiplication.

- If $A = \begin{pmatrix} 2 & -1 \\ 3 & 2 \end{pmatrix}$ and $B = \begin{pmatrix} 1 & 4 & -2 \\ 3 & 0 & 2 \end{pmatrix}$, then

$$AB = \begin{pmatrix} 2 & -1 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 1 & 4 & -2 \\ 3 & 0 & 2 \end{pmatrix} = \begin{pmatrix} -1 & 8 & -6 \\ 9 & 12 & -2 \end{pmatrix}.$$

In this case, BA is not defined: B is a 2×3 matrix, and A is a 2×2 matrix; and so each row of B has three entries, while each column of A has two entries.

- If $A = \begin{pmatrix} 1 & 0 \\ 2 & 3 \end{pmatrix}$, $B = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$, $C = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix}$, and $D = \begin{pmatrix} 1 & 0 \\ 2 & 2 \\ 1 & 1 \end{pmatrix}$, then we have the following:

- $AB = \begin{pmatrix} 0 & 1 \\ 0 & 5 \end{pmatrix}$
- $BA = \begin{pmatrix} 2 & 3 \\ 2 & 3 \end{pmatrix}$
- $AC = \begin{pmatrix} 1 & -1 & 1 \\ 5 & -2 & -1 \end{pmatrix}$
- CA is not defined.
- $CD = \begin{pmatrix} 0 & -1 \\ 0 & -1 \end{pmatrix}$
- $DC = \begin{pmatrix} 1 & 0 \\ 2 & 2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 \\ 4 & -2 & 0 \\ 2 & -1 & 0 \end{pmatrix}$

This example illustrates several things about the product of two matrices:

- $AB \neq BA$, and so matrix multiplication is *not* commutative in general;
- AC is defined, but CA is not defined, so both need not be defined, and in particular both are not defined unless $m = p$ in Definition 8; and
- if both are defined, then their dimensions need not be the same and are not unless $m = n = p$.

Here is another example of the product of two matrices.

- Given $A = \begin{pmatrix} 1 & 2 \end{pmatrix}$ and $B = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$,

- $AB = \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 4 \end{pmatrix} = 11.$

This is just the dot product we saw before, also known as the *inner product*.

- $BA = \begin{pmatrix} 3 \\ 4 \end{pmatrix} \begin{pmatrix} 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ 4 & 8 \end{pmatrix}.$

This is sometimes known as the *outer product* of the vectors A and B (since these matrices can be viewed as vectors in this example).

Matrix multiplication takes a particularly simple form when one of the matrices involved in the multiplication is one of the standard basis vectors. For example, the i^{th} column of a matrix A is Ae_i , e.g,

$$\begin{pmatrix} 3 & -2 & 0 \\ 2 & 1 & 2 \\ 0 & 4 & 3 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -2 \\ 1 \\ 4 \\ 0 \end{pmatrix}.$$

This holds for $i \in \{1, 2, 3\}$. Thus, if we consider the matrix $(e_1 \ e_1 \ e_3)$ in which the i^{th} vertical column is e_i , then we have

$$\begin{pmatrix} 3 & -2 & 0 \\ 2 & 1 & 2 \\ 0 & 4 & 3 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 3 & -2 & 0 \\ 2 & 1 & 2 \\ 0 & 4 & 3 \\ 1 & 0 & 2 \end{pmatrix}. \quad (3.3)$$

That is, the matrix is left unchanged.

Relatedly, the i^{th} column of AB is Ab_i , where b_i is the i^{th} column of B . For example, consider the matrix product:

$$\begin{pmatrix} 2 & -1 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 1 & 4 & -2 \\ 3 & 0 & 2 \end{pmatrix} = \begin{pmatrix} -1 & 8 & -6 \\ 9 & 12 & -2 \end{pmatrix}.$$

We can view this matrix multiplication in terms of post-multiplying the first matrix by the columns of the second matrix, in which case we get columns of the product matrix.

$$\bullet \begin{pmatrix} 2 & -1 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} -1 \\ 9 \end{pmatrix}$$

$$\bullet \begin{pmatrix} 2 & -1 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 4 \\ 0 \end{pmatrix} = \begin{pmatrix} 8 \\ 12 \end{pmatrix}$$

$$\bullet \begin{pmatrix} 2 & -1 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} -2 \\ 2 \end{pmatrix} = \begin{pmatrix} -6 \\ -2 \end{pmatrix}$$

Alternatively, we can view this matrix multiplication in terms of pre-multiplying the second matrix by the rows of the first matrix, in which case we get rows of the product matrix.

$$\bullet (2 \ -1) \begin{pmatrix} 1 & 4 & -2 \\ 3 & 0 & 2 \end{pmatrix} = (-1 \ 8 \ -6)$$

$$\bullet (3 \ 2) \begin{pmatrix} 1 & 4 & -2 \\ 3 & 0 & 2 \end{pmatrix} = (9 \ 12 \ -2)$$

We saw above that matrix multiplication is not commutative. It is, however, associative, meaning that we can drop parentheses and not worry about the order of multiplication. Here is the result. We will actually go through the proof to practice working with matrix-matrix multiplication.

Claim 4 Let A be an $m \times n$ matrix, B be an $n \times p$ matrix, and C a $p \times q$ matrix, so that $(AB)C$ and $A(BC)$ are defined. Then $(AB)C = A(BC)$.

Proof: Recall that $(AB)C$ is a matrix with mq elements, indexed as $\alpha \in \{1, \dots, m\}$ and $\beta \in \{1, \dots, q\}$. Let's consider the $(\alpha\beta)$ element:

$$((AB)C)_{\alpha\beta} = \sum_{\ell=1}^p (AB)_{\alpha\ell} C_{\ell\beta} \quad (3.4)$$

$$= \sum_{\ell=1}^p \left(\sum_{k=1}^n A_{\alpha k} B_{k\ell} \right) C_{\ell\beta} \quad (3.5)$$

$$= \sum_{\ell=1}^p \sum_{k=1}^n A_{\alpha k} B_{k\ell} C_{\ell\beta} \quad (3.6)$$

$$= \sum_{k=1}^n \sum_{\ell=1}^p A_{\alpha k} B_{k\ell} C_{\ell\beta} \quad (3.7)$$

$$= \sum_{k=1}^n A_{\alpha k} \left(\sum_{\ell=1}^p B_{k\ell} C_{\ell\beta} \right) \quad (3.8)$$

$$= \sum_{k=1}^n A_{\alpha k} (BC)_{k\beta} \quad (3.9)$$

$$= A(BC). \quad (3.10)$$

Here, Equations (3.4) and (3.5) follow by expanding out the matrix multiplication; Equations (3.6), (3.7), and (3.8) follow by re-ordering the terms of the sum since addition of real numbers is associative; and Equations (3.9) and (3.10) follow by the definition of matrix multiplication.

◊

Problem. Show that the set of $m \times n$ real-valued matrices, sometimes denoted $\mathbb{R}^{m \times n}$ is itself a vector space, i.e., that it satisfies the (addition and multiplication by scalar) conditions in the definition of a vector space.

3.2.4 A complementary perspective on matrix multiplication

We have described matrix multiplication very operationally, i.e., in terms of element-wise operations on the input matrices to get an element of the output matrix. Here, we will describe a complementary perspective. This complementary perspective allows you to think about rows and columns of A and B in a slightly different way, and thus it is helpful to understand better what is going on, but it is equivalent to the first way we presented matrix multiplication.

Before doing that, note the following. The approach we described above is operational in the following sense: given matrices A and B of appropriate dimensions, an element of the product matrix $C = AB$ is the dot product of a row of A and a column of B . A special case of this is if B is a column vector, in which case the dot product of the rows of A with B gives the elements of the column vector C ; or if A is a row vector, in which case the dot product of A with the columns of B gives the elements of row vector C . A special case of this is if A is a row vector and B is a column vector, in which case $C \in \mathbb{R}^1 = \mathbb{R}$ is just the dot product of A and B .

Let's not view matrices in terms of their elements, but instead let's view them more explicitly in terms of their rows and columns, and then define operations on them.

Let's instead view an $m \times n$ matrix

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & & & \\ \vdots & & & \vdots \\ A_{m1} & \cdots & & A_{mn} \end{pmatrix}.$$

as a row vector

$$A = (A_{:,1} \ A_{:,2} \ \cdots \ A_{:,n}),$$

each element of which is a column vector

$$A_{:,k} = \begin{pmatrix} A_{1k} \\ A_{2k} \\ \vdots \\ A_{mk} \end{pmatrix}.$$

Let's also view an $n \times p$ matrix

$$B = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1p} \\ B_{21} & & & \\ \vdots & & & \vdots \\ B_{n1} & \cdots & & B_{np} \end{pmatrix}.$$

as a column vector

$$B = \begin{pmatrix} B_{1,:} \\ B_{2,:} \\ \vdots \\ B_{m,:} \end{pmatrix},$$

each element of which is a row vector

$$B_{k,:} = (B_{k1} \ B_{k2} \ \cdots \ B_{kn}).$$

Then, the product matrix $C = AB$ is just the dot product of A viewed as a row vector and B viewed as a column vector, i.e., it is just the sum of things. Importantly, each thing in this sum is the product of a column of A , i.e., an $m \times 1$ vector, and a row of B , i.e., a $1 \times n$ vector, i.e., it is an $m \times n$ matrix.

This perspective may at first seem a little unusual, but it makes sense operationally, i.e., we can just define the various operations algebraically and see that it works, which makes it easier for a computer to do, and it also makes sense when matrices are viewed as linear transformations, i.e., in a way that makes it easier for people to think about. We turn to that topic next.

3.3 Functions, linear functions, and linear transformations

3.3.1 Functions and transformations

So far, we have talked about matrices as describing points in high-dimensional Euclidean spaces, and we have given some sense that they are related to transformations. Let's next talk in more detail about viewing matrices as transformations.

Recall the following definition.

Definition 9 A function is a rule that consists of three things: two sets, called the domain and range, and a rule that associates to each element in the domain one element in the range.

Sometimes this is written as “Let $f : X \rightarrow Y$ be a function, where X is the domain, Y is the range, and f specifies the rule.” For example, a simple example of a function is $f(x) = \sin(x)$ or $f(x) = ax^2$ or $f(x) = ax$ or $f(x) = \log(x)$, where in each case the domain X is (perhaps some subset of) \mathbb{R} and the range is (perhaps some subset of) \mathbb{R} .

We are going to be interested in functions that have more general domains and ranges. In particular, we will be interested in when the domains and ranges are \mathbb{R}^m or \mathbb{R}^n . We have already seen several examples of

this. For example, the random walk transition matrix took as input a vector in \mathbb{R}^n and returned as output a (usually different) vector in \mathbb{R}^n that describes how mass on the nodes changes after one step of the random walk. Alternatively, if we fix a , then the dot product $a \cdot x = f(x)$ transforms a vector $x \in \mathbb{R}^n$ into a number $a \cdot x \in \mathbb{R}$. There are lots of functions we could have between vector spaces, just as there are lots of functions from \mathbb{R} to \mathbb{R} , but the following will be a particularly important class of functions.

Definition 10 Let V be a vector space such as \mathbb{R}^n or a subspace of \mathbb{R}^n . Then a linear function is a mapping f such that for $x, y \in V$ and $\alpha \in \mathbb{R}$ we have that $f(x + y) = f(x) + f(y)$ and $f(\alpha x) = \alpha f(x)$.

This definition applies much more generally, but here we are interested in \mathbb{R}^n or subspaces of \mathbb{R}^n .

Here are examples of functions that are *not* linear functions.

- $y_1 = (x_1 - 2)^2 + 3$ is *not* a linear function.
- $y_1 = \sin(x_1)$ is *not* a linear function.
- $y_1 = ax_1 + b$, where $a, b \in \mathbb{R}$, which may be written in a more familiar manner as $f(x) = ax + b$, is *not* a linear function, in general. This is potentially a source of confusion, since that is the equation of a line on the two-dimensional plane. It's not a linear function, however, by Definition 10. We want to use Definition 10 since we want to relate the idea of a linear function to linear subspaces, and a requirement for that will be that the origin maps to the origin. Note that this is not true if $b \neq 0$.

The class of functions such as $y_1 = ax_1 + b$ for $b \neq 0$ is a sufficiently important class of functions, however, that it has a special name. It's called an *affine function*, meaning basically a linear function plus a constant offset. Note that b could be anything, but in data science b often corresponds to something like the mean, and thus subtracting it corresponds to preprocessing that is mean-centering. In this case, linear functions are mean centered, while affine functions are not mean centered. Both are of interest, and they are clearly related, but they are different. This notion of affine functions also generalizes to higher dimensional spaces (which we will get back to later).

- $y = a_{11}x_1^2 + a_{12}x_1x_2 + a_{22}x_2^2$ and $y = a_{11}x_1^2 + a_{12}x_1x_2 + a_{22}x_2^2 + b_1x_1 + b_2x_2 + c$, where all variables are real numbers, are *not* linear functions.

Here are examples of functions that are linear functions.

- $y_1 = ax_1$, which may be written in a more familiar manner as $f(x) = ax$, is a linear function. This is a special case of the example of an affine function, $y_1 = ax_1 + b$, but with $b = 0$.
- The function that takes as input the vector $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ and returns as output the vector

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{pmatrix} \quad (3.11)$$

is a linear function. Note that by writing the RHS of Equation (3.11) as a matrix-vector product, this can be written as

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

- The composition of two (and thus more than two) linear functions is a linear function. For example, if $f(x) = ax$ and $g(x) = bx$ (where x, a, b, f, g are all real numbers in \mathbb{R}), then $f(g(x)) = a(bx) = (ab)x$, which is a linear function. Alternatively, consider the linear function that takes as input the vector $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ and returns as output the vector

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} b_{11}y_1 + b_{12}y_2 \\ b_{21}y_1 + b_{22}y_2 \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}. \quad (3.12)$$

If we consider the function that first applies the linear function of Equation (3.11) and then applies the linear function of Equation (3.12), then we have

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

If this is written out, it's easy to show that this is a linear function.

Note that you might have been thinking of the function in Definition 10 as having a domain \mathbb{R}^n and range \mathbb{R} , but we could easily call that f_1 and have f_2, \dots, f_n , in which case we have both the domain and range \mathbb{R}^n . The example given in Equation (3.11) is an example of this, with $n = 2$. (Also, the random walk transition matrix was a function $\mathbb{R}^n \rightarrow \mathbb{R}^n$.) We could even have the range be \mathbb{R}^m , for $m \neq n$.

3.3.2 Connection between functions and matrices

We have talked about high dimensional vector spaces, and we have given some indication that those spaces have some counterintuitive properties, relative to \mathbb{R}^2 . We also know that, while some functions on the plane are fairly simple, e.g., $y_1 = ax_1$, others can be much more complicated. Thus, we might ask whether it is possible to represent in a concise way a function between two high-dimensional Euclidean spaces, say \mathbb{R}^m and \mathbb{R}^n .

- In general, the answer is NO.
- If the function is a linear function, then the answer is YES. In addition, the representation of this linear function is as a matrix. (In particular, the elements of a matrix describe how one elementary basis vector in the domain gets mapped to the elementary basis vectors in the range.)

Here is a fact: if we have a function $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$, then if it is a linear function it can be fully described by mn numbers, call them A_{ij} , for $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$. This should *not* be obvious, and we will describe in more detail next time. Given this connection, since we can do all sorts of things with functions, it *suggests* that we can do more complex operations that flat tables and have these operations be more meaningful. For example, to quantify the similarity in the word counts in the last 10 chapters of the book you saw in the first class.

We will spend a lot of time on this.

3.3.3 Transformations and matrices as transformations

So far, we have been thinking of matrices as sets of points in \mathbb{R}^n , and we have started discussing operations on matrices. For example, given two vectors we can operate on them to compute a dot product and from that an angle, which gives a measure of how close the two vectors are on the unit sphere; and given one vector we can operate on it with a norm, which gives a notion of size. We know that the set of matrices forms a vector space, and so we can add them, multiply them by a scalar, etc.; and we also have the operation of matrix multiplication, which gives the set of matrices additional structure.

Now, let's turn to the question: What is this additional structure? Relatedly, what is this matrix multiplication doing? Relatedly, what does it "mean"? Relatedly, why is it so useful? We will consider the answers to these questions, but notice that we have two examples of this so far.

- **Dot product.** We have viewed this in a couple of ways: as computing a function of two vectors; and as doing a matrix-matrix multiplication. Now, let's view $\text{dot}(x, y) = y^T x$ as a function that fixes y and takes x as an input. Then $f(x) = f_y(x) = y^T x$, and this is a function that takes as input a vector $x \in \mathbb{R}^n$ and returns as output a number that is an element of \mathbb{R} that equals $\sum_{i=1}^n x_i y_i$, where recall y is assumed to be fixed.

- **Random walk transition matrix.** Here, we take as input a vector $x \in \mathbb{R}^n$ and return a vector $y \in \mathbb{R}^n$, where $y_i = \sum_{j=1}^n A_{ij}x_j$. Of course, we could send y in as input and iterate the process, but regardless it is a function that transforms input vector in \mathbb{R}^n and returns a vector in \mathbb{R}^n .

This perspective as viewing a matrix in terms of transformations or as a function is true more generally: an $m \times n$ matrix can be used to characterize a certain class of functions (linear transformations) from \mathbb{R}^m to \mathbb{R}^n ; and, conversely, every linear transformation can be written as a matrix..

Viewing matrices as linear transformations (a special kind of function) is the central notion of linear algebra, and it allows us to see matrices and matrix operations as more than just subscript chasing. Moreover, not only is this very common and very powerful, but it explains the perhaps-initially non-obvious definition of matrix multiplication: defined this way, applying a function twice (as in iterating the random walk matrix) corresponds exactly to doing a matrix-matrix multiplication of the two matrices.

As an example, consider $A = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 0 & 1 \end{pmatrix}$, which can be viewed as a transformation with domain \mathbb{R}^3 and range \mathbb{R}^2 . For example, if $v = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \in \mathbb{R}^3$, then

$$w = Av = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 5 \\ 2 \end{pmatrix} \in \mathbb{R}^2.$$

The $m \times n$ matrix A transforms vectors from \mathbb{R}^n to \mathbb{R}^m according to the rule $w_i = \sum_{j=1}^n A_{ij}v_j$, for $i = 1, 2$ (which, recall, is a linear function).

More abstractly, here is the definition of a linear transformation.

Definition 11 A linear transformation $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a mapping s.t. for all scalars $a \in \mathbb{R}$ and vectors $v, w \in \mathbb{R}^n$: (1) $T(u + v) = T(v) + T(v)$; and (2) $T(av) = aT(v)$.

This should be familiar, since we had similar expressions before. The point here is that we are talking about general functions/transformation, without any reference to matrices/vectors/etc., but we will soon see that there is a very close connection.

Here is a “gotcha” about which we will need to be careful, since it is often a source of confusion initially. Question: What happens if we apply T to the 0 vector. The answer is that $T(0) = 0$. Be careful, though, since the two different 0s here mean different things: the 0 inside the parentheses has n elements, all of which are 0, i.e., it lives in \mathbb{R}^n , i.e., that $0 \in \mathbb{R}^n$; and the 0 to the right of the equals sign has m elements, all of which are 0, i.e., it lives in \mathbb{R}^m , i.e., that $0 \in \mathbb{R}^m$. Recall also that this means that a linear transformation goes through the origin.

Why is this? More abstractly, it is often convenient to do math by looking at some abstract space, say a high-dimensional Euclidean space like \mathbb{R}^n or a subspace of \mathbb{R}^n , that has addition and scalar multiplication as structure, and then ask what properties are “preserved” by certain classes of transformations. Here, linear transformations preserve structure in the following sense: one can first add two vectors or multiply by a scalar, then do the mapping, or first do the mapping and then add the mapped vectors or multiply them by a scalar. The offset that affine transformations messes that up, so it is typically dealt with separately, and not included in the definition.

Definition 12 An affine transformation is a transformation that is a linear transform plus a constant offset.

In particular, this means that the function would be a linear transformation if it were modified by taking the image of the origin and transforming it back to the origin.

Note that affine transformations give outputs that are first degree polynomials in the inputs, but they could have a term that is a zero order polynomial, i.e., a constant, while linear transformation have only first order terms, e.g., the origin has been shifted and/or variables redefined. In \mathbb{R} , this is just another way of saying that $y_1 = ax_1 + b$, for $b \neq 0$, is not a linear function, but $y_1 = ax_1$ is a linear function. This holds true more generally.

Here is an example:

$$\begin{aligned} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &\rightarrow \begin{pmatrix} x_1 - x_2 + 2 \\ 2x_1 + x_2 + 1 \end{pmatrix} \text{ is affine} \\ &= \begin{pmatrix} x_1 - x_2 \\ 2x_1 + x_2 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \end{aligned}$$

i.e., $x \rightarrow Ax + b$, where the first term is the linear transformation and the second term is the affine offset. Here the constant offset $b = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \in \mathbb{R}^2$ is a vector offset and not a scalar offset.

Here are two important theorems that we won't prove but that the above discussion suggests.

Theorem 1 (Any matrix gives a linear transformation) *Let A be an $m \times n$ matrix. Then, A defines a linear transformation $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ by matrix multiplication: $T(v) = Av$, where v is a column vector.*

Theorem 2 (Any linear transformation gives a matrix) *Every linear transformation $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is given by an $m \times n$ matrix, call it A . The functional form if given by $T(v) = Av$, i.e., a matrix -vector multiplication, where the i^{th} column of A is $T(e_i)$.*

Taken together, Theorem 1 and Theorem 2 say that matrices and linear transformations are essentially the same thing, in the following two complementary senses.

- Every $m \times n$ matrix corresponds to a linear transformation from \mathbb{R}^m and \mathbb{R}^n .
- Every linear transformation from \mathbb{R}^m and \mathbb{R}^n is given by an $m \times n$ matrix.

For completeness, we note the following results, which states that if we have a linear transformation corresponding to a matrix, then the inverse linear transformation corresponds to the inverse matrix.

Theorem 3 (Invertability of linear transformation is equivalent to invertability of matrix) *The linear transformation $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is invertible iff the $m \times n$ matrix A associated with it is invertible, and $T^{-1} = A^{-1}$. Fact: to have an inverse, we need that $m = n$.*

That is an overview of viewing a single matrix as a linear transformation. There are many more results along these lines, and they form the basis for more advanced presentations of linear algebra. We will mention only the following, which has to do with the initially-counterintuitive definition of matrix-matrix multiplication given in Definition 8.

If we then compose two linear transformations, then we get a linear transformation. (We may show that on a homework.) Since the composition is a linear transformation, we can represent it as a matrix. The following theorem says that that composed transformation is given as a matrix by the matrix product of the two matrices corresponding to the two transformations, where the matrix product is given by Definition 8. We won't prove it, but it is a key result, in the linear transformation theory underlying matrix operations used in data science and more generally.

Theorem 4 (Composability of linear transformations is equivalent to matrix multiplication) *Let $S : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $T : \mathbb{R}^m \rightarrow \mathbb{R}^\ell$ be linear transformations, given by matrices A and B , respectively. Then, the linear transformation $S \circ T$ (where \circ represents functional composition) is given by $A \cdot B$ (where \cdot represents the matrix product).*

3.4 Special types of matrices

3.4.1 Transpose of a matrix

Given a matrix A , here is a related matrix that is often of interest.

Definition 13 *Given an $m \times n$ matrix A , the transpose A^T of A is the matrix formed by interchanging the rows and columns, i.e., $(A^T)_{ij} = A_{ji}$.*

We have already seen this. For example, given an n -vector A , represented as a column vector, $A = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$, which an $n \times 1$ matrix, the transpose, $A^T = (a_1 \ a_2 \ a_3)$, is a row vector, or equivalently a $1 \times n$ matrix.

Here is another example. If $A = \begin{pmatrix} 1 & 4 & -2 \\ 3 & 0 & 2 \end{pmatrix}$, then $A^T = \begin{pmatrix} 1 & 3 \\ 4 & 0 \\ -2 & 2 \end{pmatrix}$. In this case A and A^T are of

different dimensions. But, since one is $m \times n$ and the other is $n \times m$, it is always the case that AA^T and A^TA are defined. Here, AA^T is a 2×2 matrix, and A^TA is a 3×3 matrix. (Matrices of the form AA^T and A^TA are important matrices, and we will get back to them later.)

Here are two things that are good to know about transposes.

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$

3.4.2 Inverse of a matrix

Motivation and definition. Given a matrix $A \in \mathbb{R}^{m \times n}$, one often wants to find a matrix, let's call it $A^{-1} \in \mathbb{R}^{n \times m}$, such that $AA^{-1} = I_m$ and/or $A^{-1}A = I_n$. Call these a right inverse and a left inverse, respectively. We won't go into the details here, but for our purposes inverses only exist for square matrices, and in those cases if a left inverse exists, then it is a right inverse, and vice versa, and so we will only refer to it as an inverse. In addition, for us, if we are talking about inverses, then $m = n$.

This approach to matrices is of particular interest from a linear equation perspective, since in that case a (textbook, but not good in practice) way to solve $y = Ax$ for x is to premultiply each side of the equation by A^{-1} to get $x = A^{-1}Ax = A^{-1}y$. Even though this linear equation perspective isn't the approach we emphasize, we will discuss some of those issues in later chapters. More generally, inverses are very important, and they can be used to introduce other important ideas. While we are talking about inverses, we don't know yet that they even exist, so let's talk about that.

Definition 14 *An invertible matrix is a matrix with an inverse.*

When inverses exist and how to compute them. Not every matrix has an inverse, and for matrices that do, computing the inverse is non-trivial (and almost never actually needed in practice—we will get to this later). Before discussing them, it is good to know, e.g., whether they exist.

Question. When does such an inverse matrix exist, i.e., when is a matrix invertible?

Answer. Here is the answer (or at least the beginning of it).

- $m = n = 1$: We can write a 1×1 matrix as $A = (a)$. In this case $A^{-1} = \frac{1}{a}$, which is defined for all $a \neq 0$. So, for 1×1 matrices, i.e., real numbers, such a number exists for every $a \in \mathbb{R}$ s.t. $a \neq 0$. This is particularly simple, and the situation is considerably more subtle for larger matrices.
- $m = n = 2$: We can write a 2×2 matrix as

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad (3.13)$$

in which case

$$A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}, \quad (3.14)$$

which is defined only when $ad - bc \neq 0$, and otherwise the inverse is not defined. To check that the matrix given in Equation (3.14) is the inverse of the matrix given in Equation (3.13), and vice versa, observe that

$$AA^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \frac{1}{ad - bc} = \begin{pmatrix} ad - bc & cd - cd \\ ad - bc & ab - cd \end{pmatrix} \frac{1}{ad - bc} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and that

$$A^{-1}A = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} ad - bc & ac - ac \\ bd - bd & ad - bc \end{pmatrix} \frac{1}{ad - bc} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

To understand the role of $ad - bc$ in whether the inverse exists, observe that if A is the all-zeros matrix, then $a = b = c = d = 0$, in which case $ad - bd = 0$ and so the inverse doesn't exist. This is the 2×2 generalization of the 1×1 case when the single number equals 0. But we can have $ab - cd = 0$ in another case as well, i.e., when one column of A is a scalar multiple of the other column. Let's call the two columns of A v and w . Then $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$ and $w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$, where $w_i = \alpha v_i$ for some scalar α . In this case, $ad - bc = v_1w_2 - w_1v_2 = v_1(\alpha v_2) - (\alpha v_1)v_2 = 0$. (We could run this in reverse to show that this is necessary and sufficient.)

So, the 2×2 matrix A doesn't have an inverse in two cases: (1) when it is the all-zeros matrix, in which case it doesn't have any non-trivial information and it sends all input vectors to the zero vector; and (2) when its two columns are the same, up to scaling, in which case you might imagine that it is missing some information. Note that this degeneracy corresponds to multiplication by a scalar, i.e., one column is a scalar multiple of the other column. Note also that if $\alpha = 0$, then the second column is the all-zeros column, and it still holds that the matrix is not invertible. By the way, this thing is called the determinant, and for the matrix A in Equation (3.13), it is sometimes denoted as

$$\det(A) = \left| \begin{array}{cc} a & b \\ c & d \end{array} \right| = ad - bc,$$

i.e., with horizontal bars.

- $m = n \geq 3$: Both of the previous cases appear in higher dimensions, but in higher dimensions a more nontrivial sort of degeneracy appears, namely that one column can be expressed as a sum or linear combination of two of the other columns. For example, the matrix could consist of three columns, call them u and v and w , and it is possible that no column is a scalar multiple of another column, but we can find numbers $\alpha, \beta \in \mathbb{R}$ such that $u = \alpha v + \beta w$. Because in higher dimensions there are many ways to do this, this is the most important reason for lack of invertibility of larger matrices. To make some of the above discussion more precise, we'll need a few notions having to do with linear dependence and spanning subspaces that we will get to next time, and this will be the foundations for linear algebra.

Here is a basic result about inverses.

Claim 5 *If A and B are invertible, then AB is invertible, and $(AB)^{-1} = B^{-1}A^{-1}$.*

Proof: Let's show that $B^{-1}A^{-1}$ gives the identity when acting on the left and right.

$$\begin{aligned} AB(AB)^{-1} &= ABB^{-1}A^{-1} = AIA^{-1} = AA^{-1} = I \\ (AB)^{-1}AB &= B^{-1}A^{-1}AB = B^{-1}IB = B^{-1}B = I \end{aligned}$$

Note that in the above chain of equalities we didn't specify the order of operations since we know that matrix-matrix multiplication is commutative. \diamond

Note the simplicity of this result—it didn't involve any subscript chasing or determinants or anything messy (like what is about to follow).

Connection with determinants. By the way, one of the main reasons for introducing the concepts of linear dependence and spanning subspaces and so on (which we will get to very soon) is that we want a relatively-clean theory to express the ideas and intuitions that appear above in the 2×2 and the 3×3 case, for arbitrary $n \times n$ matrices, and doing it in this basically brute force way gets very messy very quickly. To see this, consider the 3×3 case. In general, the equation of the inverse of a matrix A is

$$A^{-1} = \frac{1}{\det(A)} \text{adj}(A),$$

where $\det(A)$ and $\text{adj}(A)$ are the determinant and adjoint of the matrix A . Computing the inverse in this way is *not* a good idea except in very very special cases, and so we aren't even going to present the general definition of $\det(A)$ and $\text{adj}(A)$. For the case of 2×2 matrices, however, it is given by Equation (3.14); and for the case of the 3×3 matrix

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix},$$

things can be defined recursively. In particular, the determinant of this 3×3 matrix is given by

$$\begin{aligned} \det(A) &= \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} \\ &= a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix} \\ &= a(ei - fh) - b(di - fg) + c(dh - eg) \\ &= aei + bfg + cdh - afh - bdi - ceg. \end{aligned}$$

Then, if we were to write down the formula for $\text{adj}(A)$ (which we won't), then pre-multiplying by the inverse of $\det(A)$ would give A^{-1} , as could be confirmed by just pre-multiplying and/or post-multiplying A by this and confirm that you get the 3×3 identity matrix. This, of course, assumes that $\det(A) \neq 0$. What does the condition $\det(A) \neq 0$ mean?

- In the 2×2 case, we saw that $\det(A) = 0$ meant that one column could be computed from the other column by performing a scalar multiplication.
- In the 3×3 case, $\det(A) = 0$ means that one column can be computed from the other two by some combination of scalar multiplication and vector addition, i.e., the two basic operations in which we are interested. (This may be a homework problem—it will involve practicing the basic operations as well as motivating why we need a more general way to deal with these basic operations.)

(a) Determinant of a 2×2 matrix. (b) Determinant of a 3×3 matrix.Figure 3.9: Illustration of the determinant of a 2×2 matrix and a 3×3 matrix.

- In the $n \times n$ case, it means the same thing, but things get hopelessly messy and very non-intuitive—if we deal with everything in this element-wise way. So linear algebra will be about coming up with a simpler way to do all this. It will involve a certain level of abstraction (e.g., linear dependence and spanning subspaces and the like are defined abstractly and shown to correspond to intuitive things in \mathbb{R}^2 and \mathbb{R}^3), but most of the ideas can be understood by extrapolating from \mathbb{R}^2 and \mathbb{R}^3 , and the gain is that you can work with data vectors in \mathbb{R}^n , for arbitrary n , in a clean and uniform way.

Let's move onto that. Before doing that, though, to understand determinants more geometrically, consider Figure 9.4.

- For a 2×2 matrix, if we consider the parallelogram defined by the two columns (or rows) of the matrix, then the determinant equals the area of the parallelogram; clearly, this equals zero if the two columns are linearly dependent, i.e., if one is a scalar multiple of the other, and thus they point in the same direction, in which case the parallelogram has no height.
- For a 3×3 matrix, if we consider the parallelepiped defined by the three columns (or rows) of the matrix, then the determinant equals the volume of the parallelepiped; clearly, this equals zero if the three columns are linearly dependent, i.e., if one is a scalar multiple of another or if one can be obtained by a linear combination of two others, in which case the parallelepiped has no height.

Analogous statements hold in higher dimensions, but they turn out to be less useful to work with than some of the linear algebraic concepts we will turn to soon.

3.4.3 Symmetric, triangular, diagonal, and Identity matrices

Symmetric matrices. Symmetric matrices are important in general but are very important in data science. Here is the definition.

Definition 15 A symmetric matrix is a matrix that equals its transpose, i.e., $A_{ij} = A_{ij}^T = A_{ji}$.

For example, the matrix

$$\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$$

is a symmetric matrix, while the matrix

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

is not a symmetric matrix. (Matrices that are not symmetric are sometimes called non-symmetric matrices.)

In particular, if an $m \times n$ matrix is symmetric, then $m = n$, i.e., the matrix is a square matrix.

Remark. Symmetric matrices seem like a very special case of matrices, e.g., since term-document matrices are not symmetric. As we will see, nearly all of the matrix computations performed in data science can be thought about and understood in terms of symmetric matrices. That is, either the matrices of interest are symmetric, or there are related symmetric matrices “under the hood.” (We will see examples of this later.) Later, we will describe very strong properties that symmetric matrices have that general matrices do not have that makes them very well-suited for understanding a lot of the mathematics underlying data science and that makes them useful for modeling data.

Triangular matrices. A type of matrix that is important for slightly less obvious reason are triangular matrices. Here is the definition.

Definition 16 An upper/lower triangular matrix is a square matrix with non-zero entries only on or below/above the diagonal.

For example, the matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

is neither upper-triangular nor lower-triangular, the the matrix

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

is upper triangular but not lower triangular, and the matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

is both upper triangular and lower triangular.

Diagonal matrices. Matrices that are both upper triangular and lower triangular are sufficiently important that they have a special name. Here is the definition.

Definition 17 A diagonal matrix is a square matrix that is both upper triangular and lower triangular, i.e., that has non-zero entries only on the diagonal.

We saw an example of this before when we discussed the diagonal degree matrix of the adjacency matrix of a graph.

Diagonal matrices have many nice properties. For example,

$$\begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}^k = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix} \quad \dots \quad \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix} = \begin{pmatrix} A_{11}^k & 0 \\ 0 & A_{22}^k \end{pmatrix}.$$

Thus, in this case, matrix multiplication, and in particular multiplying a matrix by itself takes a particularly simple form. In particular, let’s say for some reason that we can normalize things such that $A_{11} = 1$ (we’ll see later examples of this) and $A_{22} < A_{11}$. In this case,

$$\begin{pmatrix} 1 & 0 \\ 0 & A_{22} \end{pmatrix}^k = \begin{pmatrix} 1 & 0 \\ 0 & A_{22}^k \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

as k increases. (We’ll get back to the significance of this later.)

A special type of diagonal matrix is an Identity matrix.

Definition 18 The Identity matrix, denoted I or I_n is the $n \times n$ matrix with 1 along the diagonal and 0 off diagonal.

For, example, $I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

The Identity matrix is called the identity matrix since it is an identity for the operation of matrix multiplication. That is, if A is an $m \times n$ matrix, then $I_m A = A I_n = A$. We saw an example of this in Equation (3.3).

3.5 Examples of matrices as transformations

We saw that every matrix represents a linear transformation, and vice versa. In spite of that, it can be difficult to tell what exactly an arbitrary matrix is “doing” when it is just presented as an array of numbers. Fortunately, there are some basic examples of matrices that are relatively simple to think about, in the sense that their associated transformations are relatively simple to understand. In some cases, these basic examples form the “building blocks” of arbitrary matrices. When that happens, one can express the arbitrary matrix in terms of those building blocks; these are often called *matrix decompositions*. Matrix decompositions are useful for two things: first, they can help to understand structure in data (by describing what a matrix is “doing,” and thus what is happening in data that are being represented by the matrix, in terms of simpler operations); and second, they can help to perform computations faster (by describing a difficult-to-work-with matrix in terms of simpler easier-to-implement parts). We will describe several examples of this in later chapters. For now, we will give several examples.

Here are several examples of matrices as transformations.

- **Identity:** $I_n = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$. See Figure 3.10(a). This is the trivial transformation that doesn’t do anything.
- **Scaling:** $A = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}$. This takes any input vector and multiplies it by $a \in \mathbb{R}$. For example,

$$\begin{aligned} Ae_1 &= A \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ 0 \end{pmatrix} \\ Ae_2 &= A \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ a \end{pmatrix}, \end{aligned}$$

Since we can write an arbitrary vector $x \in \mathbb{R}^2$ in terms of elementary vectors, we have that

$$\begin{aligned} x &= x_1 e_1 + x_2 e_2 \\ &= x_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \end{aligned}$$

In addition, we have that

$$Ax = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} ax_1 \\ ax_2 \end{pmatrix} = a \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = ax.$$

See Figure 3.10(b) for an illustration of uniform scaling.

- **Stretching:** $A = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$. This is like scaling, except that the scaling is by a different amount in each direction. For example, since

$$\begin{aligned} Ae_1 &= A \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ 0 \end{pmatrix} \\ Ae_2 &= A \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ b \end{pmatrix}, \end{aligned}$$

i.e., the different e_i directions are stretched by different amounts, we have that

$$Ax = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} ax_1 \\ bx_2 \end{pmatrix},$$

i.e., an arbitrary vector is stretched by different amounts, depending on how long it is along each e_i . See Figure 3.10(c) for an illustration; in particular, note that this transformation maps a circle into an ellipse where the axes are aligned along the coordinate directions.

Note that although we are calling this stretching, if one of the entries is negative, it might be a reflection. For example, if $A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, then

$$Ax = A \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ -x_2 \end{pmatrix}.$$

This is a reflection about the line $x_1 = 0$.

- **Rotation:** $A = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$ It takes a minute to see that this involves rotating by an angle of θ . In particular

$$\begin{aligned} Ae_1 &= \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix} \\ Ae_2 &= \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -\sin(\theta) \\ \cos(\theta) \end{pmatrix} \end{aligned}$$

See Figure 3.11, which illustrates how this rotation rotates an ellipse from one with its axes along the canonical axes to one with its axes rotated by an angle of θ .

Things to note about this rotation.

- For this matrix, $A^T A = I$ and $AA^T = I$, and thus $A^T = A^{-1}$. This is an example of a matrix known as an orthogonal matrix. (We may show this in a HW.)
- For this matrix, $A_{\theta_1} A_{\theta_2} = A_{\theta_1 + \theta_2} = A_{\theta_2} A_{\theta_1}$. You should be able to convince yourself of this. (We may show this in a HW.) This holds here since both rotations are about the same axis (e.g., implicitly the x_3 axis, if we view the (x_1, x_2) plane as sitting in \mathbb{R}^3). This is not true in general for products of orthogonal matrices—we might show that in a homework—but it holds for orthogonal transformations about the same axis.

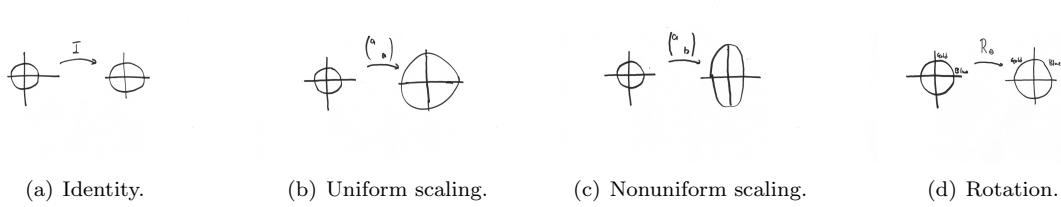


Figure 3.10: Illustration of several basic types of transformations.

- Here are other orthogonal transformations that you can easily visualize.

$$\begin{aligned}
 A &= \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} && \text{rotate counter-clockwise by 90 degrees} \\
 A &= \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} && \text{rotate by 180 degrees} \\
 A &= \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} && \text{rotate counter-clockwise by 270 degrees} \\
 A &= \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} && \text{reflection in the } x_1 \text{ axis} \\
 A &= \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} && \text{reflection in the } x_2 \text{ axis} \\
 A &= \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} && \text{reflection through origin, i.e., both axes} \\
 A &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} && \text{reflection about the line } x_1 = x_2
 \end{aligned}$$

Going beyond a single transformation, here are several examples of matrix compositions.

- Given an $n \times n$ square matrix A , apply A multiple times: $AA \dots A = A^k \in \mathbb{R}^{n \times n}$.
- Stretch, then rotate by $\pi/4$. In this case, the matrix is

$$A = \begin{pmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 2\cos(\pi/4) & -\sin(\pi/4) \\ 2\sin(\pi/4) & \cos(\pi/4) \end{pmatrix}$$

and when applied to a vector we get

$$y = Ax = \begin{pmatrix} \cdot & \cdot \\ \cdot & \cdot \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2\cos(\pi/4)x_1 - \sin(\pi/4)x_2 \\ 2\sin(\pi/4)x_1 + \cos(\pi/4)x_2 \end{pmatrix} \in \mathbb{R}^2.$$

By the way, this is “stretch, then rotate” if we apply this from the left to a column vector, i.e., if we pre-multiply the column vector. (It would be the other way around if we operated from the right on a row vector, as sometimes happens.)

- Rotate by $\pi/4$, then stretch. In this case, the matrix is

$$B = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{pmatrix} = \begin{pmatrix} 2\cos(\pi/4) & -2\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{pmatrix}$$

and when applied to a vector we get

$$y = Bx = \begin{pmatrix} \cdot & \cdot \\ \cdot & \cdot \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2\cos(\pi/4)x_1 - 2\sin(\pi/4)x_2 \\ \sin(\pi/4)x_1 + \cos(\pi/4)x_2 \end{pmatrix} \in \mathbb{R}^2.$$

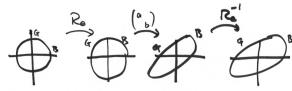


Figure 3.11: Illustration of the transformation shown in Eqn. (3.15).

- Rotate by $\pi/4$, then stretch, then rotate back by $\pi/4$. In this case, the matrix is

$$C = \begin{pmatrix} \cos(\pi/4) & \sin(\pi/4) \\ -\sin(\pi/4) & \cos(\pi/4) \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{pmatrix} = \begin{pmatrix} \cdot & \cdot \\ \cdot & \cdot \end{pmatrix} \quad (3.15)$$

and when applied to a vector we get

$$y = Cx = \begin{pmatrix} \cdot & \cdot \\ \cdot & \cdot \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2.$$

See Figure 3.11 for an illustration of the transformation given in Eqn. (3.15).

As mentioned previously, there are of course many other types of matrix products one could consider. We will see later that products of the form given in Eqn. (3.15)—or actually that generalize this very special case—are particularly important in data science. In particular, note that if we define

$$V = \begin{pmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{pmatrix} \quad \text{and} \quad \Lambda = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix},$$

then V is an orthogonal matrix, and we can write Eqn. (3.15) as

$$C = V^T \Lambda V.$$

This provides a decomposition of C into the product of three simpler matrices, which we will see have very nice interpretations for data science and more generally. Although we derived this by composing several seemingly-arbitrary operations, this is the first example of the spectral theorem, a very important and much more general result that we will get back to later.

3.6 Problems

3.6.1 Pencil-and-paper Problems

1. Show that Definition 3 and Definition 4 provide equivalent definitions of a vector space. Do this by showing that if a set V satisfies the main condition of Definition 3, then it satisfies the main condition of Definition 4, *and vice versa*.
2. Prove that \mathbb{R}^n is a vector space, for all integers $n \geq 1$. That is, show that if we start with vectors $x, y \in \mathbb{R}^n$ and numbers $a, b \in \mathbb{R}$, then if we perform vector addition and multiplication of a vector by a scalar, then we obtain another vector in \mathbb{R}^n .
3. Perform the following matrix multiplications when possible.

$$(a) \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \begin{pmatrix} 7 & 8 \\ 9 & 0 \\ 1 & 2 \end{pmatrix}$$

(b) $\begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 1 & 4 \\ -1 & 3 \\ -2 & 2 \end{pmatrix}$

(c) $\begin{pmatrix} 1 & -1 & 1 \\ -1 & 0 & 2 \\ -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & -1 \\ -1 & 1 & 2 \\ 2 & 0 & -2 \end{pmatrix}$

(d) $\begin{pmatrix} 7 & 1 \\ -1 & 0 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} 5 \\ -4 \end{pmatrix}$

(e) $\begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 1 & 4 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 3 \end{pmatrix}$

(f) $\begin{pmatrix} 0 & 2 & 1 \\ 1 & 3 & 2 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 3 & 5 \end{pmatrix}$

4. Consider the following two matrices:

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 3 & 1 & -1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 2 & 5 & 1 \\ 1 & 4 & 2 \\ 1 & 3 & 3 \end{pmatrix}.$$

- (a) Compute the third column of AB *without* computing the entire matrix AB .
 (b) Compute the second row of AB *without* computing the entire matrix AB .

5. Confirm by matrix multiplication that the inverse of

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{is} \quad A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

6. Prove that the matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is not invertible if $ad - bc = 0$.

7. Show that if A is any matrix, then $A^T A$ and AA^T are both symmetric.

8. Let $T : \mathbb{R}^3 \rightarrow \mathbb{R}^4$ be a linear transformation such that

$$T \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 3 \\ 4 \\ 1 \end{pmatrix}, T \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ 2 \\ 1 \end{pmatrix}, T \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \\ 6 \\ 2 \end{pmatrix}, T \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ 5 \\ 1 \end{pmatrix},$$

$$T \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 3 \\ 2 \end{pmatrix}, T \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ 3 \\ 0 \end{pmatrix}, T \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -4 \\ 0 \\ 1 \\ -1 \end{pmatrix}, T \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ 1 \\ 1 \end{pmatrix}.$$

How much of this information do you need to determine the matrix of T ? What is that matrix?

9. Let T be a transformation such that evaluated on the following five vectors it gives

$$T \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}, T \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, T \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, T \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \\ 3 \end{pmatrix}, T \begin{pmatrix} 2 \\ -1 \\ 4 \end{pmatrix} = \begin{pmatrix} 7 \\ 0 \\ 4 \end{pmatrix}.$$

Is T linear? Justify your answer.

10. (*Probably skip, maybe too advanced.*)
 (Hubbard 1.2.17(a)-(d))
 Hubbard: Exercise 1.2.17(a)-(d).

11. (*Probably skip.*)

(Hubbard 1.2.18)

Consider the adjacency matrix A corresponding to the cube graph (see Figure ??).

- (a) Compute A , A^2 , A^3 , and A^4 , and check directly that $(A^2)(A^2) = (A^3)A = A^4$.
- (b) The diagonal entries of A^4 should all be 21; count the number of walks of length 4 from a vertex to itself directly.
- (c) For this same matrix A , some entries of A^n are always 0 when n is even, and other entries (e.g., the diagonal entries) are always 0 when n is odd. Explain why.
- (d) Is this phenomena true for the adjacency matrix of a triangle graph (see Figure ??) or the adjacency matrix of a square graph (see Figure ??). Explain why or why not.

12. (*Probably skip. Not in 2016.*)

(Hubbard 1.3.2)

Recall that a $m \times n$ matrix has m horizontal rows (and is thus m tall) and n vertical columns (and is thus n wide). For the following linear transformations, what must be the dimension of the corresponding matrix?

- (a) $T : \mathbb{R}^2 \rightarrow \mathbb{R}^3$
- (b) $T : \mathbb{R}^3 \rightarrow \mathbb{R}^3$
- (c) $T : \mathbb{R}^4 \rightarrow \mathbb{R}^2$
- (d) $T : \mathbb{R}^4 \rightarrow \mathbb{R}^1$
- (e) $T : \mathbb{R}^1 \rightarrow \mathbb{R}^4$
- (f) $T : \mathbb{R}^1 \rightarrow \mathbb{R}^1$

13. Is there a linear transformation $T : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ such that

$$T \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix}, T \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 4 \end{pmatrix}, T \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 3 \end{pmatrix}?$$

If so, what is the matrix?

14. (a) What is the matrix of the linear transformation $S : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ corresponding to a reflection in the plane of the equation $x_1 = x_2$?
- (b) What is the matrix $T : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ corresponding to reflection in the plane $x_2 = x_3$?
- (c) What is the matrix $S \cdot T$? What is the matrix $T \cdot S$?
- (d) What is the relationship between $[S \cdot T]$ and $[T \cdot S]$?

15. Consider the transformation

$$R = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix},$$

corresponding to a rotation by θ counterclockwise around the origin. Use compositions of this transformation to derive the fundamental theorems of trigonometry:

$$\begin{aligned} \cos(\theta_1 + \theta_2) &= \cos(\theta_1)\cos(\theta_2) - \sin(\theta_1)\sin(\theta_2) \\ \sin(\theta_1 + \theta_2) &= \sin(\theta_1)\cos(\theta_2) + \cos(\theta_1)\sin(\theta_2). \end{aligned}$$

16. XXX. Give some more baby examples of spaces/subspaces.

17. XXX. Give some more baby examples of matrix multiplication.

18. XXX. Explicit numerical illustration of post multiplication by upper triangular matrix, as a forward pointer to QR.

19. (a) Let A and B be diagonal matrices. Show that $AB = BA$.
- (b) Let $U \in \mathbb{R}^{n \times n}$ be a matrix with orthonormal columns, i.e., such that $U^T U = I$, and let $D_1, D_2 \in \mathbb{R}^{n \times n}$ be diagonal matrices. Define $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times n}$ as $A = UD_1U^T$ and $B = UD_2U^T$. Show that $AB = BA$.
- (c) Let A be a 3×3 diagonal matrix, which is not the all-zeros matrix; and let B be a 3×3 matrix that is not diagonal. Give an example of an (A, B) pair such that $AB = BA$. Give an example of an (A, B) pair such that $AB \neq BA$.
20. XXX. Give a non-trivial explicit numerical example of simultaneously-diagonalizable 3×3 matrices A and B whose product commute, with a forward pointer to the example later, and tell them to compute AB and BA to confirm they commute.
21. Work through the jupyter notebook `linear-algebra-nb2.ipynb`, which can be downloaded from Piazza.

3.6.2 Implementations and Applications of the Theory

1. Let $x, y, z, u, v, w \in \mathbb{R}^{20}$ be the vectors:
- (A) A discrete sinusoid on the interval $[0, 2\pi]$ (when measured in radians, $\theta \in [0, 360]$ when measured in degrees), with frequency 2π .
 - (B) A discrete sinusoid on the interval $[0, 2\pi]$, with frequency 2π , with a small amount of noise added.
 - (C) A discrete sinusoid on the interval $[0, 2\pi]$, with frequency 4π .
 - (D) A discrete exponentially decaying function on the interval $[0, 2\pi]$, with decay parameter π or $\pi/2$.
 - (E) A vector in which each of its entries drawn as a random number $x \in [-1, 1]$.
 - (F) A different vector in which each of its entries drawn as a random number $x \in [-1, 1]$.

Do the following.

- (a) Compute the dot product between all pairs of vectors, including each vector with itself.
- (b) Also, divide each vector by its Euclidean norm, and compute the angle between each pair of vectors. (Be careful about the units with which you are measuring the angle.)
- (c) What do you notice about the angle between different pairs of vectors? For example, consider (A), and sort all six vectors ((A), (B), (C), (D), (E), (F)) in decreasing order, base on cosine similarity. Can you explain the ordering in terms of the entries of the vectors?
2. Let A be the matrix with x as its first column, y as its second column, z as its third column, u as its fourth column, v as its fifth column, and w as its sixth column.
- (a) Construct A , and print it out.
 - (b) Compute $A^T A$ and AA^T , and print it out.
 - (c) Let D be the diagonal matrix with $D_{11} = 1/\|x\|_2$, $D_{22} = 1/\|y\|_2$, $D_{33} = 1/\|z\|_2$, $D_{44} = 1/\|u\|_2$, $D_{55} = 1/\|v\|_2$, and $D_{66} = 1/\|w\|_2$. Compute AD and DA .
 - (d) Which of these has an interpretation in terms of the original columns of A ? What is that interpretation, e.g., in light of the previous problem?
 - (e) Compute $(AD)^T(AD)$, and compare with the results of the previous problem and/or say what the entries are.
3. Let's consider the matrix multiplication of a random walk transition matrix pre-multiplying different vectors. To do so, let A be the adjacency matrix of the graph considered in Chapter 1, let D be the diagonal degree matrix, and consider $A = AD^{-1}$.

- (a) Start with a vector x_0 , with 1 on (say) the 4th entry, and 0 on the other entries. Compute $x_{t+1} = Ax_t$, for $t = \{1, \dots, t\}$, and output x_{10} .
- (b) Consider the vector y_0 , in which 1/4 is on each of the four entries. Compute $y_{t+1} = Ay_t$, for $t = \{1, \dots, t\}$, and output y_{10} .
- (c) Compute A^t , for $t = \{1, \dots, t\}$. Then compute and output $(A^t)x_0$ as well as $(A^t)y_0$.
4. Given a column vector $x \in \mathbb{R}^n$, we are going to consider different ways to compute $xx^T x \in \mathbb{R}^n$. Let $x \in \mathbb{R}^n$ be the all-ones column vector.
- (a) For $i \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, let $n = 10^i$. For each value of n , by hand, compute $xx^T x$. (Obviously don't write down every entry. Instead, write the answer in terms of a simple-to-state vector.)
- (b) Write a python function to compute the result as $x(x^T x)$, i.e., first compute $x^T x$ and then pre-multiply this intermediate result by x . Then, write a python function to compute the result as $(xx^T)x$, i.e., first compute xx^T and then post-multiply this intermediate result by x .
- (c) For $i \in \{1, 2, 3, 4\}$, let $n = 10^i$, and compute $xx^T x$ in both ways. Are the answers the same or different? Answer this without printing out the vectors and looking at them, e.g., instead by computing the norm of the difference.
- (d) Next, let $i \in \{5, 6, 7, \dots\}$, let $n = 10^i$. For each of the two different ways to compute $xx^T x$, in python, compute it for as large a value of i as you can. For each method, when you are no longer able to compute the result, explain what prevents you from going larger. Explain why the maximum attainable values of i in each case are the same or different.
5. (PYTHON) LATER. Illustrate examples of 2×2 matrices illustrating pictorially matrix multiplication, including of basic matrices. Consider rotations, inversions, diagonal scaling, general matrices, and plot before and after in some color-coded way that students can play with. XXX. NOTE: ANTHONY HAS SOME STUFF ON THIS, WE JUST NEED TO DO IT CLEANER. XXX. Take some material/questions from Anthony's NBs from last year. XXX. CAN I DO THIS WIHT 3X3. As for the visualization question, I'd like to ask a few questions that force them to see that matrices applied to the standard basis vectors can allow them to (1) scale axes uniformly, (2) scale axes non-uniformly, (3) reflect through a line, (4) rotate, and (5) project onto a line, i.e., a subspace of the plane. For the projection, we should ask them to compute the mass along the line and the line perpendicular to the line. I'd like it to be visually nice, and I'd like it to be something they can play with, e.g., changing entries in the input matrices. Also inversions and compositions of matrices, and the latter may mean applying scaling, etc., to bases other than the canonical basis. XXX. THIS IS HERE FOR SCALING AND OTHER TYPES OF MATRICES, AND IT SHOULD PLANT SEEDS FOR EIGENVECTORS LATER.

Chapter 4

Geometry: angles, spans, bases, and projections

In this chapter, we'll describe geometric properties of \mathbb{R}^n . This includes angles and perpendicularity; linear combinations, spans, and linear dependence/independence; basis vectors, including orthogonal/orthonormal basis vectors; and projections onto basis vectors. These ideas are central to linear algebra. Part of this is since these fundamental linear algebra ideas provide a natural formal way to generalize many of the ideas and intuitions we have been discussing from the more familiar \mathbb{R}^2 and \mathbb{R}^3 up to the less familiar \mathbb{R}^n .

As we will see, several of these ideas have a large algebraic component, in the sense that to implement them on problems of interest typically requires many many relatively-straightforward algebraic steps. This is the sort of thing at which computers excel. On the other hand, to focus on the algebraic steps is to miss the trees for the forest. To understand why these linear algebra methods are so useful in data science (and beyond) requires only a little bit of algebra, but it does also require a good understanding Euclidean geometry, and in particular how basic ideas of the Euclidean geometry of \mathbb{R}^2 , i.e., the familiar planar geometry, generalize to \mathbb{R}^n .

4.1 Geometry of \mathbb{R}^n : dot products, angles, and perpendicularity

4.1.1 Dot products

Recall that the dot product on \mathbb{R}^n is a generalization of the notion of the dot product on the plane \mathbb{R}^2 , and it gives us the *geometric* notions of lengths and angles for vectors in \mathbb{R}^n .

We have hinted at and alluded to a lot of the following, but let's make it explicit for \mathbb{R}^n .

Definition 19 *The dot product $x \cdot y$ of two vectors $x, y \in \mathbb{R}^n$ is defined as*

$$x \cdot y = x_1y_1 + x_2y_2 + \cdots + x_ny_n = \sum_{i=1}^n x_iy_i.$$

Remark. As defined, the dot product is a multiplication between two vectors that returns as output a number. If we view these two (column) vectors as matrices, then they are $n \times 1$ matrices. Unless $n = 1$, neither the matrix-matrix multiplication xy nor the matrix-matrix multiplication yx are defined. (The dimensions don't match up.) The dot product can be expressed, however, in terms of matrix-matrix multiplication by taking transposes. If $x \in \mathbb{R}^n$ is a column vector, then viewed as a matrix $x \in \mathbb{R}^{n \times 1}$, in which case we can

define the transpose $x^T \in \mathbb{R}^{1 \times n}$. In this case, the matrix-matrix product

$$x^T y = \sum_{i=1}^n x_i y_i = y^T x \quad (= x \cdot y)$$

is just the dot product. So, this is a very special case of a matrix-matrix multiplication. (In this special case, it is in fact commutative; and in this special case, it is also distributive, i.e., $(x+z)^T y = x^T y + z^T y$ and $x^T (y+z) = x^T y + x^T z$.) This is sometimes called the *inner product* or the *standard inner product*. Informally, this is called an inner product since the “inner dimension” (i.e., n) is “dotted out” and one ends up with a number.

Remark. We will see later that we can use matrices of a special form to define a *generalized inner product*, which has all the properties of this standard inner product, except that it is “rotated” and/or “stretched.” A simple example of this is

$$x^T \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} y = x_1 y_1 + 2 x_2 y_2$$

This may be of interest if we can rotate it and stretch it in a way that conforms better to the properties of the data.

Remark. Given two vectors $x, y \in \mathbb{R}^n$, if we view them as $n \times 1$ matrices and consider transposes, then we can compute the matrix-matrix product in two different ways, yielding a number (in $\mathbb{R} = \mathbb{R}^{1 \times 1}$) or an $n \times n$ matrix (in $\mathbb{R}^{n \times n}$). For example, $x^T y$ or xy^T . The former is a number, i.e., an element of \mathbb{R} , and the latter is an $n \times n$ matrix. We saw the former above, it is called an inner product. The latter is also of interest, and it is called the *outer product* of x and y . We know that matrix-matrix multiplication is not commutative, and here order matters, i.e., in general $xy^T \neq yx^T$. That’s fine, and should be expected, since x and x^T are not the same, when viewed as matrices, and ditto for y and y^T . (From the matrix-matrix multiplication perspective, the exception is the inner product, which commutes since the result is a number.) Informally, this other operation is called an “outer product” since the higher dimension (i.e., n) “sticks out” to be multiplied by a vector in \mathbb{R}^n .

Remark. Given any $m \times n$ matrix X , let’s

- denote the i^{th} row of X by $X_{i:}$, and
- denote the j^{th} column of X by $X_{:j}$.

Let’s say that A is an $m \times n$ matrix and B is an $n \times p$ matrix. Then, the (ij) entry of the matrix-matrix product AB is given by

$$(AB)_{ij} = A_{i:} \cdot B_{:j} = A_{i:}^T B_{:j} = B_{:j}^T A_{i:},$$

i.e., it is given by the inner product between the i^{th} row of A and the j^{th} column of B . There are mp such entries in the product matrix. See Figure 3.8 for an illustration. (Here, of course, we are overloading notation and viewing, e.g., $A_{i:}$ as a vector or matrix, depending on context.) On the other hand, the entire matrix-matrix product can be expressed as

$$AB = \sum_{k=1}^n A_{:k} B_{k:}^T$$

i.e., as the sum of the outer products between the k^{th} column of A and the corresponding k^{th} row of B . Observe that each $A_{:k} B_{k:}^T$ is an $m \times p$ matrix, for each $k \in \{1, \dots, n\}$, and we can add them up elementwise to get the matrix AB which is also an $m \times p$ matrix.

As we will see, the inner product is important for many reasons. One reason is that it has close connections with a particular vector norm.

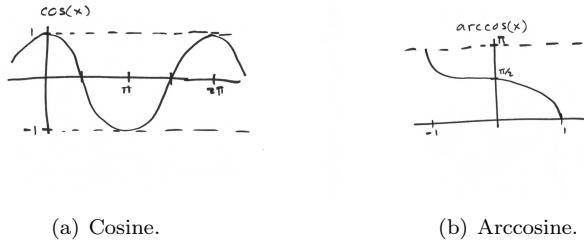


Figure 4.1: Illustration of cosine and arccosine functions.

Definition 20 Given $x \in \mathbb{R}^n$, its length or norm or Euclidean norm is given as

$$\|x\|_2 = (x \cdot x)^{1/2} = (x^T x)^{1/2} = \left(\sum_{i=1}^n x_i^2 \right)^{1/2}.$$

Remark. As we have discussed, there are other notions of norms, e.g., L_1 and L_∞ and others that we did not discuss, but the L_2 or Euclidean norm is so ubiquitous since it gives the most useful *geometry*. That geometry is directly related to the fact that it can be expressed as an inner product, or equivalently as a matrix-matrix multiplication. That is the reason it of central interest in linear algebra, machine learning, data science, etc.

4.1.2 Angles

Recall that for $x, y \in \mathbb{R}^2$ or \mathbb{R}^3 , we have a notion of an angle between x and y , and this angle can be computed from the dot product between x and y and Euclidean norm of x and y as

$$x \cdot y = \|x\|_2 \|y\|_2 \cos(\theta). \quad (4.1)$$

We want to generalize this to \mathbb{R}^n , i.e., to the notion of an angle between two vectors x and y in \mathbb{R}^n .

On the one hand, the generalization is straightforward: simply use Equation (4.1), where the vectors are now in \mathbb{R}^n . The equations are the same, and the generalization works. Doing so, however, does require establishing one slightly subtle thing, i.e., establishing that

$$\frac{x \cdot y}{\|x\|_2 \|y\|_2} \in [-1, 1], \quad (4.2)$$

for arbitrary vectors x and y . The reason we need to establish this is so that we can take the arccos of this expression, i.e., so that there is in fact an angle θ , the cosine of which equals it. See Figure 4.1 for an illustration.

Expression (4.2) is true, for arbitrary vectors x and y in \mathbb{R}^n , and it is known as the Cauchy-Schwartz Inequality. It is an extremely important result, largely because, although we show it for vectors in \mathbb{R}^n , it actually holds much more generally. We will actually prove this special case using elementary methods.

Theorem 5 (Cauchy-Schwartz Inequality) Given vectors $x, y \in \mathbb{R}^n$, it follows that

- $|x \cdot y| \leq \|x\|_2 \|y\|_2$
- Equality holds iff $x = \alpha y$, for $\alpha \in \mathbb{R}$.

Proof: Consider the function

$$\begin{aligned}
 f(t) &= |x + ty|^2 \\
 &= (x + ty)^T (x + ty) \\
 &= x^T x + 2tx^T y + t^2 y^T y \\
 &= \|y\|_2^2 t^2 + 2x^T y t + \|x\|_2^2 \\
 &= at^2 + bt + c \quad (\text{a quadratic formula in } t) \\
 &\geq 0,
 \end{aligned}$$

where the expression is non-negative since it is a quantity squared. In particular, the quadratic (in t) function $f(t)$ satisfies $f(t) \geq 0$, which means that the graph of f does not cross the t axis, i.e., it does not change sign and become negative.

Recall the quadratic formula which can be used to characterize the solution to quadratic equations and which says that solutions are of the form

$$t = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

There are three cases to consider, depending on the value of the discriminant $b^2 - 4ac$.

1. **Positive Discriminant:** If the discriminant > 0 , then the equation will have two distinct roots, i.e., the function will achieve both positive and negative values, and the function will cross the t axis.
2. **Zero Discriminant:** If the discriminant $= 0$, then the equation has a double root, and the function will achieve only positive and zero values or only negative and zero values.
3. **Negative Discriminant:** If the discriminant < 0 , then there are imaginary roots, i.e., there are no real roots, and the function doesn't cross or touch the t axis, and thus it is only strictly positive or strictly negative.

In the case of $f(t)$ given above, we noted that it is a non-negative function. In this case, the discriminant is

$$(2x^T y)^2 - 4\|y\|_2^2\|x\|_2^2,$$

and the function doesn't cross the t axis, and so the discriminant is not positive. Thus,

$$4(x^T y)^2 - 4\|x\|_2^2\|y\|_2^2 \leq 0,$$

and so $|x^T y| \leq \|x\|_2\|y\|_2$. This establishes the first part.

The second part of the Cauchy-Schwartz inequality is $|x \cdot y| = \|x\|_2\|y\|_2$ iff $x = \alpha y$. There are two directions here. First, assume that $y = \alpha x$, in which case we have that

$$\begin{aligned}
 |x \cdot y| &= |x \cdot (\alpha x)| \\
 &= |\alpha| |x \cdot x| \\
 &= |\alpha| \|x\|_2^2 \\
 &= (\alpha \|x\|_2) \|x\|_2 \\
 &= \|y\|_2 \|x\|_2.
 \end{aligned}$$

Second, assume that $|x \cdot y| = \|x\|_2\|y\|_2$, in which case the discriminant $= 0 = (2x^T y)^2 - 4\|y\|_2^2\|x\|_2^2$. So, the quadratic equation has a single root t_0 as $|x + t_0 y|^2 = 0$, from which it follows that $x = -t_0 y$. \diamond

This theorem shows that $|x \cdot y| \leq \|x\|_2\|y\|_2$, from which it follows that

$$-1 \leq \frac{x \cdot y}{\|x\|_2\|y\|_2} \leq 1,$$

for all vectors x and y . So, the arccosine of this expression exists, for all vectors x and y . This means that there is number, which we can interpret as an angle θ , such that this is the cosine if it. That is, we can *define* this as the angle between two vectors in \mathbb{R}^n .

Definition 21 Given $x, y \in \mathbb{R}^n$, the angle θ between these two vectors is $\theta = \arccos\left(\frac{x \cdot y}{\|x\|_2 \|y\|_2}\right)$, which is an angle θ such that $0 \leq \theta \leq \pi$.

Observe what we have done. In \mathbb{R}^2 , where we have the well-known intuition of an angle between two vectors x and y , we showed that the idea of an angle could be related to a formula that depended on dot products involving those vectors. In \mathbb{R}^n , where we have less intuition, but where from Theorem 5 we know that the same formula satisfies a certain property that the cosine of an angle also satisfies, we used this formula to define the angle between two vectors x and y .

Given this definition, we might wonder how similar or different are vectors in \mathbb{R}^2 and \mathbb{R}^n . There are a number of ways to answer this, and we'll consider several of them in one of the homeworks. Here is one.

Question. What is the angle between the diagonal of the unit cube in the positive orthant and the vector e_1 ?

Answer: For \mathbb{R}^2 , it is 45 degrees or $\pi/4$ radians. (The analogous question for \mathbb{R}^3 is ca. 54.7 degrees or 0.955 radians.) To get the answer, recall that the unit cube is given by $0 \leq x_i \leq 1$, for x_i , for $i \in \{1, \dots, n\}$. Also,

$$d = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \text{and} \quad e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

and so $\|d\|_1 = \sqrt{n}$ and $\|e_1\|_2 = 1$. Thus, $\theta = \arccos\left(\frac{d \cdot e_1}{\|d\|_2 \|e_1\|_2}\right) = \arccos\left(\frac{1}{\sqrt{n}}\right)$, which decreases with increasing n . (We will probably show this with computational results in one of the homeworks.)

Question. What is the angle between the diagonal of the unit cube in the positive orthant and a “typical” vector, or between two typical vectors?

Answer: To answer this question, we need to be precise about what we mean by this, e.g., what we mean by “typical.” For example, we might choose a point/vector randomly or a pair of points/vectors randomly and try to figure out some properties of it/them. We will get to this later. As we will see, just as with the answer to the previous question, the answer depends strongly on the value of n in \mathbb{R}^n .

4.1.3 Orthogonality between two vectors in \mathbb{R}^n

Given the notion of dot product between two vectors, we can *define* what it means for two vectors to be perpendicular or orthogonal, which generalizes the notion for \mathbb{R}^2 . Here is the definition.

Definition 22 Two vectors x, y are orthogonal or perpendicular if $x \cdot y = 0$.

Remark. That is, in \mathbb{R}^n , two vectors are orthogonal/perpendicular if the angle θ between them is 90° or $\pi/2$ radians.

Remark. Clearly, Definition 22 is the same as $x^T y = y^T x = 0$. This is true whether or not the vectors are unit length.

Remark. This definition does not depend on whether or not the vectors are unit-length. (The particular value of the angle does, but whether or not the vectors are orthogonal does not.) Thus, the following should also be clear. Given vectors x, y , they are orthogonal if the unit-length vector in the direction of x , call it u , and the unit-length vector in the direction of y , call it v , are orthogonal.



(a) Roughly linearly dependent data. (b) Roughly linearly independent data.

Figure 4.2: Examples of data in \mathbb{R}^2 that are roughly linearly dependent and roughly independent.

The answer to the above question (in the previous subsection) shows that the angle between a canonical axis and the vector pointing to the corner of the unit square in the positive orthant is 45 degrees or $\pi/4$ radians in \mathbb{R}^2 and gradually increases toward 90 degrees or $\pi/2$ radians, as n gets large, i.e., the vectors become closer and closer to being perpendicular. We will explore this computationally in one of the homeworks.

Problem. As an example of this, given a vector $v \in \mathbb{R}^n$, let's define $v^\perp \subset \mathbb{R}^n$ to be the set of vectors $w \in \mathbb{R}^n$ such that $v \cdot w = 0$. It is easy to show that this is a subspace. On \mathbb{R}^2 , this is just a one-dimensional line perpendicular to v ; on \mathbb{R}^3 , is a two-dimensional plane perpendicular to v ; and in \mathbb{R}^n , it is a subspace of dimension $n - 1$ that is oriented to be perpendicular to v . (Conversely, we could have asked for the set of vectors perpendicular to two or more vectors. This is algebraically more complex, but the ideas are similar, and we will get to this later. In \mathbb{R}^3 , this would be a one-dimensional subspace, but in \mathbb{R}^4 , this would be a two-dimensional subspace, meaning that we have two different two-dimensional subspaces in \mathbb{R}^4 . This is the typical situation in higher dimensions, and this can be made precise using the linear algebra ideas we will discuss.)

4.2 Linear combinations, spans, and linear dependence/independence

See Figure 4.2. We would like to be able to say—given everything we have done with linear combinations, linear subspaces, etc., that all of the points in Figure 4.2(a) are in some sense similar or the same or almost the same qua linear algebra, while the points in Figure 4.2(b) are more different qua linear algebra. For example, we would like to quantify how in the former case the data are roughly the same up to scalar multiplications, and thus how one could approximately generate each data point from the others by performing a scalar multiplication of any one of them. Actually, we would like to be able to do this for data points in \mathbb{R}^n , where we can't visualize the data and where we want to say that data points are roughly the same qua linear algebra in the sense that one could generate each data point from the others by performing vector addition as well as scalar multiplication. See Figure 4.3 for an illustration in three dimensions. Let's consider more precisely linear algebraic ideas that permit one to do that.

Linear independence captures the idea that the vectors do not contain redundant information in the sense that you can compute one from the others with the linear operations of scalar multiplication and vector addition. Conversely, linear dependence captures the idea that the vectors contain redundant information in the same sense. In \mathbb{R}^2 , this is less interesting, since it boils down to the idea that if you have 2 vectors, it is not the case that one is a scalar multiple of the other. In \mathbb{R}^3 , the situation is much more interesting: the idea is that if you have 3 vectors in \mathbb{R}^3 , then it is not the case that you can compute one of them from the other two by taking scalar multiples and vector additions of the other two. (See Figure 4.3, and compare with Figure 4.2.) The idea generalizes to \mathbb{R}^n , and it is much more interesting and useful for \mathbb{R}^n , $n \geq 3$, and it is a key idea in linear algebra. In the following, we will make this precise with the ideas of linear combination, span, linear dependence, etc., but the examples you should have in the back of your mind is that we want to explain the differences in the data in Figure 4.2(a) versus Figure 4.2(b) as well as in Figure 4.3(a) versus Figure 4.3(b).



(a) Roughly linearly dependent data. (b) Roughly linearly independent data.

Figure 4.3: Examples of data in \mathbb{R}^3 that are roughly linearly dependent and roughly independent.

4.2.1 Linear combinations

To develop this, let's start with the following notion of linear combination.

Definition 23 If v_1, \dots, v_k is a collection of vectors in \mathbb{R}^n , and if $a_1, \dots, a_k \in \mathbb{R}$, then a linear combination of $\{v_i\}_{i=1}^k$ is a vector $w \in \mathbb{R}^n$ s.t.

$$w = \sum_{i=1}^k a_i v_i.$$

The flavor of the operations in this definition should be familiar, as they are the familiar addition of vectors and multiplication of a vector by a scalar, but a somewhat different statement is being made here than before. In particular, a vector is a linear combination of other vectors if it can be computed from those other vectors by applying repeatedly the basic operations of scalar multiplication and vector addition.

Remark. Note that we have *not* specified anything about the relationship between k and n : k could be less than, equal to, or larger than n .

Examples. Here are some examples illustrating Definition 23.

- Let $v = \begin{pmatrix} 3 \\ 4 \end{pmatrix} \in \mathbb{R}^2$. Then:
 - $v = 3 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 4 \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, and thus v is a linear combination of e_1 and e_2 .
 - $\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 3 \\ 4 \end{pmatrix} - \frac{4}{3} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, and thus e_1 is a linear combination of v and e_2 .
 - $\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 3 \\ 4 \end{pmatrix} - \frac{3}{4} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, and thus e_2 is a linear combination of v and e_1 .
- $\begin{pmatrix} 3 \\ 4 \end{pmatrix} = 3 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \alpha \begin{pmatrix} 0 \\ 1 \end{pmatrix} + (2 - \frac{\alpha}{2}) \begin{pmatrix} 0 \\ 2 \end{pmatrix}$, where α is *any* real number in \mathbb{R} .
 - The reason for this redundancy is that $\begin{pmatrix} 0 \\ 2 \end{pmatrix}$ is a linear combination of $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$, and vice versa.
 - This illustrates the situation when we have more vectors than the number of dimensions.
- Let $v = \begin{pmatrix} 0 \\ 3 \\ 4 \end{pmatrix} \in \mathbb{R}^3$. Then, $v = 3 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + 4 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$, and thus v is a linear combination of e_2 and e_3 .
 - (Similarly for e_2 being a linear combination of v and e_3 , and e_3 being a linear combination of v and e_2).

Here, the (x_2, x_3) plane as a subset of \mathbb{R}^3 is a (relatively uninteresting) subspace of \mathbb{R}^3 . This example does illustrate, however, the situation when we have fewer vectors than the number of dimensions.

- Let A be an $m \times n$ matrix, and let x be an n -dimensional column vector. Then Ax is a linear combination of the *columns* of A . Consider the following example:

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 5 \\ 6 \end{pmatrix} = \begin{pmatrix} 5+12 \\ 15+24 \end{pmatrix} = \begin{pmatrix} 17 \\ 39 \end{pmatrix}.$$

If we view this multiplication (post-multiplication of a matrix A by a column vector x) as taking the linear combination of the columns of A , then we get:

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 5 \\ 6 \end{pmatrix} = 5 \begin{pmatrix} 1 \\ 3 \end{pmatrix} + 6 \begin{pmatrix} 2 \\ 4 \end{pmatrix} = \begin{pmatrix} 5 \\ 15 \end{pmatrix} + \begin{pmatrix} 12 \\ 24 \end{pmatrix} = \begin{pmatrix} 17 \\ 39 \end{pmatrix}.$$

- Let A be an $m \times n$ matrix, and let x be an m -dimensional column vector. Then $x^T A$ is a linear combination of the *rows* of A . Consider the following example:

$$(7 \ 8) \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = (7+24 \ 14+32) = (31 \ 46).$$

If we view this multiplication (pre-multiplication of a matrix A by a row vector x^T) as taking the linear combination of the rows of A , then we get:

$$(7 \ 8) \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = 7(1 \ 2) + 8(3 \ 4) = (7 \ 14) + (24 \ 32) = (31 \ 46).$$

Remark. The last two examples—of how matrix-vector multiplication just computes linear combinations of columns (if post-multiplying by the vector) or rows (if pre-multiplying by the vector)—shows that matrix-matrix multiplication isn’t really a “new” thing but instead can be understood in terms of the basic operations of linear combinations, i.e., of vector addition and scalar multiplication, and it is because matrix-matrix products correspond to compositions of linear functions.

4.2.2 Span

The previous notion of linear combination had to do with whether a given vector could be described by a set of vectors with the operations of scalar multiplications and vector additions. We often want to go “in the other direction” and ask: if we have a set of vectors, then what is the set of vectors that can be computed from them with the operations of scalar multiplications and vector additions. (In the previous sentence, when we say “described” by them, “described” is nothing more than an informal way to say computed with these operations, i.e., it doesn’t mean describing in terms of the data or the processes that generated the data.) This gets us to the following notion of span. Again, the operations of addition of vectors and multiplication of a vector by a scalar should be familiar, but a somewhat different statement is being made here than before.

Definition 24 Given a set of vectors, $v_1, \dots, v_k \in \mathbb{R}^n$, the span of that set of vectors is the set of vectors that can be computed as linear combinations of the form

$$\text{Span}(v_1, \dots, v_k) = a_1 v_1 + \dots + a_k v_k = \sum_{i=1}^k a_i v_i,$$

where $a_i \in \mathbb{R}$, for $i = 1, \dots, k$.

Remark. Again, note that we have *not* specified anything about the relationship between k and n : k could be less than, equal to, or larger than n .

Examples. Here are some examples illustrating Definition 24.

- The span of $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is the set of all vectors of the form $\begin{pmatrix} \alpha \\ 0 \end{pmatrix}$, for $\alpha \in \mathbb{R}$; and the span of $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$ is the set of all vectors of the form $\begin{pmatrix} \alpha \\ 2\alpha \end{pmatrix}$, for $\alpha \in \mathbb{R}$. Similarly, the span of $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and $\begin{pmatrix} -3 \\ -6 \end{pmatrix}$ is the set of all vectors of the form $\begin{pmatrix} \alpha \\ 2\alpha \end{pmatrix}$, for $\alpha \in \mathbb{R}$. These sets are all lines through the origin on \mathbb{R}^2 , and thus subspaces of \mathbb{R}^2 , which we will see is true more generally.

- Let e_1 and e_2 be the coordinate vectors for \mathbb{R}^2 , i.e., $e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Then, $\text{Span}(e_1, e_2) = \mathbb{R}^2$. The reason is that any vector $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ can be written as a linear combination of e_1 and e_2 , i.e.,

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = x_1 e_1 + x_2 e_2.$$

- Similarly, $\text{Span}(e_1, e_2, e_3) = \mathbb{R}^3$.
- $\text{Span}\left(\frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}\right) = \mathbb{R}^2$. There are two ways to see this. First, one could take an arbitrary vector in \mathbb{R}^2 and express it in terms of these two vectors. There are ways to do this, and we will get to them soon. Second, one could note that we can express this basis in terms of the standard basis, and we can express any vector in terms of the standard basis, and we can combine those two operations.

- On the other hand, if we view e_1 and e_2 as vectors in \mathbb{R}^3 , i.e., $e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ and $e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, i.e., we don't include e_3 , then $\text{Span}(e_1, e_2) \neq \mathbb{R}^3$. The reason is that no vector x with non-zero x_3 component can be computed from these two vectors. However, we have mentioned informally that and we will soon see more precisely that $\text{Span}(e_1, e_2)$, i.e., the $x_3 = 0$ two-dimensional plane, is a subspace of \mathbb{R}^3 . There are more non-trivial subspaces that we will get to, e.g., other two-dimensional planes that go through the origin that are not axis-aligned.

- $\text{Span}\left(\begin{pmatrix} 3 \\ 4 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \mathbb{R}^2$. The reason for this is that any vector $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$ can be written as

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \alpha \begin{pmatrix} 3 \\ 4 \end{pmatrix} + \beta \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

If this is not obvious, then express $\begin{pmatrix} 3 \\ 4 \end{pmatrix}$ in terms of the standard basis vectors and combine with this expression.

- $\text{Span}\left(\begin{pmatrix} 3 \\ 4 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}\right) \neq \mathbb{R}^3$, since any vector with nonzero x_3 value cannot be computed from these vectors.

- A somewhat less trivial example is that $\text{Span}\left(\begin{pmatrix} 3 \\ 4 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}\right) \neq \mathbb{R}^3$. Here, the vector $\begin{pmatrix} 3 \\ 4 \\ 1 \end{pmatrix}$ (and many others) cannot be computed as linear combinations from these vectors.

- Let A be an $m \times n$ matrix, and let x vary over all possible n -dimensional column vector. Then, the span of the *columns* of A is given by

$$\{Ax : x \in \mathbb{R}^n\}.$$

In particular, if $A_{:j}$ denotes the j^{th} column of A , then this set is all vectors of the form

$$\sum_{i=1}^n x_j A_{:j},$$

as x is varied over all of \mathbb{R}^n .

- Let A be an $m \times n$ matrix, and let y vary over all possible m -dimensional column vector. Then, the span of the *rows* of A is given by

$$\{y^T A : y \in \mathbb{R}^m\}.$$

In particular, if $A_{i:}$ denotes the i^{th} row of A , then this set is all vectors of the form

$$\sum_{i=1}^m y_i A_{i:},$$

as y is varied over all of \mathbb{R}^m .

Remark. The last two examples are harder to visualize, but they are much more important than the easier-to-visualize examples. The reason that we are building up this linear algebra machinery is to make it easier to reason about these two examples and many related examples that arise in data science.

Problem. Given vectors v_1, \dots, v_k are vectors in \mathbb{R}^n , let $V = \text{Span}(v_1, \dots, v_k)$. Prove that V is a subspace of \mathbb{R}^n .

Remark. It is a fact that we will not prove that, in addition, it is the smallest subspace of \mathbb{R}^n that contains v_1, \dots, v_n .

Problem. Given vectors v_1, \dots, v_k are vectors in \mathbb{R}^n , let V^\perp be the set of vectors that are perpendicular to $V = \text{Span}(v_1, \dots, v_k)$. Prove that V^\perp is a subspace of \mathbb{R}^n .

4.2.3 Linear dependence and independence

We have seen that, in \mathbb{R}^3 :

$$\begin{aligned} \text{Span}(e_1, e_2, e_3) &= \mathbb{R}^3 \\ \text{Span}\left(e_1, e_2, e_3, \begin{pmatrix} 3 \\ 4 \\ 0 \end{pmatrix}\right) &= \mathbb{R}^3 \\ \text{Span}\left(e_1, e_2, \begin{pmatrix} 3 \\ 4 \\ 0 \end{pmatrix}\right) &\neq \mathbb{R}^3 \end{aligned}$$

Informally, we need the vectors to be “different,” and in the latter case the information to compute $\begin{pmatrix} 3 \\ 4 \\ 0 \end{pmatrix}$

is already “in” e_1 and e_2 , in the sense that we can compute the vector $\begin{pmatrix} 3 \\ 4 \\ 0 \end{pmatrix}$ from e_1 and e_2 via the operations of addition of vectors and multiplication of a vector by a scalar. A similar statement holds for linear combinations of rows or columns of a matrix that are harder to visualize. We can make this notion of “different” precise with the idea of linear independence.

Definition 25 *The vectors v_1, \dots, v_k are linearly independent if there is at most one way of writing a vector w as a linear combination of v_1, \dots, v_k . That is,*

$$w = \sum_{i=1}^k \alpha_i v_i = \sum_{i=1}^k \beta_i v_i \Rightarrow \alpha_i = \beta_i, \quad \text{for all } i.$$

If a set of vectors is not linearly independent, then they are linearly dependent.

Examples. Here are some examples illustrating Definition 25.

- $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ are linearly independent, since there is only one way to write $\begin{pmatrix} 3 \\ 4 \end{pmatrix}$ as a linear combination of them. Similarly for the other combinations.
- $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$, and $\begin{pmatrix} 0 \\ 2 \end{pmatrix}$ are not linearly independent. For example,

$$\begin{aligned} \begin{pmatrix} 3 \\ 2 \end{pmatrix} &= 3\begin{pmatrix} 1 \\ 0 \end{pmatrix} + 2\begin{pmatrix} 0 \\ 1 \end{pmatrix} + 0\begin{pmatrix} 0 \\ 2 \end{pmatrix} \\ &= 3\begin{pmatrix} 1 \\ 0 \end{pmatrix} + 0\begin{pmatrix} 0 \\ 1 \end{pmatrix} + 1\begin{pmatrix} 0 \\ 2 \end{pmatrix} \end{aligned}$$

- Similarly, $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$, and $\begin{pmatrix} 3 \\ 4 \end{pmatrix}$ are not linearly independent. For example,

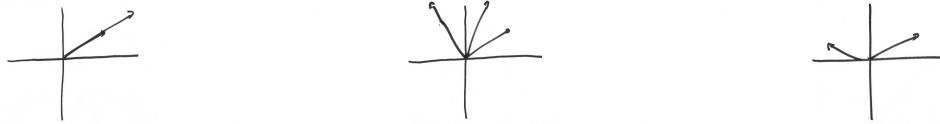
$$\begin{aligned} \begin{pmatrix} 3 \\ 2 \end{pmatrix} &= 3\begin{pmatrix} 1 \\ 0 \end{pmatrix} + 2\begin{pmatrix} 0 \\ 1 \end{pmatrix} + 0\begin{pmatrix} 3 \\ 4 \end{pmatrix} \\ &= 0\begin{pmatrix} 1 \\ 0 \end{pmatrix} - 2\begin{pmatrix} 0 \\ 1 \end{pmatrix} + 1\begin{pmatrix} 3 \\ 4 \end{pmatrix} \end{aligned}$$

There are, of course, many other solutions here, since $\begin{pmatrix} 3 \\ 4 \end{pmatrix} = 3\begin{pmatrix} 1 \\ 0 \end{pmatrix} + 4\begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

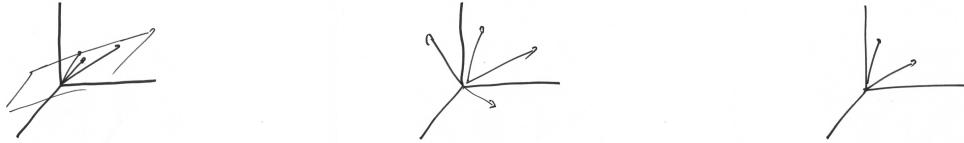
We mentioned this informally above, but now that we have the notions of linear independence and subspaces of \mathbb{R}^n , we re-mention the geometric interpretation of linear independence. See Figure 4.4 and Figure 4.5. Note that this is almost “trivial” for \mathbb{R}^2 , somewhat less so for \mathbb{R}^3 ; but it is one of the most important ways to think about \mathbb{R}^n .

- One vector is linearly independent if it isn’t 0, the all-zeros vector. This is true, but it is not particularly interesting (since there are no non-trivial scalar multiples and linear combinations to consider).
- Two vectors are linearly independent if they don’t lie on the same one-dimensional line, i.e., if one is not a scalar multiple of the other. This is more interesting, but not extrememly interesting (since it only involves scalar multiples and not (non-trivial) linear combinations).
- Three vectors are linearly independent if they don’t lie on the same two-dimensional plane, i.e., if one cannot be computed as a linear combination of the other two (or of one, meaning that it a scalar multiple of that one).
- Four vectors are linearly independent if one cannot be computed as a linear combination of the other three (or two or ...), meaning that they don’t all lie in the same three-dimensional subspace.
- Five vectors ...

In the above, a three-dimensional subspace is the span of any 3 linearly independent vectors, whether or not they are $\{e_i\}_{i=1}^3$. Similarly, a two-dimensional subspace is the span of any 2 linearly independent vectors, whether or not they are $\{e_i\}_{i=1}^2$. Similarly, a one-dimensional subspace is the span of any 1 linearly independent vector, whether or not it is $\{e_i\}_{i=1}^1$ (although note that here the span means just taking scalar multiplications, since in the one-dimensional case, vector addition degenerates to scalar multiplication).



(a) Linear dependence of two vectors. (b) Linear dependence of three vectors. (c) Linear independence of two vectors.

Figure 4.4: Examples of linear dependence and independence in \mathbb{R}^2 .

(a) Linear dependence of three vectors. (b) Linear dependence of four vectors. (c) Linear independence of two vectors.

Figure 4.5: Examples of linear dependence and independence in \mathbb{R}^3 .

As one gets more vectors, the intuition can get tricky, since it can be difficult to visualize high-dimensional spaces, but the definitions and ideas go through very cleanly; and having a good understanding of these ideas in \mathbb{R}^2 and \mathbb{R}^3 , and how they are different in \mathbb{R}^2 and \mathbb{R}^3 , goes a long way to helping the intuition in higher dimensions.

By the way, here is another definition of linear independence.

Definition 26 A set of k vectors is linearly independent iff the only solution to $a_1v_1 + \dots + a_kv_k = 0$ is $a_1 = \dots = a_k = 0$. Otherwise, they are linearly dependent.

In addition, here is yet another definition of linear independence.

Definition 27 The vectors $\{v_i\}$ are linearly independent if none of the v_i are a linear combination of the others. Otherwise, they are linearly dependent.

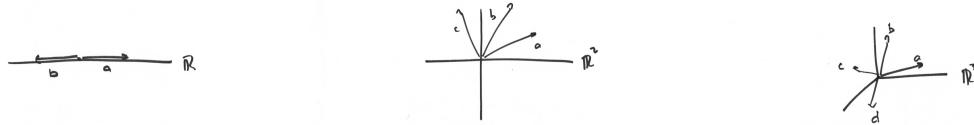
It is a fact that we won't prove that all of these definitions of linear independence are equivalent. We aren't going to go through all the details to prove the equivalence of these definitions, but you should develop some working knowledge of their similarities and work with whichever is the easiest.

Here is an important question: How many vectors in \mathbb{R}^n can be linearly independent? To get an idea, see Figure 4.6, which gives an example for \mathbb{R}^k , where $k = 1, 2, 3$, there can be k linearly independent vectors and that $k+1$ vectors are not linearly independent.

Here is the theorem that generalizes that.

Theorem 6 In \mathbb{R}^n :

- Any set of $n+1$ vectors are never linearly independent.



(a) Linear dependence/independence in \mathbb{R} . (b) Linear dependence/independence in \mathbb{R}^2 . (c) Linear dependence/independence in \mathbb{R}^3 .

Figure 4.6: Examples of the number of linearly dependent/independent vectors in \mathbb{R} , \mathbb{R}^2 , and \mathbb{R}^3 .

- Any set of $n - 1$ vectors never span all of \mathbb{R}^n .

Examples. Here are some examples of this.

- $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ don't span \mathbb{R}^3 . Similarly, $\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$ don't span \mathbb{R}^3 . In both cases, the span of these two linearly independent vectors is a two-dimensional plane corresponding to $x_3 = 0$, and so the span of these two vectors is a two-dimensional subspace of \mathbb{R}^3 .
- Similarly, the two vectors $\begin{pmatrix} 1 \\ 7 \\ 4 \end{pmatrix}$ and $\begin{pmatrix} 3 \\ 8 \\ -2 \end{pmatrix}$ don't span \mathbb{R}^3 . They are linearly independent, and their span is a two-dimensional subspace: it is some other two-dimensional plane at some non-trivial angle with respect to the canonical axes.
- $\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$, $\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$, $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$, and $\begin{pmatrix} 17 \\ 12 \\ -2 \end{pmatrix}$ are not linearly independent, but their span is all of \mathbb{R}^3 .
- $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, $\begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}$, $\begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix}$, and $\begin{pmatrix} 17 \\ 12 \\ -2 \end{pmatrix}$ are not linearly independent, and their span is not all of \mathbb{R}^3 , but instead a two-dimensional subspace of \mathbb{R}^3 .

This last example is particularly interesting since it illustrates that we can have more than n vectors in \mathbb{R}^n that span a subspace of dimension less than n , which is a common situation in data science.

4.2.4 Testing for linear dependence and independence

Advanced comment. This subsection contains some more advanced forward-pointing comments. Skim it, but don't worry if it doesn't all make sense yet.

The notions of linear combinations, span, and linear dependence/independence are related but distinct. To see the relationships, look carefully about what is assumed in each definition, and what is being specified by each definition.

- The definition of linear combination in Definition 23 says that if we are given several vectors (and several coefficients), then another vector is a linear combination of the given set if that other vector

can be written in a certain way. That is, a given vector is a linear combination of other vectors if it can be written in a certain way. (That certain way may or may not be unique—the definition does not specify that.)

- The definition of span in Definition 24 says that if we are given several vectors, then all of the vectors that can be written in a certain way, for some values of coefficients, is the span. So, in particular, the span of a set of vectors is not (in general, unless, e.g., the set consists of just the zero vector) a single vector, but instead it is an entire set of vectors, that can be expressed in terms of the given set, for some value of the coefficients.
- The definitions of linearly dependence/independence in Definition 25 (or 26 or 27) say that if we are given several vectors, then that set of vectors is linearly dependent or linearly independent, depending on how those vectors can be expressed in terms of each other, for some unspecified values of coefficients.

Among other things, a vector in the span of a set of vectors is a linear combination of that set and is linearly dependent on that set, while a vector not in the span of a set of vectors is not a linear combination of that set and is linearly independent of that set. (Some of these claims can be good HW problems.) We will see more examples of how these distinct notions are related in later chapters. For now, we simply want to note the following.

If we are given two vectors u and v , and if we then compute a vector in their span, say $w = 2u + 3v$, then the three vectors u, v, w are not linearly independent. (Obviously, since $2u + 3v - w = 0$.) In many cases, the more interesting question is the following: say that we are given three vectors, u, v, w , then are they linearly dependent or linearly independent?

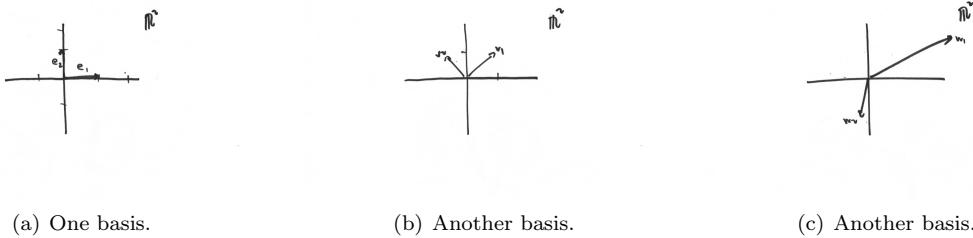
The way to answer this question is the following. If we view each vector as a column vector, we can define a matrix, call it A , such that the first/second/third column of A is $u/v/w$. (If the number of vectors is the same as the dimension of each vector, then A is square, otherwise it is rectangular, either with more rows than columns or with more columns than rows.) Then, we can consider post-multiplying A by some vector $x \in \mathbb{R}^3$. That is, we can compute Ax . Clearly, $Ax = x_1u + x_2v + x_3w$. (If that is not clear, then pause and confirm it.) Then, the question of whether u, v, w are linearly dependent or linearly independent is the same as the question of whether there is a vector x such that $Ax = 0$, or whether $Ax = b \neq 0$, for all vectors x . (That was a mouthful, and it is important enough that we will get back to it in some detail in a later chapter, but take a minute to re-read it.)

This “inverse problem” (which is basically the linear equation solving perspective) is a harder problem than the “forward problem” (which is basically the linear transformation perspective) of computing a given vector in the span of other vectors, and to solve it needs a bit more machinery. (We are calling this an inverse problem since if A and b are in \mathbb{R} , then the solution is $x = A^{-1}b$ assuming that $A \neq 0$. As we will see eventually, a similar statement holds true more generally, but the situation is more complex.) Moreover, there are some subtleties, depending on whether there is noise in the data and/or rounding error due to the finite precision of the computer. That is, there is a mathematically well-defined answer, but that answer is sometimes not robust to noise in the data or noise in roundoff. It is, however, a very important in practice in data science. We will get back to how to deal with this in a few chapters, but it is good to keep it in mind for now.

4.3 Bases, orthogonal bases, and orthonormal bases

4.3.1 Basis vectors

The idea of a basis and of basis vectors generalizes the idea of a set of standard/canonical vectors, which we know can be used to describe other vectors. For example, any vector in the plane \mathbb{R}^2 can be written

Figure 4.7: Three different bases for \mathbb{R}^2 .

uniquely as a linear combination of

$$e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Relatedly, choosing a basis for all of \mathbb{R}^n or for a lower-dimensional subspace of \mathbb{R}^n generalizes the idea of choosing the standard basis as the set of axes on \mathbb{R}^2 or \mathbb{R}^3 with which to describe other vectors. The idea of standard vectors and of using those vectors as basis vectors is so “obvious” that these vectors are sometimes just used and not explicitly thought about as a basis. While they are “nice” in some ways (e.g., they have only one non-zero entry which equals one, they are perpendicular to each other, etc.), they are not unique. For example, any point in the plane \mathbb{R}^2 can also be written uniquely as a linear combination of

$$\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

In addition, and important for data science, the standard basis vectors are often not the best basis to use to describe the properties of a given data set. Indeed, we will see later that a given data set often implicitly defines a basis that is the best basis (and much better than the standard basis) with which to describe that data set.

To begin to understand bases, see Figure 4.7 for an illustration of different axes that we could choose to represent a point on \mathbb{R}^2 . The first is the usual X-Y axes; the second is that rotated by $\pi/4$ radians; and the third is some funny set of axes, but the point is that we can uniquely represent each point in \mathbb{R}^2 with it. In each case, the direction of the basis vectors gives the direction of the axes, and the length of the basis vectors gives the units. The point is that *every* point on the plane \mathbb{R}^2 can be written *uniquely* in terms of the two basis vectors.

The word “every” and “uniquely” were important in the last claim. If it is not the case that we can write every point in a space uniquely in terms of a set of vectors, then that set of vectors does not provide a basis for that space. See Figure 4.8 for two problematic cases that come up when discussing bases.

- In the first case, we have three vectors in \mathbb{R}^2 , and so given a point we can’t represent it in a unique way.
- In the second case, we have two vectors, but they point along the same direction, i.e., they are linearly dependent, and so points along that line can’t be represented uniquely, and also there are lots of points, i.e., all those not exactly on the line, that we can’t represent as linear combinations of those two vectors.

In both of those cases, for two different reasons, we do *not* have a basis for \mathbb{R}^2 .

Here is the definition of the basis of a subspace of a vector space. Note that if $V = \mathbb{R}^n$ in this definition, i.e., if the subspace is all of \mathbb{R}^n , then this definition provides a definition for a basis of the vector space \mathbb{R}^n .



(a) Too many vectors.

(b) Not enough dimensions.

Figure 4.8: Two issues that come up with bases.

Definition 28 Let $V \subset \mathbb{R}^n$ be a subspace. A set of vectors, $v_1, \dots, v_k \in V$ is called a basis of V if it satisfies any of the following three equivalent conditions.

- The set is a maximally linearly independent set, i.e., it is linearly independent, and if we add one more vector from V to it, then it will not be linearly independent.
- The set is a minimal spanning set, i.e., it spans V and if we remove one vector from it, then it will no longer span V .
- The set is a linearly independent set spanning V .

This definition of a basis of a vector space is designed so that the two problematic cases illustrated in Figure 4.8 do not arise. In particular, given a basis for V , there is a unique way to represent each vector $v \in V$: every vector $v \in V$ can be expressed as a linear combination of elements of V , and there is a unique way to do so.

Examples. Here are examples.

- The standard basis, e_1 , e_2 , and e_3 , is the basis that is given to you that is often not the most convenient, but it is a basis for \mathbb{R}^3 .
- $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$ span \mathbb{R}^2 and provide a basis for \mathbb{R}^2 . Ditto for $\begin{pmatrix} 2 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 1/2 \\ -3 \end{pmatrix}$, as well as other pairs of vectors that are not scalar multiples of each other. Each such pair provides a basis for \mathbb{R}^2 .
- Consider the solution to the following linear equation

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

which we can write as $x_3 = -x_1 - x_2$. We claim that this is a subspace of \mathbb{R}^3 , call it V . Then,

$$\left\{ \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \right\}$$

is a basis for V . Also,

$$\left\{ \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} \right\}$$

is a basis for V . Also,

$$\left\{ \begin{pmatrix} -1/2 \\ -1/2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \right\}$$

is a basis for V . Also, any set of 2 vectors that are linearly independent, i.e., s.t. the sum of the entries is 0, is a basis for V .

Note that if we add any of the following vectors

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}$$

then we get a basis for \mathbb{R}^3 . But if we don not consider any of these vectors, and we only consider linear combinations of the first two, then we have a basis for the subspace orthogonal to the line we started with.

Given this definition of a basis of a subspace, we can define the dimension of a subspace.

Definition 29 Let $V \subset \mathbb{R}^n$ be a subspace. The dimension of V is the number of vectors in any basis of that subspace.

Remark. Clearly, this definition holds for the subspace consisting of just the origin as well as the subspace consisting of a line through the origin as well as all of \mathbb{R}^n . The point here is that it holds more generally.

Remark. It is a fact that we won't prove that, although there are many different bases for a given subspace, the number of vectors in any of those bases is the same.

4.3.2 Orthogonal and orthonormal bases

We saw in Figure 4.7 several examples of bases. The first two were “nice,” in the sense that the two axes were unit length and perpendicular to each other, and the third seemed a little strange and perhaps harder to work with. The third is a legitimate basis, and it can be useful, but it can also be quite “difficult” to work with, both in terms of human understanding as well as when algorithms are implemented on a computer.

Given their importance, the first two examples in Figure 4.7 have a special name, and one is often interested in computing bases that look like them. By “look like,” we do not mean in the sense that they correspond to the canonical vectors, but instead we mean in the sense that they are unit length and perpendicular/orthogonal to each other. Bases that have this property are sometimes easier to understand in data science, and they are generally much easier to work with when algorithms are implemented on a computer.

Definition 30 A basis v_1, \dots, v_k for a subspace $V \subset \mathbb{R}^n$ is orthonormal if:

- each basis vector has unit length; and
- each basis vector is orthogonal to all others.

Remark. Sometimes a basis is called orthogonal if it satisfies the orthogonality requirement but not the normality requirement, and sometime the term orthogonal is used to refer to what we are calling orthonormal. People aren't always consistent about this, so when in doubt, ask.

These two conditions in this definition are important, and they can be written in a compact way. Let's actually do this in two related ways: first, in terms of the vectors $\{v_i\}_{i=1}^k$ themselves; and second, in terms of an $n \times k$ matrix constructed to have these vectors as its columns.

To express these two orthonormality conditions in terms of conditions on vectors, note that

- the two vectors v_i and v_j are orthogonal to each other if $v_i^T v_j = 0$; and
- the vector v_i is unit length if $\|v_i\|_2 = \sqrt{v_i^T v_i} = 1$.

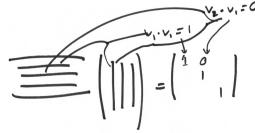


Figure 4.9: Illustration of the multiplication of a matrix consisting of orthonormal columns with its transpose.

If we use the relatively common notation that

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases},$$

then these two conditions can be written compactly (i.e., if we don't want to have to specify for all $i \neq j$ and for all i separately) as

$$v_i^T v_j = \delta_{ij}, \quad \text{for } i, j \in \{1, \dots, k\}.$$

Alternatively, to express these two orthonormality conditions in terms of conditions on matrices, let's define an $n \times k$ matrix A to be

$$A = (v_1 \ v_2 \ \cdots \ v_k).$$

(Note what we did: we defined a matrix A that looks like a $1 \times k$ matrix, but recall that each element of that matrix is a vector, i.e., a $n \times 1$ matrix, and thus A is an $n \times n$ matrix, the i^{th} column of which is the vector v_i .) Next, let's consider the matrix $A^T A$. Observe that $A^T A$ is a $k \times k$ matrix. Let's ask what information is contained in the (ij) element of this matrix, i.e., in the element $(A^T A)_{ij}$? See Figure 4.9 for an pictorial illustration of this. That is,

$$(A^T A)_{ij} = v_i^T v_j, \tag{4.3}$$

which—in this case—equals 1 or 0, depending on whether or not $i = j$. That is, in this case

$$(A^T A)_{ij} = \delta_{ij},$$

and we can write the matrix-matrix product as

$$A^T A = I_k. \tag{4.4}$$

This is the matrix way to express that a matrix has columns that form an orthonormal basis.

Remark. Matrices satisfying the condition in Equation (4.4) are known as *orthogonal* or *orthonormal* matrices, are they are often denoted by Q or V or U . You should keep in mind that sometimes this terminology is restricted to square matrices. If in doubt, ask; and it's best to be careful about what exactly the dimensions of such matrices are.

Aside. Even though we introduced the matrix $A^T A$ in the context of determining whether or not it equals an identity, to determine whether or not A contains orthonormal columns, matrices of this form (i.e., a matrix multiplied by the transpose of itself) are of more general interest. The reason is that each element of such matrices contain the dot product between two columns of A , i.e., as in Equation (4.3). We will get back to the usefulness of such matrices later.

Remark. Hopefully, it is obvious that the standard basis, $\{e_i\}_{i=1}^n$ is orthonormal. If not, then observe that $e_i^T e_i = 1$, for all $i \in [n]$; and $e_i^T e_j = 0$, if $i \neq j$. Alternatively, if the n standard basis vectors $\{e_i\}_{i=1}^n$ are encoded as the columns of a matrix $A = (e_1 \ e_2 \ \cdots \ e_n)$, then $A = I_n$. Since $I_n^T = I_n$, we have that $A^T A = I_n$.

Remark. Although we described orthogonal matrices in terms of their element-wise construction (computed either as a product of two column vectors or as a product of two matrices), there is another related way to interpret them. For this, let's not consider a subspace $V \subset \mathbb{R}^n$. Instead, let's consider all of \mathbb{R}^n , in which case $k = n$. Given a matrix $A \in \mathbb{R}^{n \times n}$, recall that the inverse matrix A^{-1} is the matrix such that $A^{-1}A = AA^{-1} = I_n$. So, orthogonal matrices are matrices A such that $A^{-1} = A^T$.

We'll conclude by noting that orthogonal matrices are “nice” for many reasons. Here is one, which can be described in terms of sets of orthogonal vectors or the columns of an orthogonal matrix.

Theorem 7 *An orthogonal set of non-zero vectors v_1, \dots, v_k is linearly independent.*

Combined with Definition 28, we see that an orthogonal set of non-zero vectors defines a basis for their span. This is true for k vectors, and it is also true for n vectors. In the latter case, we have a basis for all of \mathbb{R}^n . In particular, this says that *any* $n \times n$ orthogonal matrix—and not just the standard basis vectors—provides a basis for \mathbb{R}^n .

Computing orthonormal bases is *very* important and *very* useful in data science and beyond. Conceptually, this consists of two parts: first, computing a basis such that the basis vectors are orthogonal to each other; and second, making sure that the basis vectors in that basis are also unit-length/normal. (Sometimes, as we will see below, in an algorithm, these two steps happen together.) The latter is easy, while the former is somewhat harder. Next, we turn to each in turn.

4.3.3 Computing an orthonormal basis from an orthogonal basis

Let's say that we have a basis, the elements of which are orthogonal to each other, and we want to compute an orthonormal basis, i.e., one in which the elements are orthogonal to each other and are also normalized. (Here, we will be working with the Euclidean norm.) This is easy—just compute the norm of each vector, and divide by the norm to compute the normalized version of that vector.

Let's be a little more precise about that and put that into matrix language. (Although the basic idea is easy, putting this into matrix language will help us put an easy idea into the less-familiar form of matrix-matrix multiplication, and so it's a chance to practice that, before we get to less easy things.)

In the notation we used before, this means that we have a set of vectors, $\{v_1, \dots, v_k\}$, each of which is an element of \mathbb{R}^n , such that any two different vectors v_i and v_j in that set are orthogonal to each other, i.e., such that $v_i^T v_j = 0$ if $i \neq j$. Note that we have not said anything about the values of $v_i^T v_i$, for any $i \in [k]$. To organize this information into a matrix, let's define an $n \times k$ matrix A to be

$$A = \begin{pmatrix} v_1 & v_2 & \cdots & v_k \end{pmatrix}.$$

The requirement that the basis vectors be orthogonal to each other means that

$$A^T A = D,$$

where D is a $k \times k$ diagonal matrix, all the diagonal entries of which are positive, i.e., non-zero and non-negative. (It is diagonal since the off-diagonal elements are the dot products between different basis vectors, which equal zero, since they are perpendicular/orthogonal; and the diagonal entries are all non-zero since each vector in the basis is non-zero and thus has a non-zero non-negative norm.)

If we denote the i^{th} diagonal entry of D by $d_i = (v_i^T v_i)^{1/2}$, then we can define the diagonal matrix in which the i^{th} diagonal equals $1/d_i$. Let's call this matrix D^{-1} . (We are calling it D^{-1} since it is the inverse of D , in the sense that $D^{-1}D = DD^{-1} = I_k$. A good problem is to confirm that that is true.) Let's now consider the matrix defined by the product of A and D^{-1} , i.e., $\tilde{A} = AD^{-1}$. In terms of the columns of A and the elements of D^{-1} , this matrix has the form

$$\tilde{A} = AD^{-1} = \begin{pmatrix} \frac{1}{d_1} v_1 & \frac{1}{d_2} v_2 & \cdots & \frac{1}{d_k} v_k \end{pmatrix}.$$

(A good problem is to confirm that that is true.) Then,

$$\tilde{A}^T \tilde{A} = I.$$

That is, the columns of \tilde{A} are orthogonal and normal, and thus they form an orthonormal basis.

Note that we could have $k < n$ and we could have $k = n$ in this discussion, but we could not have $k > n$, since we can't have $k > n$ basis vectors in \mathbb{R}^n .

Examples. Here are examples.

- Consider

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

In this case

$$D = \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix}$$

and

$$\tilde{A} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This just says that if we have a stretched version of the standard basis vectors, which are orthogonal but not normal, then we can unstretch them and get back the standard basis vectors, which form an orthonormal basis.

- Consider

$$A = \begin{pmatrix} 2 & 3 \\ 2 & -3 \end{pmatrix}.$$

In this case

$$D = \begin{pmatrix} 2\sqrt{2} & 0 \\ 0 & 3\sqrt{2} \end{pmatrix}$$

and

$$\tilde{A} = \begin{pmatrix} 2 & 3 \\ 2 & -3 \end{pmatrix} \begin{pmatrix} \frac{1}{2\sqrt{2}} & 0 \\ 0 & \frac{1}{3\sqrt{2}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

This is just the rotated version of the standard basis vectors we discussed above, which itself is an orthonormal basis.

- Consider

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

We saw this example before when we considered the set of vectors orthogonal to the line $x_1 + x_2 + x_3 = 0$. This is the first column, and the next two columns are two vectors that are orthogonal to that line. In this case,

$$D = \begin{pmatrix} \sqrt{3} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix},$$

meaning that the columns of A are orthogonal to each other but not normalized, and

$$\tilde{A} = AD^{-1} = \begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{2} & 0 \\ 1/\sqrt{3} & -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{3} & 0 & -1/\sqrt{2} \end{pmatrix}.$$

The columns of \tilde{A} are unit-length and pair-wise orthogonal, as can be verified by computing $\tilde{A}^T \tilde{A}$.

Problem. Show that any orthogonal basis can be converted into an orthonormal basis by normalizing each basis vector by its Euclidean norm, and show that this can be accomplished with a matrix-matrix multiplication. In particular, if A is an $n \times k$ matrix with pairwise orthogonal but not necessarily normal columns, then describe a way to convert this to an $n \times k$ orthonormal matrix V , the columns of which are orthogonal and normal, by taking the product of two matrices, one of which is A . (We outlined this above.)

4.3.4 Computing an orthonormal basis from any set of columns

A harder but much more important problem is to compute from any set of column vectors an orthonormal basis for that set of columns, i.e., for the span of that set of columns. (BTW, there is nothing special about column vectors. This procedure could be applied to row vectors. To see that, note that if you have a set of row vectors, then form a matrix in which those row vectors are the rows, then take the transpose of that matrix, and apply the following procedure to the transposed matrix, the columns of which are the transpose of the original row vectors.)

There are several related ways to do this. Here, we will describe a procedure known as the *Gram-Schmidt* process that forms something called a *QR decomposition*. Like many things in linear algebra, there is an algebraic perspective to this (which is often easily automated on a computer), and there is a geometric perspective to this (which is typically more intuitive for humans). In this section, we will describe the algebraic perspective; and in the next section, we will describe the geometric perspective.

Let's say that we have an $n \times k$ matrix A , which in terms of its k columns can be represented as

$$A = (\begin{array}{cccc} a_1 & a_2 & \cdots & a_k \end{array}).$$

For now, let's say that $n \geq k$. If $n = k$, then A is a square matrix; and if $n \geq k$, then it is a “tall” rectangular matrix. (Much of what we say will generalize to $n < k$, in which case the matrix is rectangular the other way, and we will get to that later.) What this assumption means is the following: if we think of the rows of A as data points, each of which is an element of \mathbb{R}^k , then we have $n \geq k$ data points. This corresponds to the picture, e.g., of having 2 or more points on the 2-dimensional plane (as opposed to having 2 points in \mathbb{R}^n , for $n \geq 3$). Aside from that, though, now the matrix is arbitrary, i.e., there is no assumption that $A^T A$ is diagonal or the identity or anything like that.

Here is the Gram-Schmidt process.

1. $u_1 = a_1$, and then $q_1 = \frac{1}{\|u_1\|_2} u_1$
2. $u_2 = a_2 - (a_2 \cdot q_1)q_1$, and then $q_2 = \frac{1}{\|u_2\|_2} u_2$
3. \dots
4. $u_{\xi+1} = a_{\xi+1} - (a_{\xi+1} \cdot q_1)q_1 - \cdots - (a_{\xi+1} \cdot q_\xi)q_\xi$, and then $q_{\xi+1} = \frac{1}{\|u_{\xi+1}\|_2} u_{\xi+1}$

That is, for each vector in turn, take the vector and subtract off a term that is in the direction of each previous vector, and then normalize the resulting vector.

What this Gram-Schmidt procedure outputs is a set of vectors $\{q_i\}$, for $1 \leq i \leq k$, each of which is in \mathbb{R}^n (since the columns of A were in \mathbb{R}^n), as well as a bunch of coefficients $\{a_i \cdot q_j\}$, for $1 \leq i \leq j \leq k$. (Be sure to parse that last inequality: the k refers to the dimension of A , which is an $n \times k$ matrix, and the i and j are indices, and the reason $1 \leq i \leq j$ is that in the Gram-Schmidt process we normalized and subtracted things off of previous vectors.) These coefficients can be organized into a matrix, and since we have only defined them for $1 \leq i \leq j \leq k$, let's set the other elements (those for which $i > j$) equal to zero.

Given this information the QR decomposition or QR factorization is

$$A = \begin{pmatrix} a_1 & a_2 & \cdots & a_k \end{pmatrix} = \begin{pmatrix} q_1 & q_2 & \cdots & q_k \end{pmatrix} \begin{pmatrix} a_1 \cdot q_1 & a_2 \cdot q_1 & \cdots & a_k \cdot q_1 \\ 0 & \ddots & & \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & a_k \cdot q_k \end{pmatrix} = QR. \quad (4.5)$$

Remark. It can be shown that the matrix Q is orthonormal, i.e., for an $n \times k$ matrix A , the $n \times k$ matrix Q is such that $Q^T Q = I_k$. This is easier to see from the geometric perspective, so we will show that in the Section 4.4. Importantly, the $k \times k$ matrix $Q^T Q$ (which encodes the orthonormality information) is different than the $n \times n$ matrix QQ^T . This latter matrix also has a nice geometric interpretation, which we will also discuss.

Remark. Equation (4.5) is known as the QR decomposition or QR factorization. It is our first example of a matrix factorization, but we will see others. By *matrix factorization*, we mean that we can express a matrix A in another form that is convenient for something. Here, we are expressing a matrix A as the product of two matrices, the first of which is orthonormal, the second of which is upper triangular.

Remark. Matrix factorizations are important in data science and applied mathematics more generally. Basically, they are of interest since they are useful for two things: first, to express a matrix in a form on which it is easier to perform computations; and second, to highlight structure that may be present in data. The former gets more attention in traditional numerical classes, but the latter is probably more important in data science.

Remark. You should think of the elements of the R matrix as undoing all the operations in the Gram-Schmidt process. Those on the diagonal of R correspond to the normalization, and those off the diagonal correspond to the terms that are subtracted from each a_j to get each u_j .

Examples. Here are some examples.

- Consider

$$A = \begin{pmatrix} -1 & 3 \\ 1 & 5 \end{pmatrix}. \quad (4.6)$$

In this case,

$$\begin{aligned} u_1 &= \begin{pmatrix} -1 \\ 1 \end{pmatrix} \\ q_1 &= \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \\ u_2 &= \begin{pmatrix} 3 \\ 5 \end{pmatrix} - \left(\frac{1}{\sqrt{2}} \begin{pmatrix} 3 & 5 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right) \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 4 \end{pmatrix} \\ q_2 &= \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \end{aligned}$$

So, the matrix Q is

$$Q = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix},$$

i.e., it is just the rotated version of the standard basis that we saw before. From this, we can compute the matrix R as

$$\begin{aligned} R &= \begin{pmatrix} a_1 \cdot q_1 & a_2 \cdot q_1 \\ 0 & a_2 \cdot q_2 \end{pmatrix} \\ &= \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ 0 & 4\sqrt{2} \end{pmatrix} \\ &= \sqrt{2} \begin{pmatrix} 1 & 1 \\ 0 & 4 \end{pmatrix}. \end{aligned}$$

And, to confirm

$$QR = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} \sqrt{2} \begin{pmatrix} 1 & 1 \\ 0 & 4 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 4 \end{pmatrix} = \begin{pmatrix} -1 & 3 \\ 1 & 5 \end{pmatrix} = A.$$

- Consider

$$A = \begin{pmatrix} 3 & -1 \\ 5 & 1 \end{pmatrix}, \quad (4.7)$$

which is just the previous matrix with the columns in a different order. In this case,

$$\begin{aligned} u_1 &= \begin{pmatrix} 3 \\ 5 \end{pmatrix} \\ q_1 &= \frac{1}{\sqrt{34}} \begin{pmatrix} 3 \\ 5 \end{pmatrix} \\ u_2 &= \begin{pmatrix} -1 \\ 1 \end{pmatrix} - \left(\frac{1}{\sqrt{34}} (-1 \ 1) \begin{pmatrix} 3 \\ 5 \end{pmatrix} \right) \frac{1}{\sqrt{34}} \begin{pmatrix} 3 \\ 5 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix} - \frac{2}{34} \begin{pmatrix} 3 \\ 5 \end{pmatrix} = \begin{pmatrix} \frac{-20}{17} \\ \frac{12}{17} \end{pmatrix} \\ q_2 &= \frac{1}{\sqrt{34}} \begin{pmatrix} -5 \\ 3 \end{pmatrix}. \end{aligned}$$

So, the matrix Q is

$$Q = \frac{1}{\sqrt{34}} \begin{pmatrix} 3 & -5 \\ 5 & 3 \end{pmatrix}.$$

(From this, we can compute the matrix R , as before, and then the matrix $A = QR$; verify this if you like.) The point here is that the two matrices lead to two different Q matrices. Each Q matrix consists of a different orthonormal basis. If we wanted exactly the same bases, we would have had to do something different, but if we wanted some orthonormal basis, then we have achieved that.

- Consider

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \quad (4.8)$$

in which the columns of A are $a_1 = (1 \ 1 \ 0)^T$, $a_2 = (1 \ 0 \ 1)^T$, and $a_3 = (0 \ 1 \ 1)^T$. In

this case,

$$\begin{aligned}
u_1 &= a_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \\
q_1 &= \frac{1}{\|u_1\|_2} u_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{pmatrix} \\
u_2 &= a_2 - (a_2 \cdot q_1)q_1 \\
&= \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} - \left(\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right) \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ 1 \end{pmatrix} \\
q_2 &= \frac{1}{\|u_2\|_2} u_2 = \frac{1}{\sqrt{3/2}} \begin{pmatrix} 1/2 \\ -1/2 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{6} \\ -1/\sqrt{6} \\ 2/\sqrt{6} \end{pmatrix} \\
u_3 &= a_3 - (a_3 \cdot q_1)q_1 - (a_3 \cdot q_2)q_2 \\
&= \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} - \frac{1}{\sqrt{2}} \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{pmatrix} - \frac{1}{\sqrt{6}} \begin{pmatrix} 1/\sqrt{6} \\ -1/\sqrt{6} \\ 2/\sqrt{6} \end{pmatrix} = \begin{pmatrix} -2/3 \\ 2/3 \\ 2/3 \end{pmatrix} \\
q_3 &= \frac{1}{\|u_3\|_2} u_3 = \begin{pmatrix} -1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{pmatrix}.
\end{aligned}$$

So, the matrix Q is

$$Q = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{6} & -1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 0 & 2/\sqrt{6} & 1/\sqrt{3} \end{pmatrix},$$

and, from this, we can compute the matrix R as

$$\begin{aligned}
R &= \begin{pmatrix} a_1 \cdot q_1 & a_2 \cdot q_1 & a_3 \cdot q_1 \\ 0 & a_2 \cdot q_2 & a_3 \cdot q_2 \\ 0 & 0 & a_3 \cdot q_3 \end{pmatrix} \\
&= \begin{pmatrix} 2/\sqrt{2} & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 3/\sqrt{6} & 1/\sqrt{6} \\ 0 & 0 & 2/\sqrt{3} \end{pmatrix}.
\end{aligned}$$

And, if these are multiplied out, you can confirm that $QR = A$.

- Consider

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad (4.9)$$

This is a different matrix than given in Eqn. (4.8), e.g., it has different dimensions, but it is related: its two columns are the same as the first two columns of the matrix given in Eqn. (4.8). What effect does that have on the QR decomposition? Well, we can run through the same procedure as before, except that we don't compute any u_3 or q_3 since there is no a_3 . If we do that, then we get that the matrix Q is

$$Q = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{2} & -1/\sqrt{6} \\ 0 & 2/\sqrt{6} \end{pmatrix},$$

and the matrix R is

$$R = \begin{pmatrix} 2/\sqrt{2} & 1/\sqrt{2} \\ 0 & 3/\sqrt{6} \\ 0 & 0 \end{pmatrix}.$$

And, if these are multiplied out, you can confirm that $QR = A$.

Remark. In all these examples, we have kept track of both Q and R and then confirmed that $QR = A$. Alternatively, you could only compute Q and use the fact that it is orthogonal to compute R as follows:

$$Q^T A = Q^T QR = IR = R.$$

It's good to check, so it's best to do the first way until you are comfortable, but the second way is also good to know.

4.3.5 Orthogonality more generally

Orthogonality between two things is an important notion that holds more generally than just between two vectors in \mathbb{R}^n . Here, we give an overview.

Orthogonality between two one-dimensional subspaces. Two one-dimensional subspaces, i.e., two lines through the origin, are orthogonal if every vector in one subspace is orthogonal to every vector in the other subspace. This boils down to the question of whether any particular vector, i.e., a basis, in one subspace is orthogonal to any particular vector, i.e., a basis, in the other subspace.

Orthogonality between a vector (or one-dimensional subspace) and a subspace. A vector is said to be orthogonal to a subspace if the vector is orthogonal to every vector in the subspace. This boils down to the question of whether the vector is orthogonal to any basis of the subspace. If x is the vector, written as a $n \times 1$ matrix, and B is an $n \times d$ matrix whose columns consist of a basis for that subspace, then this means that $x^T B = 0$. Clearly, the same holds for orthogonality between a one-dimensional subspace and any other subspace.

Orthogonality between two subspaces. Two subspaces are said to be orthogonal to each other if every vector in one subspace is orthogonal to every vector in the other subspace. This boils down to the question of whether every vector in a basis for the one subspace is orthogonal to a basis for the other subspace. If A is an $n \times d_1$ matrix whose columns consist of a basis for the first subspace, and If B is an $n \times d_2$ matrix whose columns consist of a basis for the second subspace, then this means that $A^T B = 0$. (As with vectors, these matrices don't need to be normalized or orthonormal to answer the question of whether two subspaces are orthogonal, but they certainly can be.)

(If you didn't get all of that, that is okay, as we'll be spending a lot more time on it.)

4.4 Projections

Consider Figure 4.10, which illustrates the idea of the projection of a vector onto another vector. In particular, Figure 4.10(a) shows a vector $b \in \mathbb{R}^2$ as well as the projection of b onto each of the two vectors in the standard basis

$$e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$



(a) Projection of a vector onto the standard basis vectors. (b) Projection of a vector onto another vector. (c) Projection of a vector when it lies in the span. (d) Projection of a vector when it is perpendicular to the span.

Figure 4.10: Several illustrations of a projection.

For example, the projection of the vector $b = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$ onto e_1 is the vector $\begin{pmatrix} 2 \\ 0 \end{pmatrix}$, and the projection of it onto the vector e_2 is the vector $\begin{pmatrix} 0 \\ 3 \end{pmatrix}$. Figure 4.10(b) shows the projection of b onto another vector. Informally, a projection takes a point and finds the closest point in the span of what is being projected onto, and it returns that closest point. The question we want to consider is how to project onto other vectors that point in some arbitrary direction of \mathbb{R}^n , and more generally how to project onto other subspaces.

4.4.1 Projecting onto a vector, i.e., onto a one-dimensional subspace

Let's start off with a simple case and then generalize.

Projecting a point onto a line, when both are on the plane. Let's start off with a relatively simple example, which is illustrated in Figure 4.10(b). Let's say that we have a vector $b \in \mathbb{R}^2$ and a line through the origin determined by a vector $a \in \mathbb{R}^2$. That is, we have a vector a , and we are interested in the subspace defined by it, i.e., in $\text{Span}(a)$. Let's call $p \in \mathbb{R}^2$ the vector on the line, i.e., in the subspace spanned by a , that is closest to b . Then, how do we find the vector p ?

Observe from Figure 4.10(b) that this vector p is at the intersection of $\text{Span}(a)$ and the line through b that is orthogonal to $\text{Span}(a)$. Observe also that if we think of p as an approximation of b , then $b = e + p$, and we can think of e as the error in the approximation, where $e \in \text{Span}(a^\perp)$, where a^\perp is the line perpendicular to a .

We could try to find p using trigonometry or calculus, but let's use linear algebra. Since $p \in \text{Span}(a)$, i.e., since $\text{Span}(a)$ is one-dimensional, p is on the line through a , we know that $p = xa$, for some $x \in \mathbb{R}$. We also know that a is perpendicular to $e = b - xa$, i.e.,

$$a^T (b - xa) = 0.$$

(This just says that $a^T e = 0$.) From this, we have that $a^T ax = a^T b$, from which it follows that $x = \frac{a^T b}{a^T a}$, from which it follows that

$$p = ax = a \frac{a^T b}{a^T a} \quad (4.10)$$

From this expression, changing a by, e.g., $a \rightarrow 2a$ does not change p (as it should not, since we are interested in $\text{Span}(a)$, and that does not change if we simply multiply a by a number). On the other hand, changing b by, e.g., $b \rightarrow 2b$ does change p (as it should, since we have a different point that we are projecting).

Let's look at Equation (4.10) in more detail. We can write this as

$$p = \frac{1}{a^T a} a(a^T b) = \frac{1}{a^T a} aa^T b = \frac{1}{a^T a} (aa^T)b, \quad (4.11)$$

where we have put the vector a in the numerator. Two things to note.

- The denominator $a^T a$ is dot product and thus a number. If $\|a\|_2 = 1$, then $a^T a = 1$, and we could ignore the denominator. That seems like a big assumption, but we are not interested in a , but instead in $\text{Span}(a)$, and so we could replace a with a normalized version of a . If we did that, then Eqn. (4.11) would become

$$p = aa^T b.$$

This expression has exactly the same form as the terms that we subtracted off from each column vector in each step of the QR factorization. (Thus, what the QR decomposition was doing geometrically was to consider each column, and project away the part of that column in the span of each previous column, for each previous column.)

- The remainder of Equation (4.11), on the other hand is the product of three matrices: a 2×1 matrix multiplied by a 1×2 matrix multiplied by a 2×1 matrix, assuming that a is a column vector, i.e., a 2×1 matrix. Since matrix multiplication is associative, we could do the second multiplication first, as we implicitly did in putting $a^T b$ in the numerator of Equation (4.10). Alternatively, we could multiply aa^T first, and this gives a 2×2 matrix. Thus, we can write, Equation (4.11) as

$$p = Pb,$$

where

$$P = \frac{aa^T}{a^T a} \quad (4.12)$$

is a projection matrix onto $\text{Span}(a)$.

That is, this projection is a linear transformation, and thus it can be expressed as a matrix, a projection matrix, that takes as input any vector $b \in \mathbb{R}^2$ and via a matrix-vector multiplication returns a vector $p \in \mathbb{R}^2$ that is the closest point in $\text{Span}(a)$ to b .

There are two special examples of vectors b to consider:

- If $b \in \text{Span}(a)$, then $b = \alpha a$ for some $\alpha \in \mathbb{R}$. In that case, then

$$p = Pb = \frac{1}{a^T a} aa^T b = \frac{1}{a^T a} aa^T \alpha a = \frac{a^T a}{a^T a} a \alpha = b.$$

That is, in this case, p is unchanged.

- If b is perpendicular to $\text{Span}(a)$, then $a^T b = 0 \in \mathbb{R}$, and thus

$$p = Pb = \frac{1}{a^T a} aa^T b = 0,$$

where in this expression $0 \in \mathbb{R}^2$. That is, in this case, p gets mapped to the 0 vector or origin.

These two cases are illustrated in Figure 4.10(c) and Figure 4.10(d). Later, we will see that these properties generalize to projections more generally.

Generalization to \mathbb{R}^n . Observe that nothing we said in the above derivation depends on the vectors a or b or p being in \mathbb{R}^2 , i.e., everything goes through if they are vectors in \mathbb{R}^n . If you don't see that, then ignore the figure, and run through the derivation again, except that replace "2" with "n" everywhere. In that case, of course, the matrix P is an $n \times n$ matrix, rather than a 2×2 matrix, but everything else holds true.

Two expressions for a projection. We have seen that the projection of a point onto a line in \mathbb{R}^n is a linear transformation that can be written as a matrix as

$$P_a = \frac{aa^T}{a^T a}. \quad (4.13)$$

You might wonder what exactly is the role of $a^T a$ in the denominator. It arises since the vector a may not be of unit norm (in the L_2 norm). But we know that the line defined by a is the same as the line defined by the unit-norm vector in the direction of a , i.e., $\text{Span}(a) = \text{Span}(u)$. Thus, the projection onto a should be the same as the projection onto the unit-norm vector in the direction of a . Thus, if a is not of unit L_2 -norm, then let's define $u = \frac{1}{\|a\|_2}a$, which is of unit L_2 -norm. In that case, if we run through the derivation again, then we get

$$P_a = uu^T \quad (4.14)$$

(Of course, if a has unit L_2 norm, then both expressions are the same.) What Eqn. (4.14) says is that to compute the projection matrix onto the span of a , simply work with a unit vector u in the same direction and compute the outer product of u with itself.

Some other facts about projections. Here are a few other facts about these projections that hold more generally. Some of these we will discuss in more detail later.

- $P = P^T$, i.e., P is a symmetric matrix.
- $P^2 = P$. That is, if you take a projection and apply it to a vector, and then apply it to the resulting vector, then the output is the same as if you only applied the projection once.
- If Q is a projection onto $v^\perp \subset \mathbb{R}^2$ that is the set of vectors $w \in \mathbb{R}^2$ such that $v^T w = 0$, i.e., which is the span of the line perpendicular to v , then $PQ = QP = 0$, where 0 is the all-zeros matrix.
- The decomposition $b = p + e$ of b into vectors p and e such that $p^T e = 0$ is basically an example of the Pythagorean Theorem.
- Finally, and more generally, if P is a projection matrix (onto the span of one column or onto the span of multiple columns), then it is easy to get the projection matrix onto the set of vectors perpendicular to that, call it Q : $Q = I - P$. (One could compute a basis for the perpendicular space, but that is in general a harder operation, and we have not discussed it yet.)

Projections more generally.

Advanced comment. Here, we have defined projections from a point to a line, or from a vector to a one-dimensional subspace. This provides the most immediate connection to what is done in planar Euclidean geometry. More generally, the basic idea of a projection is the following: given a point and some set, find the “best” or “closest” point in that set to the point. Clearly, what we have said is an example of this. But we could also define the projection of a point onto an affine set (i.e., a set which if shifted to the origin would be a subspace), or onto the unit ball of some norm, or onto all sorts of other things. Since we are interested in linear algebra, we will consider the generalization that is the projection of a point onto a higher-dimensional subspace.

4.4.2 Projecting onto higher-dimensional subspaces

Let's review what we have.

- **Angle and orthogonality between vectors.** We have defined the idea of the angle between two vectors as well as the notion of two vectors being orthogonal. This generalizes the idea of angle/perpendicularity on the Euclidean plane to arbitrary vectors in \mathbb{R}^n .



(a) Projecting a vector onto a plane in \mathbb{R}^3 .
(b) Projecting a vector onto a k -dimensional subspace V of \mathbb{R}^n .

Figure 4.11: Visualizing higher dimensional projections.

- **Direction versus magnitude.** Among other things, this notion of angle does not depend on the particular vector but only the direction in which the vector points. In particular, if we are interested in whether two vectors are orthogonal, then we can consider the unit-length vector in the direction of each, and if the dot product between those two unit-length vectors is zero, then the two vectors are orthogonal.
- **Vectors and one-dimensional subspaces.** Any vector, and thus in particular any unit length vector, can be used to define a simple subspace. In this simplest non-trivial case, this is a line through the origin in the plane. In the next simplest non-trivial case, this is a line through the origin in three-dimensional space. If the vectors are in \mathbb{R}^n , then this is a line through the origin in \mathbb{R}^n . More generally, and more precisely, the subspace is the set of all vectors that can be written as linear combinations of (since we have only one vector, this is basically just scalar multiples of) the original vector.
- **Orthogonality of one-dimensional subspaces.** This implies that the angle between two one-dimensional subspaces, and in particular whether those two one-dimensional subspaces are orthogonal, does not depend on any particular vector in those subspaces. It can, however, be computed from the unit length vectors (or some other vectors/bases) in those subspaces.
- **Subspaces orthogonal to one-dimensional subspaces.** The set of vectors orthogonal to a given vector is a subspace. In the simplest non-trivial case, on the plane, this is the one-dimensional line through the origin that is orthogonal to a given line through the origin. In the next simplest non-trivial case, in three-dimensional space, this is the two-dimensional plane through the origin that is orthogonal to a given line through the origin. A similar statement holds for any line/subspace through the origin in \mathbb{R}^n .
- **Projecting onto one-dimensional subspaces.** We can take any point and project it onto a line through the origin, i.e., onto a one-dimensional subspace that is the span of a single vector. This amounts to finding a point on that line/subspace that is closest to the original point. The difference between that point and the original point is orthogonal to the subspace that is being projected onto, and it is in the subspace orthogonal to that subspace.
- **First generalization of the Pythagorean Theorem.** A generalization of the Pythagorean Theorem holds. Recall that if we are considering the special base of a line through the origin in the plane, the we get exactly the traditional Pythagorean Theorem. More generally, if we project a vector in \mathbb{R}^n onto a one-dimensional subspace through the origin, then we get two vectors, one in the one-dimensional subspace through the origin, and one in the subspace perpendicular to that one-dimensional subspace. The sum of the Euclidean norm of the two vectors equals the original vector.
- **Higher-dimensional subspaces.** If we have an $m \times n$ matrix A , where $m > n$, i.e., that is a “tall” matrix, then the span of the columns of A form a subspace of \mathbb{R}^n , and the dimension of this subspace is m (or less, but we didn’t discuss that in detail yet). If we consider the QR decomposition of A , then we can express A as $A = QR$, where Q is an orthonormal matrix, i.e., a matrix such that $Q^T Q = I$,

and the subspace of \mathbb{R}^n that is the span of the columns of Q is the same as the subspace of \mathbb{R}^n that is the span of the columns of A . (Recall that this is since we can express the columns of Q in terms of linear combinations of the columns of A , and vice versa.)

All of this leads to the questions:

- Can we project onto higher-dimensional subspaces?
- If yes, then how do we do it?
- Once we do it, does a generalization of the Pythagorean Theorem (or other “nice” properties) hold?

The answers are:

- Yes.
- With a generalization of the procedure from Section 4.4.1, where the “lines” are replaced with “subspaces.” There are better and worse ways to do it, depending on whether you want rough insight, the method that is best to implement on a computer, etc. We will describe one below (as well as a few of the gotchas later).
- Yes.

Let’s describe the procedure.

Consider a vector in \mathbb{R}^n , and let’s say that we want to project onto the span of k vectors in \mathbb{R}^n . See Figure 4.11. To do that, let’s consider the $n \times k$ matrix $A \in \mathbb{R}^{n \times k}$, defined as

$$A = (a_1 \ a_2 \ \cdots \ a_k),$$

and project a vector p onto $\text{Span}(a_1, a_2, \dots, a_k)$. We note that this is equivalent to considering the QR decomposition of A , where we can express A as

$$Q = (q_1 \ q_2 \ \cdots \ q_k),$$

in which case we want to project the vector p onto $\text{Span}(q_1, q_2, \dots, q_k)$. If it helps, think of $n = 3$ and $k = 2$, in which case we are considering a vector $p \in \mathbb{R}^3$ and projecting it onto the plane that is the set of linear combinations of the columns of the matrix $A = (a_1 \ a_2)$, which is the same as the set of linear combinations of the columns of the matrix $Q = (q_1 \ q_2)$. (This special case is illustrated in Figure 4.11(a); the more general case is illustrated more abstractly in Figure 4.11(b).) In this case, if we start with the vector b , then after the projection we get the vector

$$p = A (A^T A)^{-1} A^T b = Q Q^T b = P b.$$

Here, the projection matrix P is written in one of two ways.

- If we express it in terms of A and the columns of A , then we have

$$P = A (A^T A)^{-1} A^T. \quad (4.15)$$

This expression generalizes Eqn. (4.13), where the $a^T a \in \mathbb{R}$ in the denominator is replaced with $(A^T A)^{-1}$. Since in general matrix multiplication does not commute, we need to be careful where exactly this expression appears (and this is clear from the derivation of this expression which we will not do in detail now).

- If we express it in terms of Q and the columns of Q , then we have

$$P = QQ^T. \quad (4.16)$$

This expression generalizes Eqn. (4.14), where instead of an outer product $uu^T \in \mathbb{R}^{n \times n}$, we have a matrix-matrix product $UU^T \in \mathbb{R}^{n \times n}$. Also, note that Eqn. (4.15) reduces to Eqn. (4.16) in the special case that A is an orthogonal matrix.

We won't derive this expression now, but note how it generalizes Equation (4.10) and Equation (4.12) in a nice way. We'll revisit this later. The reason this is of interest is the following. Recall that, given vectors $v_1, \dots, v_k \in \mathbb{R}^n$, then $V = \text{Span}(v_1, \dots, v_k)$ is a subspace of \mathbb{R}^n ; and if V^\perp is the set of vectors that are perpendicular to V , then V^\perp is also a subspace of \mathbb{R}^n . But, by definition every vector in V is orthogonal to every vector in V^\perp , in a way that directly generalizes how in the two-dimensional case vectors in $\text{Span}(a)$ are orthogonal to vectors orthogonal to $\text{Span}(a)$. In the same way as we used projections to decompose the vector $b = p + e$ into two vectors, one of which was in $\text{Span}(a)$ and one of which was perpendicular to $\text{Span}(a)$, we can use projections to decompose a vector $p \in \mathbb{R}^n$ into two orthogonal parts, one of which is in V and one of which is in V^\perp . This provides a generalization of the Pythagorean Theorem to \mathbb{R}^n . Although difficult to visualize, this is illustrated pictorially in Figure 4.11(b).

(If you didn't get all of that, that is okay, as we'll be spending a lot more time on it.)

4.5 Problems

4.5.1 Pencil-and-paper Problems

1. Let $a = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$ be a unit vector in \mathbb{R}^3 , so that $a_1^2 + a_2^2 + a_3^2 = 1$.
 - Let $v \in \mathbb{R}^3$, and show that the transformation T_a defined by $T_a(v) = v - 2(a^T v)a$ is a linear transformation $\mathbb{R}^3 \rightarrow \mathbb{R}^3$.
 - What is $T_a(a)$? If v is orthogonal to a , then what is $T_a(v)$? Can you give a name to the transformation T_a ?
 - Write out the matrix M corresponding to the transformation T_a (in terms of a_1, a_2, a_3). What can you say about M^2 .
2. (Probably skip.)
 (Hubbard 1.4.21)
 For the two matrices and the vector

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix}, B = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, c = \begin{pmatrix} 1 \\ 3 \end{pmatrix}.$$
 - Compute $\|A\|$, $\|B\|$, and $\|c\|$.
 - Confirm that $\|AB\| \leq \|A\|\|B\|$, $\|Ac\| \leq \|A\|\|c\|$, and $\|Bc\| \leq \|B\|\|c\|$.
3. Let $v \in \mathbb{R}^n$ be a nonzero vector, and denote by $v^\perp \subset \mathbb{R}^n$ be the set of vectors $w \in \mathbb{R}^n$ such that $v \cdot w = 0$.
 - Show that v^\perp is a subspace of \mathbb{R}^n .
 - Given any vector $a \in \mathbb{R}^n$, show that the vector $a - \frac{a \cdot v}{\|v\|^2}v$ is an element of v^\perp .

- (c) Define the projection of a onto v^\perp by the formula

$$P_{v^\perp}(a) = a - \frac{a \cdot v}{|v|^2} v.$$

Show that there is a unique number $t = t(a)$ such that $(a + tv) \in v^\perp$, and show that

$$a + tv = P_{v^\perp}(a)$$

4. (SOURCE: HUBBARD 1.4.25.)
XXX. TO DO.

5. (SOURCE: HUBBARD 1.4.26.)
XXX. TO DO.

6. (SOURCE: HUBBARD 2.4.1.)

Show that the standard basis vectors are linearly independent.

7. The vectors $\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ form an orthogonal basis for \mathbb{R}^2 . Use this to form an orthonormal basis for \mathbb{R}^2 .

8. (a) For what values of α are the three vectors $\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ \alpha \end{pmatrix}$ linearly dependent.

(b) Show that for each such α the three vectors lie in the same two-dimensional plane, and give an equation of the plane.

9. (SOURCE: ADAPTED FROM HUBBARD 2.4.7.)

Let A be an $n \times n$ matrix. Show that A is orthogonal if and only if $A^T A = I$. In this case, show that both its rows as well as its columns form an orthonormal basis for \mathbb{R}^n , i.e., that $AA^T = I$ also.

10. (HW3)

(Hubbard 2.4.2)

(Not in 2016.)

- (a) Do the vectors $u = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, v = \begin{pmatrix} -2 \\ 1 \\ 2 \end{pmatrix}, w = \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}$ form a basis for \mathbb{R}^3 ? If so, is this basis orthogonal?

- (b) Is the vector $\begin{pmatrix} 4 \\ 1 \\ 2 \end{pmatrix}$ in the span of $\begin{pmatrix} 4 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 0 \\ 4 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix}$? Is it in the span of $\begin{pmatrix} 4 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 0 \\ 4 \end{pmatrix}, \begin{pmatrix} 5 \\ 1 \\ 4.5 \end{pmatrix}$?

11. Let $v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $v_2 = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$; let x and y be the coordinates of a vector with respect to the standard/canonical basis $\{e_1, e_2\}$; and let u and v be the coordinates of that vector with respect to the basis $\{v_1, v_2\}$.

- (a) Write the equations to translate from (x, y) to (u, v) and back.

- (b) Use these equations to write the vector that is expressed in the standard basis as $\begin{pmatrix} 3 \\ -5 \end{pmatrix}$ in terms of the basis v_1 and v_2 .

12. (*Probably skip.*)

(Hubbard 2.4.14)

Let $A_t = \begin{pmatrix} 2 & t \\ 0 & 2 \end{pmatrix}$, and denote by $\text{Mat}(2, 2)$ the vector space consisting of 2×2 matrices, with the usual matrix-matrix addition and matrix multiplication by scalar operations.

- (a) Are the elements I , A_t , A_t^2 , and A_t^3 linearly independent in $\text{Mat}(2, 2)$? What is the dimension of the subspace $V_t \subset \text{Mat}(2, 2)$ which they span? (The answer will depend on t .)
- (b) Show that the set W_t of matrices $B \in \text{Mat}(2, 2)$ which satisfy $A_t B = B A_t$ is a subspace of $\text{Mat}(2, 2)$. What is the dimension of this subspace? (Again, the answer will depend on t .)
- (c) Show that $V_t \subset W$. For what values of t are they equal?

13. Let $V = \{x \in \mathbb{R}^3 : x_1 + x_2 + 2x_3 = 0, 2x_1 + x_2 + x_3 = 0\}$. Find a vector $v \in \mathbb{R}^3$ such that $V = \text{Span}(v)$.

14. XXX. THIS PROBLEM MAY HAVE BEEN TOO HARD. I MAY NEED TO DROP OR ADD MORE WARMUP IN THE TEXT.

Let $V = \text{Span} \left(\begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \right)$.

- (a) Find a, b, c such that $V = \{x \in \mathbb{R}^3 : ax_1 + bx_2 + cx_3 = 0\}$.

Let $V_1 = \{x \in \mathbb{R}^3 : x_1 + x_2 + 2x_3 = 0, 2x_1 + ax_2 + bx_3 = 0\}$. Let $V_2 = \text{Span} \left(\begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ c \\ d \end{pmatrix} \right)$.

- (b) For which a, b is V_1 a straight line.
- (c) For which c, d is V_2 a plane.
- (d) For which a, b, c, d is it the case that $V_1 \subset V_2$.
- (e) For which a, b, c, d is it the case that $V_2 \subset V_1$.
- (f) For which a, b, c, d is it the case that $V_2 = V_1$.

In the above, for two sets Ω_1 and Ω_2 , the notation $\Omega_1 \subset \Omega_2$ means that every element of Ω_1 is also an element of Ω_2 . For example, a line in the plane is a subset of the plane, but not vice versa; and a line in the plane is not a subset of the unit ball in the plane, and vice versa.

15. XXX. THIS PROBLEM MAY HAVE BEEN TOO HARD. I MAY NEED TO DROP OR ADD MORE WARMUP IN THE TEXT. ALSO, I MAY WANT TO DROP PART A WHICH SORT OF IMPLIES THE REST.

Let $V = \text{Span}(v_1, v_2, v_3, v_4)$, where

$$v_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 3 \end{pmatrix}, v_2 = \begin{pmatrix} 4 \\ 0 \\ 2 \\ 0 \end{pmatrix}, v_3 = \begin{pmatrix} 0 \\ 3 \\ 1 \\ 2 \end{pmatrix}, v_4 = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix}.$$

- (a) Find all $\alpha_1, \dots, \alpha_4$ such that $V \neq \mathbb{R}^4$.
- (b) Find all $\alpha_1, \dots, \alpha_4$ such that $v_4 \in \text{Span}(v_1, v_2, v_3)$.
- (c) Find all $\alpha_1, \dots, \alpha_4$ such that $v_4 \in \text{Span}(v_1, v_2)$.
- (d) Find all $\alpha_1, \dots, \alpha_4$ such that v_1, v_2, v_4 are linearly dependent.
- (e) Find all $\alpha_1, \dots, \alpha_4$ such that v_1, v_3, v_4 are linearly dependent.
- (f) Find all $\alpha_1, \dots, \alpha_4$ such that v_2, v_3, v_4 are linearly dependent.

16. (a) Compute the QR decomposition of

$$A = \begin{pmatrix} 1 & -1 \\ 1 & 4 \\ 1 & 4 \\ 1 & -1 \end{pmatrix}.$$

Verify that Q is orthogonal, R is upper triangular, and that $A = QR$. Let P_Q and P_{Q^\perp} be the projection matrices onto the subspace defined by the span of Q and the subspace orthogonal to the span of Q , respectively. Consider the vector

$$x = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 2 \end{pmatrix},$$

and compute $P_Q x$ and $P_{Q^\perp} x$. Verify that $\|x\|_2^2 = \|P_Q x\|_2^2 + \|P_{Q^\perp} x\|_2^2$. (This is a generalization of the Pythagorean Theorem to \mathbb{R}^4 .)

- (b) Compute the QR decomposition of

$$A = \begin{pmatrix} -1 & 1 \\ 4 & 1 \\ 4 & 1 \\ -1 & 1 \end{pmatrix},$$

and verify that, although the matrices Q are different, their spans define the same subspace of \mathbb{R}^4 .

17. XXX. ONE DIM PROJECTION.

18. Work through the jupyter notebook `linear-algebra-nb3.ipynb`, which can be downloaded from Piazza.

4.5.2 Implementations and Applications of the Theory

1. Let $n \geq 2$, and, for each n , consider the following three vectors:

- $u = e_1$, i.e., the first standard basis vector.
- $v = e_1 + e_2$, i.e., the sum of the first two standard basis vectors.
- $w = \sum_{i=1}^n e_i$, i.e., the all-ones vector.

Plot, as a function of n , for $2 \leq n \leq 100$, the angle between each of the pairs of these vectors.

2. Consider the 20×6 matrix with rows consisting of sinusoids, exponentials, etc. that we considered in the last homework.

- (a) In python, compute a QR decomposition, call it QR , and verify that $Q^T Q = I$.
- (b) Then move the first column to the last, and compute the QR decomposition, call it US , and verify that $U^T U = I$.
- (c) Confirm that $Q \neq U$.
- (d) Finally, consider some vector, e.g., the vector with $x_i = i$, call it x , and confirm that $QQ^T x = UU^T x$. Show by computing both QQ^T and UU^T that the reason for this is that $QQ^T = UU^T$.
- (e) Compute $x_Q = QQ^T x$ and $x_{Q^\perp} = (I - QQ^T)x$, and verify that $\|x\|_2^2 = \|x_Q\|_2^2 + \|x_{Q^\perp}\|_2^2$. This is a generalization of the Pythagorean Theorem.

3. (PYTHON) Implement the product of three (fairly rectangular, e.g., column-row) matrices a smart way and a dumb way, to show that one is very wasteful. (We may be able to show a bad way to implement projection matrices and/or a good way to project onto one or a small number of dimensions.) E.g., do the previous multiplication for a 100×4 and a 1000×4 matrix of sinusoids/exponentials. XXX. E.G., TAKE THE Q MATRIX FROM QR ON MATRIX OF SINUSOIDS, ETC. GOAL: SHOW THAT THE ORDER MATTERS BEYOND JUST MULTIPLYING VECTORS, AS A FORWARD POINTER TO PROJECTION MATRICES, OR MAYBE DO IT FOR PROJECTION MATRICES, AND AS A FOLLOW UP OT SIMILAR QUESTION FOR VECTORS.

Part III

Basic Probability: A Way to Understand High-Dimensional Spaces

Chapter 5

Introduction to probability

5.1 Overview of the chapter

In this chapter, we are going to switch gears and talk about probability and random variables and related topics. This will seem quite different than the linear algebra we have been discussing (and in certain ways it is), but there are important connections between the two topics (that we will get to in a few chapters).

If statistics is the science of data, then probability is that area of mathematics/statistics that is designed to deal with uncertainty. In particular, these subjects deal with efforts to:

- Understand limitations that arise from measurement inaccuracies.
- Find trends and patterns in noisy data.
- Test hypotheses and models with noisy data.
- Estimate confidence levels for future predictions with noisy data.
- And so on.

There are several ways in which ideas having to do with noise and probability arise in data science.

- **Data have noise.** There is the obvious way, i.e., what is measured is often some combination of signal that is of interest and artifactual noise that is not of interest, and so it is of interest to develop a rule to determine whether what is seen is “real” or “noise”? (We saw an example of this in the first notebook, where we visualized correlated noise and naively it looked like there might be signal there.)
- **As an algorithmic resource.** We can put noise/randomness “inside” an algorithm that computes something we want. For example, we can compute π in many ways, but a particularly simple way is by throwing darts at a two-dimensional box and counting the fraction of times that the inscribed two-dimensional ball is hit. In this case, the noise/randomness is being used as a computational/algorithmic resource. (We saw this in a recent notebook problem, and we will see it again in a notebook problem soon.)
- **Asking many questions of the data.** Let’s say that we have a rule that says what we see is real/correct, and we apply that rule not to 1 but instead to 10 or 100 or 1000 or ... data sets, or variants of the original data set, and on one data set the rule says that what I see is real/correct. Then, the question is: is what we are seeing is real/correct, or could we in some sense be fitting to the noise in all those other data sets? This question arises if we want to test many hypotheses in a given data set and also when we generate many artificial data sets from a given data set to establish

things like confidence intervals. (We mentioned this when we described whether having one of a huge number of studies that tried find a signal report a signal was reliable, and it is related to something more general called multiple hypothesis testing.)

- **The curse/blessing of dimensionality.** We have described linear algebra as a way to generalize many of the ideas/intuitions that we have from the two-dimensional Euclidean plane and three-dimensional Euclidean space to \mathbb{R}^n , for $n \gg 3$. While this is true, we will see soon that many peculiar and counterintuitive things happen in high-dimensional Euclidean spaces. This so-called *curse of dimensionality* is associated with a related phenomenon known as *measure concentration*. This phenomenon can most easily be understood in terms of the vary basic probabilistic process of flipping coins that come up Heads or Tails randomly, but understanding this phenomenon is crucial to understanding when the basic intuitions from low-dimensional Euclidean spaces break down in high-dimensional Euclidean spaces.

Remark. With respect to the algorithmic resource example given above, note that there are better (in the sense of being more accurate for a given amount of effort) ways to compute π , and throwing darts at a two-dimensional box is a particularly simple way since it is very easy to see if we are inside the two-dimensional ball. The point, however, is that there are many cases where the “ball” is hard to compute and/or one derives some rule that is intractable to compute. In these cases, things may not be so simple, and it is very common to put randomness inside the algorithm. It should not be obvious at this point, but the mathematics that says how long one needs to run those algorithms in order to get a reliable answer, e.g., how many darts need to be thrown to get three digits of precision, is the same as the mathematics that answers the above more obviously data-motivated questions about how to deal with noise in data, and this is also the same mathematics that is needed to understand the peculiarities and counterintuitive properties of high-dimensional Euclidean spaces. This will be one of the points that we will cover in the next few weeks.

5.2 Simple models for understanding probability

Recall that algebra and geometry start with axioms that abstract ideas from the world but are then of interest in their own right. So too, probability theory starts with axioms. Initially, these axioms might seem strange, but after thinking about them and using them a bit, it should be clear that they also abstract common and intuitive ideas from the world. Before stating the axioms, let’s start with a few running examples that should provide intuition about where the axioms come from and how they might be used.

5.2.1 Flipping coins, rolling dice, and throwing darts

Flipping coins. We will start with a canonical motivating example, that of flipping a coin. In the simplest version, we have a “fair” coin, and when we flip it, there are two possible outcomes: Heads (H), and Tails (T). By a “fair” coin, we mean that when we flip it, we expect the following.

- If we flip the coin once, then it is equally likely to land on H or T.
- If we flip the coin many times, then it will land on H roughly half of the time, and it will land on T roughly half of the time.

It is important to observe that these two statements are quite distinct. In particular, the first is a statement of the outcome of one coin flip, while the second is a statement about many coin flips. They are, not surprisingly, related, and we will discuss some of those connections, but keep in mind that they are two distinct statements. (BTW, this is related to the frequentist versus Bayesian approaches to statistics, if you are familiar with that, but those are topics we won’t discuss here.)

Rolling dice. A second example is provided by rolling a dice. In the usual setup, the dice is six-sided (where the sides are taken to be in $\{1, 2, 3, 4, 5, 6\}$) and “fair.” In this case when we roll it, we expect the following.

- If we roll it once, then it is equally likely to be a 1 or a 2 or a 3 or a 4 or a 5 or a 6.
- If we roll it many times, then we will obtain any given side roughly $1/6$ of the time.

Again, observe that these two statements are quite distinct.

More generally, the dice could have N sides, where N is any positive integer, e.g., we could choose $N = 365$ if we want to choose a random day of the year (and we ignore leap year issues). Of course, $N = 1$ isn’t so interesting, so we want to choose $N > 1$; but even $N = 2$ is interesting, since $N = 2$ should be “the same” as flipping a coin. Alternatively, rolling an N sided dice is “the same” as flipping a coin that happens to have N sides.

Throwing darts. A third example is provided by throwing darts at a two-dimensional box and asking about the fraction of times that the maximally-inscribed two-dimensional ball is hit. In this case, we expect the following.

- If we throw one dart, we are more likely to hit the ball than miss the ball, in the sense that the chance/probability that we hit the ball should be $\pi/4 \approx 0.785$.
- If we throw many darts, then we will hit the ball roughly $\pi/4$ fraction of the time, and we will miss the ball roughly $1 - \pi/4$ fraction of the time.

Again, these two statements are quite distinct.

Note that this is basically “the same” as flipping an “unfair” coin, in which the chance of flipping a H is $\pi/4$ and the chance of flipping a T is $1 - \pi/4$.

Connections between these running examples. If those statements about the similarities between flipping coins and rolling dice and throwing darts aren’t clear, think about them some more, and keep them in mind as we provide more examples and more formal discussion below, since it is important to have an understanding of those similarities.

Say that we are given a coin, and we want to test the hypothesis that it is fair. How would we do that? Clearly, if we flip the coin once, we will get H or T. WLOG, let’s say it is H. Observe that that doesn’t distinguish between (1) a fair coin and (2) an unfair coin that always comes up H, i.e., H with probability 1 and T with probability 0 and (3) an unfair coin that comes up H with probability $\pi/4$ or $1 - \pi/4$ or something else. So, let’s flip it 10 times. If we get (say) 7 heads, then we can discount the possibility that the coin will come up with H with probability 1, but what can we say about whether the coin is fair?

Our intuition is that, if the coin is fair, then we should get “close to” 5 heads, but is 7 close enough to 5? Moreover, even if the coin is fair, then it is certainly possible that if we flip it 10 times, then it will come up H all 10 times. Most people would say that they think it is pretty unlikely to come up H 10 times, thereby providing evidence that the coin is NOT fair. But what exactly is the chance that if you flip a fair coin 10 times, then it will come up H 10 times? And what exactly is the chance that you will get 100 H in a row if you do 100 flips? Flipping a fair coin 100 times and getting 100 H in a row seems less likely or more improbable than flipping a fair coin 10 times and getting 10 H in a row, but how much more unlikely? Relatedly, what if there is a freshman class of 50 or 500 or 5000 or 5,000,000 people, and they all do 10 (or 100) coin flips, then what is the chance that someone, i.e., at least one person in the class, gets 10 (or 100) H in a row?

There are precise answers to each of these questions, and they are central to a lot of what goes on in probability as well as the connections between probability and high-dimensional linear algebra, and so we will describe basic probability ideas with an eye to answering these questions.

5.2.2 A warm up

Before getting technical, let's start with a warm-up. This will involve thinking in a bit more detail about what we expect from flipping a fair coin once, or a few times, or many times. For example, what exactly do we mean that a fair coin is fair, and what exactly would we expect when we flip it once, or a few times, or many times, if it is fair versus if it is not fair? Similar ideas hold for the other running examples we discussed, and these ideas are the things that we want to formalize in the next few chapters (since similar phenomena hold for \mathbb{R}^n , at least when $n \gg 1$).

In particular:

- The observation that 50% of the flips will be H is “false” for 1 coin flip (either you get one H, i.e., 100% H, or you get zero H, i.e., 0% H and 100% T).
- It is not too unlikely to be “false” for 2 flips (you might get H then T or T then H, in which case you have 50% H and 50% T, but you might also get H then H or T then T, in which case you don't have 50% H).
- Taken literally, it is “false” for 10^6 flips. That is, while it should not be obvious yet, it turns out to be *very* unlikely that you will get *exactly* $(\frac{1}{2}) \times 10^6$ H.

In some sense, however, it feels like this observation should be “more true” for 10^6 flips than for 1 flip or 10 flips. There are various ways to make this precise, and this will be central to what we do.

Before we do many flips of a fair coin, let's start with two flips. In this case, there are four possibilities:

- An H on the first flip, then an H on the second flip.
- An H on the first flip, then a T on the second flip.
- A T on the first flip, then an H on the second flip.
- A T on the first flip, then a T on the second flip.

Note that, depending on what we care about, we may have been careful to specify the order, i.e., what we got on the first flip and what we got on the second flip. Respecting the order, we have the following four possibilities:

- $\{HH, HT, TH, TT\}$.

Given this, there are a number of questions we might want to ask:

- What is the probability of HT?
- What is the probability of “one of each,” i.e., either HT or TH?
- What is the probability of 2 H?
- What is the probability that the first flip is H?
- What is the probability of at least k heads? (This is the same as the previous question if $k = 2$, but it gets more interesting here if we do 3, 4, ... flips, and if we let k be some arbitrary number in $\{0, 1, \dots, n-1, n\}$.)

In some cases, e.g., if we want to know the total number of H, then the order isn't *directly* important. It is, however, often *indirectly* important. In other cases, e.g., if we want to know the chance that the first flip is H, the order is *directly* important. In any case, there is a *very* big difference between worrying about the order and not worrying about the order, and while it may seem pedantic to worry about such a seemingly-minor point, you should pay attention to it.

Probability theory asks questions such as: what is the chance of a given outcome from a set of possibilities? E.g., 2 H, that the sum of two fair 6 sided dice is greater than 10, etc. To be more precise, we need to be mathematically careful about what we mean by "outcome," "possibilities," and related words. Let's do that.

5.3 Foundations of Probability

5.3.1 Sample spaces and events

There are two basic notions of interest here: sample spaces and events. Let's define each in turn.

We'll start with sample spaces, which are basically the "universe" of things under consideration.

Definition 31 *A sample space is the set of all possible outcomes of a particular "experiment" of interest.*

It is important to note that sample spaces aren't "given" to you. Instead, they must be "chosen" by you. Indeed, they are chosen based on what problem you want to solve—in many cases, the hard part is figuring this out. It is also a source of confusion for students who aren't careful about worrying whether the order of things matters.

Given a sample space, which is the set of all possible outcomes, we want to define the set of outcomes in which we are interested. For example, we may be interested only in the case that all coin flips were H, and we may not be interested in sub-cases when that is not true. Or, for some reason, we may be interested in the case when the third flip was H or the first and second flips were a T. To determine the set of outcomes in which we are interested, we have the following definition for events.

Definition 32 *An event is a subset of a sample space, i.e., it is a subset of the set of all possible outcomes of an experiment.*

It is important to keep distinct the ideas of sample spaces and events. Understanding the following question and its answer should help with this.

Question. If a sample space, call it Ω , has size N , then what is the number of possible events? The answer is: 2^N . Here is the proof of that. Given a sample space, an event can be represented by a list of elements of the sample space that are included in the event. Alternatively, it could be represented by a list of elements of the sample space that are not included in the event. Alternatively, it could be represented by a list specifying for every element of the sample space whether or not that element is included in the event, e.g., by a string of length N , where we encode that the element is included in the event with a 1 and that the element is not included in the event with a 0. Here are several possible such strings:

- 100110...01
- 110010...00
- 111111...11

The question then becomes: what are the number of possible such strings? To answer this, let's say that we have a string of length N , each element of which can be a 0 or a 1. To determine how many possible strings

we could have, observe the first element could be a 0 or a 1, the second element could be a 0 or a 1, the third element could be a 0 or a 1, and so on. Thus, the total number of possibilities is

$$2 \times 2 \times 2 \times 2 \times 2 \times 2 \times \dots \times 2 \times 2 = 2^N.$$

In particular,

- if the sample space contains $N = 10$ elements, then the number of possible events is $2^{10} \approx 10^3$;
- if the sample space contains $N = 20$ elements, then the number of possible events is $2^{20} \approx 10^6$;
- if the sample space contains $N = 30$ elements, then the number of possible events is $2^{30} \approx 10^9$;
- if the sample space contains $N = 40$ elements, then the number of possible events is $2^{40} \approx 10^{12}$;
- and so on.

That is, the number of elements in a sample space, i.e., events, is *much* larger than the size of the sample space. Keeping this in mind sometimes helps keep track of what is a sample space and what are events.

Examples (of sample spaces). Here are some examples of sample spaces.

- **Single flip of a coin.** Let's say we flip 1 coin once. Then, possible sample spaces are:

- Get H; get T; get H or T; get neither H or T.

Clearly, some of these sample spaces may be more or less interesting than others.

- **Two flips of a coin.** Let's say we flip 2 coins, each once, or we flip 1 coin twice. Then, here are some possible sample spaces:

- Possible ordered flips. In this case, there are 4 elements of the sample space, which are HH, HT, TH, TT.
- Possible unordered flips. In this case, there are 3 elements of the sample space, which are HH, HT, TT. (Note that this HT is different than the previous HT in the previous bullet. Since we are not worrying about the order here, here it means roll H then T or T then H, whereas before when we considered the order, it meant only H then T.)
- The number of heads. In this case, there are 3 elements of the sample space, which are 0, 1, and 2. (Clearly, this sample space has similarities with the previous sample space.)
- And so on.

In the above, by “ordered,” we mean that if we flip 1 coin twice, then we keep track of the temporal order in which we flipped them, and if we flip 2 coins, each once, then we number the coins arbitrarily and use that numbering as the order.

Question. For each of these three sample spaces, what is the probability of each of the 4/3/3 elements?

- **Two rolls of a dice.** Let's say we roll 2 6-sided dice. Then, here are possible sample spaces:

- Possible ordered rolls:

$$(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), \dots, (6, 3), (6, 4), (6, 5), (6, 6)$$

where, here, e.g., (1, 2) is different than (2, 1), since they are ordered.

- Possible unordered rolls:

$$(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 2), (2, 3), \dots, (4, 6), (5, 5), (5, 6), (6, 6)$$

where, here, e.g., (1, 2) means 1 then 2 or 2 then 1, and thus there is no (2, 1). (That is, this unordered (1, 2) is different than the previous ordered (1, 2).)

- The sum of the two rolls: in this case, the possibilities are:

$$2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$$

- $A = \{\text{The sum of the two rolls } \leq 9\}$, i.e., all pairs of rolls s.t. the sum ≤ 9 . (So, in particular, this includes $(1, 1)$ and it excludes $(6, 6)$.)
- $B = \{\text{The sum of the two rolls } \geq 10\}$. Note that these last two are legitimate, but more complex, but often we are interested in these—as we shall see, such more complex sample spaces are often computed from simpler sample spaces.

- **Multiple throws of a dart.** If we throw darts at a ball in a box, or if we throw a single dart multiple times, then here are possible sample spaces:

- The Sequence of Hit/Miss, indicating whether or not the dart thrown at the box hit the ball.
- The Sum of the total number of Hits.
- And so on.

- **Other examples.** If we choose “random” point on the plane, then here are possible sample spaces:

- Histogram of bins of where the points are.
- And so on.

Examples (of events). Here are some examples of events.

- **Single flip of a coin.** When there is a 2 element sample space—call it H,T—then there are $2^2 = 4$ events. They are: neither, one, other, both, i.e., $\{HH, HT, TH, TT\}$.
- **Two flips of a coin.** If the sample space $\Omega = \{HH, HT, TH, TT\}$, then there are $2^4 = 16$ possible events:

- (HH) : flip H both times
- $(HH), (HT)$: flip H on the first flip
- $(HH), (HT), (TH)$: flip at least one H
- $(HH), (HT), (TT)$: flip H on first flip or not at all
- Etc.
- \emptyset
- $\{(HH), (HT), (TH), (TT)\}$

It may seem strange to include the empty set \emptyset as a possible event. This corresponds to never succeeding, i.e., the successes are the empty set. This suggests connections to naïve set theory, which we will get to below.

- **Other examples.** Similar results hold for the other examples given above.

Remark. Note that for some of these events the order of the elements matters, e.g., the event $((HH), (HT))$, and for some other events, the order does not matter, e.g., $((HH), (HT), (TH))$. Both are events, by this definition, and you can imagine cases where one or the other is of interest. (In fact, we saw such an example above, when we worried about the order.)

Remark. Events that are singletons, i.e., that consist of only one element of the sample space, are sometimes called *basic events* or *elementary events*.

5.3.2 Some basic set theory

Given the above Definition 31 and Definition 32 of sample spaces and events, we can lay out the basics of probability. Although probability will have strong connections with linear algebra, especially as used in data science, some of it—in particular, its axiomatic basis—is more combinatorial. To describe this, we need some set theory notions, and thus we will take a brief detour into set theory. Set theory is a complicated area, especially the foundations of set theory, but we will not get into those issues, and instead we will adopt the so-called naïve set theory perspective.

Here are some basic definitions to start.

Definition 33 *The empty set, denoted \emptyset , is the set with no elements.*

Definition 34 *If A and B are sets, e.g., possible events from a sample space, then the union $A \cup B$ consists of the elements that appear in A or B or both; and the intersection $A \cap B$ consists of the elements that appear in both A and B .*

Examples (of sets). Here are some examples.

- If $A = \{x \in \mathbb{R} : x \geq 0\}$ and $B = \{x \in \mathbb{R} : x \leq 2\}$, then $A \cup B = \mathbb{R}$, and $A \cap B = [0, 2]$.
- If $A = \{1, 2, 3\}$ and $B = \{3, 4, 5\}$, then $A \cup B = \{1, 2, 3, 4, 5\}$ and $A \cap B = \{3\}$.
- If $A = \{1, 2, 3\}$ and $B = \{4, 5, 6\}$, then $A \cup B = \{1, 2, 3, 4, 5, 6\}$ and $A \cap B = \emptyset$.
- If we roll a 6-sided dice twice, and $A = \{\text{Sum of rolls } \geq 5\}$ and $B = \{\text{Sum of rolls } \leq 6\}$, then $A \cup B = \{\text{All pairs of rolls}\}$ and $A \cap B = \{\text{Sum of rolls is 5 or 6}\}$.

We are often interested in various notions of containment, e.g, whether one set is contained in another, whether two sets overlap, etc. Here are some definitions related to that.

Definition 35 *Two sets A and B are disjoint if there are no elements in common between A and B , i.e., if $A \cap B = \emptyset$.*

Definition 36 *Given two sets, A and B , A is a subset of B , denoted $A \subseteq B$ if $x \in A \rightarrow x \in B$, i.e., if A is entirely contained within B . In this case, B is said to be a superset of A , denoted $B \supseteq A$.*

Examples (of containment). Here are some examples.

- If $A = \{2, 3\}$ and $B = \{2, 3, 4, 5\}$, then $A \subset B$.
- If $A = \{1, 2, 3\}$ and $B = \{2, 3, 4, 5\}$, then $A \not\subset B$.
- If Ω is a sample space and A is an event, then $A \subseteq \Omega$.
- $A \subseteq A$.

Note that, according to Definition 36, a set A is a subset of itself, i.e., $A \subseteq A$. (Clearly, $x \in A \rightarrow x \in A$.) If A is a subset of B , and we want to emphasize that $A \neq B$, then we typically write $A \subset B$, in which case we say that A is a proper subset of B .

An important notion is the complement of a set. Here is the definition.

Definition 37 *Given a set S , if $A \subseteq S$, then the complement of A , denoted A^C is the set that contains the elements of S that are not in A .*

Note that complement is defined relative to some parent set.

Examples (of complementation). Here are some examples.

- If $S = \{1, 2, 3, 4, 5\}$ and $A = \{1, 2\}$, then $A^C = \{3, 4, 5\}$.
- If $S = \{1, 2, 3, 4, 5\}$ and $A = \{1, 2, 3, 4, 5\}$, then $A^C = \emptyset$.
- If $S = \{1, 2, 3, 4, 5\}$ and $A = \emptyset$, then $A^C = \{1, 2, 3, 4, 5\}$.

Here are two easy facts to prove:

- $A \cup A^C = S$, and
- $A \cap A^C = \emptyset$.

Definition 38 *If an element e is in A , then $e \in A$; otherwise $e \notin A$.*

Given two sets, sometimes it is necessary to talk about elements in one set that are not in the other set. Sometimes this is called the *relative complement* or *set theoretic difference*.

Definition 39 *Given sets A and B , the relative complement of A in B , also known as the set-theoretic difference of B and A , denoted $B \setminus A$, is the set of elements in B but not in A .*

Examples (of relative complementation). Here are some examples.

- If $A = \{1, 2, 3, 4, 5\}$ and $B = \{4, 5, 6, 7, 8\}$, then $A \setminus B = \{1, 2, 3\}$.
- If $A = \{1, 2, 3, 4, 5\}$ and $B = \{4, 5, 6, 7, 8\}$, then $B \setminus A = \{6, 7, 8\}$.

Pictorially, these ideas are often illustrated with so-called Venn diagrams. See Figure 5.1.

5.3.3 Basics of probability

Basic ideas we want to generalize. Given a sample space Ω , a probability function assigns probabilities to various subsets of the sample space, i.e., to various events. We will use the notation $\mathbf{Pr}[\cdot]$ to denote the probability of something, and we will define it more precisely below. Informally, though, we have the following notions, and we want a probability function to generalize these notions.

- **Single flip of a coin.**

If the sample space is $\Omega = \{H, T\}$, then we might say that $\mathbf{Pr}[H] = \frac{1}{2}$ and $\mathbf{Pr}[T] = \frac{1}{2}$, in which case we have a fair coin. Alternatively, with the same sample space, we might say that $\mathbf{Pr}[H] = \frac{2}{3}$ and $\mathbf{Pr}[T] = \frac{1}{3}$, in which case the coin is biased towards H. Both are legitimate, and which one we use will depend on the application.

Question. When might we want to model something with a biased coin?

- **Single roll of a dice.**

If the sample space is $\{1, 2, 3, 4, 5, 6\}$, i.e., we are rolling a 6-sided dice, then we might have $\mathbf{Pr}[x_i] = \frac{1}{6}, \forall i \in [6]$, in which case we have a fair dice. Alternatively, we could choose $\mathbf{Pr}[x_i = 1] = \frac{1}{2}$, $\mathbf{Pr}[x_i = 2] = \frac{1}{4}$, and $\mathbf{Pr}[x_i] = \frac{1}{16}, \forall i \in \{3, 4, 5, 6\}$. Of course, there are many other possibilities.

- **Single throw of a dart.**

If the sample space Ω is the two-dimensional unit box, and the two sets of interest are the maximally-inscribed ball, call it A , and the complement of the maximally-inscribed ball, A^C , then $\mathbf{Pr}[A] = \pi/4$ and $\mathbf{Pr}[A^C] = 1 - \pi/4$.

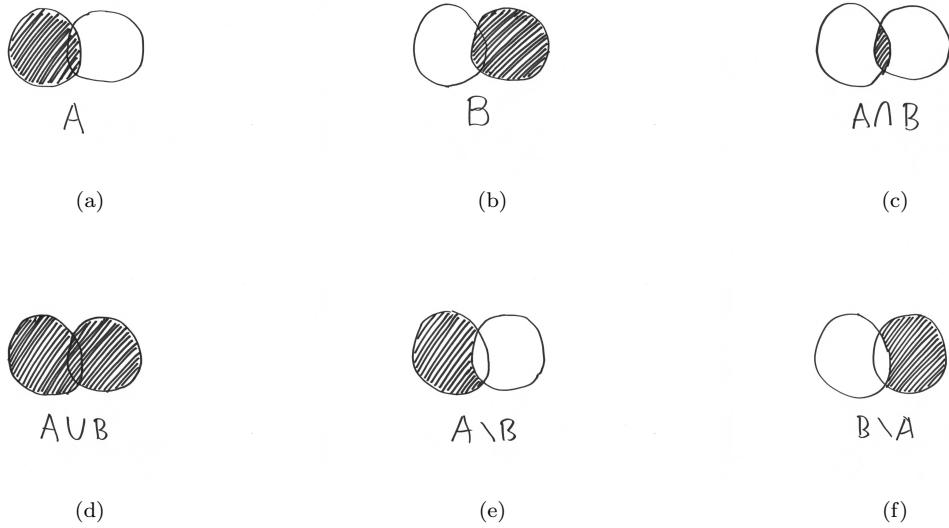


Figure 5.1: Venn diagrams.

A probability function is simply a function that takes as input events, i.e., subsets of a sample space, and returns out as output a number, and that satisfies additional requirements that generalize these intuitive conditions we have been discussing.

More precise definition. Let's be more mathematically precise about what is a probability function. Here is the definition.

Definition 40 Given a sample space Ω , a probability function $\mathbf{Pr}[\cdot]$ is an assignment of a number to various subsets of Ω s.t.

1. $0 \leq \mathbf{Pr}[A] \leq 1$,
2. $\mathbf{Pr}[\Omega] = 1$, and
3. $\mathbf{Pr}[A \cup B] = \mathbf{Pr}[A] + \mathbf{Pr}[B]$, when $A \cap B = \emptyset$.

That is, a probability function is just a function on a sample space, i.e., with domain the possible events of the sample space, that satisfies some additional constraints.

Examples. Here are a few examples.

- If we are flipping a fair coin, the $\mathbf{Pr}[H \cup T] = \mathbf{Pr}[H] + \mathbf{Pr}[T] = \frac{1}{2} + \frac{1}{2} = 1$.
- If we roll a fair dice, then $\mathbf{Pr}[\text{Even}] = \mathbf{Pr}[2 \cup 4 \cup 6] = \mathbf{Pr}[2] + \mathbf{Pr}[4] + \mathbf{Pr}[6] = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$; and $\mathbf{Pr}[\geq 4] = \mathbf{Pr}[4 \cup 5 \cup 6] = \mathbf{Pr}[4] + \mathbf{Pr}[5] + \mathbf{Pr}[6] = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$.
- On the other hand, $\mathbf{Pr}[\text{Even} \cup \geq 4] \neq \mathbf{Pr}[\text{Even}] + \mathbf{Pr}[\geq 4] = \frac{1}{2} + \frac{1}{2} = 1$. Even without figuring out what the probability of this event is, we know it can't be 1, since rolling a 1 or 3 not included, and each of those have a non-0 probability.

Before moving on, let's discuss the requirements in Definition 40 a bit more. In words, what these three requirements say is that (1) the probability of something is a nonnegative number between 0 and 1, (2) the

total probability of anything happening is 1, and (3) if two events do not overlap, then the probability of one or the other is equal to the sum of the probabilities.

- **Zero-one requirement.** If an experiment is carried out a large number of times, with identical conditions each time, the $\Pr[A]$ is the prediction for the *fraction* of experiments with outcome A . In particular, the fraction must be ≥ 0 and ≤ 1 , and the fraction of times that some event happens is 1 since there is always some outcome.

- **Zero-one requirement, cont.** In some sense, the more important thing here is the ≥ 0 requirement, since one can normalize the function to sum to 1.

(That being said, you shouldn't think of the normalization constant as "trivial." In many cases, it is of interest, and it can be difficult to compute.)

- **Total probability requirement.** The total probability of something happening should be one.

(This seems "obvious," but being precise can be helpful later. For example, let's consider the following "promise problem." I promise you that some event, call it A , has happened, and I ask you what is the probability of some other event B . Since you know that A has happened, you can't simply consider $\Pr[B]$; and you can't simply consider $\Pr[A \cap B]$, since that doesn't sum to 1, since we know that A^C which in general has some probability mass, has not happened. Normalizing the right way, so that we satisfy the total probability requirement, will be the right way to deal with these so-called conditional probabilities.)

- **Summation of probability of disjoint events requirement.** This disjoint-additivity requirement captures the idea that the probability of an event is the sum of the probabilities of constituent events. In this light, the condition that $A \cap B = \emptyset$ prevents "overcounting," in the sense that there isn't an event that belongs to both sets. For example, consider the question of $\Pr[\text{Even} \cup \geq 4]$ above. Well, since $\{\text{Even}\} = \{2, 4, 6\}$ and $\{\geq 4\} = \{4, 5, 6\}$, if we were to add the probabilities of these two events, we would be "double counting" 4 and 6. (Dealing with more complex events is a lot of what we do in probability, and so we will get back to it below.)

Question. A question students often ask is: who determines the probabilities?

Answer. The answer is: You do.

Here are several examples to illustrate what we mean by this.

- Flipping coins and rolling dice are of interest more generally than in magic shows and board games since they can be used to model uncertainty very generally, and depending on the application at hand, you should choose the probabilities. For example, if there is a 1% chance that you will get a certain disease, then that corresponds to flipping a coin with the probability of H equal to 1% and the probability of T equal to 99%. The choice of $\frac{1}{2}$ for each of H/T for coins and $\frac{1}{6}$ for dice simply corresponds to a uniform distribution, which is reasonable for coins and dice from experience with them.
- In many cases, we assume the probabilities over a sample space (of singleton events, and thus of non-singleton events) are uniform, basically since we don't know what else to assume. E.g., a fair coin, a fair dice, etc.
- While we can have biased coins, where nonuniform probabilities arise for some reason, and we have seen several examples of that, here is a more important way in which nonuniform probabilities arise.

Let's say we flip a fair coin some number n of times, and we are interested in something like the total number of Hs that arise. We can start with a sample space, call it Ω_{primary} , that is the set of all sequences of flips, and if we don't know anything about the coin, we may assume that all sequences of flips are equally. Then, we can construct a second sample space, call it Ω_{derived} , which is the number of possible Hs. If we do n flips, then this second sample space is the set of integers

$$\{0, 1, 2, \dots, n - 1, n\}.$$

Importantly, even if we assume a probability distribution that is uniform over Ω_{primary} , the probability distribution over Ω_{derived} will not be uniform. In general, it will be *extremely* nonuniform. That is, if you assume that the coin is fair, then you are assuming that the probabilities on Ω_{derived} are nonuniform. Determining the probabilities of events on Ω_{derived} can be quite challenging, but it is very important, and a large part of the art of applying probability in data science is doing this properly. In particular, understanding this connection is important in general, and it is important for understanding how probability is useful for understanding high-dimensional linear algebra in data science. We will get to this in more detail below.

Discrete versus continuous probabilities. Let's briefly discuss an important technical issue that you should know about but that won't matter much for us.

In data science, we typically deal with *discrete probabilities* and *discrete random variables*.

Definition 41 A discrete random variable is one that can assume only a finite or countably infinite number of values.

This should be contrasted with continuous probability and continuous random variables that can assume uncountably infinite number of values in a continuous interval. An example of a continuous random variable is the following: let $\Omega = [0, 1] \subset \mathbb{R}$, and consider the uniform distribution over Ω . In this case, the probability of choosing a number in $[0.3, 0.4]$ should be 0.1, and it is, but what is the probability that should be assigned to choosing a given number, e.g. 0.5 or $\pi/4$? Most work in abstract probability theory involves continuous sample spaces, e.g., \mathbb{R} , and thus it has a very measure theoretic flavor. Basically, the issue is that if you choose a uniform measure over, e.g., $[0, 1]$, then $\Pr[x] = 0$, where x is any number in $[0, 1]$, while $\Pr[(x + \Delta x)] \neq 0$, for $\Delta x > 0$. In this case, a lot of time is spent worrying about measure zero events, various related notions of convergence, etc.

The vast majority of these measure theoretic issues are *not* of any interest in data science, and this we will focus on discrete probability. This is *much* easier, but this will still highlight all of the basic ideas. In particular,

- We won't spend too much time on CLT and LLN (if you know them), but instead we will focus on finite sample versions of them. We will mention them, and we will highlight the relevant parts of them that can be seen in finite sample versions of them.
- We won't spend much time on convergence of random walks, but instead we will discuss the convergence rate, i.e., how long it takes to get to near the stationary distribution. As we will see, this is related to an eigenvalue of a matrix associated with a graph.
- We will focus mostly on discrete distributions. There will be a few cases, however, where we will use important continuous probability distributions, e.g., the normal distribution and the uniform distribution. In these cases, we will adopt a form of naïve continuous probability theory. This basically amounts to considering only the probabilities of not-too-small intervals, in which case we basically think of the continuous space as being binned, in which case we reduce to discrete probability theory.

There are many technical differences between discrete and continuous probability theory, but there are qualitative similarities, and so we will point out where those differences happen, so that you have a rough idea of where the gotchas are.

Some additional immediate results. Given what we have said, and in particular Definition 40, most of discrete probability theory and practice is “just” counting. Students often think that counting is “easy,” presumably since they have been doing it since kindergarten. However:

- Counting some things, e.g., things with combinatorial constraints, can be very hard, i.e., intractable, in realistic settings for even fairly small problem.
- Integration is basically just weighted counting (with limits and infinitesimals, in the continuous case, when done on paper, as opposed to the computer). For example, when we threw darts and tried to hit the ball inside the box, we were counting to approximate an integral.

As such, we will use the basic definitions above and apply them in various ways.

Being precise about some of these set theoretic issues that we have been discussing will permit us to compute a lot of non-obvious things. For example, if our sample space is finite and we have assigned probabilities for all the one-element subsets of the sample space, then we can compute probabilities of all events by invoking the Definition 40—just sum the probabilities that are assigned to elements of that subset. Here are some facts that make it easier to do that.

- $\Pr[\emptyset] = 0$
- $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$ (for any A and B)
- $\Pr[A] \leq \Pr[B]$ if $A \subseteq B$
- $\Pr[B] = \Pr[B \cap A] + \Pr[B \cap A^C]$
- $\Pr[A^C] = 1 - \Pr[A]$

These are easy to prove from the axioms. (You will do it in a homework.) If you think carefully about each one, you should see that they are simply a more abstract statement of things that should be “obvious” to you when, e.g., you roll two dice.

That’s all there is to probability (in the sense that the axioms of algebra or the axioms of geometry are all there is to those areas). The rest is just applying these basic ideas.

As an example of this last claim, one can derive the so-called *union bound*: if we have events A_1, A_2, \dots, A_n , then

$$\Pr[\bigcup_{i=1}^n A_i] \leq \sum_{i=1}^n \Pr[A_i].$$

Observe what this says: the union bound says that the probability that at least one of the events happens is no greater than the sum of the probabilities of the individual events. The union bound can be proven by induction, but it can also be proven directly without induction by using the axioms of probability. (We won’t do it here.) We did want to mention it though since, although it can be loose, basically due to the “double counting” issue, the union bound is often a very simple and useful way to get bounds in probability.

5.3.4 Mass/Volume as an intuition

Here is an intuition that some people find helpful: think of probability as the *mass* of a body in three-dimensional space. To see why this might help, recall that we defined a probability distribution/measure as a completely additive, nonnegative set function. This definition may seem strange and far removed from experience. But in everyday experience, we know functions like this.

- Volume of a region of space: volume is nonnegative, the volume of two disjoint regions is the sum of the volume of each region, and the volume of a body sums to 1—if we choose units/normalization such that it sums to 1 (e.g., you may weigh 143 pounds, but your weight is 1, if the units of measurement is 143 pounds), which we are free to do in this analogy.

- Mass of a body: mass is nonnegative; the mass of two disjoint bodies is the sum of the masses of each body (this is an analogy, so ignore funny stuff like chemical reactions or relativistic quantum mechanics); and the mass of a body sums to 1—again, assuming units s.t. it sums to 1, which we are free to do.

As an aside, we often think of mass as the mass of a body in two-dimensional or three-dimensional space, but we can try to think of mass more abstractly, e.g., as the mass of things that do not live in a Euclidean space, where there is a geometry, but instead where there is just a set relationship between the things. BTW, if those things are connected pair-wise, then we call those connections edges, in which case we have a graph, with vertices (things) and edges (connections between pairs of things). We will get back to this.

BTW, the usual Venn diagram has a bit of both of these perspectives. They are used to illustrate something, e.g., set overlap and set containment, typically on a two-dimensional piece of paper, but the notions of distances, angles, etc. that are typical of two-dimensional Euclidean spaces are not assumed to hold.

If a set S is visualized as a set of points in \mathbb{R}^2 , as with Venn diagrams, but without distances, angles, etc., then the probability of any event/set may be visualized as a mass associated with the set of points.

This suggests two things.

- If the sum isn't 1, then we can normalize to make it a probability distribution. (This is true very generally—recall in the first class when we said that we can divide by the node degree.)
- If we want to do analysis s.t. a given event holds, we can divide by that probability and still have a probability distribution and everything goes through. (This will turn out to be very useful.)

5.4 Conditional Probabilities

5.4.1 Conditional Probability

Conditional probability is a very useful concept that generalizes the following simple idea: depending on what we know, or what we see happen, we may choose different probabilities for various uncertain events. For example, if we flip a fair coin twice, and if we observe that the first flip is Heads, then the probability of getting two Heads (which in that case is 50%) is very different than if we observe that the first flip is a Tails (in which case it is 0%), and both of those are different than the probability of getting two Heads before we observe any flips (in which case it is 25%). That is, conditioned on some piece of information, in this case that the first flip is a Heads, then our probabilities of various events is modified. So, conditional probabilities are themselves just probabilities—that in some sense incorporate other information that is available.

To formalize this, let's say that we have a sample space Ω with a probability function $\Pr[\cdot]$, in accordance with Definition 40. That is, $\Pr[\cdot]$ assigns probability mass to the elements of the power set of Ω . Suppose also that A and B are events, i.e., subsets of Ω , and that we have knowledge that the event represented by B has already occurred. Then, we want to know the probability of the event A . We will call this a *conditional probability of A , given B* , to be defined more precisely below, and we will denote this by $\Pr[A|B]$. Importantly, this is usually different than $\Pr[A]$. That is, the probability of event A is in general different if we condition on event B than if we do not condition on event B .

Example (of conditional probability). Here is an example illustrating the basic idea of probabilities being different, depending on whether or not an event has occurred.

- If we flip 2 coins, and we observe that we have H on the first flip, then we can ask what is $\Pr[0 \text{ heads}]$, $\Pr[1 \text{ head}]$, and $\Pr[2 \text{ heads}]$? Before we saw the first H, then the answers would be $\frac{1}{4}$ (from TT), $\frac{1}{2}$ (from HT and TH), and $\frac{1}{4}$ (from HH), respectively. After we see the first H, then we know that TH and TT don't occur, so the probabilities become 0, $\frac{1}{2}$ (from HT), and $\frac{1}{2}$ (from HH), respectively.

Numbers of H	Prob before 1st flip	Prob after 1st flip
0	1/4	0
1	1/2	1/2
2	1/4	1/2

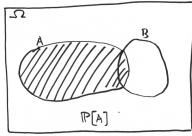
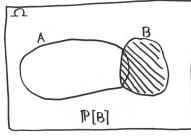
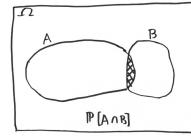
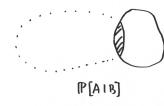
(a) $\Pr[A]$.(b) $\Pr[B]$.(c) $\Pr[A \cap B]$.(d) $\Pr[A|B]$.

Figure 5.2: Illustration of conditional probability.

Here is the more formal definition of conditional probability that generalizes this idea.

Definition 42 *The conditional probability of event A given event B is*

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}.$$

In words, the definition of conditional probability states that the probability of the outcome A , given that the outcome is known to be in B , is equal to the probability of the outcome being in both, divided by the probability of being in B .

A second (equivalent) way to describe this is that we are told that event B has happened, in which case the probability of event A is the fraction of event B 's probability that is accounted for by the elements that are in both A and B .

See Figure 5.2 for an illustration of conditional probability.

Examples (of conditional probability). Here are a few examples illustrating conditional probabilities.

- Let's roll a fair dice, in which case there are six outcomes: $\{1, 2, 3, 4, 5, 6\}$. Then, $\Pr[2] = \frac{1}{6}$ and $\Pr[2|\text{Even}] = \frac{1/6}{1/2} = \frac{1}{3}$. Alternatively, $\Pr[1] = \frac{1}{6}$ and $\Pr[1|\text{Even}] = \frac{0}{1/2} = 0$.
- Alternatively, we could roll an unfair dice. Let's refer to as a so-called “pathological” dice one with the following nonuniform probabilities:

$$\begin{aligned}\Pr[1] &= \frac{1}{21} \\ \Pr[2] &= \frac{2}{21} \\ \Pr[3] &= \frac{3}{21} \\ \Pr[4] &= \frac{4}{21} \\ \Pr[5] &= \frac{5}{21} \\ \Pr[6] &= \frac{6}{21}\end{aligned}$$

That is, this pathological dice is one for which, for $i \in \{1, 2, 3, 4, 5, 6\}$, one rolls an i with probability:

$$\Pr[i] = \frac{i}{\sum_{i=1}^6 i} = \frac{i}{21}.$$

If we let $A = \{1, 2, 4\}$ and $B = \{2, 4, 6\}$, then we have that

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{6/21}{12/21} = \frac{1}{2},$$

while $\Pr[A] = \frac{7}{21} = \frac{1}{3}$. Alternatively

$$\Pr[B|A] = \frac{\Pr[A \cap B]}{\Pr[A]} = \frac{6/21}{7/21} = \frac{6}{7},$$

while $\Pr[B] = \frac{12}{21}$.

Conditional probability is a very useful notion since it can be used in many different ways.

- **Independence of events.** It can be used to define independence of events. See Section 5.4.2.
- **Bayes' Theorem.** It can be used to derive Bayes' Theorem. See Section 5.4.3.
- **Compute complex probabilities.** It can be used to compute more complex probabilities by splitting the sample space Ω into different parts, each of which is easier to work with. See Section 5.4.4.
- And so on.

5.4.2 Independence

A first use of conditional probabilities is to define the notion of independence of two events. Informally, independence means that two events do not depend on each other. To make this precise, in statistics and probability, independence is defined in terms of conditional probabilities.

Here are two identical definitions for when two events are independent.

Definition 43 *An event A is independent of an event B if $\Pr[A|B] = \Pr[A]$.*

In words, this says that event A is independent of event B if the probability of A is the same whether or not we have knowledge of B , i.e., whether or not we condition on the event B .

Definition 44 *An event A is independent of an event B if $\Pr[A \cap B] = \Pr[A]\Pr[B]$.*

In words, this says that event A is independent of event B if the probability of both events happening equals the product of the probability of each event happening.

To show that Definition 43 and Definition 44 are equivalent, recall that $\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}$. Thus, we have that

$$\Pr[A \cap B] = \Pr[B]\Pr[A|B] = \Pr[A]\Pr[B],$$

where the last equality holds if $\Pr[A|B] = \Pr[A]$. This makes it clear that event A is independent of event B iff event B is independent of event A .

Examples (of independence). Here are some examples illustrating independence and non-independence.

- Consider

$$\Omega = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\},$$

i.e., where we do 3 coin flips and the ordering of the three flips matters. Let A be the event that there are 3 Hs, and let B be the event that the first flip is H. Then, $\Pr[A] = \frac{1}{8}$, $\Pr[B] = \frac{1}{2}$, and $\Pr[A \cap B] = \frac{1}{8}$. In addition, $\Pr[A|B] = \frac{1/8}{1/2} = \frac{1}{4}$. Clearly, $\Pr[A] \neq \Pr[A|B]$, and so the events are not independent. This is what intuition would suggest, as we could not obtain 3 H if the first flip was a T.

- Let Ω be the same, let A be the event that there are 2 Hs, and let B be the event that the first flip is H. Then, $\Pr[A] = \frac{3}{8}$, $\Pr[B] = \frac{1}{2}$, and $\Pr[A \cap B] = \frac{1}{4}$. In addition, $\Pr[A|B] = \frac{3/8}{1/2} = \frac{3}{4}$. Since $\Pr[A] \neq \Pr[A|B]$, the events are not independent. Again, this is what intuition would suggest, as whether or not we obtain 2 H depends on whether the first flip is an H.
- Let Ω be the same, and let A be the event that there is a H in the first position and B be the event that there is a H in the third position. Then, $\Pr[A] = \frac{1}{2}$, $\Pr[B] = \frac{1}{2}$, $\Pr[A \cap B] = \frac{1}{4}$. Clearly, $\Pr[A] = \Pr[A|B]$, and so the events are independent. Again, this is what intuition would suggest, as each coin flip was originally independent.
- Let Ω be the same, and let A be the event that there is a H in the first position, and B the event that there is a tails in the first position. Then, $\Pr[A] = \frac{1}{2} = \Pr[B]$, and $\Pr[A \cap B] = 0 \neq \frac{1}{4} = \Pr[A]\Pr[B]$, so the events are not independent.
- Pathological dice. Let $A = \{1, 2, 4\}$ and $B = \{2, 4, 6\}$. Then $\Pr[A] = \frac{7}{21}$ and $\Pr[B] = \{2, 4, 6\}$. But $\Pr[A \cap B] = \frac{6}{21} \neq \Pr[A]\Pr[B]$, so they are not independent. Equivalently, $\Pr[A|B] = \frac{6/21}{12/21} = \frac{1}{2} \neq \Pr[A]$, so they are not independent.

5.4.3 Bayes' Theorem

A second use of conditional probabilities is to derive something known as Bayes' Theorem. Bayes' Theorem is an important and general result that is one of the many uses of conditional probabilities. It is also the source of a lot of confusion, especially in some of the more applied uses of probabilities in data science. The basic question addressed by Bayes' Theorem is: given, $\Pr[A|B]$, what if anything can we say about $\Pr[B|A]$? From Definition 42, these two quantities are clearly different:

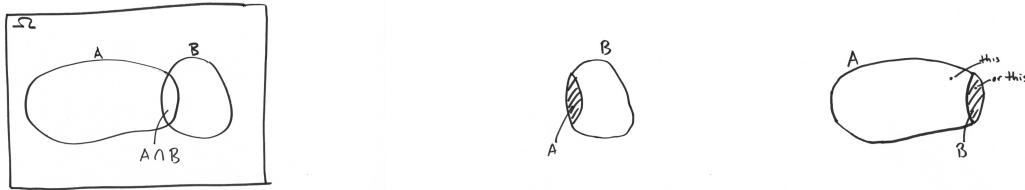
$$\begin{aligned}\Pr[A|B] &= \frac{\Pr[A \cap B]}{\Pr[B]} \\ \Pr[B|A] &= \frac{\Pr[B \cap A]}{\Pr[A]}.\end{aligned}$$

In particular, since $\Pr[A \cap B] = \Pr[B \cap A]$, these two expressions are only the same if $\Pr[A] = \Pr[B]$, which presumably is rarely the case, since in general A and B can be completely different and unrelated events.

Far from being of hypothetical interest, and in spite of the apparent simplicity/obviousness of this observation, there is an enormous amount of confusion about the relationship between these two quantities. Indeed, this is one of the most common source of errors about which you regularly hear on the radio, TV, paper, etc.

This all may sound abstract, so you may ask: Who cares? Let's be specific. Here is an answer.

- Consider the case where you don't feel well and you go to the doctor's office to take a test to determine whether you are sick. Often you know something of the form "If a person is sick, then that person is probably going to have a positive diagnostic test result." So, you take the test and the answer comes back positive (or negative). What does this result say about whether or not you are sick? That is, if



(a) Nothing is known about events A and B . (b) Event B has happened. (c) Event A has happened.

Figure 5.3: Illustration of the probabilities in Bayes' Theorem.

you get a bad (i.e., positive for being sick) test result, then are you sick or probably sick or not sick or probably not sick? Alternatively, if you get a good (i.e., negative for being sick) test result, then are you sick or probably sick or probably not sick or not sick?

- In addition, if the answer is that you are probably but not definitely sick and in need of treatment, which is presumably why you thought you should go to the doctor's office in the first place, or probably but not definitely not sick, then:
 - Should you undergo a course of treatment?
 - What if anything does this say about whether or not this test should be applied to the entire population e.g., as a preventive measure?
 - In both of those cases, how does the answer change if, e.g., “sick” means having an aggressive form of brain cancer, as opposed to a minor flu?

It turns out that the answer to these questions depends very sensitively on $\Pr[A]$ and $\Pr[B]$. These quantities are sometimes known, but often they are not known. In addition, they are typically very different than each other.

Here is the basic setup, illustrated in Figure 5.3. We have a sample space Ω and events A and B , illustrated in Figure 5.3(a). Let's say:

1. we know $\Pr[A|B]$, as illustrated in Figure 5.3(b),
2. we don't know whether event B has happened, but we do observe that a particular event A has happened, as illustrated in Figure 5.3(c), and
3. we want to know the probability of B , i.e., $\Pr[B|A]$, also illustrated in Figure 5.3(c).

The question is: how do we compute $\Pr[B|A]$?

The answer to this question is sufficiently important that it is given the name *Bayes' Theorem*. Here we state it, and we also give its relatively straightforward proof.

Theorem 8 (Bayes' Theorem (usual form))

$$\Pr[B|A] = \frac{\Pr[A|B]\Pr[B]}{\Pr[A]} \tag{5.1}$$

Proof: Recall that

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}$$

and also that

$$\Pr[B|A] = \frac{\Pr[B \cap A]}{\Pr[A]}.$$

Then, by multiplying both sides by the respective denominators, it follows that

$$\Pr[A|B] \Pr[B] = \Pr[A \cap B] = \Pr[B|A] \Pr[A]. \quad (5.2)$$

From this, Equation (5.1) follows. \diamond

(Note that, in this usual formulation of Bayes' Theorem, there are probabilities in the denominator, which means that we are implicitly assuming that they are non-zero. This is typically not a problem in discrete probability, but keep in mind if you study continuous probability that it is a gotcha, and so in that case you have to be careful about it.)

This result is very important, but it can be somewhat counterintuitive. In addition, students often have confusion about where to put the various factors (e.g., should $\Pr[B]$ or $\Pr[A]$ be in the denominator of Equation (5.1)?), if they try to remember Bayes' Theorem in the form given in Equation (5.1). On the other hand, the more symmetric expression given in Equation (5.2) can be easier to remember, and from it the form given in Equation (5.1) immediately follows. I'd suggest remembering/deriving it in that way. To highlight this, I'll restate the symmetric form of Bayes' Theorem here.

Theorem 9 (Bayes' Theorem (symmetric form))

$$\Pr[A|B] \Pr[B] = \Pr[A \cap B] = \Pr[B|A] \Pr[A] \quad (5.3)$$

To use this symmetric form of Bayes' Theorem, just remember our original goal: we want to compute $\Pr[A|B]$ or $\Pr[B|A]$, and so we just need to isolate that factor on one side of the equation.

Examples (of Bayes' Theorem). Here are some examples of this result that are quantitative versions of the healthy/sick example given above.

- **Pathological dice.** Consider the pathological dice, and let $A = \{2, 4\}$ and $B = \{2, 4, 6\}$. Then, from the definition of conditional probability, we have that

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{6/21}{12/21} = 1/2$$

and also that

$$\Pr[B|A] = \frac{\Pr[A \cap B]}{\Pr[A]} = \frac{6/21}{6/21} = 1.$$

(Note that the second expression, i.e., the one for $\Pr[B|A]$ should be obvious, since given that we see a 2 or a 4, then we certainly see a 2 or a 4 or a 6. We included this for completeness and as a sanity check.) Let's check this against Bayes' Theorem:

$$\Pr[A|B] = \Pr[B|A] \frac{\Pr[A]}{\Pr[B]} = 1 \frac{6/21}{12/21} = \frac{1}{2},$$

and also

$$\Pr[B|A] = \Pr[A|B] \frac{\Pr[B]}{\Pr[A]} = \frac{1}{2} \frac{12/21}{6/21} = 1,$$

and clearly both of these check out as expected.

- **Are you really sick?** Let's consider the sickness example again, where the real sickness is a flu infection and the symptom we see is that we have a high temperature. Let's assume the following

probabilities:

$$\begin{aligned}\Pr [\text{Temperature} \mid \text{Infection}] &= 95\% = \frac{19}{20} \\ \Pr [\text{Temperature}] &= 5\% = \frac{1}{20} \\ \Pr [\text{Infection}] &= 1\% = \frac{1}{100}.\end{aligned}$$

In this case, if the doctor gives you a test and the test says that you have a high temperature (i.e., positive for sick), then what is the chance that you have an infection? Here is the answer.

$$\begin{aligned}\Pr [\text{Infection} \mid \text{Temperature}] &= \Pr [\text{Temperature} \mid \text{Infection}] \frac{\Pr [\text{Infection}]}{\Pr [\text{Temperature}]} \\ &= \frac{19}{20} \frac{1}{1} \frac{1}{100} \\ &= \frac{19}{100}.\end{aligned}$$

Only 19%, i.e., meaning more likely than not you are actually healthy. Of course, in this example, the the particular numbers used were just pulled out of the air, and in general we do not know them. The point is that if the treatment is not too bad, e.g., side effects are short-term and mild, as with childhood antibiotics or a decongestant, then you may go ahead with treatment; but if the treatment is severe, e.g., chemotherapy or brain surgery, then it means that you are considering to go ahead with a severe treatment even though most likely, i.e., with 81% chance, you are healthy. What would you do in this case?

- **Are you really sick?** This is similar to the previous example, but with the numbers slightly different, so you can start to see the effect of playing with the numbers. Here, we consider the case of a potentially-mutated protein and whether or not a patient is sick with a disease. Let's say

$$\begin{aligned}\Pr [A] &= \Pr [\text{MutatedProtein}] = 1\% = \frac{1}{100} \\ \Pr [B|A] &= \Pr [\text{Sick} \mid \text{MutatedProtein}] = 20\% = \frac{1}{5} \\ \Pr [B] &= \Pr [\text{Sick}] = 5\% = \frac{1}{20}.\end{aligned}$$

Let's now say that you are sick, e.g., you can't digest food. Then what is the chance that you have a mutated protein? Here is the answer:

$$\Pr [A|B] = \Pr [B|A] \frac{\Pr [A]}{\Pr [B]} = \frac{20}{100} \frac{1}{100} \frac{100}{5} = \frac{4}{100}.$$

That is, not too likely. How confident are you? Should you go ahead with a treatment that targets a protein? Again, it depends on side effects, etc.

- **Are you really guilty?** Here is another example, which looks superficially different since it involves taking blood from a crime scene, but which is really very similar. Here the goal is to determine whether or not is person is guilty of a crime, and there is a piece of evidence that a particular DNA SNP was found at the crime scene which was a match for the suspect. Let's say

$$\begin{aligned}\Pr [A] &= \Pr [\text{particularSNP}] = 10\% = \frac{1}{10} \\ \Pr [B] &= \Pr [\text{commit crime}] = 5\% = \frac{1}{20} \\ \Pr [A|B] &= \Pr [\text{particular SNP} \mid \text{commit crime}] = 20\% = \frac{1}{5}.\end{aligned}$$

Let's also say that an individual is charged with a crime, and that a given SNP is found at the crime scene. The basic question is: is the person (or some other person with this SNP) guilty of the crime? What we can compute is

$$\Pr[B|A] = \Pr[A|B] \frac{\Pr[B]}{\Pr[A]} = \frac{20}{100} \frac{5}{100} \frac{100}{10} = \frac{1}{10} = 10\%.$$

Given this confidence, if you were on the jury, should the person be found guilty?

A few final points.

- Although the examples were quite different, they all boiled down to applying the same Bayes' Theorem since we wanted to "go in the other direction."
- In all these examples, these probabilities are either made up or taken from biased samples, so they are often not known reliably. As we can see by playing with the numbers, if we change any of those probabilities by a factor of 2 or 10, then the answers to our questions change by the corresponding factor, i.e., they are quite sensitive to our ignorance.
- Determining these probabilities reliably is a different issue than the basic results of Bayes' Theorem. So, don't confuse challenges associated with determining them with the basic fact that "going in the other direction" depends sensitively on them and can lead to counterintuitive conclusions in very practical applications. The counterintuitiveness is simply exacerbated by the difficulty in determining them.
- Even if not from genetics, many of these examples have a "genotype/phenotype" form—that is, we observe something like a phenotype (that we can "see"), and we want to infer something like is seen in a phenotype (that we can't "see").

5.4.4 Computing more complex probabilities

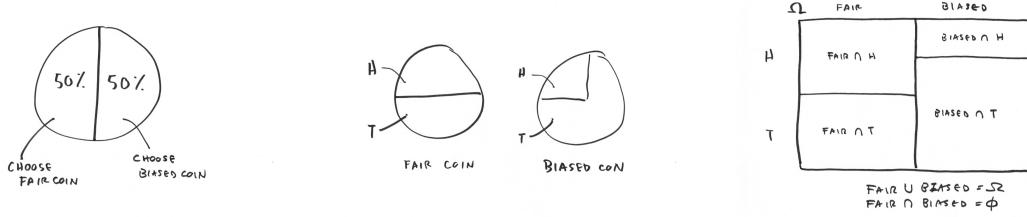
A third use of conditional probabilities involves using them in judicious ways to compute more complex probabilities, i.e., the probabilities for more complex events for which computing the probability is more difficult.

Here is a motivating example. Say that I have two coins: one is fair, meaning 50% H and 50% T; and the other is biased, meaning, e.g., that it is 25% H and 75% T. I don't know which is which, but I choose one and I flip it. One question is: what is the probability that I see H? More generally: how would one compute such a "complex" probability? (By complex, here, I mean it isn't just a simple coin flip, but instead it depends on several steps, perhaps with some constraints, dependencies, etc. Clearly, this is a simple example of a complex probability.)

A general way to compute such complex probabilities is to split the overall problem into smaller or simpler problems and compute the probabilities for each. In this example, let's split it into whether or not we chose the fair or biased coin. We get the following:

$$\begin{aligned}\Pr[H] &= \Pr[\text{choose fair coin}] \Pr[\text{that coin is H}] + \Pr[\text{choose biased coin}] \Pr[\text{that coin H}] \\ &= \frac{1}{2} \frac{1}{2} + \frac{1}{2} \frac{1}{4} \\ &= \frac{1}{4} + \frac{1}{8} \\ &= \frac{3}{8}.\end{aligned}$$

Let's look at what we did. See Figure 5.4 for an illustration. Basically, we split up the problem of computing the overall probability into the sum of disjoint events (i.e., that we choose the fair or biased coin), where



(a) Probability of choosing fair or biased coin.
(b) Probability of choosing heads for each of the fair and biased coin.
(c) Computing complex probabilities with Equation (5.4).

Figure 5.4: Illustration of using conditional probabilities to compute more complex probabilities.

for each of the disjoint events, we know the conditional probability, and then for each part we computed the probabilities separately.

More generally, we get the following.

Theorem 10 Let Ω be the sample space, and let $B_i, i = 1, \dots, N$ be events such that the following two properties hold:

$$\begin{aligned} \text{AllCovering Property: } & \Omega = B_1 \cup B_2 \cup \dots \cup B_N \\ \text{Disjointness Property: } & B_i \cap B_j = \emptyset. \end{aligned}$$

Then, we have the following

$$\begin{aligned} \Pr[A] &= \Pr[A|B_1]\Pr[B_1] + \Pr[A|B_2]\Pr[B_2] + \dots + \Pr[A|B_N]\Pr[B_N] \\ &= \sum_{i=1}^N \Pr[A|B_i]\Pr[B_i]. \end{aligned} \tag{5.4}$$

Proof: Since $\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}$, we have that $\Pr[A \cap B] = \Pr[A|B]\Pr[B]$. The RHS of Equation (5.4) becomes

$$\Pr[A] = \Pr[A \cap B_1] + \Pr[A \cap B_2] + \dots + \Pr[A \cap B_N].$$

The theorem follows since the set of events $\{B_i\}_{i=1}^N$ satisfies the AllCovering Property and the Disjointness Property. \diamond

Remark. Note that we have to assume that $B_i \cap B_j = \emptyset$, since otherwise we will “overcount.” We saw this before.

Remark. In the above example, $N = 2$, and B_1 was the event that we chose a fair coin, and B_2 was the event that we chose a biased coin. Of course, we could have partitioned along the other variable, i.e., we could have chosen B_1 to be that we flipped H and B_2 to be that we choose T—that would have satisfied the two conditions of Theorem 10, but it wouldn’t have made the computations much easier. A lesson is that part of the “art” of applying these conditional probability methods to computing complex probabilities is choosing the right partition. Indeed, that is often the hard part of solving the problem.

5.5 Problems

5.5.1 Pencil-and-paper Problems

1. (Mitzenmacher-Upfal 1.1)

Assume that we have flipped a fair coin ten times. Find the probability of the following events.

- (a) The number of **heads** (H) and the number of **tails** (T) are equal.
 (b) There are more H than T .
 (c) The i^{th} flip and the $(11 - i)^{th}$ flip are the same, for $i = 1, 2, 3, 4, 5$.
 (d) The first four flips are H .
 (e) We flip at least four consecutive H .
2. (Mitzenmacher-Upfal 1.2)
 Assume that we roll two standard six-sided dice. Find the probability of the following events, assuming that the outcomes of the rolls are independent.
- (a) The two dice show the same number.
 (b) The number that appears on the first die is larger than the number on the second die.
 (c) The sum of the dice is even.
 (d) The product of the dice is a perfect square.
3. (Mitzenmacher-Upfal 1.10)
 Assume that I have a fair coin and a two-headed coin. I choose one of the two coins randomly with equal probability and flip it. Given that the flip was H , what is the probability that I flipped the two-headed coin?
4. (Mitzenmacher-Upfal 2.1)
 Suppose that we roll a fair k -sided die with numbers 1 through k on the faces. If X is the number that appears, then what is $\mathbf{E}[X]$?
5. (a) Suppose that we roll a standard fair die 10 times. Let X be the sum of the numbers that appear over the 10 rolls. Use Markov's Inequality to bound $\mathbf{Pr}[|X - 35| \geq 5]$. Simulate this process and comment on how loose or tight is this bound.
 (b) Suppose that we roll a standard fair die 100 times. Let X be the sum of the numbers that appear over the 100 rolls. Use Chebychev's Inequality to bound $\mathbf{Pr}[|X - 350| \geq 50]$. Again, based on your simulation of this process, comment on how loose or tight is this bound.
6. We keep tossing a fair coin until it lands Head up.
- (a) What is the sample space? Also, describe it in words.
 (b) What is the probability of only tossing the coin once?
 (c) What is the probability of only tossing the coin twice?
 (d) What is the probability of only tossing the coin ten times? (Give both an exact expression as well as the numeric value to the nearest tenth of a percent.)
7. Consider the experiment in which one tosses a fair die and counts the dots on the side facing up.
- (a) What is the sample space of this experiment? (Your answer should list the elements (outcomes) in the sample space.)
 (b) What is the event A corresponding to “an even number of dots were counted”? (Your answer should list the elements of the set A , i.e., the outcomes contained in the event A .)
 (c) List the outcomes contained in the event A^C . Also characterize the event A^C in words.
8. A fair die is tossed twice and the numbers n_i of dots facing up are noted, ($i \in \{1, 2\}$).
- (a) What is the sample space?
 (b) What are the elements of the event A corresponding to “total number of dots is odd”?
 (c) What are the elements of the event B corresponding to “both tosses are odd.”
 (d) What are the elements of the event $A \cap B^C$. Also describe this event in words.

- (e) Let C correspond to the event “numbers of dots observed in the two tosses differ by 1”. (Be sure to state precisely how you have defined C .) Find $A \cap C$.
9. Consider the experiment in which one tosses two dice and records the total number of dots facing up.
- What is the sample space?
 - What are the elements of the event A corresponding to “total number of dots counted is even”?
 - Express each of the outcomes in this sample space in terms of the elements of the sample space of the previous problem. (Think of the two throws of the one die in the previous problem as being throws of two separate dice.)
10. Consider tossing a die and recording the number N_1 of dots facing up, then choosing an integer N_2 between 1 and N_1 uniformly at random (meaning that each integer is equally likely to be chosen).
- Find the sample space.
 - Find the set of outcomes in the event “die shows four dots facing up.”
 - Find the set of outcomes in the event “ $N_2 = 3$ ”.
 - Find the set of outcomes in the event “ $N_2 = 6$ ”.
11. (a) If A and B are independent, then show that A^c and B , A and B^c , A^c and B^c are also independent.
(b) Flip 10 biased coins. Their outcomes are independent with the i^{th} coin turning up Heads with probability p_i . Find $\Pr[1\text{st coin H} \cup 2\text{nd coin T} \cup 7\text{th and 9th coin H}]$.
12. (Taubes 2.6.1, S16)
Suppose an experiment has three possible outcomes, labeled 1, 2, and 3. Suppose, in addition, that you do the experiment three successive times.
- Give the sample space for the possible outcomes of the three experiments.
 - Write down the subset of your sample space that correspond to the event that outcome 1 occurs in the second experiment.
 - Write down the subset of your sample space that corresponds to the event that outcome 1 appears in at least one experiment.
 - Write down the subset of your sample space that corresponds to the event that outcome 1 appears at least twice.
 - Under the assumption that each element in your sample space has equal probability, give the probabilities for the events that are described in parts (b), (c) and (d) above.
13. (Taubes 3.7.1, S16)
Consider the sample space
- $$\Omega = \{+, +-, +-, -, ++, -+, -, --, ---\},$$
- which could correspond, e.g., to three measurements of whether your heart rate was above, e.g., 70 bpm. If Ω represents the outcomes for three pulse rate measurements of a given individual, it is perhaps more realistic than assuming a uniform distribution over Ω to take the following nonuniform probability function: The function P assigns probability $1/3$ to $++$ and to $--$ while assigning $1/18$ to each of the remaining elements. Given this probability function:
- Is the event that $+$ appears first independent of the event that $+$ appears last?
 - Is the event that $+$ appears second independent for the event that $+$ appears last?
 - What is the conditional probability that $+$ appears first given that $-$ appears second?
14. A number from the three element set $\{-1, 0, 1\}$ is selected uniformly at random; thus each of -1 , 0 , or 1 has probability $1/3$ of appearing. This operation is repeated twice and so generates an ordered set (i_1, i_2) where i_1 can be any one of -1 , 0 or 1 , and likewise i_2 . Assume that these two selections are done independently so that the event that i_2 has any given value is independent from the value of i_1 .

- (a) Write down the sample space that corresponds to the possible pairs $\{i_1, i_2\}$.
- (b) Let X denote the random variable that assigns $i_1 + i_2$ to any given (i_1, i_2) in the sample space. Write down the probabilities $\Pr[X = x]$ for the various possible values of x .
- (c) Compute the mean and standard deviation of X .
- (d) Let Y denote the random variable that assigns $|i_1| + |i_2|$ to any given (i_1, i_2) . Write down the probabilities $\Pr[Y = y]$ for the various possible values of y
- (e) Compute the mean and standard deviation of Y .
- (f) Compute the correlation matrix for the pair (X, Y) .
- (g) Which pairs of (x, y) with x a possible value for X and y a possible value for Y are such that the event $\{X = x\}$ is independent from the event $\{Y = y\}$?
15. (Taubes 3.7.6)

For certain types of cancer early detection is the key to successful treatment. For prostate cancer, the National Cancer Institute suggests screening of patients using the Serum Prostate-Specific Antigen (PSA) Test. There is, however, controversy due to the lack of evidence showing that early detection of prostate cancer and aggressive treatment of early cancers actually reduces mortality. The treatment is not without risks, and it can often lead to complications like impotence and incontinence.

The main question we would like to address here is:

- If a patient receives a positive test for prostate cancer, what is the probability that he truly has cancer?

Here is some terminology.

- The *sensitivity* of a test is the probability of a positive test when the patient has the disease, i.e., it is the conditional probability of a positive test given that the disease is present.
- The *specificity* of a test is the probability of a negative result when the patient does not have the disease, i.e., it is the conditional probability of a negative test given that the disease is not present.

The standard PSA test to detect early stage cancer has Sensitivity = 0.71 and Specificity = 0.91; and that roughly 0.7% of the male population is diagnosed with prostate cancer each year.

To answer our main question, let A denote the event that a person has a positive test, and let B denote the event that a person has prostate cancer. So, our main question is to determine the probability of B given A , i.e., $\Pr[B|A]$. The data above gives $\Pr[A|B] = 0.71$ and $\Pr[B] = 0.007$ and $\Pr[A^C|B^C] = 0.91$.

$$(a) \text{ Why is } \Pr[B|A] = \frac{0.71 \times 0.007}{\Pr[A]} \approx \frac{0.005}{\Pr[A]}?$$

If you answered this, then the task is to find $\Pr[A]$.

- (b) Why is $\Pr[A] = \Pr[A \cap B] + \Pr[A \cap B^C]$?
- (c) Why is $\Pr[A \cap B^C] = \Pr[B^C] - \Pr[A^C \cap B^C]$?
- (d) Why is $\Pr[B^C] = 1 - \Pr[B]$?

If you answeres these questions, you discovered that $\Pr[A] = \Pr[A \cap B] + 1 - \Pr[B] - \Pr[A^C \cap B^C]$, and thus, using what you just derived, that $\Pr[A] = 0.993 + \Pr[A \cap B] - \Pr[A^C \cap B^C]$.

- (e) Why is $\Pr[A \cap B] = 0.71 \times 0.007 \approx 0.005$?
- (f) Why is $\Pr[A^C \cap B^C] = 0.91 \times 0.993 \approx 0.904$?

If you answered all these sub-questions, then you have found out that $\Pr[A] \approx 0.094$, and thus the answer to out main question is that the probability is ≈ 0.054 .

16. (a) A coin is tossed three times. Consider the following events.

- i. A : Heads on the first toss.
- ii. B : Tails on the second.
- iii. C : Heads on the third toss.
- iv. D : All three outcomes the same (HHH or TTT).
- v. E : Exactly one head turns up.

(b) Which of the following pairs of these events are independent?

- i. A , B
- ii. A , D
- iii. A , E
- iv. D , E

(c) Which of the following triples of these events are independent?

- i. A , B , C
- ii. A , B , D
- iii. C , D , E

17. Work through the jupyter notebook `prob-nb1.ipynb`, which can be downloaded from Piazza.

18. Work through the jupyter notebook `prob-nb2.ipynb`, which can be downloaded from Piazza.

5.5.2 Implementations and Applications of the Theory

1. **Properties of rolling two-sided dice.** We are interested in the number of times a Heads is obtained when a fair two-sided dice is rolled several times. Let's pretend we don't know that this can be computed exactly from the binomial coefficients, and let's compute it by simulating this process many times.

- (a) Consider rolling the two-sided dice $n = 10$ times. Give your best estimate of the probability that you roll: exactly $n/2$ Heads; approximately (e.g., within 3% of) $n/2$ Heads; greater than $3n/4$ Heads; exactly n Heads; and approximately (e.g., within 3% of) n Heads. Do this by simulating the rolling of the dice $n = 10$ times, again and again, until you are confident in the answer.
- (b) Compute the same quantities when the two-sided dice is rolled $n = 10^2$ times.
- (c) Compute the same quantities when the two-sided dice is rolled $n = 10^6$ times.
- (d) Do this for other values of n , including $n = 10^i$, for $i \in \{1, 2, 3, 4, 5, 6\}$, and plot as a function of n the probability that you roll exactly $n/2$ Heads.

Optional: when you can, confirm that your estimates are correct by computing the exact answers via the binomial coefficients.

2. **Histogram count of of rolling multi-sided dice.** We are interested in rolling multi-sided dice.

- (a) Write a function to roll a fair k -sided dice n times, where k and n are inputs to the function, and the function returns an array of length k with the number of times each of the k sides appeared in the n rolls.
- (b) Evaluate this for $k = 2$, and plot the histogram count of Heads versus Tails for different values of n , e.g., for $n = 10$ and $n = 100$, as in the previous question. Do this several times, and comment on how the variability of this histogram from 50-50 percent is related to the width of the histograms in the total number of Heads in the previous question.
- (c) Demonstrate that, as n is increased, the values in the histogram become more and more similar in magnitude, i.e., closer and closer to 50-50 percent.
- (d) Do the same for a fair 6-sided dice.

- (e) Do the same for a fair 100-sided dice.

3. **Generating nonuniform probabilities.** We are interested in generating dice with a nonuniform probability distribution over sides. Here is a simple way to do it. Assume that we have a vector $x \in \mathbb{R}^k$, where every element $x_i \geq 0$, and say that we want to simulate a dice that returns i with a probability $p_i = \frac{x_i}{\sum_{i=1}^k x_i}$, i.e., proportional to x_i . Then, we can do the following.

- Let $y_0 = 0$, and for $j = 1, \dots, k$, let $y_j = \sum_{i=1}^j p_i$.
- Generate a uniform random number in $z \in [0, 1]$.
- For $j = 1, \dots, k$, if $y_{j-1} \leq z \leq y_j$, then the dice returns the j^{th} value.

It can be shown that this process returns a number $i \in \{1, \dots, k\}$, with probability p_i .

- (a) Simulate this process n times for the uniform distribution when $k = 6$.
- (b) Simulate this process n times for the 6-sided pathological dice.
- (c) Simulate this process n times for the 100-sided dice, when $x_i = i$ (i.e., the probability of higher-number faces is higher).
- (d) Simulate this process n times for the 100-sided dice, when $x_i = \sqrt{\left(\frac{1}{2}\right)^2 - \left(i - \frac{1}{2}\right)^2}$ (i.e., the probability of higher-number faces first gets larger and then gets smaller, according to this semi-circular formula).

In each case, construct histograms of the number of times that each value of $i \in \{1, \dots, k\}$ appears; and show that, as the number of trials n is increased, the histogram becomes closer and closer to p_i .

4. **Rejection sampling.** XXX. NOT READY YET, MAYBE LATER. Something about sampling from nonuniform distribution with rejection sampling. Maybe this can also lead to Bayes theorem result. XXX. FLESH THIS OUT, FOR LATER WHEN I DO MCMC.

Chapter 6

Random variables and their properties

6.1 Random Variables

Although we didn't define precisely what exactly is an "experiment," we gave several examples, and we said that a sample space is the set of all possible outcomes of that "experiment" and an event is a subset of that sample space. Ideally, the different outcomes of a given "experiment" have measurable properties that distinguish them. (Otherwise, the experiment isn't going to be too informative.) Here are two examples.

- If the sample space is the ordered list of outcomes of some number of flips of a coin, then a measurable property could be the total number of Hs, in which case we might be interested in the number of sequences of flips that have that given number of Hs. Alternatively, we might be interested in the number of the sequences that have greater than or equal to a given number of Hs.
- If the sample space is the ordered list of outcomes of whether a thrown dart hits the target, then a measurable property could be the total number of H(it)s, in which case we might be interested in the number of the sequences that have that given number of Hs. This is something that can often be related to quantities of interest, e.g., the volume of a ball (an ℓ_2 ball) in a box (an ℓ_∞ ball) in \mathbb{R}^n .

Defining these measurable properties is often just a matter of defining the right probability space, e.g., the sum of H, the number of Hs in the last 10 trials, etc. (BTW, that is why it is worth being a little pedantic about probability spaces and setups related to them.) Informally, different events have different probabilities, and so the values of these measurements may also be different and "random." These measurable properties take an input, and they return an output, and so they are just functions. A "random variable," mathematically, is just a more precise way to study this in more generality.

6.1.1 Random variables are just functions

Let's start with some examples of random variables.

Examples (of random variables).

- **Coin flipping.** Let's flip a fair coin $n = 100$ times. A random variable could be the number of Hs. For example, it could take the value 0 or 100 but likely it is somewhere in between. As stated, $s \in \{T, H\}^{100}$, but of course we can let $s \in \{-1, +1\}^{100}$ or $s \in \{0, 1\}^{100}$. In any case, it is the ordered outcome of 100 flips, encoded by some variables. (We will work with numerical variables since we want to compute sums of things like the number of Hs.)

Then, given s , we can let $X(s)$ be a function of s that is a random variable, i.e., that is a function that takes as input an $s \in \Omega$. Examples of such a random variable include the following:

- $X(s) = \sum_{i=1}^{100} s_i$
- $X(s) = \sum_{i=1}^{100} s_i^2$
- $X(s) = \sum_{i=1}^{100} (s_i - \bar{s})^2$, where $\bar{s} = \frac{1}{100} \sum_{i=1}^{100} s_i$.

These three examples have a very natural interpretation, as we will see soon. Of course, $X(s)$ could be all sorts of other things.

- **Dice rolling.** Let's do k rolls of a 365 sided dice, i.e., $s \in [365]^k$, where we have used the notation $[n] = \{1, \dots, n\}$. Then, here are two random variables.

$$\begin{aligned} - s \rightarrow X(s) &= \begin{cases} 1 & \text{if at least 2 digits are repeated} \\ 0 & \text{otherwise} \end{cases} \\ - s \rightarrow X(s) &= \begin{cases} 1 & \text{if every digit is chosen } \geq \text{once} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Note that you have seen these before (in the main class). The first is the so-called *birthday paradox*. The second is the so-called *coupon collecting problem*.

Given these examples, here is our definition of a random variable.

Definition 45 A random variable *is a function on the sample space Ω , i.e., it is a function that assigns a real number to each and every elements $s \in \Omega$.*

Remark. Clearly, for the example above of the number of Hs in n flips of a coin, the sum, the sum of squares, the sum of squares of the difference with the mean, and many other things, are all examples of random variables. Ditto for dice rolling and many other examples.

Remark. We have said that the range is a real number, i.e., an element of \mathbb{R} , but it could just as well be a set of real numbers, e.g., an element of \mathbb{R}^n , a positive integer, an $m \times n$ matrix, or all sorts of other things.

Remark. Note that there is nothing really random about a random variable—we are given a set that we call a sample space and a function on that sample space. The randomness comes with how we interpret it and use it and the questions we ask, etc.

6.1.2 Two different definitions of the probability of an event

A probability function is a function, and thus it has a domain as well as a range. There are two related ways to compute the probability of an event: one involves summing over elements of the domain; and the other involves summing over elements of the range. The former involves working with the original sample space, and the latter involves defining and working with a new “derived” sample space. There are important connections between the domain and the range, and these connections hold quite generally for many problems of interest in data science, and so we will describe both ways.

Probability via summing over the domain. Recall that an event is a subset of a sample space. An event is either a singleton, or it is not. In the latter case (unless it is the null event consisting of the empty set, in which case it has probability zero), it consists of several singletons, and by the definition of a probability function, the probability of that event is equal to the sum of the probabilities of the singletons that make it up. Saying this somewhat more formally, given the definition of a random variable in Definition 45, here is one way to define the probability of an event (that a random variable takes a given value).

Definition 46 Suppose that Ω is a sample space and $\Pr[\cdot]$ is a probability function on Ω . If X is a random variable and r is a possible value of X , then

$$\Pr[X = r] = \sum_{s \in \Omega : X(s) = r} \Pr[s].$$

That is, in words, $\Pr[X = r]$ is the probability of the subset of Ω such that $s = r$, which is the sum of the probabilities of the elements in $s \in \Omega$ such that $X(s) = r$.

Remark. Again, note that there is nothing really random about this. It is a mathematical idealization that formalizes the idea of randomness and that has an interpretation of randomness when we interpret sample spaces, events, etc., in terms of randomness.

Examples (of probabilities of random variables). Here are several examples.

- **Flipping a fair coin twice.** If we flip a fair coin 2 times, and if we let $X(s) = \sum_{i=1}^2 s_i$, where $s \in \{0, 1\}^2$, then it is easy to see that

$$\Pr[X = 0] = \Pr[X = 2] = \frac{1}{2^2} = \frac{1}{4},$$

and it is also easy to see that

$$\Pr[X = 1] = \frac{1}{2}.$$

- **Flipping a fair coin 3 times.** If we flip a fair coin 3 times, and if we let $X(s) = \sum_{i=1}^3 s_i$, where $s \in \{0, 1\}^3$, then it is easy to see that

$$\Pr[X = 0] = \Pr[X = 3] = \frac{1}{2^3} = \frac{1}{8},$$

and it is also easy to see that

$$\Pr[X = 1] = \Pr[X = 2] = \frac{3}{2^3} = \frac{3}{8}.$$

- **Flipping a fair coin 100 times.** If we flip a fair coin 100 times, and if we let $X(s) = \sum_{i=1}^{100} s_i$, where $s \in \{0, 1\}^{100}$, then it is easy to see that

$$\Pr[X = 0] = \Pr[X = 100] = \frac{1}{2^{100}}, \tag{6.1}$$

and it is only slightly harder to see that

$$\Pr[X = 1] = \Pr[X = 99] = 100 \frac{1}{2^{100}}. \tag{6.2}$$

Quantities such as those given in Equation (6.1) and Equation (6.2) are not so interesting, and they are not hard to compute. We usually are much more interested in things like

$$\Pr[X = 37] \quad \text{or} \quad \Pr[X \geq 37] \quad \text{or} \quad \Pr[44 \leq X \leq 56].$$

Expressions of these forms are *much* more difficult to compute exactly using Definition 46. While these quantities in which we are interested are usually much more difficult to compute exactly, it is often possible to compute approximations to these quantities and/or bounds on these quantities. That is, it is often possible to say that the probability of these events is greater than or less than some other more-easy-to-compute quantity. A large part of probability theory involves computing such bounds.

Let's step back and consider how we might want to compute something like $\Pr[X = 37]$ in the last example. The way we have been describing it so far is the following: consider a sample space Ω , e.g., all possible sequences of ordered sets of 100 flips of coins, and add up the probabilities of the elements $s \in \Omega$ that satisfied a condition, e.g., that $X(s) = 37$. This basically involves summing over the elements in the domain of X . Alternatively, we can also sum over elements in the range of X . (In many cases, this is easier.)

Probability via summing over the range. To compute probabilities by summing over elements in the range of X , note that the assignment

$$r \rightarrow \mathbf{Pr}[X = r]$$

of a nonnegative number to each of the possible values of X defines a probability function on the set of all possible values of X . (By values, this means elements of the range of X .) Observe that this essentially involves defining a new sample space, and choosing a new probability distribution over that new sample space, as follows.

1. Let Ω_X be the new sample space, whose elements consist of the elements of the range of X .
2. Let $\mathbf{Pr}^X[r]$ be the new probability function on Ω_X , defined as follows. If $r \in \Omega_X$, define the probability of the (now singleton) event $X = r$ to be

$$\mathbf{Pr}^X[X = r] = \sum_{s \in \Omega : X(s) = r} \mathbf{Pr}[s].$$

(That is, the subscript on Ω indicates that it is a new derived sample space related to the random variable X ; and the superscript on $\mathbf{Pr}[\cdot]$ indicates that it is a probability distribution for a derived sample space related to the random variable X .)

(BTW, this construction of a new sample space Ω_X and a new probability function $\mathbf{Pr}^X[r]$, based on a random variable of interest, is very common and very useful in data science.)

Given this change, here is another way to define the probability of an event (that a random variable takes a given value).

Definition 47 Suppose that Ω is a sample space and $\mathbf{Pr}[\cdot]$ is a probability function on Ω , and assume that we define Ω_X and $\mathbf{Pr}^X[r]$ as above. If X is a random variable and r is a possible value of X , then

$$\mathbf{Pr}[X = r] = \mathbf{Pr}^X[X = r]$$

Clearly, these two definitions are equivalent. The point of discussing both of them in some detail is two-fold:

- First, it is often easier to work with the sample space defined on the range of a random variable of interest (e.g., there are many fewer values of the range than of the original sample space).
- Second, there are two sample spaces and two probability distributions, which can get confusing, and it is important to keep them distinct in your mind.

Importantly, while this seems abstract, we have actually seen it before. Recall the following examples.

Examples (of defining range-based sample spaces). Here are several examples.

- **Flipping 2 fair coins.** Let's flip 2 fair coins, in which the ordered set of outcomes is $\{HH, HT, TH, TT\}$. If we define a new sample space Ω_X to be equal to the number of heads, corresponding to $X(s) = \sum_{i=1}^2 s_i$ (i.e., that is the set of outcomes we are interested in and we don't really care about the number/order of flips), then the set of possibilities in the range of X is $\{0, 1, 2\}$, and we know the probability of each of those. That is, the new sample space $\Omega_X = \{0, 1, 2\}$, the elements of which have the following probability:

Element	Probability
0	1/4
1	1/2
2	1/4

- **Flipping 3 fair coins.** Let's next flip and sum up 3 fair coins. Here,

$$\begin{aligned}\Omega &= \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\} \\ \Omega_X &= \{0, 1, 2, 3\}.\end{aligned}$$

While the elements of Ω have uniform probabilities, i.e., $\Pr[s] = \frac{1}{|\Omega|} = \frac{1}{8}, \forall s \in \Omega$, the elements of Ω_X have the following non-uniform probabilities.

Element	Probability
0	1/8
1	1/2
2	1/2
3	1/8

- **Flipping n fair coins.** Let's flip a fair coin n times. Then, here are two important questions

Question	Answer
What is the size of Ω ?	2^n
What is the size of Ω_X ?	n

In general, the size of the “primary” sample space Ω is something like 2^n , while the size of the “derived” sample space Ω_X is much MUCH smaller, something like n^2 or n (as it is here).

While we are talking about the previous examples in terms of coin flips, a similar phenomena holds much more generally.

To see this, let's start by looking at dice.

More examples (of defining range-based sample spaces). Here are a few more examples.

- **Rolling 2 dice.** Sum of two dice. Here,

$$\begin{aligned}\Omega &= \{(11), (12), (13), \dots, (21), (22), \dots, (66)\} \\ \Omega_X &= \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}.\end{aligned}$$

Thus, in this case, $|\Omega| = 36$, while $|\Omega_X| = 11$. In addition, the elements of Ω_X have the following probabilities: See Figure 6.1(a).

Element	Probability
2	1/36
3	2/36 = 1/18
4	3/36 = 1/12
5	4/36 = 1/9
6	5/36
7	6/36 = 1/6
8	5/36
9	4/36 = 1/9
10	3/36 = 1/12
11	2/36 = 1/18
12	1/36

- **Rolling 3 dice.** Sum of three dice. Here,

$$\begin{aligned}\Omega &= \{(111), (112), (121), (122), (211), \dots, (665), (666)\} \\ \Omega_X &= \{3, 4, \dots, 18\}.\end{aligned}$$

Thus, in this case, $|\Omega| = 216$, while $|\Omega_X| = 16$. It is more involved to write down the probabilities, but see the figure. See Figure 6.1(b).

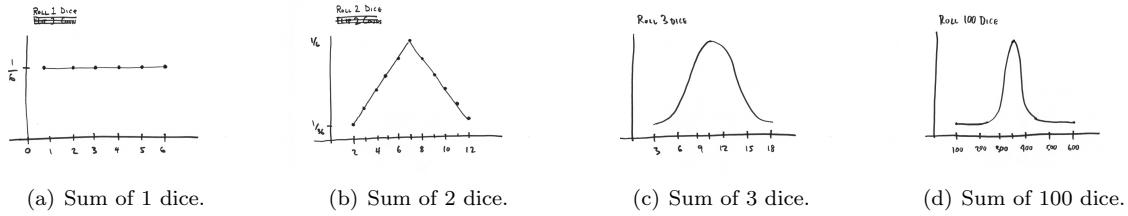


Figure 6.1: Probability of given sum from rolling dice.

- **Rolling n dice.** Sum of $n = 100$ dice. Here we can't even write down Ω in any reasonable amount of space, $\Omega_X = \{100, 101, \dots, 600\}$, and the probabilities are shown in the figure. See Figure 6.1(c).

6.1.3 Measure concentration: a first example

These examples show a general pattern. The size of Ω is MUCH larger than the size of Ω_X . Fortunately, we are typically interested in measuring things only on the much smaller Ω_X . There will be many consequences of this, but among them is that if we want to compute probabilities for Ω_X , then there will be a lot of “averaging” going on.

We saw this with two coins: there are two ways to throw one H and one T, but there is only one way to throw two Hs, and there is only one way to throw two Ts. This isn’t particularly pronounced or extremely interesting for two coins, but the effect is slightly more pronounced for three coins, and the effect becomes MUCH more pronounced and important as the number of coins gets larger. While there is only one way to get 100 Hs and 0 Ts in 100 flips, there are a HUGE number of ways to get 50 Hs and 50 Ts with 100 flips, and only slightly fewer HUGE number of ways to get 49 Hs and 51 Ts with 100 flips. In particular, if we think of it in terms of a mass interpretation, most of the “mass” (or “measure,” as this is the term used in more advanced classes) of the probability distribution becomes “concentrated” on a small number of possible configurations. Here, this means $50 \pm \alpha$ Hs, where $\alpha \ll 50$.

Coin flipping is of interest since it is the simplest way to illustrate this phenomena of *measure concentration*, but this phenomenon is ubiquitous. In particular, this phenomenon is central to what we are doing, and it is central to high-dimensional linear algebra that underlies a lot of the mathematics of data science. (Recall term-document matrices, where each document is described by a vector of terms—the document doesn’t contain much content, if it has only two or three terms in it.) There are a number of ways to quantify the idea of measure concentration and make it formal, e.g., for a fixed number of coin flips versus in the limit as the number of coin flips goes to ∞ , for more complex functions than coin flips, etc. More advanced classes will focus on these.

Here, our goal is to understand the basic intuition in a semi-formal manner. Technicalities aside, this is the basic reason that we think that a fair coin should end up getting ca. 50% H. In addition, while it may not be obvious, it is also the basic reason for the peculiar properties about throwing darts at high-dimensional boxes that we will see in a homework. That is, there are (as we will eventually see in more detail) strong connections between the “concentration” going on with the sum of 100 dice rolls or coin flips and the properties of high-dimensional spaces and high dimensional balls in high-dimensional boxes we saw in the homework. Said another way: to the extent that high-dimensional Euclidean spaces are very different than the more familiar two-dimensional plane or three-dimensional space, thinking about the properties of flipping coins and throwing darts is a good way to understand them.

6.1.4 Other stuff on random variables

Let’s be a little more formal with this.

Definition of random variable and discrete random variable. Here are two definitions.

Definition 48 A random variable X on a sample space Ω is a real-valued function of Ω , i.e., $X : \Omega \rightarrow \mathbb{R}$.

Definition 49 A discrete random variable is a random variable that takes on only a finite or countably infinite number of values.

Example. As an example of this, let's assume that physical time is infinitely divisible to an element of \mathbb{R} , but that we measure time with a clock that is precise only to 0.01 seconds. Then, if runners run a race, the “exact” winning time (which could be any element of \mathbb{R}) is a continuous random variable, but the “official” winning time (as measured by the clock) is a discrete random variable.

Remark. Students sometimes think that finite \rightarrow infinite is a big deal, but mathematically nearly “everything goes through” relatively easily. That is, most of the results and intuitions extend without much change. The really hard part is going countably-infinite \rightarrow uncountably-infinite. In particular, this happens when real numbers are involved, e.g., with continuous probability distributions. In this case, one deals with subsets of the real numbers \mathbb{R} , rather than subsets of the integers \mathbb{Z} . The basic problem is that to choose any given real number the natural uniform probability is basically zero for singletons, and so we need to specify what are the possible sets to consider and to define probabilities for, e.g., small line segments. This leads to sigma fields and all sorts of measure theoretic issues. Fortunately, almost none of that is relevant for data science applications; and in those applications, you can basically get all the basic ideas by considering discrete probability and discrete random variables. We will focus on that in this class. There will be a few exceptions, e.g., we will discuss the so-called normal distribution, and for those we will assume that things are smooth enough and that we only measure thing at a not-too-fine level of granularity. This basically amounts to ignoring all those measure theoretic issues and falling back on the discrete probability ideas.

- For a discrete random variable, the event $X = a$ includes all the basic events of the sample space where $X = a$, i.e.,

$$\{s \in \Omega | X(s) = a\}.$$

In this case,

$$\Pr[X = a] = \sum_{s \in \Omega : X(s) = a} \Pr[s],$$

which is what we saw before.

- For a continuous random variable, things get technically much more complicated, but those technical issues are of less practical relevance, and so we won't cover them. If you are dealing with a continuous random variable, the uniform distribution on the unit interval or a normal/Gaussian distribution, think of it as an n -sided coin, with a uniform or nonuniform probability distribution, and look at binned histograms, and what we are discussing will hold for our purposes.

Independence of random variables. Recall that we have a definition on independence between events. The definition of independence between events can be used to define a notion of independence between two random variables.

Definition 50 Two random variables X and Y are independent iff

$$\Pr[(X = x) \cap (Y = y)] = \Pr[X = x] \cdot \Pr[Y = y]$$

holds for all x and y .

Note that this is a quite strong requirement, since it must hold for all $x, y \in \Omega$.

We will see later what this notion of independence means in terms of matrices later.

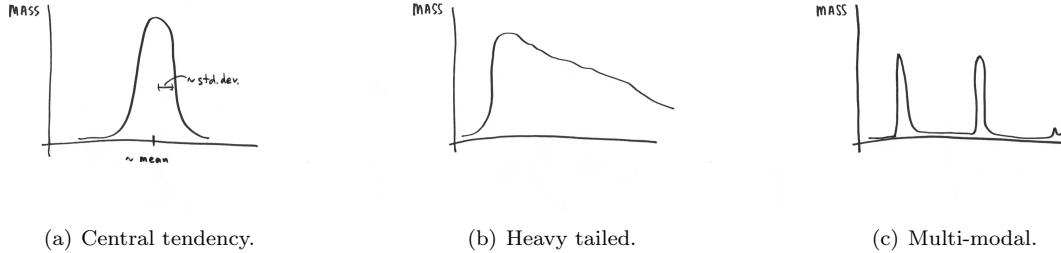


Figure 6.2: Illustration of several types of distributions a random variable can have.

6.2 Moments of random variables

Given a random variable, we typically want to compute things about it, e.g., in order to learn about it and its properties. The moments of a random variable provide a systematic way to compute things about it. What we might want to compute depends very much on the properties of the random variable. Here is a common property that many random variables of interest have.

- The distribution of many random variables looks something like that of Figure 6.2(a), in that there is a sort of central tendency that leads most of the probability mass to be near a single value.

An example of a random variable with this property is the height of human adults: adult humans are about 5 or 6 feet tall; some are 4 feet tall, and some are 7 feet tall; but no adult human is 0.5 feet tall or 50 feet tall. That is, no adult human has a height that is very far from the typical value for everyone else.

For a random variable with this property, statistics called the mean and variance (both are defined below) give a measure of the typical value and the variability of the random variable about that typical value.

Before describing the properties of mean and variance in detail, it is important to keep in mind that these statistics are not always particularly useful. When these statistics are not particularly useful, it is often the case that the distribution of that random variable does not exhibit the sort of central tendency that is illustrated in Figure 6.2(a). Here are two examples of this.

- Some random variables look like Figure 6.2(b), where there is a large amount of probability mass very far from the mean. An example of a random variable with this property is the net wealth of human beings—the mean changes dramatically when a millionaire walks in the room and even more dramatically when a billionaire walks in the room—or the frequency distribution of words in natural languages or many other such things.
- Other random variables look like Figure 6.2(c), where there is no central tendency. An example of a random variable with this property is the number of X chromosomes that human beings have—most people have 1 or 2, but no one has 1.5. It is not particularly meaningful to say that the average human has 1.5 X chromosomes.

For random variables that exhibit these latter properties, one must use caution, and there are other more advanced methods that may be more appropriate. There are many other statistics of interest in statistics (not the two very different uses of the word) and data science for these and other cases. We won't cover them in this class. (We will, however, be interested in methods to help gauge how reliable or unreliable are these measures of central tendency.)

6.2.1 Mean/expectation

Let's start with the mean, also known as the expectation or expected value. The mean of a random variable is a very basic property, of interest both in and of itself and since other things of interest can be computed from it. It can be defined in one of two related ways: one involves a sum over the domain of a random variable; and the other involves a sum over the range of a random variable.

Here is one definition of the mean/expectation of a random variable.

Definition 51 *Let X be a random variable on a sample space Ω that assigns a number to each element of Ω . Then the mean or expectation of X is*

$$\mu[X] = \mathbf{E}[X] = \sum_{s \in \Omega} X(s) \mathbf{Pr}[s],$$

where the summation is taken over all values in the domain of X .

By this definition, the mean is the average over all values in the range of the random variable, where the average is weighted depending on the probability distribution function. In particular, the mean is *not* $\frac{1}{N} \sum_{s \in \Omega} X(s)$, unless all the elements of Ω all have the same probability, i.e., unless $\mathbf{Pr}[s_i] = \frac{1}{N}$, for all i . While this is often but not always true for Ω , it is typically false for Ω_X , and it can also be false for Ω . For example, recall the pathological dice. In general, we will be particularly interested in cases where it is not uniform.

Here is another definition of the mean/expectation of a random variable.

Definition 52 *The mean or expectation of a discrete random variable is*

$$\mu[X] = \mathbf{E}[X] = \sum_i \mathbf{Pr}[X = i] \cdot i$$

where the summation is taken over all values in the range of X .

Remark. Note that the summation in Definition 52 assumes that we have indexed the random variable by the integers (which we can do since it is a discrete random variable). We could, alternatively, only sum values in the range, in which case we would get an expression for the mean of the form

$$\mu[X] = \mathbf{E}[X] = \sum_i \mathbf{Pr}[X = x_i] \cdot x_i.$$

Remark. As an aside, if $X(s)$ equals 1 or 0, depending on whether or not $s \in A$, for some event $A \subseteq \Omega$, i.e., if it is the so-called indicator function of the event A , then $\mathbf{E}[X]$, i.e., the expectation of the random variable X , is equal to $\mathbf{Pr}[A]$, i.e., the probability of the event A .

Remark. As an aside, the expectation is finite if

$$\sum_i |i| \mathbf{Pr}[X = i]$$

converges, and otherwise it is unbounded. This doesn't matter as much if the state space is finite, but it can be an issue if the state space is not finite, e.g., countably infinite (in which case we are still dealing with discrete probabilities). This is similar to the convergence criterion for integrals, familiar from calculus. We will also use this definition with higher moments below. For example, here is an example of a discrete random variable that is unbounded: Take X to be 2^i with probability $\frac{1}{2^i}$, for all $i \in \mathbb{Z}^+$. Then,

$$\mathbf{E}[X] = \sum_{i=1}^{\infty} \frac{1}{2^i} 2^i = \sum_{i=1}^{\infty} 1 = \infty$$

This does not happen if we have a finite range. This example is at the heart of the so-called *gambler's ruin problem*.

Example. If X is the sum of 2 dice, then we can compute the expectation in one of two ways, which basically amounts to summing over the domain or summing over the range.

- Here is the first way to compute the expectation of X .

$$\begin{aligned}
 \mathbf{E}[X] &= \sum_{i=1}^{36} X(i) \mathbf{Pr}[i] \\
 &= \sum_{i=1}^{36} X(i) \frac{1}{36} \\
 &= \frac{1}{36} (2 + 3 + 4 + 5 + 6 + 7 + \\
 &\quad 3 + 4 + 5 + 6 + 7 + 8 + \\
 &\quad 4 + 5 + 6 + 7 + 8 + 9 + \\
 &\quad 5 + 6 + 7 + 8 + 9 + 10 + \\
 &\quad 6 + 7 + 8 + 9 + 10 + 11 + \\
 &\quad 7 + 8 + 9 + 10 + 11 + 12) \\
 &= 7/2
 \end{aligned}$$

- Here is the second way to compute the expectation of X .

$$\begin{aligned}
 \mathbf{E}[X] &= \sum_{i=2}^{12} i \mathbf{Pr}[i] \\
 &= \frac{1}{36} 2 + \frac{2}{36} 3 + \frac{3}{36} 4 + \frac{4}{36} 5 + \frac{5}{36} 6 + \frac{6}{36} 7 + \frac{5}{36} 8 + \frac{4}{36} 9 + \frac{3}{36} 10 + \frac{2}{36} 11 + \frac{1}{36} 12 \\
 &= 7/2
 \end{aligned}$$

Note that the two approaches lead to the same answer, as they should.

Remark. Although we have not introduced the mean this way, it is worth noting that we can relate these ideas to the dot product of two vectors, as we discussed in linear algebra. In particular, if $|\Omega| = N$, then let's view $X(s)$ as a vector in \mathbb{R}^N , and let's view $\frac{1}{N}$ as a vector in \mathbb{R}^N equal to the number $\frac{1}{N}$ times the all-ones vector. Then

$$\mu[X] = \frac{1}{N} \sum_{i=1}^N X(s_i) = \mathbf{dot}\left(X(s), \frac{1}{N}\right). \quad (6.3)$$

Observe that this is of the form of a dot product from linear algebra. This is more than a formal connection—important use of this is made in many areas of data science. More generally,

$$\mu = \mathbf{dot}_{\mathbf{Pr}[S]}(X(s), 1) = \sum_{s \in S} X(s) \mathbf{Pr}[s]. \quad (6.4)$$

Equation (6.3) is a special case of Equation (6.4), with $\mathbf{Pr}[s] = \frac{1}{N}$, for all $s \in \Omega$. This is a weighted generalization of the dot product. (It is also a dot product—we will get back to this later.)

6.2.2 Variance and other moments

Let's proceed beyond the mean to define some other things.

Definition 53 *The k^{th} moment of a random variable X is $\mathbf{E}[X^k]$.*

Moments of a random variable are of interest for many reasons, and they are defined for all $k \in \mathbb{Z}^+$. The most important moments are $k = 1$, i.e., the mean, and $k = 2$, which (almost) gives the notion of variance. Variance is of interest since it is one measure of the variability of a random variable about its mean. Here is the definition of variance.

Definition 54 *The variance of a random variable X is*

$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2.$$

The standard deviation of a random variable X is

$$\sigma[X] = \sqrt{\mathbf{Var}[X]}.$$

To see that the two expressions for the variance are the same, observe

$$\begin{aligned}\mathbf{E}[(X - \mathbf{E}[X])^2] &= \mathbf{E}[X^2 - 2X\mathbf{E}[X] + (\mathbf{E}[X])^2] \\ &= \mathbf{E}[X^2] - 2\mathbf{E}[X\mathbf{E}[X]] + (\mathbf{E}[X])^2 \\ &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2.\end{aligned}$$

Sometimes people try to use the latter expression but can't remember whether it is $\mathbf{E}[X^2] - (\mathbf{E}[X])^2$ or $(\mathbf{E}[X])^2 - \mathbf{E}[X^2]$. If you have a hard time remembering that, then just use the first expression to derive the latter, noting that first expression is squared, so the order doesn't matter and that it is always a non-negative quantity.

Since the standard deviation is the square root of the variance and has the notation σ , sometimes the variance is denoted σ^2 .

Examples (of variance). Here are several examples.

- **Fair dice.** Consider a fair dice, i.e., 6-sided, with $\mathbf{Pr}[i] = \frac{1}{6}, \forall i \in [6]$. Then,

$$\mu = \mathbf{E}[X] = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = \frac{7}{2} = 3.5,$$

and

$$\begin{aligned}\sigma^2 &= \frac{1}{6}(2.5^2 + 1.5^2 + 0.5^2 + 0.5^2 + 1.5^2 + 2.5^2) \\ &= \frac{1}{6}\left(\frac{25}{4} + \frac{9}{4} + \frac{1}{4} + \frac{1}{4} + \frac{9}{4} + \frac{25}{4}\right) \\ &= \frac{170}{6 \cdot 4} \\ &= \frac{35}{12} \\ &\approx 3.\end{aligned}$$

- **Mediumly pathological dice.** Consider a dice with the following probabilities: $\mathbf{Pr}[1] = \mathbf{Pr}[6] = \frac{1}{2}$ and $\mathbf{Pr}[i] = 0$, for $i \in \{2, 3, 4, 5\}$. Then

$$\mu = \mathbf{E}[X] = \frac{1}{2}1 + 0 + \frac{1}{2}6 = \frac{7}{2}$$

and

$$\sigma^2 = \frac{1}{2}2.5^2 + 0 + \frac{1}{2}2.5^2 = \frac{25}{4} \approx 6.$$

The latter has the same mean but larger standard deviation, as you might expect since we don't allow all the middle values.

6.2.3 Two basic properties of expectations of random variables

Given sample spaces, random variables, etc., we have defined the expectation of a random variable. Expectations of random variables have two properties that are very important in and of themselves and that are also very important in that they will permit us to draw a connection between probability theory and linear algebra. These two properties of expectations are so important and their proofs are so simple that we will state and prove them here. To prove them, we will take advantage of writing the expectation of a discrete random variable X as

$$\mathbf{E}[X] = \sum_i i \mathbf{Pr}[X = i],$$

where the summation is over all the values in the range of X .

Theorem 11 (Linearity of Expectation) *For any finite collection of discrete random variables, X_1, X_2, \dots, X_n with finite expectation, it holds that*

$$\mathbf{E} \left[\sum_{i=1}^n X_i \right] = \sum_{i=1}^n \mathbf{E}[X_i].$$

Proof: Consider two random variables X and Y . We have the following.

$$\begin{aligned} \mathbf{E}[X + Y] &= \sum_{i=1}^n \sum_{j=1}^n (i + j) \mathbf{Pr}[(X = i) \cap (Y = j)] \\ &= \sum_{i=1}^n \sum_{j=1}^n i \mathbf{Pr}[(X = i) \cap (Y = j)] + \sum_{i=1}^n \sum_{j=1}^n j \mathbf{Pr}[(X = i) \cap (Y = j)] \\ &= \sum_{i=1}^n i \sum_{j=1}^n \mathbf{Pr}[(X = i) \cap (Y = j)] + \sum_{j=1}^n j \sum_{i=1}^n \mathbf{Pr}[(X = i) \cap (Y = j)] \\ &= \sum_{i=1}^n i \mathbf{Pr}[X = i] + \sum_{j=1}^n j \mathbf{Pr}[Y = j] \tag{6.5} \\ &= \mathbf{E}[X] + \mathbf{E}[Y] \tag{6.6} \end{aligned}$$

Here, Equation (6.5) follows by the Law of Total Probability, and Equation (6.6) follows from the definition of expectation. The theorem follows by applying the argument iteratively. \diamond

Theorem 12 (Scalar Multiplication of Expectation) *For any constant $c \in \mathbb{R}$ and any discrete random variable X with finite expectation, it holds that*

$$\mathbf{E}[cX] = c\mathbf{E}[X].$$

Proof: This is obvious if $c = 0$, so let's assume that $c \neq 0$. In this case, we have the following.

$$\begin{aligned} \mathbf{E}[cX] &= \sum_{j=1}^n j \mathbf{Pr}[cX = j] \\ &= c \sum_{j=1}^n \frac{j}{c} \mathbf{Pr}\left[X = \frac{j}{c}\right] \\ &= c \sum_k k \mathbf{Pr}[X = k] \\ &= c\mathbf{E}[X] \end{aligned}$$

◊

Remark. Both Theorem 11 and Theorem 12 hold for *any* set of random variables. Below we will see some statements that hold only for independent random variables or uncorrelated random variables, etc., and it can be confusing sometimes to know exactly when such statements hold. Here it is simple: these two results hold for *any* (dependent or independent, correlated or uncorrelated, etc.) set of random variables.

Remark. Although we have stated Theorem 11 and Theorem 12 in terms of a random variable with n components, one shouldn't assume that n is finite. The same expressions hold if $n = \infty$, as long as we are working with discrete random variables.

Remark. Theorem 11 is about taking the sum of two (or by extension more than two) random variables, and Theorem 12 is about multiplying a random variable by a scalar. Clearly, Theorem 11 and Theorem 12 can be combined. This should remind you of the two basic operations of linear algebra, i.e., taking linear combinations of two vectors and multiplication of a vector by a scalar. This is more than a coincidence. We will use these two results below to make a strong connection between the probability theory perspective and the linear algebra perspective.

An example of addition and scalar multiplication of expectations. To illustrate these two properties of expectation (linear combination and multiplication by a scalar), consider the following random variables associated with the standard six-sided dice.

Examples (of expectation of sums independent/dependent random variables). Let's consider the standard six-sided dice, and let's roll it twice, and let's let:

$$\begin{aligned} X_i &= \text{the outcome of the } i^{\text{th}} \text{ roll} \\ X &= X_1 + X_2 \\ Y &= X_1 + X_1^2. \end{aligned}$$

- **Independent random variables.** Let's start with $X = X_1 + X_2$. This is the sum of two independent random variables. In this case, we have the following:

$$\begin{aligned} \mathbf{E}[X_i] &= \frac{1}{6} \sum_{i=1}^6 i = \frac{21}{6} = \frac{7}{2} \quad \text{for } i = 1, 2 \\ \mathbf{E}[X] &= \mathbf{E}[X_1] + \mathbf{E}[X_2] = 7. \end{aligned}$$

Here, the expectation of the sum equals the sum of the expectations.

- **Dependent random variables.** Let's next consider $Y = X_1 + X_1^2$. This is the sum of two random variables that are not independent. In this case, we have the following:

$$\begin{aligned} \mathbf{E}[Y] &= \frac{1}{6} \sum_{i=1}^6 Y_i = \frac{1}{6} (2 + 6 + 12 + 20 + 30 + 42) = \frac{1}{6} 112 = \frac{56}{3} \\ \mathbf{E}[X_1] &= \frac{7}{2} \\ \mathbf{E}[X_1^2] &= \frac{1}{6} \sum_{i=1}^6 i^2 = \frac{1}{6} (1 + 4 + 9 + 16 + 25 + 36) = \frac{1}{6} 91 \\ \mathbf{E}[X_1 + X_1^2] &= \frac{1}{6} (21 + 91) = \frac{112}{6} = \frac{56}{3}. \end{aligned}$$

Here, too, the expectation of the sum equals the sum of the expectations.

6.2.4 Conditional expectation

Conditional expectations are just expectations conditioned on a random variable or an event happening (exactly analogously to conditional probabilities). As such, they are expectations, and properties of expectations (such as the linear combination and multiplication by a scalar) also hold for them. In the same way that the notion of conditional probability can be useful (indeed, we gave several examples of this), so too is the notion of conditional expectation. Here, we define this notion and make that explicit.

Here is an example to keep in mind. The mean or expected value of the age of a group of people is different, depending on whether you consider everyone in the country or whether you condition on the event that the people are currently students at a given university or are currently residents of a given retirement home.

We start with the definition for the conditional expectation of a random variable given that another random variable takes a given value.

Definition 55 *Given a random variable Y and a random variable Z , the conditional expectation of Y given Z is*

$$\mathbf{E}[Y|Z = z] = \sum_y y \mathbf{Pr}[Y = y|Z = z].$$

So, the conditional expectation is the weighted sum, where the value is weighted by the conditional probability, rather than the probability.

Given this, and the notion of the conditional probability of a random variable given an event, we can define the conditional expectation of a random variable given an event.

Definition 56 *Given a random variable Y and an event E , the conditional expectation of Y given event E is*

$$\mathbf{E}[Y|E] = \sum_y y \mathbf{Pr}[Y = y|E].$$

That is, this conditional expectation is the weighted sum, where the value is weighted by the conditional probability given the event E .

Examples (of conditional expectation). Let's roll two fair dice. Let X_i be the value on the i^{th} roll, for $i = 1, 2$, and let $X = X_1 + X_2$.

- Then, we can compute the expectation of X , given that we see a particular value on the first roll, i.e., $\mathbf{E}[X|X_1 = x_1^*]$. Before we compute anything, let's ask what our intuition would suggest the results of our calculations should be. Our intuition would suggest that if the first roll is lower than average (respectively, higher than average), then the sum of the two rolls is expected to be lower (respectively, higher) than the unconditional expectation or average of the two rolls. Let's see what we get.

$$\begin{aligned} \mathbf{E}[X] &= 7 \\ \mathbf{E}[X|X_1 = 2] &= \sum_x x \mathbf{Pr}[X = x|X_1 = 2] \\ &= \sum_{i=1}^8 i \frac{1}{6} \\ &= \frac{1}{6}(3 + 4 + 5 + 6 + 7 + 8) = \frac{33}{6} = \frac{11}{2} < 7, \text{ as expected.} \end{aligned}$$

$$\begin{aligned}
\mathbf{E}[X|X_1 = 5] &= \sum_x x \mathbf{Pr}[X = x | X_1 = 5] \\
&= \sum_{i=6}^{11} i \frac{1}{6} \\
&= \frac{1}{6}(6 + 7 + 8 + 9 + 10 + 11) = \frac{51}{6} = \frac{17}{2} > 7, \text{ as expected.}
\end{aligned}$$

Not surprisingly, the conditional expectation is greater than or less than the full expectation, depending on whether or not the first roll was greater than or less than the expected value of one roll.

- We can also “go in the other direction.” That is, we can also compute the expectation of one roll, given that the sum of the two rolls is a given value. Here, our intuition would suggest that if the sum of the two rolls is greater than (respectively, less than) the unconditional expectation of two rolls, then the expected value of the first roll conditioned on this information is greater than (respectively, less than) the expected value of one roll. Let’s see what we get. The basic calculation is similar, but it is slightly different, so we do it here.

$$\begin{aligned}
\mathbf{E}[X_1 | X = 5] &= \sum_{x=1}^4 x \mathbf{Pr}[X_1 = x | X = 5] \\
&= \sum_{x=1}^4 x \frac{\mathbf{Pr}[X_1 = x \cap X = 5]}{\mathbf{Pr}[X = 5]} \\
&= \sum_{x=1}^4 x \frac{1/36}{4/36} \\
&= \frac{1}{4}(1 + 2 + 3 + 4) = \frac{10}{4} = \frac{5}{2} < \frac{7}{2}, \text{ as expected} \\
\mathbf{E}[X_1 | X = 8] &= \sum_{x=2}^7 x \mathbf{Pr}[X_1 = x | X = 8] \\
&= \sum_{x=2}^7 x \frac{\mathbf{Pr}[X_1 = x \cap X = 8]}{\mathbf{Pr}[X = 8]} \\
&= \sum_{x=2}^7 x \frac{1/36}{5/36} \\
&= \frac{1}{5}(2 + 3 + 4 + 5 + 6 + 7) = \frac{27}{5} > \frac{7}{2}, \text{ as expected}
\end{aligned}$$

Again, note that the expectation of X_1 is higher or lower than the unconditional expectation, depending on whether the revealed value of X is higher or lower than its expected value.

Both the addition of expectations and also the multiplication of expectations by a scalar extend to conditional expectations. (This is since conditional expectations are expectations.) We won’t go into the details of proving this, but for completeness we will state these two results here.

Lemma 1 *For any finite collection of discrete random variables X_1, \dots, X_n with finite expectation, and for any random variable Y , we have that*

$$\mathbf{E}\left[\sum_{i=1}^n X_i | Y = y\right] = \sum_{i=1}^n \mathbf{E}[X_i | Y = y].$$

Lemma 2 *For any discrete random variables X and Y with finite expectation and any scalar $c \in \mathbb{R}$, we have that*

$$\mathbf{E}[cX | Y = y] = c\mathbf{E}[X | Y = y].$$

Both of these results have been stated for conditioning on the value of a random variable, but they also extend to conditioning on an event.

6.2.5 How good are your estimates of the mean and the variance?

What we have been discussing holds for discrete random variables, and it can be generalized to continuous random variables. Here are two types of continuous random variables that are important.

- **Uniform distribution.** Here, the random variable is drawn uniformly over the interval $[a, b]$. A special case of this is the uniform distribution over the unit interval, i.e., $[0, 1]$.
- **Normal/Gaussian distribution.** Here, the random variable is drawn according to the normal/Gaussian distribution with mean μ and variance σ^2 (or standard deviation σ). A special case of this is the standard normal/Gaussian distribution with mean 0 and variance 1 (or standard deviation 1).

We will discuss both of these in more detail later. For now, we want to mention an important “use case” of means and variances (or standard deviations) for which there is an important subtlety—very important in practice—that can be quite confusing and thus that we want to explain.

The definitions of mean and standard deviation given in Definitions 51/52 and 54, respectively, are “exact,” in the sense that for a discrete distribution, that is what they are, and they are “easy” to compute. They are often used as a proxy to estimate the mean and standard deviation of something else, however, and this is where the subtlety arises.

Let’s say that we sample from some distribution (such as the uniform or normal/Gaussian or some other distribution such as a biased coin that happens to have a probability of Hs of $\frac{\pi}{4}$ that we don’t know but are trying to determine); and let’s say that we do this some number, say, $n = 100$, of times.

- Let’s call that distribution from which we are sampling the *ground truth distribution*. This is often a continuous distribution, but it could be a discrete distribution, and it has a mean μ_{gt} and standard deviation σ_{gt} .
- Let’s call the sample of size $n = 100$ the *empirical distribution*. This is a discrete distribution, and it has a mean μ_{emp} and standard deviation σ_{emp} .

A common situation is that we want to make some claim about the ground truth distribution. For example, we are often interested in knowing μ_{gt} and/or σ_{gt}^2 , but we can’t compute it since we don’t know the ground truth distribution. (A subtlety here is that it may or may not have a mean or standard deviation, but this is a subtlety which we will ignore, i.e., we will assume that it does and that both are well-defined.) A challenge is that we only have the empirical distribution, i.e., the only insight we have into the ground truth distribution is via the empirical distribution.

The usual solution is the following.

- **Approximation of the mean.** Compute μ_{emp} , and say $\mu_{gt} \approx \mu_{emp}$.
- **Approximation of the standard deviation.** Compute σ_{emp} , and say $\sigma_{gt} \approx \sigma_{emp}$.

Doing this is a reasonable thing to do, and it is intuitive. Moreover, there is good statistical theory (that we will not discuss) to explain why this and variants of this are the right thing to do. A question then arises:

- How good are these approximations? That is, how approximate is the “ \approx ” step?

By this, we mean the generalization of the following. If we flip a fair coin once, then we will not get a good estimate of the probability of Hs; and if we flip a fair coin twice or three times, then there is a relatively

large probability that we will not get a good estimate of the probability of Hs; but if we flip a fair coin many many times, i.e., as the number n of flips gets large, then we expect that the fraction of Hs should be pretty close to 50%. There are two consequences of this (that we will quantify in more detail in the next chapter).

- The variability around 50%, e.g., as measured by the standard deviation of the actual flips, should be pretty small.
- This empirical fraction of Hs could be used as an estimate of the actual probability of Hs if we didn't know that the coin was fair.

So, too, similar claims hold when we are trying to approximate μ_{gt} and σ_{gt} .

In more detail, we can define a random variable X to be the difference between something (e.g., the mean or standard deviation) computed on the ground truth distribution and a related something (e.g., the mean or standard deviation) computed on the empirical distribution; and then we can measure how good is the estimate by computing the standard deviation of that random variable X . Of course, that new random variable X that measures how good are our estimates itself depends on the ground truth distribution, and so we have to estimate it. We won't derive this here, but here is a summary of the bottom line.

- **Quality of estimate of the mean.**

$$\left(\mathbf{E} [(\mu_{emp} - \mu_{gt})^2] \right)^{1/2} \approx \frac{\sigma_{emp}}{\sqrt{n}} = \frac{\text{"function of the second moment"}}{\sqrt{n}}$$

This is called the *standard error of the mean*. It equals the standard deviation divided by \sqrt{n} . A few points:

- The \sqrt{n} arises since if we compute the variance of $\mu_{gt} - \mu_{emp}$, i.e., without the square root, then that would be approximated by the empirical variance σ_{emp}^2 divided by n .
- The quality of the estimate of the mean depends on a higher moment, in this case the second moment.
- Due to the \sqrt{n} in the denominator, this estimate gets better and better as n increases, i.e., our estimate of the mean gets better and better as n increases.

- **Quality of estimate of the standard deviation.**

$$\left(\mathbf{E} [(\sigma_{emp} - \sigma_{gt})^2] \right)^{1/2} \approx \frac{\text{"function of the fourth moment"}}{\sqrt{n}}$$

This is called the *standard error of the standard deviation*. A few points:

- The \sqrt{n} arises again.
- The quality of the estimate of the standard deviation depends on a higher moment, in this case the fourth moment.
- Due to the \sqrt{n} in the denominator, this estimate gets better and better as n increases, i.e., our estimate of the standard deviation also gets better and better as n increases, but since it depends on the fourth moment and not the second moment, the estimate for the standard deviation is worse than for the mean.

Students are often confused since σ_{emp} can be used to estimate two things: the standard deviation σ_{gt} ; and also how good is the estimate of the mean, i.e., the standard error of the mean. (The similar names don't help to unconfuse things.) If you understand what each is measuring, though, you should see the difference and how each should behave, e.g., as n changes. Among other things:

- Your estimate of standard deviation should not go to 0 as n increases.

- Your estimate of the standard error of the mean should go to 0 as n increases.

The derivations are more involved than we want to go into here, but that's a simple way to remember that the \sqrt{n} is in the denominator of the one and not the other. Don't get confused with the two different uses of σ_{emp} .

6.3 More complex combinations: Covariances and correlations

We saw that the expectation operation had “nice” properties that mapped very well to linear algebra (i.e., that the expectation of a sum of random variables equals the sum of the expectations of the random variables, and also that the expectation of the scalar multiple of a random variable equals the same scalar multiple of the expectation of the random variable). One might wonder whether the same holds for the variance. The answer is no, that both are different. Rather than just remembering the expression, it's relatively easy to derive the expression, and we'll get terms that we can then understand in retrospect. Let's do that.

Let's start with the variance of the sum of two arbitrary random variables, call them X and Y . In particular, since X and Y are arbitrary random variables, there is no assumption that they are independent or have any particular relationship to each other. As we have said, that doesn't matter if we are interested in expectations. Let's see what happens if we are interested in variances.

Variances and covariances. Let's first consider the variance of the sum of two random variables.

Lemma 3 *Let X and Y be two random variables. Then:*

$$\mathbf{Var}[X + Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}[X, Y],$$

where $\mathbf{Cov}[X, Y]$ is the covariance between X and Y .

Proof: We apply linearity of expectations and other basic properties as follows:

$$\begin{aligned} \mathbf{Var}[X + Y] &= \mathbf{E}[(X + Y - \mathbf{E}[X + Y])^2] \\ &= \mathbf{E}[(X + Y - \mathbf{E}[X] - \mathbf{E}[Y])^2] \\ &= \mathbf{E}[(X - \mathbf{E}[X] + Y - \mathbf{E}[Y])^2] \\ &= \mathbf{E}[(X - \mathbf{E}[X])^2 + (Y - \mathbf{E}[Y])^2 + 2(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] \\ &= \mathbf{E}[(X - \mathbf{E}[X])^2] + \mathbf{E}[(Y - \mathbf{E}[Y])^2] + 2\mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] \\ &= \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}[X, Y] \end{aligned}$$

◊

Here, we have used a new quantity, the covariance between two random variables X and Y . This is an important notion, so let's define it.

Definition 57 *Given random variables X and Y , the covariance between X and Y is*

$$\mathbf{Cov}[X, Y] = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])].$$

Remark. The covariance is a measure of how two variables covary, and it is the generalization (up to scaling) of the r quantity that you learned about in the main class.

Remark. The covariance looks a little like the variance. The difference is that it is the expectation of the product of two (in general different) mean-centered random variables, rather than the expectation of the product of a mean-centered random variable with itself. Of course, one could consider the case that $Y = X$. If $X = Y$, then $\text{Cov}[X, Y] = \text{Var}[X]$.

Remark. The derivation above in the proof of Lemma 3 can be generalized to the case of more than two random variables, call them X_1, X_2, \dots, X_n . In this case, we get the following expression:

$$\text{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \text{Var}[X_i] + 2 \sum_{i=1}^n \sum_{j>i} \text{Cov}[X_i, X_j].$$

The right hand side of this expression involves numbers $\text{Var}[X_i]$, for $i \in [n]$, and $\text{Cov}[X_i, X_j]$, for $i, j \in [n]$. This gives us $n + \binom{n}{2} = n + \frac{n(n-1)}{2} = \frac{n(n+1)}{2}$ numbers, indexed by i and j . It should not be obvious at this point, but we can very fruitfully organize this information into a matrix—called a *covariance matrix*. For example:

- If we have 2 random variables, X_1 and X_2 , then we get a 2×2 matrix with the elements:

$$\begin{pmatrix} \text{Var}[X_1] & \text{Cov}[X_1, X_2] \\ \text{Cov}[X_2, X_1] & \text{Var}[X_2] \end{pmatrix}.$$

- If we have 3 random variables, X_1 , X_2 , and X_3 , then we get a 3×3 matrix with the elements:

$$\begin{pmatrix} \text{Var}[X_1] & \text{Cov}[X_1, X_2] & \text{Cov}[X_1, X_3] \\ \text{Cov}[X_2, X_1] & \text{Var}[X_2] & \text{Cov}[X_2, X_3] \\ \text{Cov}[X_3, X_1] & \text{Cov}[X_3, X_2] & \text{Var}[X_3] \end{pmatrix}.$$

- If we have 10,000 random variables, $X_1, X_2, \dots, X_{10,000}$, then we get a $10,000 \times 10,000$ matrix.

As with inner/outer products used to construct projections, and many other things we saw before, it is much easier to operate on these matrices directly (by understanding their geometry and performing algebraic operations on the entire matrix) than deal with the all of the elements and all of the subscripts, and linear algebra has tools to do this.

Remark. If, in addition, we scale the elements of the covariance matrix in the right way, then we obtain a *correlation matrix*. We will get back to this soon, and we will spend a lot more time on these matrices in a few classes.

Let's next consider the variance of the product of a random variable with a scalar.

Lemma 4 *For all $c \in \mathbb{R}$ and discrete random variables X , we have that*

$$\text{Var}[cX] = c^2 \text{Var}[X].$$

Proof:

$$\begin{aligned} \text{Var}[cX] &= \mathbf{E}[(cX)^2] - (\mathbf{E}[cX])^2 \\ &= \mathbf{E}[c^2 X^2] - (c\mathbf{E}[X])^2 \\ &= c^2 \mathbf{E}[X^2] - c^2 (\mathbf{E}[X])^2 \\ &= c^2 \text{Var}[X] \end{aligned}$$

◊

Remark. As a special case of this lemma, we could have chosen $c = 2$. Since $2X = X + X$, this would be equivalent to the previous result about the sum of two random variables (that clearly are not independent). In particular:

$$\mathbf{Var}[2X] = \mathbf{Var}[X + X] = \mathbf{Var}[X] + \mathbf{Var}[X] + 2\mathbf{Cov}[X, X] = 4\mathbf{Var}[X].$$

A few points to note about this.

- This is analogous to what we know about vectors. If we take a vector, call it x , and we scale it by a scale $c = 2$, then we get the same result as when we consider the sum $x + x$. For example, $\|x + x\|_2 = \|2x\|_2 = 2\|x\|_2$, and $\|x + x\|_2^2 = \|2x\|_2^2 = 4\|x\|_2^2$.
- Linearity does not hold in general for the variance of a random variable. This should not be surprising, since the variance involves a second moment, and since linearity does hold for the first moment of a (mean-centered) random variable.

While the variance of the sum of random variables does not in general equal the sum of the variances of those random variables, in certain cases, however, these two quantities are equal. Those are cases where the cross terms, i.e., the covariances, are all equal to zero.

This has a connection with the notion of independence that we saw before. Recall that the random variables X and Y are independent iff $\forall x, y$ we have that

$$\mathbf{Pr}[(X = x) \cap (Y = y)] = \mathbf{Pr}[X = x] \times \mathbf{Pr}[Y = y].$$

It is important to keep distinct the ideas of two random variables being independent versus two random variables being uncorrelated (having correlations/covariances equal to zero). The two concepts are different, and students are sometimes confused.

If random variables are independent—which is a strong condition and thus which we can't in general assume to be true—then things simplify. For example, for independent random variables, we have the following result.

Theorem 13 *If X and Y are independent random variables, then*

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y].$$

Proof:

$$\begin{aligned}\mathbf{E}[XY] &= \sum_i \sum_j (i \cdot j) \mathbf{Pr}[(X = i) \cap (Y = j)] \\ &= \sum_i \sum_j i \cdot j \mathbf{Pr}[X = i] \mathbf{Pr}[Y = j] \\ &= \sum_i i \mathbf{Pr}[X = i] \cdot \sum_j j \mathbf{Pr}[Y = j] \\ &= \mathbf{E}[X] \cdot \mathbf{E}[Y]\end{aligned}$$

(Note that, in the summations in the first line, the sum over i (respectively, j) are sums over all values in the range of X (respectively, Y).)

◊

So, if two random variables are independent, which is a rather strong condition, then things are “nice” in this sense (and other senses). If the random variables are not independent, then things are not nice (in the sense that the expectation of the product equals the product of the expectations).

Examples (of expectation of products of independent/dependent random variables). Consider flipping two coins that are either independent or dependent.

- **Independent random variables.** In the first case, let Y and Z be two coin flips, in which the two coins are independent and flipped separately, and each with

$$\begin{cases} 1 & \text{H} \\ 0 & \text{T} \end{cases} .$$

If they are independent coin flips, then $\mathbf{E}[Y] = \mathbf{E}[Z] = \frac{1}{2}$. In addition, the random variable YZ takes the values

$$\begin{cases} 1 & \text{with probability } 1/4 \\ 0 & \text{otherwise} \end{cases} ,$$

in which case $\mathbf{E}[Y \cdot Z] = \frac{1}{4}1 + \frac{3}{4}0 = \frac{1}{4} = \mathbf{E}[Y]\mathbf{E}[Z]$. Thus, for these independent random variables, the product of the expectations equals the expectation of the product.

- **Dependent random variables.** In the second case, let Y and Z be two coin flips, in which the two coins are “tied together,” in the sense that both are H or both are T. Then we still have that $\mathbf{E}[Y] = \mathbf{E}[Z] = \frac{1}{2}$, but the random variable YZ now takes the values

$$\begin{cases} 1 & \text{with probability } 1/2 \\ 0 & \text{otherwise} \end{cases} ,$$

and so, $\mathbf{E}[Y \cdot Z] = \frac{1}{2} \neq \mathbf{E}[Y]\mathbf{E}[Z]$. Thus, for these dependent random variables, the product of the expectations does not equal the expectation of the product.

If two random variables are independent, then the covariance is zero and the expression for the sum of the variances simplifies to be the variance of the sums. For the case of two random variables, this gives the following.

Lemma 5 *Let X and Y be two independent random variables. Then $\mathbf{Cov}[X, Y] = 0$ and*

$$\mathbf{Var}[X + Y] = \mathbf{Var}[X] + \mathbf{Var}[Y].$$

Proof:

$$\begin{aligned} \mathbf{Cov}[X, Y] &= \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] \\ &= \mathbf{E}[X - \mathbf{E}[X]]\mathbf{E}[Y - \mathbf{E}[Y]] \\ &= 0 \end{aligned}$$

The first line is from the definition of covariance; the second line follows from Theorem 13; and the third line follows since $\mathbf{E}[X - \mathbf{E}[X]] = \mathbf{E}[X] - \mathbf{E}[\mathbf{E}[X]] = 0$.

◊

Remark. This result also generalizes to the product of more than two random variables, in which case all of the cross terms that were put into the covariance matrix equal zero, i.e., the covariance matrix is a diagonal matrix, with the variances along the diagonal. We will get back to covariance matrices and what more they can tell us about information in data later.

Remark. The converse of Lemma 5 is *not* true. Just because two random variables are uncorrelated, one cannot conclude that they are independent.

Correlations. To wrap up, let’s make explicit the following. In the same way as one may have a random variable X , and one may want to work with a mean-centered version of it, call it

$$X' = X - \mathbf{E}[X],$$

or with a mean-centered and variance-normalized version of it, call it

$$X'' = \frac{X'}{\sigma[X']} = \frac{X - \mathbf{E}[X]}{\sigma[X]},$$

so too it is often convenient to work with a variance-normalized version of the covariance (which, note, is already mean-centered). This is a sufficiently-important notion that it has its own name.

Definition 58 Given random variables X and Y , the correlation between X and Y is

$$\text{Corr}[X, Y] = \frac{\text{Cov}[X, Y]}{\sigma[X]\sigma[Y]}.$$

Remark. As with the covariance, if we consider the case of more than two random variables, call them X_1, X_2, \dots, X_n , then we obtain $n + \binom{n}{2} = \frac{n(n+1)}{2}$ numbers, indexed by i and j . This too can very fruitfully be organized into a matrix—called a *correlation matrix*.

- If we have 2 random variables, X_1 and X_2 , then we get a 2×2 matrix with the elements:

$$\begin{pmatrix} 1 & \text{Corr}[X_1, X_2] \\ \text{Corr}[X_2, X_1] & 1 \end{pmatrix}.$$

- If we have 3 random variables, X_1 , X_2 , and X_3 , then we get a 3×3 matrix with the elements:

$$\begin{pmatrix} 1 & \text{Corr}[X_1, X_2] & \text{Corr}[X_1, X_3] \\ \text{Corr}[X_2, X_1] & 1 & \text{Corr}[X_2, X_3] \\ \text{Corr}[X_3, X_1] & \text{Corr}[X_3, X_2] & 1 \end{pmatrix}.$$

- And so on.

In these cases, the diagonal elements of this matrix equal 1 since the correlation of a vector with itself equals 1, and the off-diagonal elements are numbers between -1 and 1 which measure the correlation between X_i and X_j .

Generalization to random vectors in \mathbb{R}^n . If $X \in \mathbb{R}^n$ is a column vector that is a random vector, i.e., where each element is a random scalar, then we can define the mean elementwise, and we can define the variance (of the random variable X that is a vector), which itself is a variance-covariance matrix, as:

$$\mathbf{Var}[X] = \mathbf{E}\left[(X - \mathbf{E}[X])(X - \mathbf{E}[X])^T\right].$$

This quantity, being an outer product, is an $n \times n$ matrix, and it is a matrix with particularly “nice” properties that we will discuss in some detail later. (Thus, here, the square in the definition of the variance of a scalar-valued random variable generalizes to an outer product in the case of the variance of a vector-valued random variable. This is a definition, but it captures a lot more information than the inner product, which one might assume provides an alternative definition, basically since the inner product gives just the sum of variances, and no covariances.) Among other things, one can show that

$$\begin{aligned} \mathbf{Var}[AX] &= A\mathbf{Var}[A]^T \\ \mathbf{E}[X^TAX] &= \mathbf{Tr}(AV) + \mu^T A \mu, \end{aligned}$$

where $\mu = \mathbf{E}[X] \in \mathbb{R}^n$ and $V = \mathbf{Var}[X] \in \mathbb{R}^{n \times n}$. (This generalization and these quantities will be particularly important when we discuss PCA.)

6.4 Problems

6.4.1 Pencil-and-paper Problems

1. Suppose that we independently roll two standard six-sided dice. Let X_1 be the number that shows on the first die, X_2 be the number that shows on the second die, and let $X = X_1 + X_2$.
 - (a) What is $\mathbf{E}[X|X_1 \text{ is even}]$?
 - (b) What is $\mathbf{E}[X|X_1 = X_2]$?
 - (c) What is $\mathbf{E}[X_1|X = 9]$?
 - (d) What is $\mathbf{E}[X_1 - X_2|X = k]$, for k in the range $[2, 12]$?
2. Let X be a number chosen uniformly at random from $[1, \dots, n]$. What is $\mathbf{E}[X]$ and $\mathbf{Var}[X]$?
3. Let X be a number chosen uniformly at random from $[-k, \dots, k]$. What is $\mathbf{E}[X]$ and $\mathbf{Var}[X]$?
4. Prove that, for any real number c and any discrete random variable X , that $\mathbf{Var}[cX] = c^2\mathbf{Var}[X]$.
5. Given any two random variables X and Y , by the linearity of expectations we have that $\mathbf{E}[X - Y] = \mathbf{E}[X] - \mathbf{E}[Y]$. Prove that, when X and Y are independent, $\mathbf{Var}[X - Y] = \mathbf{Var}[X] + \mathbf{Var}[Y]$.
6. Work through the jupyter notebook `prob-nb3.ipynb`, which can be downloaded from Piazza.

6.4.2 Implementations and Applications of the Theory

1. **Variances of distributions versus variances of estimates of distributions.** This problem will illustrate that we can compute the variance of a probability distribution as well as the variance of our estimate of, e.g., the mean of that distribution, and the two behave differently as the number of trials increases. Consider the following probability distributions:

P3 : X_i is drawn from a uniform distribution on the unit interval,

First, by hand compute the mean μ and the standard deviation σ , but for the following let's pretend that we don't know μ and σ .

- (a) Compute an estimate of μ by drawing $n = 10$ samples from *P3*. Do this for $n = 10^i$, for $i = 1, \dots, 6$, and plot the results as a function of i .
- (b) Compute an estimate of σ for $n = 10^i$, for $i = 1, \dots, 6$, using the formula for sample variance, and plot the results as a function of i .
- (c) Comment on what these two quantities converge to as n increases, and comment on how quickly these two quantities converge to that value.
- (d) Now consider your first estimate for the mean, i.e., the $n = 10$ numbers you used to estimate the mean, and view that as a set of ten numbers. Compute the mean and variance of that set of numbers. Do the same for $n = 10^i$, for $i = 1, \dots, 6$, and plot the results as a function of i . What do these two quantities converge to as n is increased?

Do the same for the following probability distributions:

P4 : X_i is drawn from a normal/Gaussian distribution with mean μ and variance σ^2 ,

where again let's pretend that we don't know μ and σ .

2. **Finding low probability events.** Consider a Bernoulli random variable with a given value of the parameter p . (This is just the example of throwing darts, where the probability of a hit is p .) We are interested in roughly how many trials are required before we see a hit, or relatedly what is the chance that we see no hits, as a p gets small.

- (a) Let $p = 10^{-1}$, and consider $n = 1$ trial. By simulating this process many times, compute an estimate of the fraction of times you see a hit in $n = 1$ trial.
 - (b) Do the same for $n = 2^i$ trials, where $i = 1, \dots, 20$. Plot the results as a function of i .
 - (c) Do the same for $p = 10^{-2}$.
 - (d) Do the same for $p = 10^{-3}$.
 - (e) Do the same for $p = 10^{-4}$.
 - (f) What do you observe about the shape of these curves as n changes for a given value of p ?
 - (g) Can you suggest a rule of thumb for the number n of trials you need to do to observe of to fail to observe a hit for a Bernoulli random variable with a given value of the parameter p ?
3. **Empirically estimating small means.** Consider a Bernoulli random variable, where the probability of success is $p = \frac{1}{2^i}$, where $i = 1, 2, 3, \dots, 10$, and pretend that you don't know the value of p . For each value of p , by simulating this random variable, we want to get an estimate of p .
- (a) For $i = 1$, by simulating this $k = 100$ times, get an estimate of p . Repeat this process $m = 100$ times to get m estimates of p , and plot this histogram of estimates.
 - (b) Do the same for $i = 2, \dots, 10$.
- Observe that the width of the histogram is in some sense a measure of your confidence in the estimate. How does this width change with i ? What does this say about your confidence in your estimate when i is small, i.e., p is large? When i is large, i.e., p is small? XXX. GOOD BUT TOUCH UP A BIT.
4. **Computing correlation and covariance matrices.** XXX. I THINK THIS IS TOO HARD AT THIS POINT. Let A be the 20×4 matrix consisting of noisy sinusoids, exponentials, etc., that we considered before. Compute the correlation matrix. Compute the covariance matrix in two ways: first, by mean centering and computing the correlation matrix of the mean-centered matrix; and second, by . XXX. CAN I DO: CORRELATION OF ORTHOGONAL MATRICES (TAKEN FROM QR FROM BEFORE), AND COVARIANCE OF ORTHOGONAL MATRICES (TAKEN FROM QR FROM BEFORE).
5. XXX. PLOT HISTOGRAM FOR DIFF DSTBNS OF RV, E.G. CONTINUOUS VERSUS DISCRETE 0,1 VERSUS GAUSSIAN WITH SAME/DIFFERENT MEAN/VARIANCE. I MAY WANT TO DO THIS HERE AS A FORWARD POINTER TO MORE COMPLICATED DISTRIBUTIONS IN THE NEXT CHAPTER.

Chapter 7

Quantifying variability and concentrating measure

7.1 Large, small, and typical variability

Recap. Recall the definition of the variance and standard deviation of a random variable X .

Definition 59 *The variance of a random variable X is*

$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2,$$

and the standard deviation of a random variable X is

$$\sigma(X) = \sqrt{\mathbf{Var}[X]}.$$

Remark. The reason we mean-center a random variable X before taking its second moment in the definition of the variance is that we want to use the variance as a measure of the variation or variability of a random variable about its mean. If we do not subtract off the mean, then the second moment itself has information about both the magnitude of the random variable (e.g., as captured by the mean) as well as the variability of the random variable about the mean (e.g., as captured by the variance). Indeed, writing the second moment as

$$\mathbf{E}[X^2] = (\mathbf{E}[X])^2 + \mathbf{Var}[X]$$

makes this dual dependence explicit.

Remark. The variance of a random variable is a perfectly legitimate measure of the variability of that random variable about its mean, but it has a drawback when one wants to compare the variability to the mean, which is often the case. The reason for the drawback is that the variance has different “units” than the mean. For example, if we are measuring height, and the mean height is a number that is measured in *feet* or *meters*, i.e., with units of length, then the variance of the height is given by a number with units *feet*² or *meters*², i.e., an area. Since we often want to compare the variability to the mean, it helps to put them both in the same units, and for this we typically take the square root of the variance. This is the standard deviation, and it gives a different but related measure of the variability of the mean, expressed in the same units as the mean.

Small, typical, and large. We gave several numerical examples of the variance of random variables, and we showed that the variance is larger when the probability distribution is in some sense more “spread out.”

We did not, however, get a sense of when the value obtained was large or small, e.g., compared with what is possible or compared with what it could be, or compared with what is typical.

Let's now ask how large or small the variance can be, e.g., in terms of the number of data points or other parameters of the problem, as well as what is a "typical" value for the variance, e.g., if we have noise in the data. Here are a few examples illustrating this.

Example (of a small variance). Let $X = c$ be a constant, i.e., X takes the value c for every event in the sample space Ω . In this case, we have the following:

$$\begin{aligned}\mathbf{E}[X] &= c \\ \mathbf{Var}[X] &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = 0 \\ \sigma(X) &= 0.\end{aligned}$$

See Figure 7.1(a) for an illustration. Since the variance must be non-negative, this is a "small" variance/stddev.

Example (of large variance). Let $\mu \in \mathbb{R}$ be a number, and let X be a random variable defined as follows:

$$X = \begin{cases} k\mu & \text{with probability } \frac{1}{k} \\ 0 & \text{otherwise.} \end{cases}$$

Then, we have the following:

$$\begin{aligned}\mathbf{E}[X] &= \frac{1}{k}k\mu + \left(1 - \frac{1}{k}\right)0 = \mu \\ \mathbf{Var}[X] &= \frac{1}{k}k^2(\mathbf{E}[X])^2 + \left(1 - \frac{1}{k}\right)0 - (\mathbf{E}[X])^2 = (k-1)(\mathbf{E}[X])^2 = (k-1)\mu \\ \sigma[X] &= \sqrt{k-1}\mathbf{E}[X] = \sqrt{k-1}\mu.\end{aligned}$$

See Figure 7.1(c) for an illustration. This is a "large" variance/stddev. Indeed, we will see later that this is as large as the variance/stddev can be, when you have a given value for the mean.

Example (of a "typical" variance). The well-known Gaussian/normal distribution provides a "size scale" for typical/random variation. See Equation (7.6) below for a definition, and see Figure 7.3 below for an illustration. We will get to this in more detail below, when we discuss Chernoff bounds and the Central Limit Theorem, but we mention it here since it will give us a "size scale" to determine if the variability is large or small.

Size scale. Let's comment on what the phrase "size scale" means.

As an example of why one might want to know the size scale of the random variability of a random variable, let's say that we have some data and that we get a new data point. A question we might wonder is whether the new data point is in some sense "the same" as the data we already have or whether it is "very different" than the data we already have. (Note that we haven't said anything here about whether the same or very different is good or bad—e.g., very different might mean we have an error, or it might mean that we have discovered something new and interesting.) For example, let's say that the most recent entry in a database of people's heights is 6.5 feet, or alternatively that it is 65 feet. Both of those numbers are above the mean height of people, and we want a principled way to ask and answer the question of whether one data point is "very different" from the rest of the data or whether it is "the same" but slightly larger than the rest of the data. (While an entry of 6.5 feet might be an error, it probably just corresponds to the correct height of a person who is a little taller than average, but an entry of 65 feet is almost certainly an error.) Bounding how far a variable is from its mean and understanding how to interpret that in more general situations are important in many areas of data science.

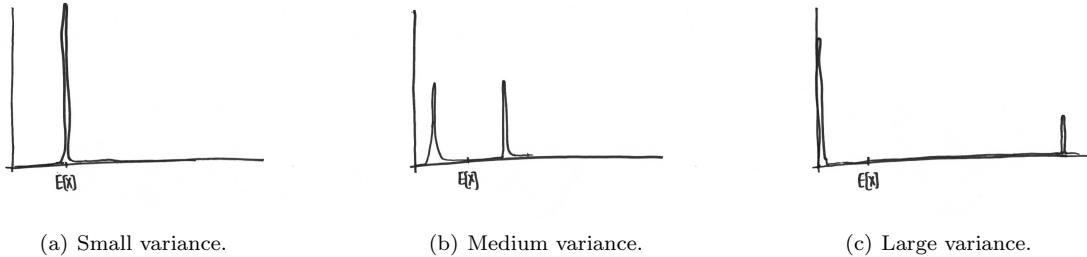


Figure 7.1: Illustration of probability distributions with different mean-versus-variance behaviors.

7.2 Bounding deviations from the mean (Part 1: weak bounds)

To start, here is the question: let's say that we have a random variable, and if all we know is a very limited amount of information about a random variable, e.g., just the mean or just the variance or just the mean and variance, then can we make any claim about how close/far any actual value is from the mean? See Figure 7.1 for several examples. In these examples, the mean is the same, but the variance is quite different. Obviously, there is a very different chance in each case that the value of a random variable drawn from that distribution will be near to or very far away from its mean value.

There are several increasingly-powerful ways to characterize the answer to this question: Markov's Inequality; Chebychev's Inequality; and Chernoff bounds. We will go through each of these in turn. We will start with Markov's Inequality and Chebychev's Inequality, both of which are simple to discuss, but both of which are rather weak.

7.2.1 Markov's Inequality.

The first of these goes by the name *Markov's Inequality*. The only information that Markov's Inequality uses about the random variable is the mean. For this reason, by itself, Markov's Inequality is very weak, in the sense that it will lead to upper bounds that are very far from optimal. It is, however, important pedagogically and as a building block for other more powerful methods. A caveat is that Markov's Inequality applies only to non-negative random variables, i.e., random variables whose range is the non-negative real numbers.

Theorem 14 (Markov's Inequality) *Let X be a non-negative random variable with expectation $\mu = \mathbf{E}[X]$. Then, $\forall a > 0$,*

$$\mathbf{Pr}[X \geq a] \leq \frac{\mathbf{E}[X]}{a}. \quad (7.1)$$

Proof: For $a > 0$, define the random variable Y to be

$$Y = \begin{cases} 1 & \text{if } X \geq a \\ 0 & \text{otherwise} \end{cases}.$$

That is, Y is the indicator vector for the event $\{X \geq a\}$. Then Y is clearly a random variable, and clearly $Y \leq \frac{1}{a}X = \frac{X}{a}$. But since Y is a 0-1 random variable, we have that

$$\mathbf{E}[Y] = 1 \cdot \mathbf{Pr}[Y = 1] + 0 \cdot \mathbf{Pr}[Y = 0] = \mathbf{Pr}[X \geq a].$$

So, it follows that

$$\mathbf{Pr}[X \geq a] = \mathbf{E}[Y] \leq \mathbf{E}\left[\frac{X}{a}\right] = \frac{\mathbf{E}[X]}{a},$$

from which the result follows. \diamond

Before proceeding, let's make sure we understand what Markov's Inequality is saying.

- First, let's parse Equation (7.1) to make sure we understand the structure of the statement. Recall that the setup is that we are dealing with a random variable X (which by assumption is non-negative) and that this random variable has mean $\mu = \mathbf{E}[X]$. This random variable is defined on a sample space Ω , and we are interested in when (i.e., for which subset of Ω , and in particular the probability of that subset) the random variable is larger than a given value $a \in \mathbb{R}$. In this case, $X \geq a$, sometimes written as $\{X = a\}$, represents an event, i.e., the subset of Ω where $X \geq a$, i.e.,

$$\{X = a\} = \{s \in \Omega | X(s) \geq a\},$$

and Equation (7.1) says that the probability of this event is “upper bounded” or “not too likely,” in the sense that it is less likely than the number $\frac{\mathbf{E}[X]}{a} \in \mathbb{R}$.

- Next, note that the nonnegativity requirement is essential, if we want to get a nontrivial result. Otherwise, we could have a random variable defined as

$$X = \begin{cases} \alpha & \text{with probability 0.5} \\ -\alpha & \text{otherwise} \end{cases},$$

where $\alpha \in \mathbb{R}$ is arbitrary. In this case, the $\mathbf{E}[X] = 0$, but any particular value of the random variable could be arbitrarily large, i.e., arbitrarily far from the mean.

- Next, let's make sure we understand whether the statement Equation (7.1) is saying anything nontrivial. (For example, if all it said was that $\Pr[X \geq a]$ was greater than -2 , that is “trivial,” in the sense that *any* probability has to be greater or equal to 0 and thus greater than -2 , and thus there is no need for all the extra assumptions that are being made.)
 - **a is independent of $\mathbf{E}[X]$.** If we choose, e.g., $a = 0.5$, then Equation (7.1) says that the probability of the event $\{X = a\}$ is less than or equal to $\frac{1}{0.5}\mathbf{E}[X] = 2\mathbf{E}[X]$. It's hard to know what to make of this statement (since it depends on $\mathbf{E}[X]$).
 - * For example, if $\mathbf{E}[X] = 0.1$, then Equation (7.1) says that the probability of the event $\{X = a\}$ is less than or equal to 0.2 , which might be meaningful.
 - * On the other hand, if X is measured in different “units,” then we could have that $\mathbf{E}[X] = 10$, in which case Equation (7.1) says that the probability of the event $\{X = a\}$ is less than or equal to 20 . In this case, the statement is trivial, in the sense that any probability has to be less than or equal to 1 .
 - **a is less than $\mathbf{E}[X]$.** If we choose, e.g., $a = 0.5\mathbf{E}[X]$, then Equation (7.1) says that the probability of the event $\{X = a\}$ is less than or equal to $\frac{1}{0.5\mathbf{E}[X]}\mathbf{E}[X] = 2$. In this case, too, the statement is trivial, in the sense that any probability has to be less than or equal to 1 .
 - **a is greater than $\mathbf{E}[X]$.** If we choose, e.g., $a = 2\mathbf{E}[X]$, then Equation (7.1) says that the probability of the event $\{X = a\}$ is less than or equal to $\frac{1}{2\mathbf{E}[X]}\mathbf{E}[X] = 0.5$. Here, the statement is not vacuous. Similarly, if $a = 10\mathbf{E}[X]$, then Equation (7.1) says that the probability of the event $\{X = a\}$ is less than or equal to $\frac{1}{10\mathbf{E}[X]}\mathbf{E}[X] = 0.1$. In this case, too, the statement is not vacuous.

There are two lessons to this discussion. First, we need to parse statements like Equation (7.1)—and other similar ones we will see below—to make sure we understand what they are saying. Second, in order to obtain nontrivial results, when we use these sorts of inequalities, we need to measure deviation in terms of some properties of the probability distribution rather than in absolute terms.

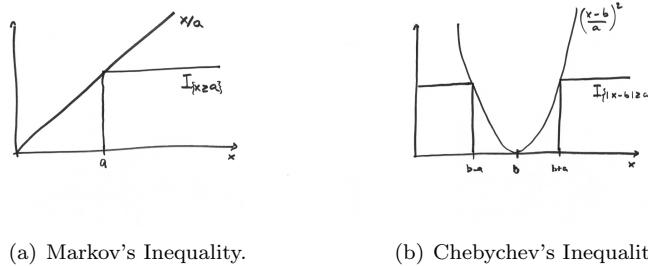


Figure 7.2: Graph illustrating the bounding relationship behind the Markov Inequality and the Chebychev Inequality.

Remark (other form of Markov's Inequality). There are several other variants/formulations of Markov's Inequality. Probably the most common is the following:

$$\Pr [X \geq \alpha \mathbf{E}[X]] \leq \frac{1}{\alpha}. \quad (7.2)$$

This form of Markov's Inequality defines the event of interest in terms of an inequality involving X and a multiple of the mean of X rather than in absolute terms, and it defines the probability of the event not holding in absolute terms rather than in terms of the mean. This and other forms of Markov's Inequality are all equivalent and thus equivalently weak, and they just amount to parameterizing the statement with slightly different variables.

Remark. For an illustration showing Markov's Inequality, see the Figure 7.2(a), where since we are dealing with a discrete random variable, we can assume that we have numbered the states in that order to get all the 0s first. The random variable Y may be viewed as a the indicator function for the set $\{X \geq a\}$, i.e., $Y = 1_{X \geq a}$. Thus, it takes values 0 or 1. But by construction, it is strictly less than than the function $\frac{1}{a}X$.

Remark. Although it is weak, Markov's Inequality is “tight” in the sense that there exist probability distributions that achieve/saturate the bound. Here is the result showing that Markov's Inequality is tight. If $\mu = \mathbf{E}[X] \leq a$, then consider the random variable defined by

$$X = \begin{cases} a & \text{with probability } \mu/a \\ 0 & \text{with probability } 1 - \mu/a \end{cases}.$$

This random variable is illustrated in Figure 7.1(c), and note that this is just the “large variance” example we gave above. Observe that there is some probability mass very far above the mean. For this random variable X , we have that

$$\mathbf{E}[X] = \frac{\mu}{a}a + \left(1 - \frac{\mu}{a}\right)0 = \mu,$$

but we also have that

$$\Pr[X \geq a] = \frac{\mu}{a} = \frac{\mathbf{E}[X]}{a}.$$

For this reason, one cannot in general get better bounds than that provided by Markov's Inequality, assuming only information about the mean. Better bounds are possible, but if one wants to obtain better bounds, then one must assume more about the probability distribution.

7.2.2 Chebychev's Inequality.

The next result is known as *Chebychev's Inequality*. Chebychev's Inequality is stronger than Markov's Inequality, both in the sense that it applies to any random variable (and not just to non-negative random variables) and also since it provides a stronger bound. It does so by making an assumption about the variance as well as the mean.

Theorem 15 (Chebychev's Inequality) *Let X be any random variable. Then $\forall a > 0$, we have that*

$$\Pr [|X - \mathbf{E}[X]| \geq a] \leq \frac{\mathbf{Var}[X]}{a^2}. \quad (7.3)$$

Proof: First observe that if we define the following two events

$$\begin{aligned} \text{Event } \mathcal{E}_1 &: |X - \mathbf{E}[X]| \geq a \\ \text{Event } \mathcal{E}_2 &: (X - \mathbf{E}[X])^2 \geq a^2, \end{aligned}$$

then the two events are identical, i.e., $\mathcal{E}_1 = \mathcal{E}_2$. Thus,

$$\Pr [|X - \mathbf{E}[X]| \geq a] = \Pr [(X - \mathbf{E}[X])^2 \geq a^2].$$

But, $(X - \mathbf{E}[X])^2$ is a non-negative random variable, and thus can apply Markov's Inequality to it to get

$$\Pr [(X - \mathbf{E}[X])^2 \geq a^2] \leq \frac{\mathbf{E}[(X - \mathbf{E}[X])^2]}{a^2} = \frac{\mathbf{Var}[X]}{a^2},$$

which establishes the result. \diamond

Before proceeding, let's parse Equation (7.3) to make sure we understand what it is saying. As with the discussion for Markov's Inequality, here we are dealing with a random variable X (but which here could have negative values) defined on a sample space Ω that has mean $\mu = \mathbf{E}[X]$ and variance $\mathbf{Var}[X]$, and we are interested in when the random variable takes values that are very far from its mean. The form of Chebychev's Inequality given in Equation (7.3) measures "very far" in absolute terms to be that $|X - \mathbf{E}[X]| \geq a$, and it says that the probability of this event is "upper bounded" or "not too likely," in the sense that it is less likely than the number $\frac{\mathbf{Var}[X]}{a^2} \in \mathbb{R}$.

Remark (other forms of Chebychev's Inequality). There are several other variants/formulations of Chebychev's Inequality. Here are several which bound the deviation of a random variable from its expectation in terms a constant factor multiplied by its expectation or standard deviation.

- $\forall t > 1$, we have:

$$\Pr [|X - \mathbf{E}[X]| \geq t\mathbf{E}[X]] \leq \frac{\mathbf{Var}[X]}{t^2(\mathbf{E}[X])^2}. \quad (7.4)$$

This form of Chebychev's Inequality is more reminiscent of Markov's Inequality, as given in Equation (7.2), in that the scale of the deviation is defined in terms of a multiple of the expectation. In this case, the probability that the event holds depends on t as well as the mean and expectation, and so the discussion above about potentially vacuous bounds holds here too.

- $\forall t > 1$, we have:

$$\Pr [|X - \mathbf{E}[X]| \geq t\sigma(X)] \leq \frac{1}{t^2}. \quad (7.5)$$

This form of Chebychev's Inequality defines the probability of the event not holding in absolute terms rather than in terms of the mean and/or variance. More importantly, this form of Chebychev's Inequality defines the scale of the deviation in terms of a multiple of the standard deviation. This is not a minor change. In many cases, $\sigma[X] \ll \mathbf{E}[X]$; and in other cases, one mean-centers the random variable X , and so $\mathbf{E}[X] = 0$. In both cases, since $\sigma[X]$ provides a measure of the variability of the data (in the same "units" as $\mathbf{E}[X]$), it is more natural to ask how far is X from its expectation on that scale.

These and other forms of Chebychev's Inequality are all equivalent and thus equivalently weak, and they just amount to parameterizing the statement with slightly different variables.

Remark. For an illustration showing Chebychev's Inequality, see the Figure 7.2(b), where since we are dealing with a discrete random variable, we can assume that we have numbered the states by the values of X . The random variable Y may be viewed as the indicator function for the set $\{|X - b| \geq a\}$, i.e., $Y = 1_{|X-b| \geq a}$. Thus, it takes values 0 or 1. But by construction, it is strictly less than than the function $(\frac{X-b}{a})^2$.

Both Markov's Inequality and Chebychev's Inequality provide bounds, and we can ask how good are they. Let's do that.

Example (of weakness of Markov's Inequality and Chebychev's Inequality). Let's flip a fair coin n times, and for $i \in [n]$, let's define the random variables

$$X_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ flip is H} \\ 0 & \text{otherwise} \end{cases}.$$

Then,

$$\begin{aligned} \mathbf{E}[X_i] &= \frac{1}{2}1 + \frac{1}{2}0 = \frac{1}{2} \\ \mathbf{E}[X_i^2] &= \frac{1}{2}1 + \frac{1}{2}0 = \frac{1}{2} \\ \mathbf{Var}[X_i] &= \mathbf{E}[X_i^2] - (\mathbf{E}[X_i])^2 = \frac{1}{2} - \frac{1}{4}. \end{aligned}$$

Next, let $X = \sum_{i=1}^n X_i$, in which case we have the following.

$$\begin{aligned} \mathbf{E}[X] &= \mathbf{E}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbf{E}[X_i] = \sum_{i=1}^n \frac{1}{2} = \frac{n}{2} \\ \mathbf{Var}[X] &= \mathbf{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbf{Var}[X_i] = \sum_{i=1}^n \frac{1}{4} = \frac{n}{4} \\ \sigma[X] &= \frac{\sqrt{n}}{2}. \end{aligned}$$

(In that, we replaced the variance of the sum by the sum of the variances since these are independent coin flips and thus independent random variables.) In particular,

$$\text{if } n = 1,000,000, \text{ then } \mathbf{E}[X] = 500,000, \text{ and } \sigma[X] = 500.$$

To ask how likely it is that $X \gg \mathbf{E}[X]$, let's ask: what is the probability that $X \geq \frac{3}{2}\mathbf{E}[X]$? (Alternatively, we could ask: what is the probability that $X \geq \mathbf{E}[X] + 10\sigma[X]$?) For example, if $n = 10^6$, then we want to know how likely is it that we flip more than 750,000 H. Our intuition suggests that it should be extremely unlikely. In principle, one could compute Pascal's triangle up to $n = 10^6$, and add up the values to to $k = 750,000$, but that is not practical, and so we will look for bounds. (These bounds might not be “tight” or “perfect,” but they will say that the probability of this event is less than or equal to some value that is easier to compute.)

One bound is provided by Markov's Inequality,

$$\Pr\left[X \geq \frac{3n}{4}\right] = \Pr\left[X \geq \frac{3}{2}\mathbf{E}[X]\right] \leq \frac{\mathbf{E}[X]}{a} = \frac{n/2}{3n/4} = \frac{2}{3}.$$

So, what should we make of this? It is certainly true as a bound. But, it is *very* weak. For example, it doesn't even depend on n . So, in particular, it holds for $n = 1$. (If you don't see that, then confirm it by checking explicitly.) Moreover, it doesn't get better as n increases, which is something we expect.

Another bound is provided by Chebychev's Inequality,

$$\Pr\left[X \geq \frac{3n}{4}\right] = \Pr\left[|X - \mathbf{E}[X]| \geq \frac{n}{4}\right] \leq \frac{\mathbf{Var}[X]}{(n/4)^2} = \frac{n/4}{(n/4)^2} = \frac{4}{n}.$$

This is much better than $2/3$, especially if n is large.

But, remember the rule of thumb:

- If an event happens with probability p , then if we sample it many fewer times than $1/p$ times, then it is very unlikely we will observe the event, and if we sample it many more times than $1/p$ times, then it is very likely that we will observe the event.

In particular, in this case, it is likely to happen if we do more than $\frac{1}{4/n}$ trials. So, e.g., if this bound provided by Chebychev's Inequality is tight, and if we perform (say) 10^7 or 10^8 trials, i.e., flip 10^6 coins repeatedly 10^7 or 10^8 times, then it is likely that we would observe the event of 750,000 Hs in one of those 10^7 or 10^8 trials. (A problem related to this may be given in the homework.) That we don't see it suggests that this bound is still very weak. We will get to how weak soon.

7.3 A baseline to aim for

When we say that Markov's Inequality or even Cheychev's Inequality is weak, what do we mean? The “small variance” example we gave above showed that there are probability distributions such that the variance is 0. So, clearly that is possible. But how realistic is that? And how common is that? And should we consider anything larger than that as “medium” or “large”? Or are there larger things that could also be considered as “small”?

This is a very practical issue. For example, we would expect that if there is some randomness or noise in the way the data are generated, then there will be at least some variability due to that, and so perhaps that should be considered as “small”? So, in those cases, what would count as “small”? E.g., how “good” or “tight” can we expect such a bound to be, e.g., if there is some/much randomness in the probability distribution that generated the data? Relatedly, can we define a “size scale” (perhaps depending on the mean or variance or number of data points or something else) such that small is smaller than that size scale and large is larger than that size scale?

7.3.1 The Gaussian/normal distribution

In general, the baseline is provided by a Gaussian/normal distribution. For the normal distribution, the sample space $\Omega = \mathbb{R}$. Thus, it is a continuous probability distribution, i.e., it is not a discrete distribution. It is, however, sufficiently important that we'll discuss it. The reason for its importance is that many probability distributions of interest roughly “look like” it. In particular, when one plots a histogram of a sampled/resampled random variable and the fluctuations decrease, then the histogram starts to look like a Gaussian/normal distribution.

The normal distribution. The normal distribution is parameterized in terms of two numbers, a mean μ and a standard deviation σ (or equivalently a variance σ^2), and it is defined to be:

$$f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \quad (7.6)$$

Often, it is written as $N(\mu, \sigma^2)$. See Figures 7.3 and 7.4 for the pdf and cdf of the normal distribution. Note in particular how the plots vary as the parameters μ and σ^2 are varied. In addition, often for simplicity one chooses $\mu = 0$ and $\sigma^2 = 1$, in which case

$$N(0, 1) = f(x|0, 1) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \quad (7.7)$$

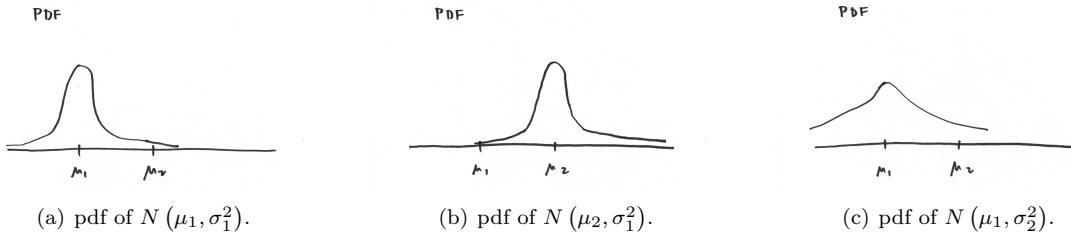


Figure 7.3: PDF of Normal distribution $N(\mu, \sigma^2)$, for different values of the mean parameter μ and variance parameter σ^2 : $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_1^2)$, and $N(\mu_1, \sigma_2^2)$, with $\mu_2 > \mu_1$ and $\sigma_2 > \sigma_1$.

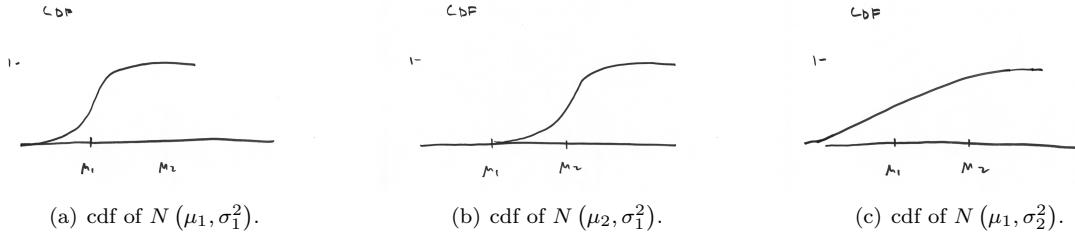


Figure 7.4: CDF of Normal distribution $N(\mu, \sigma^2)$, for different values of the mean parameter μ and variance parameter σ^2 : $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_1^2)$, and $N(\mu_1, \sigma_2^2)$, with $\mu_2 > \mu_1$ and $\sigma_2 > \sigma_1$.

Let's parse Equation (7.6) for the normal distribution and Equation (7.7) for the standard normal distribution to understand what they are saying.

- What Equation (7.7) says is that the probability for the random variable X to take a value $x \in \Omega$ is large if x is closer to 0 than roughly $\sqrt{2}$, and the probability for the random variable X to take a value $x \in \Omega$ is small if x is farther from 0 than roughly $\sqrt{2}$.
- Generalizing this, what Equation (7.6) says is that the probability for the random variable X to take a value $x \in \Omega$ is large if x is closer to μ than roughly $\sqrt{2}$, *when measured in the scale defined by σ* , and the probability for the random variable X to take a value $x \in \Omega$ is small if x is farther from μ than roughly $\sqrt{2}$ *when measured in the scale defined by σ* .

Facts about the normal distribution. Here are some facts about the Gaussian/normal distribution.

- The mean value of $N(\mu, \sigma^2)$ equals μ . That is, if X is a random variable on $\Omega = \mathbb{R}$ with $N(\mu, \sigma^2)$ as its probability distribution, which is sometimes written as $X \sim N(\mu, \sigma^2)$, then

$$\mathbf{E}[X] = \mu.$$

- The variance of $N(\mu, \sigma^2)$ equals σ^2 , and thus the standard deviation equals σ . That is, if $X \sim N(\mu, \sigma^2)$, then

$$\mathbf{Var}[X] = \sigma^2 \quad \text{and} \quad \sigma[X] = \sigma.$$

- Roughly 68.3% of the probability lies within 1σ of μ .

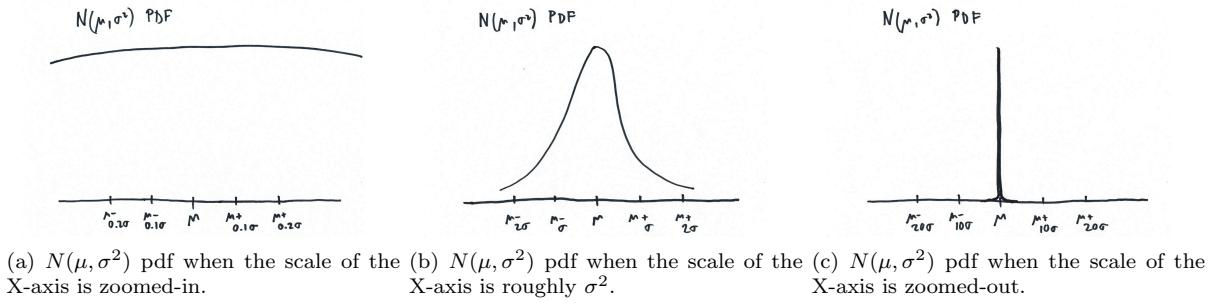


Figure 7.5: PDF of Normal distribution $N(\mu, \sigma^2)$, for different values of the scale of the X-axis.

- Roughly 95.4% of the probability lies within 2σ of μ .
- Roughly 99.7% of the probability lies within 3σ of μ .
- Roughly 99.994% of the probability lies within 4σ of μ .
- Roughly 99.99994% of the probability lies within 5σ of μ .
- Roughly 99.999998% of the probability lies within 6σ of μ .

Remark. Note the form of the last four statements: roughly $\alpha\%$ of the probability lies within $\beta\sigma$ of the mean μ . Alternatively, roughly $1 - \alpha$ of the probability lies outside $\beta\sigma$ of the mean μ . This is exactly the form of the statement provided by Chebychev's Inequality (in Equation (7.5)), and it will be exactly the form of the statement provided by Chernoff bounds (that we will get to below).

Remark. An important aspect of the Gaussian/normal distribution is that most of the mass lies very near the mean μ . When we say “near” we need to specify the “scale” or “units” with respect to which we are measuring, and here the scale is given by the standard deviation σ . This is quantified by the last few bullets above. In particular, this says that most of the probability mass lies within $\mu \pm \beta\sigma$, where $\beta > 0$ is a small constant.

Remark. For the normal distribution, the probability of a value being more than a few standard deviations from the mean is extremely small. This is very important. We will see below that certain other distributions, in particular sums of independent random variables, exhibit similar properties. We will also see that Chernoff-style bounds provide bounds that are similarly strong.

Remark. That comment about the probability being *extremely* small for a value being more than a few standard deviations from the mean for a normal distribution should be emphasized. The function $f(x) = 1/x$ or $f(x) = 1/x^2$ decay with increasing x (for $x > 0$), but they do not decay *extremely* fast. In particular, if we know the value of $f(x)$ at $x = x^*$, then the value at $x = 10x^*$ is 10% or 1% of that, and the value at $x = 100x^*$ is 0.1% or 0.01% of that, but it is not 0.00000000001%. For the normal distribution, if $\mu = 0$, then if we know the value of $f(x) = N(0, x)$ at some value of $x = x^* = \sigma^2$, then the value of $f(x)$ at $x = 10x^*$ or $x = 100x^*$ is essentially 0. In that sense, Figure 7.3 is misleading, since it plots the function where the X-axis is on a scale near σ . Let's “zoom in” and “zoom out” and plot $N(\mu, \sigma^2)$, where the X-axis is measured on smaller units and larger units. The results are shown in Figure 7.5.

- Figure 7.5(a) shows $N(\mu, \sigma^2)$ when plotted on an X-axis scale that is *much less than* σ . Here, the function is roughly constant at $\frac{1}{2\pi\sigma^2}$, which should be clear by considering a Taylor expansion.
- Figure 7.5(b) shows $N(\mu, \sigma^2)$ when plotted on an X-axis scale that is *roughly the same as* than σ . This is the familiar figure, and from this one can guess that near μ , the value is roughly flat, but this does not give a sense for how small the function is far from μ .

- Figure 7.5(c) shows $N(\mu, \sigma^2)$ when plotted on an X-axis scale that is *much larger than* σ . Here, except very near μ , the function is essentially 0—that is, nearly all the probability mass is very near μ , and it is very close to 0 otherwise.

Remark. A word of caution. In this question, μ and σ refer to the mean and variance of “some distribution,” i.e., which may or may not be a Gaussian/normal distribution. It can be any probability distribution. If that distribution happens to be a Gaussian/normal distribution, then μ and σ refer to the μ and σ of the Gaussian/normal distribution, but otherwise they are different. We’ve seen some distributions above, and we’ll see more distributions below, and we’ll compute their values of μ and σ later.

Main use case for tail bounds. Let’s go back to bounding tail events, i.e., to bounding the probability of an event that a random variable takes a value far from its mean. In general, the basic question is:

- Let $\{X_1, \dots, X_n\}$ be a random sample of size n , i.e., a sequence of i.i.d. random variables drawn from some distribution (discrete $\{-1, +1\}$ or $\{0, 1\}$, uniform on $[0, 1]$ or $[-3, 12]$, $N(0, 1)$ or $N(\mu, \sigma^2)$, or any other distribution) with expectation μ and finite variance σ^2 , and consider

$$X = S_n = \frac{1}{n} (X_1 + \dots + X_n),$$

i.e., the *sample* mean. Then, what is the behavior of S_n as n gets large?

7.3.2 Types of claims one can make: limiting versus non-limiting statements

Before we give an answer to this question, let’s consider the *types of answers* that one might get. By “types of answers,” we mean the *form* of the answer, independent of the exact quantitative answer. We will be interested in answers that are of the form provided by Markov’s Inequality and Cheychev’s Inequality, that is, answers that are of the form of bounds on the probability of a tail event for a given finite value of n , but that are much stronger than Markov’s Inequality and Cheychev’s Inequality. These will be known as Chernoff bounds, and we will get to them below.

Asymptotic statements. Before that, though, it’s good to note that there are several other *types of answers* one can offer. In particular, we’ll state here several asymptotic answers to this question that are common. (We won’t describe precisely what the various terms in these expressions mean, as they are covered in more advanced classes, but we want to mention these other types of answers for context and completeness.)

- **Weak Law of Large Numbers (WLLN).** The WLLN states that the sample average S_n converges “in probability” to the population mean μ . This means

$$\lim_{n \rightarrow \infty} \Pr [|S_n - \mu| > \epsilon] = 0,$$

which is sometimes written as $S_n \xrightarrow{P} \mu$, as $n \rightarrow \infty$.

- **Strong Law of Large Numbers (SLLN).** The SLLN states that the sample average S_n converges “almost surely” to the population mean μ . This means

$$\Pr \left[\lim_{n \rightarrow \infty} S_n = \mu \right] = 1,$$

which is sometimes written as $S_n \xrightarrow{a.s.} \mu$, as $n \rightarrow \infty$.

- **Central Limit Theorem (CLT).** The CLT describes the size and distributional form of the random fluctuations around the deterministic μ . It states that

$$\lim_{n \rightarrow \infty} \sqrt{n} \left(\left(\frac{1}{n} \sum_{i=1}^n X_i \right) - \mu \right) \xrightarrow{d} N(0, \sigma^2).$$

That is, the CLT states that $\sqrt{n}(S_n - \mu)$ converges “in distribution” to $N(0, \sigma^2)$, from which it follows that S_n converges “in distribution” to $N(\mu, \sigma^2/n)$, i.e., that $S_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

(As a more advanced aside, the phrase “in probability” above means that for a specific n , the number S_n is likely near μ , but it may happen that $|S_n - \mu| > \epsilon$ infinitely often, but at infrequent intervals; and the phrase “almost surely” means that that will not happen, i.e., with probability = 1, $\forall \epsilon$, it holds that $|S_n - \mu| \leq \epsilon$, for large enough n . These are two of the most common notions of convergence which you will see if you take more courses later, but we won’t go into them in more detail here.)

Although these are three common ways to describe the behavior of S_n as n gets large, it is important to note that the *type of guarantees* they provide is very different in form than the *type of guarantees* provided by Markov’s Inequality and Chebychev’s Inequality. In particular, they are all asymptotic results, i.e., they hold as $n \rightarrow \infty$, and they say nothing about the behavior for a fixed n . Often, they provide qualitative guidance, but they are often much less useful in finite- n setting, e.g., when one has a data set of fixed size or when one is interested in the outcome of a finite but large value of n .

One important thing that the CLT points out, however, is that even in the limit as $n \rightarrow \infty$ there is variability of S_n around μ . We have seen this before empirically. Recall that when you looked at histograms based on resampling real data, there is often a central point around which the statistics being computed concentrate, but there is still some variability around that central point.

Distributional statements and bounds. Although we aren’t going to define the more advanced “in distribution” notion in this class, there are two important points here about the CLT and the size of the variance that are implied by the “in distribution” statement of the CLT. Since we want to know how close S_n is to the mean, we want to measure the variability of this random variable in the same units as the mean μ , and thus we will state these two results for the variance in terms of the standard deviation σ .

- **Lower bound: “not less than” something.** The size of the standard deviation is not less than roughly $1/\sqrt{n}$. This is why I said that, roughly, if you do 10^6 coin flips, then you should expect a variability around 500,000 of roughly 1000. In particular, this presents a sort of “lower bound” on how precisely you can measure quantities, e.g., the mean, with a random process, and it is why your estimates of π from dart throwing at the box in \mathbb{R}^2 converged so slowly.
- **Upper bound: “not more than” something.** The size of the standard deviation is not more than roughly $1/\sqrt{n}$. That is, if we know only the mean and variance, then we have a sort of “upper bound” on how far our estimate can be from the mean—if we take many independent samples from the same distribution. Stated in terms of the CLT, this only holds in the limit, but we might hope to get a finite-sample version of it.

Remark. Viewed from this perspective, Markov’s Inequality and Chebychev’s Inequality are ways to bound the second bullet for arbitrary distribution for which we know only the mean and variance (and not just a distribution arising from taking the sum of n independent samples from an arbitrary distribution), since they permit us to bound tail events. But they are much weaker since Markov’s Inequality doesn’t depend on n and Chebychev’s Inequality depends on n only as $1/n$, while the CLT decreases *exponentially*—where exponentially means that it is EXTREMELY unlikely to happen—as a function of the distance from the mean. We can do better by using bounds known as Chernoff bounds.

7.4 Bounding deviations from the mean (Part 2: strong bounds)

Chernoff bounds are a general class of bounds that are very powerful, in that they obtain a tail dependence that is qualitatively like the Gaussian/normal distribution. Like Markov’s Inequality and Chebychev’s

Inequality, however, Chernoff bounds are finite n statements, i.e., they are not limiting or asymptotic statements like the CLT. While they require more sophisticated mathematics for general probability distributions, their basic idea can be illustrated for some simple distributions, and so we will consider one such distribution in some detail. The main point here is that they permit us to obtain CLT-like or normal-like bounds, i.e., quite tight bounds, with very high probability, for a much broader class of probability distributions. In particular, this will mean that flipping a large number of coins or rolling a large number of dice or choosing points from a high-dimensional hypercube or high-dimensional hypersphere will be qualitatively like (but quantitatively different than) the Gaussian/normal distribution.

7.4.1 Chernoff bounds

To describe Chernoff bounds, let's start with the following definition.

Definition 60 *The moment generating function of a random variable X is $M_X(t) = \mathbf{E}[\exp(tX)]$.*

Remark. As an advanced aside, here is a comment for those familiar with calculus. The function $M_X(t)$ is called a moment generating function since it is a function of the random variable X that can be used to “generate” the moments of X . This is accomplished by differentiating $M_X(t)$ with respect to t .

In the same way as we saw that having information about both the mean and variance permitted us to obtain stronger bounds than just using the mean, here we will see that by using information on higher moments we will be able to obtain much stronger bounds.

The general approach for a Chernoff bound is given as follows. The Chernoff bound for a random variable X is obtained by applying Markov's Inequality to the function e^{tX} , for some well-chosen value of t . The function e^{tX} is non-negative for any value of t , and so Markov's Inequality holds for any value of t , and then we choose the value of t that gets us the best bound or some bound that is simple and sufficient for our purposes. More precisely, this is stated in the following results.

Claim 6 *For all $t > 0$, the following holds:*

$$\Pr[X \geq a] = \Pr[e^{tX} \geq e^{ta}] \leq \frac{\mathbf{E}[e^{tX}]}{e^{ta}}$$

Corollary 1

$$\Pr[X \geq a] \leq \min_{t>0} \frac{\mathbf{E}[e^{tX}]}{e^{ta}}$$

Claim 7 *For all $t < 0$, the following holds:*

$$\Pr[X \leq a] = \Pr[e^{tX} \geq e^{ta}] \leq \frac{\mathbf{E}[e^{tX}]}{e^{ta}}$$

Corollary 2

$$\Pr[X \geq a] \leq \min_{t>0} \frac{\mathbf{E}[e^{tX}]}{e^{ta}}$$

The way these Chernoff results are used is as follows: to get bounds for a specific probability distribution, one chooses appropriate values for the parameter t . For example, the minimization over t gives the best possible bounds. Alternatively, in many cases, one can do almost as well with some other value of t that is more convenient to compute, and so instead it is more common to use that value of t .

Let's state and prove a Chernoff bound in a very simple case.

Theorem 16 (Simple Chernoff bound, one-sided) *Let X_1, \dots, X_n be independent random variables, with*

$$\Pr[X_i = 1] = \Pr[X_i = -1] = \frac{1}{2}, \quad (7.8)$$

and let $X = \sum_{i=1}^n X_i$. Then,

$$\forall a > 0 : \Pr[X \geq a] \leq e^{-a^2/2n}. \quad (7.9)$$

Proof: First, recall that the Taylor expansion of the exponential is given by

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots = \sum_{i=1}^{\infty} \frac{x^i}{i!},$$

and thus that

$$\begin{aligned} e^t &= 1 + t + \frac{t^2}{2!} + \dots + \frac{t^i}{i!} + \dots \\ e^{-t} &= 1 - t + \frac{t^2}{2!} + \dots + (-1)^i \frac{t^i}{i!} + \dots. \end{aligned}$$

Next, to let's compute $\mathbf{E}[e^{tX_i}]$, note that, for all t , we have that

$$\mathbf{E}[e^{tX_i}] = \frac{1}{2}e^t + \frac{1}{2}e^{-t}.$$

By plugging the expressions for e^t and e^{-t} into this, and observing that the terms with odd powers (of t) cancel, we get that

$$\begin{aligned} \mathbf{E}[e^{tX_i}] &= \frac{1}{2}e^t + \frac{1}{2}e^{-t} \\ &= \sum_{i \geq 0} \frac{t^{2i}}{(2i)!} \\ &= \sum_{i \geq 0} \frac{(t^2/2)^i}{i!} \\ &= e^{t^2/2}. \end{aligned}$$

So, the moment generating function is given by the following:

$$\mathbf{E}[e^{tX}] = \prod_{i=1}^n \mathbf{E}[e^{tX_i}] \leq e^{t^2 n / 2}.$$

So, the probability of the tail estimate is given by the following:

$$\Pr[X \geq a] = \Pr[e^{tX} \geq e^{ta}] \leq \frac{\mathbf{E}[e^{tX}]}{e^{ta}} \leq e^{t^2 n / 2 - ta}.$$

Now, we make use of our flexibility in the choice of t to choose $t = a/n$, from which it follows that

$$\Pr[X \geq a] \leq e^{-a^2/(2n)}$$

which proves the theorem. \diamond

Given Theorem 16, the following theorem is almost an immediate corollary.

Theorem 17 (Simple Chernoff bound, two-sided) Let X_1, \dots, X_n be independent random variables, with probability given in Equation (7.8), and let $X = \sum_{i=1}^n X_i$. Then,

$$\forall a > 0 : \quad \Pr [|X| \geq a] \leq 2e^{-a^2/2n}. \quad (7.10)$$

Proof: From Theorem 16, we have that

$$\Pr [X \geq a] \leq e^{-a^2/(2n)}.$$

By a symmetric argument, we have that

$$\Pr [X \leq -a] \leq e^{-a^2/(2n)}.$$

Next, note that the event $\{|X| \geq a\}$ is the same as the event

$$\{X \geq a\} \cup \{X \leq -a\}$$

Thus, from the union bound, the theorem follows. \diamond

7.4.2 Discussion of Chernoff bounds

Here are a few comments on these Chernoff bounds.

- First, while the event being bounded in Equation (7.9) is of the form $X \geq a$, and thus this equation looks like it has the same form as Equation (7.1), you should think of this equation as being of the form of Equation (7.5). The reason is that for the random variable X defined by Equation (7.8), we have $\mathbf{E}[X] = 0$ and $\mathbf{Var}[X] = 1$ (and thus $\sigma[X] = 1$). Thus, a corresponds to how far—in units of the standard deviation— X is from $\mathbf{E}[X]$. This is highlighted in Equation (7.10), which could alternatively be written as

$$\forall a > 0 : \quad \Pr [|X - \mathbf{E}[X]| \geq a\sigma[X]] \leq 2e^{-a^2/2n}.$$

The important point here is that the behavior of this expression (with respect to the parameter a) here is very different than the behavior of the expression given in Equation (7.4) (with respect to the parameter t).

- Second, the results provided by these Chernoff bounds are true but uninteresting if $n = 1$. That is fine, since this corresponds to the case of flipping a coin $n = 1$ times. The result is much more interesting when n gets larger, e.g., $n = 10$ or $n = 100$ or $n = 10^6$, in which case $2e^{-a^2/2n}$ is extremely small.
- Third, once n is increased to be such that $2n \gg a^2$ (which is a reasonable thing to ask for, e.g., since it corresponds to the number of times that a coin is flipped), then the probability of the event $\{X = a\}$ is exponentially small. It is this exponentially-unlikely behavior that is qualitatively like the exponentially-unlikely behavior of seeing events that are far from the mean of the normal distribution. This should be contrasted with the constant and the $1/n$ that we saw with Markov's Inequality and Chebychev's Inequality.

The general form and properties of Chernoff bounds, as provided by Theorem 16 and Theorem 17, is much more general than the random variables defined by Equation (7.8), although the statement and the proof of the results more generally can get quite complicated. We won't describe them here, but you will see a few computational examples of these on a homework, and you can find other examples in more advanced classes.

7.5 Examples of random variables

Let's consider some simple examples. These simple distributions are very common, and they can be used to model a wide range of phenomenon.

- **Bernoulli random variable.**

Motivation. Suppose that we run an experiment that succeeds with probability p and fails with probability $1 - p$ (so it is essentially a biased coin flip). For example, we can flip a biased coin many times in a row and declare success if we flip all Hs.

Definition 61 Let Y be a random variable such that

$$Y = \begin{cases} 1 & \text{if the experiment succeeds (i.e., with probability } p) \\ 0 & \text{otherwise (i.e., with probability } 1 - p) \end{cases}.$$

Then Y is a Bernoulli or indicator random variable.

Mean and Variance. To compute the mean and variance of a Bernoulli random variable, observe that

$$\begin{aligned} \mathbf{E}[Y] &= p \cdot 1 + (1 - p) \cdot 0 = p = \mathbf{Pr}[Y = 1] \\ \mathbf{Var}[Y] &= \mathbf{E}[(Y - \mathbf{E}[Y])^2] \\ &= p(1 - p)^2 + (1 - p)(-p)^2 \\ &= p(1 - 2p + p^2) + p^2(1 - p) \\ &= p - p^2 \\ &= p(1 - p). \end{aligned}$$

Application: Throwing biased darts. An application of this is throwing darts, and defining a random variable X to be 1 or 0, depending on whether a “target” is hit. We saw an example of this, when we were throwing darts at the unit L_∞ ball in \mathbb{R}^2 , and letting the target be the L_2 ball; and we'll more examples of this later, where the L_∞ ball and L_2 ball are in \mathbb{R}^n , in which case the probability of hitting the target is much less. Another example of this is flipping a coin, whether fair or not.

- **Binomial random variable.**

Motivation. Suppose that we have a sequence of n independent experiments, each of which succeeds with probability p and fails with probability $1 - p$. That is, we have a sequence of Bernoulli random variables. For this process, we want to know the number of successes in the n trials. (E.g., the number k of Hs in n coin flips.) If X is the number of successes in the n trials, then X is a random variable with the Binomial Distribution.

Definition 62 A Binomial random variable with parameters n and p , denoted $B(n, p)$, is given by

$$\mathbf{Pr}[X = j] = \binom{n}{j} p^j (1 - p)^{n-j},$$

where $\binom{n}{j} = \frac{n!}{j!(n-j)!}$ is exactly j successes and $n - j$ failures.

Remark. The expression $\binom{n}{j}$ looks complicated, but we have seen it before. For example, it appears as the coefficients in

$$(x + y)^n = \sum_{k=0}^n \alpha_k x^{n-k} y^k.$$

Relatedly, it appears in the frequency of Hs in the flips of unbiased coins. Here are the first few coefficients.

$n = 0:$	1				
$n = 1:$	1 1				
$n = 2:$	1 2 1				
$n = 3:$	1 3 3 1				
$n = 4:$	1 4 6 4 1				
$n = 5:$	1 5 10 10 5 1				

Mean and Variance. To compute the mean and variance, here is a fact.

$$\sum_{j=0}^n \Pr[X = j] = 1.$$

(You may be asked to prove this in a homework question.) Also, let's define X_i to be an indicator random variable for success in the i^{th} trial:

$$X_i = \begin{cases} 1 & \text{if the } i^{th} \text{ experiment succeeds} \\ 0 & \text{otherwise ,} \end{cases}$$

in which case $\mathbf{E}[X_i] = p$. Given this, here it is:

$$\begin{aligned} \mathbf{E}[X] &= \mathbf{E}\left[\sum_{i=1}^n X_i\right] \\ &= \sum_{i=1}^n \mathbf{E}[X_i] \\ &= \sum_{i=1}^n p \\ &= np \\ \mathbf{Var}[X] &= \mathbf{Var}\left[\sum_{i=1}^n X_i\right] \\ &= \sum_{i=1}^n \mathbf{Var}[X_i] \\ &= np(1-p). \end{aligned}$$

Note that, in both of these cases, we could have done this computation directly with summations, expressions for factorials, etc.; but using linearity of expectation and linearity of variance for independent random variables made the computation much easier.

- **Geometric random variable.**

Motivation. Suppose that we run the above experiment and it succeeds and we want to know the distribution of the number of experiments.

Definition 63 A Geometric random variable X with parameter p is given by

$$\Pr[X = n] = (1 - p)^{n-1} p.$$

As usual, it may be hard to remember this exact expression if you just try to remember it without understanding the context. It becomes much easier if you understand where it comes from: think of it as $n - 1$ independent failures, each of which gives a factor of $(1 - p)$, followed by one success, which gives a factor of p .

Here is a fact.

$$\sum_{n \geq 0} \mathbf{Pr}[X = n] = 1.$$

(You may be asked to prove this in a homework.)

Remark. Geometric random variables are *memoryless*, in that the probability that you reach your first success n trials from now is independent of the number of failures you have experienced so far. This is quantified in the following.

Claim 8 *For a geometric random variable X with parameter p and for all $n > 0$, we have*

$$\mathbf{Pr}[X = n + k | X > k] = \mathbf{Pr}[X = n].$$

Proof:

$$\begin{aligned} \mathbf{Pr}[X = n + k | X > k] &= \frac{\mathbf{Pr}[(X = n + k) \cap (X > k)]}{\mathbf{Pr}[X > k]} \\ &= \frac{\mathbf{Pr}[X = n + k]}{\mathbf{Pr}[X > k]} \\ &= \frac{(1 - p)^{n+k-1}}{\sum_{i=k}^{\infty} (1 - p)^i p} \\ &= \frac{(1 - p)^{n+k-1} p}{(1 - p)^k} \\ &= (1 - p)^{n-1} p \\ &= \mathbf{Pr}[X = n]. \end{aligned} \tag{7.11}$$

In the above, Equation (7.11) follows since, for all $0 < x < 1$, we have $\sum_{i=k}^{\infty} x^i \frac{x^k}{(1-x)}.$

◊

Mean and Variance. For a geometric random variable,

$$\begin{aligned} \mathbf{E}[X] &= \frac{1}{p} \\ \mathbf{Var}[X] &= \frac{1-p}{p^2}. \end{aligned}$$

We'll skip the proofs of these, since they are a little harder. In particular, note that the standard deviation is less than the expectation. In addition, for flipping a fair coin, it takes an average of 2 flips to see the first H, and flipping a coin with probability 2^{-10} , which is basically what you are doing when you ask for 10 Hs in a row, takes $2^{10} \approx 10^3$ flips, on average.

Application: Coupon collecting. A well-known application of geometric random variables is the coupon collecting problem, which we have mentioned before. Suppose that each box of cereal contains one of n different coupons, and that once you get at least one of every type of coupon, you win a prize. If each box contains a coupon chosen uniformly at random from the set of possible coupons, then how many boxes must you buy to win the prize? To address this, let X be the number of boxes bought until at least one of every type of coupon is obtained. Then, we want to compute $\mathbf{E}[X]$. (We'll just do this, since computing variance to get concentration is more complicated.)

How would one do this? Let's break X up into a sum of other random variables, each of which is "nicer." In particular, let X_i equal the number of boxes bought while you had exactly $i - 1$ different coupons. Then, $X = \sum_{i=1}^n X_i$. Then, each of these n random variables $X_i, i = 1, \dots, n$ is a geometric random variable. This is so since when exactly $i - 1$ coupons have been found, the probability of obtaining a new coupon is just $p_i = 1 - \frac{i-1}{n}$. So, X_i is a geometric random variable with a particular

p_i . So, $\mathbf{E}[X_i] = \frac{1}{p_i} = \frac{n}{n-i+1}$, and

$$\begin{aligned}\mathbf{E}[X] &= \mathbf{E}\left[\sum_{i=1}^n X_i\right] \\ &= \sum_{i=1}^n \mathbf{E}[X_i] \\ &= \sum_{i=1}^n \frac{n}{n-i+1} \\ &= n \sum_{i=1}^n \frac{1}{i} \\ &= n \log(n) + \Theta(n) \approx n \log(n),\end{aligned}$$

where the last line follows from the following good fact to know:

$$H_n = \text{the } i^{\text{th}} \text{ Harmonic number} = \sum_{i=1}^n \frac{1}{i} = \log(n) + \Theta(1) \quad (\log(n) \leq H_n \leq \log(n) + 1)$$

- **Normal distribution (continuous).**

Motivation. it's a bad approximation for a lot of things, but it's a good approximation for a lot of things, in particular when you do resampling like you do in the main class. It is given by

$$N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The prefactor is to make sure that it normalized to 1, i.e., that it is a probability distribution. Look inside the argument of the exponential: it is mean-centered and variance normalized, so it puts things on a common scale. By evaluating this, it is *extremely* unlikely to be more than a few standard deviations from the mean. Here, we give it for $x \in \mathbb{R}$, but we will also consider what happens if $x \in \mathbb{R}^2$ or $x \in \mathbb{R}^n$.

Mean and Variance. μ and σ .

Remark. The normal distribution can be defined for \mathbb{R}^n , and not just for \mathbb{R} . A simple example of this is given by

$$N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right).$$

We will see more non-trivial examples of this later.

Application: Lots and lots of things. Lots and lots of things.

- **Poisson distribution (continuous).**

Motivation. This distribution expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

Definition 64 A discrete Poisson random variable X with parameter μ is given by the following distribution on $j = \{0, 1, 2, \dots\}$:

$$\mathbf{Pr}[X = j] = \frac{e^{-\mu} \mu^j}{j!} = \frac{e^{-\lambda} \lambda^k}{k!}.$$

(We used two sets of letters that are sometimes used here.)

Remark. This arises in lots of cases:

- It has connections with balls and bins.
- It is the limit distribution of Binomial distribution with parameters n and p with n large and p small.
- It is used in the Poisson approximation/paradigm.
- It can be used to derive the so-called birthday paradox, which you saw in class.

Mean and Variance. $\mathbf{E}[X] = \mathbf{Var}[X] = \mu = \lambda$.

Birthday Paradox. If we have m people and n possible birthdays (e.g., m could be the number of people in the class, and n could be the number of days in the year, ignoring leap year, and assuming that births are spread out uniformly around the year), then we want to know what is the probability that everyone has different birthdays. Clearly, if $m = 1$, they it is true, and if $m \geq n$, then it is false, but what about in between? To answer this, note

$$\begin{aligned} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \left(1 - \frac{3}{n}\right) \cdots \left(1 - \frac{n-1}{n}\right) &= \prod_{j=1}^{m-1} \left(1 - \frac{j}{n}\right) \\ &\approx \prod_{j=1}^{m-1} e^{-j/n} \\ &= \exp\left(-\sum_{j=1}^{m-1} \frac{j}{n}\right) \\ &= \exp\left(\frac{-m(m-1)}{2n}\right) \\ &\approx e^{-m^2/(2n)} \end{aligned}$$

(The first approximation above came since $e^{-x} \approx 1 - x$ if $x \ll 1$, i.e., $1 - \frac{k}{n} \approx e^{-k/n}$ if $k \ll n$; and the second approximation just says that $m(m-1) \approx m^2$.) So, the 50-50 point, i.e., the number of people such that the probability for there to be an overlap in birthdays is given by the m such that $\frac{1}{2} = e^{-m^2/(2n)}$, i.e., such that $\log(2) = \frac{m^2}{2n}$, i.e., such that $m \sqrt{2n \log(2)}$, which is $m \approx 22.5$ if $n = 365$. Alternatively, if we want to be more confident, then we could choose m such that $\frac{9}{10} = e^{-m^2/(2n)}$.

There are lots of other distributions: Rademacher; Degenerate/Uniform; Discrete Uniform; etc., but we won't get to them here.

7.6 Problems

7.6.1 Pencil-and-paper Problems

1. **Aligning means and variances.** Consider the following three probability distributions:

$$P1 : \mathbb{P}[X_i = a] = \mathbb{P}[X_i = b] = \frac{1}{2},$$

$P2$: Probability of X_i is uniform on the interval $[a, b]$,

$P3$: X_i is drawn from a normal/Gaussian distribution with mean μ and variance σ^2 ,

where a and b are (as yet) unspecified parameters, and μ and σ are also unspecified parameters.

- (a) For each of these three distributions, what is the mean and variance?
- (b) For each of the first two distributions, what choice(s) of a and b would have the same mean and variance as the normal/Gaussian distribution with mean μ and variance σ^2 ? (This gives two new distributions, call them $P4$ and $P5$, where a and b take on specific values.)

- (c) Plot by hand on a common set of axes all five probability distributions.

(Hint: for a random variable X that is uniform on the interval $[a, b] \subset \mathbb{R}$, sometimes denoted $X \sim U[a, b]$, it holds that $\mathbf{E}[X] = \frac{1}{2}(a + b)$ and that $\mathbf{Var}[X] = \frac{1}{12}(b - a)^2$.)

2. Show that the two forms of Markov's Inequality, given in the notes, are equivalent.
3. Show that the three forms of Chebychev's Inequality, given in the notes, are equivalent.
4. (a) Suppose that we roll a standard fair die 10 times. Let X be the sum of the numbers that appear over the 10 rolls. Use Markov's Inequality to bound $\mathbf{Pr}[|X - 35| \geq 5]$. Simulate this process and comment on how loose or tight is this bound.
(b) Suppose that we roll a standard fair die 100 times. Let X be the sum of the numbers that appear over the 100 rolls. Use Chebychev's Inequality to bound $\mathbf{Pr}[|X - 350| \geq 50]$. Again, based on your simulation of this process, comment on how loose or tight is this bound.
5. (Mitzenmacher-Upfal 2.9)
 - (a) Suppose that we roll twice a fair k -sided die with the numbers a through k on the die's faces, obtaining values X_1 and X_2 . What is $\mathbf{E}[\max(X_1, X_2)]$? What is $\mathbf{E}[\min(X_1, X_2)]$?
 - (b) Show from the calculations in the previous sub-question that

$$\mathbf{E}[\max(X_1, X_2)] + \mathbf{E}[\min(X_1, X_2)] = \mathbf{E}[X_1] + \mathbf{E}[X_2] \quad (7.12)$$

- (c) Explain why Eqn. (7.12) must be true by using the linearity of expectations instead of a direct computation.
6. (Mitzenmacher-Upfal 3.20)
Chebychev's Inequality uses the variance of a random variable to bound its deviation from its expectation. We can also use higher moments. Suppose that we have a random variable X and an even integer k for which $\mathbf{E}[(X - \mathbf{E}[X])^k]$ is finite. Show that

$$\mathbf{Pr}\left[|X - \mathbf{E}[X]| > t \sqrt[k]{\mathbf{E}[(X - \mathbf{E}[X])^k]}\right] \leq \frac{1}{t^k}.$$

Why is it difficult to derive a similar inequality when k is odd?

7. The weak law of large numbers states that, if X_1, X_2, X_3, \dots are independent and identically distributed random variables with mean μ and standard deviation σ , then for any constant $\epsilon > 0$, we have

$$\lim_{n \rightarrow \infty} \mathbf{Pr}\left[\left|\frac{X_1 + X_2 + \dots + X_n}{n} - \mu\right| > \epsilon\right] = 0.$$

Use Chebychev's Inequality to prove the weak law of large numbers.

8. (Mitzenmacher-Upfal 2.7)
Let X and Y be independent geometric random variables, where X has parameter p and Y has parameter q .

- (a) What is $\mathbf{Pr}[X = Y]$?
- (b) What is $\mathbf{E}[\max(X, Y)]$?
- (c) What is $\mathbf{Pr}[\min(X, Y) = k]$?
- (d) What is $\mathbf{E}[X|X \leq Y]$?

Keep in mind the memoryless property of geometric random variables.

9. (Mitzenmacher-Upfal 3.10)

For geometric random variable X , find $\mathbf{E}[X^3]$ and $\mathbf{E}[X^4]$. (Use the fact that for any random variables X and Y ,

$$\mathbf{E}[X] = \sum_y \mathbf{Pr}[Y=y] \mathbf{E}[X|Y=y],$$

where the sum is over all values in the range of Y and all of the expectations exist.)

10. Let X be a non-negative integer-valued random variable with positive expectation. Prove that

$$\frac{\mathbf{E}[X]^2}{\mathbf{E}[X^2]} \leq \mathbf{Pr}[X \neq 0] \leq \mathbf{E}[X].$$

Hint: use the following special case of the Cauchy-Schwarz Inequality:

$$(a_1^2 + \cdots + a_n^2) \left(\frac{b_1^2}{a_1^2} + \cdots + \frac{b_n^2}{a_n^2} \right) \geq (b_1 + \cdots + b_n)^2$$

(First, make sure you see why this is a special case of the Cauchy-Schwarz Inequality; then apply it to get one if the inequalities of this problem.)

11. Work through the jupyter notebook `prob-nb4.ipynb`, which can be downloaded from Piazza.
 12. Work through the jupyter notebook `prob-nb5.ipynb`, which can be downloaded from Piazza.

7.6.2 Implementations and Applications of the Theory

1. Computational version of:

Aligning means and variances. Consider the following three probability distributions:

$$P1 : \mathbb{P}[X_i = a] = \mathbb{P}[X_i = b] = \frac{1}{2},$$

$P2$: Probability of X_i is uniform on the interval $[a, b]$,

$P3$: X_i is drawn from a normal/Gaussian distribution with mean μ and variance σ^2 ,

where a and b are (as yet) unspecified parameters, and μ and σ are also unspecified parameters.

- (a) For each of these three distributions, what is the mean and variance?
- (b) For each of the first two distributions, what choice(s) of a and b would have the same mean and variance as the normal/Gaussian distribution with mean μ and variance σ^2 ? (This gives two new distributions, call them $P4$ and $P5$, where a and b take on specific values.)
- (c) Plot by hand on a common set of axes all five probability distributions.

(Hint: for a random variable X that is uniform on the interval $[a, b] \subset \mathbb{R}$, sometimes denoted $X \sim U[a, b]$, it holds that $\mathbf{E}[X] = \frac{1}{2}(a+b)$ and that $\mathbf{Var}[X] = \frac{1}{12}(b-a)^2$.) XXX.

2. **Empirical coupon collecting.** We are interested in the empirical properties of the coupon collecting problem that we observe by simulating this process.

- (a) Let $n = 5$ be the number of coupons, plot your best estimate, as a function of the number of number k of coupons collected, of the probability that you win a prize.
- (b) Do the same for $n = 10$, $n = 20$, and $n = 50$.
- (c) What do you observe about the shape of this curve, as n is increased?

3. **Empirical birthday paradox.** We are interested in the empirical properties of the birthday paradox that we observe by simulating this process.

- (a) Let $n = 365$ be the number of days in a year. Plot your best estimate, as a function of the number of number m of people at the party, of the probability that there exists a pair of people with the same birthday.
- (b) Plot your best estimate, as a function of the number of number m of people at the party, of the probability that there exists three people with the same birthday.
- (c) Do the same for $n = 10$, $n = 100$, and $n = 1000$.
- (d) What do you observe about the shape of this curve, as n is increased?

4. **Estimate the probability of “tail” events.** Let X_1, \dots, X_n be independent random variables with

$$P1 : \mathbb{P}[X_i = -1] = \mathbb{P}[X_i = 1] = \frac{1}{2},$$

$$P2 : \mathbb{P}[X_i = 0] = \mathbb{P}[X_i = 1] = \frac{1}{2},$$

and, for each, let X be the number of times that 1 occurs. Let p be the probability of the event $X \geq 3n/4$. Compare theoretically, to the extent possible, the best upper/lower bounds on p that you can obtain using Markov’s Inequality, Chebyshev’s inequality, and Chernoff bounds. Then, in a notebook, simulate this process for different values of n and try to determine how close each of these bounds is to the empirically-observed results. Consider each of the following probability distributions:

- $P3$: Probability of X_i is uniform on the interval $[-1, 1]$
- $P3$: Probability of X_i is uniform on the interval $[10, 10]$
- $P3$: X_i is drawn from a normal/Gaussian distribution with mean μ and variance σ^2 .
- $P3$: X_i is drawn from a normal/Gaussian distribution with mean μ and variance $10\sigma^2$.

In a notebook, simulate this process for these distributions for different values of n , and compare with the properties you observed for the $P1$ and $P2$ distributions. XXX. I NEED TO BE A LITTLE CLEARER, COMPARE NUMBER OF HEADS AND SUM OF UNIFORM RANDOM VARIABLES WITH GAUSSIANS.

5. **Probability that the norm of a random vector is large.** XXX. NOT READY YET, MAYBE LATER. Let $x \in \mathbb{R}^n$ be a vector in which each element $x_i \in \mathbb{R}$ is a random number, distributed uniformly between 0 and 1. Observe that the Euclidean norm of such a vector is a real number such that $0 \leq \|x\|_2 \leq n$. We are interested in computing the probability that $\|x\|_2 \geq \frac{3n}{4}$.

- (a) For $n = 1$, that probability is $\frac{1}{4}$. By writing a function that computes an estimate of this probability, verify this fact computationally.
- (b) Modify the function to compute an estimate of this probability, for any particular value of n . For $n = 1, \dots, 50$, compute your best estimate of this probability, and plot the result.

XXX. I MIGHT WANT A DIFFERENT QUESTION TO ILLUSTRATE THIS QUESTION THAT TO DO A TAIL BOUND, UNLESS I MAKE THE CONNECTION WITH THE EARLIER PROBLEM MORE CLEAR, SO THIS IS A BIT OF A PLACEHOLDER.

6. **Empirical properties of different distributions.** XXX. PLOT HISTOGRAMS OF THE DISTRIBUTIONS WE DISCUSS. ALSO PLOT MEAN, MEDIAN, 90-TH PERCENTILE, ETC.

7. **Empirical properties of parameterized distributions.** XXX. PLOT PROPERITES OF MAYBE HAVE T DISTRIBUTION ALSO. XXX. MAYBE T DSTBN OR SOME OTHER HEAVY TAILED DISTRIBUTION, TO SHOW HOW GOOD OR BAD THINGS CAN BE.

8. **Empirical XXX ADITYA RAHMAN THING.** XXX. QUESTION ON ADITYA RAHMAN THING.

Chapter 8

A retrospective: Probability and high-dimensional linear algebra

8.1 Connections between probability theory and linear algebra

Probability, at least as it has so far been presented, and indeed as it is typically presented, may seem quite different than the geometry and linear algebra we discussed before. Indeed, it is very different—the axioms are very different, the types of statements one can make are different, they can be and often are used in very different ways, etc. Nevertheless, there are many strong connections between the two areas. Here, we describe the two most important.

- **Geometrizing probability.** When you work with probability, but only work with means and variances, then in some sense you are geometrizing probability and “putting it in” a Euclidean space. By that, we mean that there are natural geometric interpretations to operations like means and variances, and so one can think about probability in terms of geometric objects, with associated interpretations such as angles, orthogonality, etc. This is not the 100% “use case” of probability in data science—there are many important examples where you may be interested in other things—but this is the 99% use case.
- **Probability as a model for high-dimensional linear algebra.** When you work with linear algebra and try to reason about high-dimensional Euclidean spaces, e.g., \mathbb{R}^n , where n is more than 5 or 10, and certainly when n is more than 20 or 30, there are many subtle and counterintuitive phenomena that arise, and probability can be used as a “model” to understand these phenomena. By that, we mean that many of the properties of high-dimensional Euclidean spaces are qualitatively the same as properties of flipping n coins, throwing n darts, rolling n dice, etc. Since these latter processes are simpler and often easier to think about and analyze than high-dimensional Euclidean spaces, one can gain insight into high-dimensional Euclidean spaces from them.

Here is an important technical remark, given mostly as a caveat. Except for the uniform distribution on the unit interval and the normal/Gaussian distribution on the real line, we have been discussing discrete probability. If we try to make a precise connection between probability and high-dimensional linear algebra, then since high-dimensional Euclidean spaces are continuous objects, one must deal with continuous probability. This means that a precise statement of what we will discuss in this chapter is technically quite complex—indeed, advanced classes spend a lot of time and technical effort on this. Nevertheless, one can get the basic idea by thinking about the continuous unit interval as corresponding to a k -sided dice, where k is large enough, and realizing that the basic ideas of discrete probability go through (with considerable technical effort) to the continuous case. That is basically what you are doing when you plot the results from

Table 8.1: Correspondences between geometry and probability.

Geometry	Probability
Length squared of a vector	Variance
Length of a vector	Standard Deviation
Dot product	Covariance
Cosine of angle between two vectors	Correlation

a continuous probability distribution via a histogram; and that is sufficient to gain insight as to how these ideas appear in data science; and so we will do that.

8.1.1 A geometric approach to probability

As we have alluded to, there are strong connections between what we are discussing here about probability theory and random variables and what we saw before in linear algebra and Euclidean spaces. To summarize some of these connections, recall that Ω is the sample space of outcomes of an experiment, and a random variable is a function $X : \Omega \rightarrow \mathbb{R}$. Then,

- $\mathbf{E}[X] = \sum_{s \in \Omega} X(s) \Pr[s]$ is an expectation.
- $\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2]$ is a variance.
- $\sigma[X] = \sqrt{\mathbf{Var}[X]}$ is the standard deviation.
- $\text{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$ is the covariance.
- $\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma[X]\sigma[Y]}$ is the correlation.

Expectations of random variables obey rules that are of exactly the same form as taking linear combinations and taking scalar multiples of vectors. Thus, they can be interpreted as vectors. In addition, multiplying a random variable with itself (as in the definition of variance) or multiplying a random variable with another random variable (as in the definition of covariance) or doing this and scaling in a certain way (as in the definition of correlation) has the same form as norms and dot products and cosines of angles between vectors. We summarize these connections in Table 8.1, where we list the correspondences between geometry and probability. This connection between basic probability theory and the geometric properties of Euclidean spaces is very important in data science. Given these connections, there are many connections between probability and linear algebra.

8.1.2 A probabilistic approach to Euclidean geometry

For this, the basic intuition is that the unit interval $[0, 1]$ on \mathbb{R} is a “soft” version of the discrete set $\{0, 1\}$. If the probability distribution over each set is uniform, then they have the same mean, but they have different variances. So, let’s variance normalize one of them to have the same variance as the other. Then, if we sample from one or the other many times, then

$$[0, 1]^n \approx \{0, 1\}^n,$$

in the sense that if we consider quantities such as the mean or variance, then we should get roughly the same answer, regardless of which of the two we use. So, in particular, we can think about sampling from high-dimensional boxes such as $[0, 1]^n$ or $[-1, 1]^n$ as a soft version of flipping a coin n times or rolling a dice n times.

8.2 Properties of high dimensions versus low dimensions

If we look at points in \mathbb{R}^n , say points from the unit ℓ_2 ball or unit ℓ_∞ ball, then as n gets large, several things happen that are qualitatively different than what we see in \mathbb{R} , \mathbb{R}^2 , and \mathbb{R}^3 :

- **Distances.** The distances between pairs of points in the n -dimensional ℓ_2 ball “concentrate” near the average distance.
Informally: in high dimensions, nearly all pairs of points are approximately the same distance from each other.
- **Angles.** The angles between pairs of points (and the origin) in the n -dimensional ℓ_2 ball “concentrate” near 90° .
Informally: in high dimensions, nearly all pairs of points (viewed as vectors with respect to the origin) are approximately perpendicular to each other.
- **Probability mass.** Most of the probability mass in the n -dimensional ℓ_2 ball “concentrates” in a small shell near the surface of the ℓ_2 ball.
Informally, in high dimensions, nearly all points inside the unit ball are approximately in a shell near the surface.
- **Probability mass.** Most of the probability mass in the unit n -dimensional ℓ_∞ ball “concentrates” outside the unit n -dimensional ℓ_2 ball.
Informally, in high dimensions, nearly all of the probability mass of the unit box lies in the corners outside the unit ball, and thus the ball occupies a vanishingly small fraction of the volume of the minimum enclosing box.
- **Probability mass.** Most of the probability mass of the half-unit n -dimensional ℓ_∞ ball “concentrates” outside the unit n -dimensional ℓ_2 ball.
Informally, in high dimensions, nearly all of the probability mass of the half-unit box lies in the corners outside the unit ball.

Here is a good video: <https://www.youtube.com/watch?v=zwAD6dRSVYI&feature=youtu.be>

So, while we motivated linear algebra by showing formally that many of the properties of \mathbb{R} , \mathbb{R}^2 , and \mathbb{R}^3 (e.g., distances, angles, balls/boxes) can be generalized to \mathbb{R}^n , which is true and which underlies many of the techniques used in data science, these results show that many of the properties of distances, angles, and balls/boxes in high-dimensional spaces are *very* different than low-dimensional spaces.

Most of these properties are related to the so-called *curse of dimensionality*, which is related to the phenomenon of *measure concentration*.

Question: Why do high-dimensional Euclidean spaces have these peculiar properties?

Answer: To explain this in full detail/generality requires a lot of mathematics that goes well beyond what we will cover in this class, but this is a sufficiently ubiquitous and important phenomenon that one should have a basic informal understanding of it. To that end, here we will “explain” the basic ideas underlying why these properties hold by focusing on a few special but representative cases. In particular, given the connection we made between linear algebra and probability as well as the intuition we have that a fair coin lands on heads roughly 50% of time, we will “explain” the ideas by drawing insight from coin flipping. Importantly, this is much more than an analogy. There are strong and deep technical connections between these two areas (that we will only hint at in this class).

8.3 More on measure, concentration, and measure concentration

8.3.1 Concentration in flipping coins

Flipping coins a few or many times. The basic issue is the following. If we flip a fair coin twice:

- **Most likely fraction.** It is most likely we will get exactly 50% H. (This is as opposed to some other fraction of Hs, which in the case of two flips is just 0% or 100% H.)
 - **How likely is the most likely.** It is pretty likely that we will get exactly 50% H. (This happens 50% of the time.)
 - **How likely is very far from the most likely.** It is somewhat likely that we will get “far” from 50% H. (For two flips, this means, all H or all T, each of which happens 25% of the time, for a total of 50% of the time.)

On the other hand, if we flip a fair coin 100 times:

- **Most likely fraction.** It is most likely that we will get exactly 50% H. (That is, this is more likely than either 49% or 51% or 48% or 93% or 0%, etc.)
 - **How likely is the most likely.** It is very unlikely that we will get exactly 50% H, but it is very likely that we will get “near” or “close to” 50% H. (By this, we mean that while the probability of exactly 50% H is quite small, the sum of 50% plus 49% plus 51% plus 48%, and other values close to 50%, is quite large.) We will make “near” or “close to” more precise later.
 - **How likely is very far from the most likely.** It is very unlikely that we will get very far from 50% H. (For example, it is extremely unlikely that we will get all H or all T or 90% H, etc.)

That is, if we consider the 2^n possible ordered configurations of flips, obtained when we flip a coin n times (where for simplicity let's say that it's a fair coin so each of the 2^n possibilities is equally likely, i.e., the probability mass on the 2^n possible ordered configurations is uniform), then if we ask ourselves how likely each of the possible $n+1$ events corresponding to obtaining k Hs (which, recall, form a partition of the set of 2^n possible ordered configurations), then the probability distribution over those $n+1$ events is very very nonuniform.

Pascal's Triangle. To get a better sense of all of this, here is Pascal's Triangle, which gives the relative proportions for different numbers n of flips.

$n = 0:$	1										
$n = 1:$	1 1										
$n = 2:$	1 2 1										
$n = 3:$	1 3 3 1										
$n = 4:$	1 4 6 4 1										
$n = 5:$	1 5 10 10 5 1										
$n = 6:$	1 6 15 20 15 6 1										
$n = 7:$	1 7 21 35 35 21 7 1										
$n = 8:$	1 8 28 56 70 56 28 8 1										
$n = 9:$	1 9 36 84 126 126 84 36 9 1										
$n = 10:$	1	10	45	120	210	252	210	120	45	10	1

Remember that $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, where $n! = n(n-1)(n-2)\cdots 2 \cdot 1$, is the number of ways to choose a subset of k elements from n elements, disregarding order. Observe that, e.g., in the $n = 6$ line:

- there is only $\binom{n}{0} = 1$ way to get six Hs;
- there are $\binom{n}{1} = 6$ ways to get five Hs and one T;
- there are $\binom{n}{2} = \frac{n(n-1)}{2!} = \frac{6 \times 5}{2} = 15$ ways to get four Hs and two Ts;
- there are $\binom{n}{3} = \frac{n(n-1)(n-2)}{3!} = \frac{6 \times 5 \times 4}{6} = 20$ ways to get three Hs and three Ts;
- and so on.

Similarly, for other values of n .

This should be familiar, since Pascal's triangle determines the coefficients which arise in binomial expansions. For example, $(x+y)^2 = 1x^2 + 2xy + 1y^2$ and $(x+y)^3 = 1x^3 + 3x^2y + 3xy^2 + 1y^3$, where the coefficients in this expansion are in the $n = 2$ and $n = 3$ rows, respectively, of the triangle.

Note that the sum of the numbers in the n^{th} row equals 2^n , i.e.,

$$\sum_{i=1}^n \binom{n}{i} = 2^n. \quad (8.1)$$

That is, if one sums up the $n+1$ number of Hs, each multiplied by the number of times that number of possible ways to get that number of Hs, then one obtains 2^n , which we know is $\left(\frac{1}{2}\right)^n$ which is the total number of possible sequences of H/T.

Let's fix n , i.e., let's consider only a given number n of coin flips, and let's ask what is the probability that we get a given number of Hs. The numbers in a given row of Pascal's triangle are not probabilities—although they are not less than 0, they are in general greater than 1, and they do not sum to 1. But, if we divide the numbers in the n^{th} row by 2^n , then we obtain a probability distribution. The elements of this probability distribution equal the probability of flipping the corresponding number of Hs in n flips.

Here is the table of probabilities for n up to $n = 10$.

$n = 0:$	1					
$n = 1:$.500 .500					
$n = 2:$.250 .500 .250					
$n = 3:$.125 .375 .375 .125					
$n = 4:$.063 .250 .375 .250 .063					
$n = 5:$.031 .156 .313 .313 .156 .031					
$n = 6:$.016 .094 .234 .313 .234 .094					
$n = 7:$.008 .055 .164 .273 .273 .164 .055					
$n = 8:$.004 .031 .109 .219 .273 .219 .109					
$n = 9:$.002 .018 .070 .164 .246 .246 .164					
$n = 10:$.001 .010 .044 .117 .205 .246 .205 .117 .044					

Back to flipping coins many times. We will get back to this in more detail later, but at this point, it should be believable that as the number of flips gets large:

- **Most likely fraction.** The maximum of this distribution is at 50% Hs (if the number of flips is even) or the integer nearest 50% Hs (if the number of flips is odd).
- **How likely is the most likely.** Obtaining exactly 50% Hs (or the nearest integer if the number of flips is odd) is not going to be particularly likely, but obtaining nearly 50% Hs is going to be extremely common.
- **How likely is very far from the most likely.** Obtaining a number that is very far from 50% Hs is going to be extremely uncommon.

Remark. This phenomenon is called *measure concentration* since the probability mass (technically, it is called a measure) concentrates on a very small number of possible values. For example, for $n = 10$, if we consider each of the possible sequences of $2^{10} = 1024$ coin flips to be equally probable, i.e., to have a uniform probability distribution, then we obtain a very nonuniform distribution over $\{0, 1, \dots, 10\}$, the number of possible Hs. In this case, the most likely event (that we obtain 5 Hs) is 252 times more likely than the least likely event (that we obtain 0 Hs or that we obtain 10 Hs). The effect becomes *much* more pronounced as n gets larger.

Remark. There are many “cute” relationships involving expressions of the form $\binom{n}{k}$, and many of them have a natural interpretation. For example, Eqn. (8.1), and also Pascal’s identity:

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Importantly, there does not exist such a nice closed-form solution for the partial sums

$$\sum_{j=0}^k \binom{n}{j},$$

from which we could immediately get “tail estimates.” (Instead, what is typically done is to exploit connections between this and the normal distribution.)

8.3.2 Concentration in throwing darts

You might be wondering:

- What does this have to do with balls and boxes in high-dimensional Euclidean spaces?

Informally, the answer is that the continuous set $[0, 1]$ can be thought of as a “soft” version of the discrete set $\{0, 1\}$, in which case $[0, 1]^n \approx \{0, 1\}^n$. In somewhat more detail (but the details and precise results depend on mean centering, etc., so we will be a little cavalier about that for now, and we will get to some special cases later), the answer is the following.

Consider $x, y \in \mathbb{R}^n$, where by \mathbb{R}^n actually we mean in $[0, 1]^n$ (or $[-1, +1]^n$ or $[-\frac{1}{2}, +\frac{1}{2}]^n$). Then, consider the distance, as measured in the Euclidean norm, between those two vectors:

$$\|x - y\|_2 = \left(\sum_{i=1}^n (x_i - y_i)^2 \right)^{1/2}. \quad (8.2)$$

While this is a distance between two particular vectors, x and y , we can think of those two vectors as random vectors (i.e., random variables that are vectors and not numbers). The reason for this is that they were sampled uniformly-at-random from the ℓ_∞ box. (This is a generalization of choosing points from $[-1, 1] \subset \mathbb{R}$.) In this case, this is just a random variable that is equal to the distance between two random vectors. That is, the difference $x - y$ is itself a random vector, with random elements $z_i = x_i - y_i$.

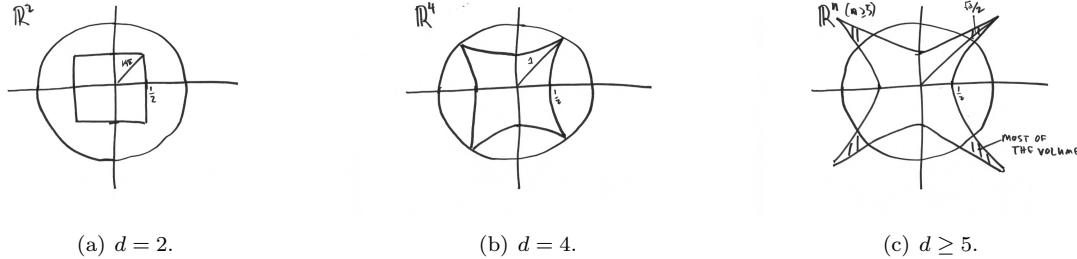


Figure 8.1: Illustration of the relationship between the half-unit box and the unit ball.

Viewed this way, the expression given on the RHS of Equation (8.2) is itself the sum of independent random variables z_i , each of which has a mean and a bounded variance (we will discuss later the significance of this comment about the mean and variance). Informally, this is “like” coin flipping, which is instead the sum of $\{0, 1\}^n$ random variables with a mean of 0.5 and a bounded variance, but instead here we are considering the sum of $[0, 1]^n$ random variables with a mean of 0.5 and a (slightly different) bounded variance. (We didn’t emphasize it, but an important aspect of flipping coins is that the mean and variance of the random variable corresponding to each flip is known and bounded. We will get back to why this is so soon.)

So, we can think of the distance between two points randomly sampled from a high-dimensional box as being like a “soft” version of coin flipping. In the same way as we can sum up $\{0, 1\}$ random variables, i.e., coin flips, to get the total number of Hs, we can sum up $[0, 1]$ random variables, i.e., continuous/soft versions of them to get things like distances between high-dimensional random vectors. That is, the equation for distance, Equation (8.2) above, is just a sum of random variables that are “soft” versions of $\{0, 1\}$ random variables. When we do this, the exact quantitative results are usually a little different (because of slight differences between the mean and/or variance, both of which we will see can be easily corrected for), but we can typically get the main quantitative insights.

In particular, this can help us explain what we saw:

- Why distances are approximately the same.
- Why angles are approximately perpendicular.
- Why most of the mass is in a shell near the surface.
- Why the center of the ball is the most probable, but is still quite rare.

The half-unit box. Let’s go into more detail about throwing darts at the half-unit box. To be a little more precise, consider

- A cube, or ℓ_∞ ball, centered at the origin with half-unit length radius, i.e., $[-\frac{1}{2}, +\frac{1}{2}]^n$, and
- A sphere, or ℓ_2 ball, centered at the origin with unit radius.

Note that this is *not* the minimum enclosing box like we used before. The minimum enclosing box was $[-1, +1]^n$, i.e., the ℓ_∞ ball with unit-length radius and twice-unit length sides. This half-unit box is half as long as the unit box *in every direction*. In two-dimensions, this is clearly completely contained in the unit ball (see Figure 8.1(a)), and we want to see what happens in higher dimensions.

In 2-D, the half-unit box is completely inside the unit sphere:

- $\text{dist}(\text{center, side along axis}) = \frac{1}{2}$

- $\text{dist}(\text{center}, \text{corner vertex}) = \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right)^{1/2} = \frac{1}{\sqrt{2}} \approx 0.707 < 1$

This is illustrated in Figure 8.1(a).

What about higher dimensions?

In 3-D, the half-unit box is still completely inside the unit sphere:

- $\text{dist}(\text{center}, \text{side along axis}) = \frac{1}{2}$
- $\text{dist}(\text{center}, \text{corner vertex}) = \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right)^{1/2} = \frac{\sqrt{3}}{2} \approx 0.866 < 1$

Note, however, since 3 dimensions is more than 2 dimensions, the distance between the center of the box and a corner vertex involves a sum of more components, and thus it is larger ($0.866 > 0.707$).

In 4-D, the sphere is still completely inside the circle, but barely:

- $\text{dist}(\text{center}, \text{side along axis}) = \frac{1}{2}$
- $\text{dist}(\text{center}, \text{corner vertex}) = \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right)^{1/2} = 1$

That is, in 4 dimensions, the distance between the center of the box and a corner vertex involves a sum of more components, 4 rather than 3, and thus it is larger still. In this case, it exactly equals 1, and thus the corner of the cube touches the sphere. This is illustrated in Figure 8.1(b).

In dimension ≥ 5 , the corners of the cube are *outside* the sphere. For example, in 5-D:

- $\text{dist}(\text{center}, \text{side along axis}) = \frac{1}{2}$
- $\text{dist}(\text{center}, \text{corner vertex}) = \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right)^{1/2} = \frac{\sqrt{5}}{2} \approx 1.118 > 1$

Thus, in 5 dimensions, there are points in the box that are not in the sphere. This is illustrated in Figure 8.1(c).

In higher dimensions, this effect gets only more prominent. For example, in n dimensions,

- $\text{dist}(\text{center}, \text{side along axis}) = \frac{1}{2}$
- $\text{dist}(\text{center}, \text{corner vertex}) = \left(\left(\sum_{i=1}^n \frac{1}{2}\right)^2 \right)^{1/2} = \frac{\sqrt{n}}{2} > 1, \text{ if } n \geq 5.$

This is because, in general, in n -dimensional space, the vertices are distance $\frac{\sqrt{n}}{2} = \left(\left(\sum_{i=1}^n \frac{1}{2}\right)^2 \right)^{1/2}$ from the origin.

Remark. Figure 8.1(b) and Figure 8.1(c) illustrate this phenomena that the high-dimensional box gets very “spiky,” in that most of the mass is in the corners. It is very important to note that while these pictures illustrate certain things, since we are showing n -dimensional balls/cubes in 2-dimensions, these pictures also mis-illustrate certain things. In particular, although the cube is “spiky,” it is also “round,” in that one can draw a line segment between any two points in the cube such that the entire line segment between those two points stays inside the cube (technically, it is *not* non-convex), and this is mis-illustrated in Figure 8.1(b) and Figure 8.1(c). Here is the lesson: visualizing high-dimensional Euclidean spaces is not easy.

Remark. We have illustrated the same phenomenon in several different ways.

- Most of the mass of the half-unit box is in the corners of the box and outside the unit ball.
- The overall mass of the unit ball is much smaller than the overall mass of the unit box.
- Most of the mass of the sphere is in a small shell near the surface of the sphere.

These are all illustrating the same peculiar properties of high-dimensional Euclidean spaces. If one of these makes more sense to you than the others, then go with whichever one makes more sense—they are, after all, two (or three) sides of the same coin ;)

8.4 Advanced Aside: Insights from calculus

If you are familiar with calculus, then the following discussion might be helpful. (If not, then ignore the following in this subsection.)

Consider the volume (V) and surface area (A) of a sphere (i.e., an ℓ_2 ball) in different dimensions n . For a given n , we have that $V(\text{sphere}) \sim r^n$ and $A(\text{sphere}) \sim r^{n-1}$. The following table gives the values for 1, 2, and 3 dimensions.

Dimension	Volume of sphere	Surface Area of sphere
1	$2r$	2
2	πr^2	$2\pi r$
3	$\frac{4}{3}\pi r^3$	$4\pi r^2$
\vdots	\vdots	\vdots

Observe that

$$V(r) = \int_0^r A(r') dr',$$

and that

$$A(r) = \frac{dV(r)}{dr}.$$

The question has to do with what is the relationship between the volume and the surface area. This is the well-known *isoperimetric problem*. One version of this problem asks: what is the shape on \mathbb{R}^2 that, for a given surface area (circumference in \mathbb{R}^2), achieves the largest volume (area in \mathbb{R}^2)? The answer is a circle. Another version of this problem asks: what is the shape on \mathbb{R}^2 that, for a given volume (area in \mathbb{R}^2) has the smallest surface area (circumference in \mathbb{R}^2)? Again, the answer is a circle. In both cases, an analogous question can be asked in \mathbb{R}^3 , and in \mathbb{R}^3 the answer is a sphere, or any other \mathbb{R}^n . Here, we are interested in a slightly different variant of this problem. We are interested in understanding how the answer changes as we increase the dimension.

Here is an intuition from calculus involving integrals. Recall that to do a two-dimensional integral, you can specify the volume increment dA in one of two ways: either as $dxdy$ or as $rdrd\theta$, in which case one writes an integral as

$$\int f(x, y) dxdy = \int f(r, \theta) rdrd\theta.$$

If $dxdy$ is the volume increment, then $dxdy$ sweeps out the same volume (“area” in two dimensions) increment, regardless of the value of x, y .

Question: How much area does $d\theta$ sweep out?

Answer: It depends on the distance from the origin or radius as r, \dots as $rdrd\theta$.

This is in \mathbb{R}^2 , where the amount of time it takes to walk around the circumference depends on the radius of the circle. In \mathbb{R}^3 , not only must we sweep around the equator, but we must sweep all the other angles. This is even bigger.

Question: How much bigger?

Answer: How much bigger is given by $r^2 d\Omega$, where $d\Omega$ is a differential angle expression that we won't discuss in detail (since we are interested in the dependence on dimension).

So, we have the following.

- \mathbb{R}^2 : Circle: the length of a string to go around the equator of the earth grows linearly with r .
- \mathbb{R}^3 : Sphere: the amount of paint to paint the entire earth grows quadratically with r .
- In \mathbb{R}^n , we must sweep out even more. How much is given by $r^{n-1} d\Omega$.

The point here is that as the dimension n increases, the amount of mass at a given distance from the origin increases as r^{n-1} , i.e., very rapidly.

8.5 Problems

8.5.1 Pencil-and-paper Problems

1. **Probability and geometry.** Suppose that a probability space Ω consists of n outcomes, $\{1, 2, \dots, n\}$, each with probability $1/n$. Then a random function X on Ω (i.e., a random variable on a probability space with n elements) can be identified with an element $X \in \mathbb{R}^n$ (i.e., a vector in an n -dimensional Euclidean space), as follows.

- Recall the definition of a vector space, in particular the “addition of two vectors” and the “multiplication of a vector by a scalar” conditions. Show that the expectation function, viewed as a function that takes as input a set of numbers and returns as output a single number, satisfies these two conditions.
- Show that

$$\mathbf{E}[X] = \frac{1}{n} (X \cdot 1) = \frac{1}{n} (X^T 1),$$

where $1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n$, and where \cdot represents the dot product operation, or equivalently where

$X^T 1$ represents the inner product between X and 1 .

- Describe in words what is $X - \mathbf{E}[X] 1$, where X and 1 are vectors in \mathbb{R}^n , and where $\mathbf{E}[X] \in \mathbb{R}$. Describe in words what is $\frac{1}{\sigma(X)}(X - \mathbf{E}[X] 1)$.
- Then, show that

$$\mathbf{Var}[X] = \frac{1}{n} \|X - \mathbf{E}[X] 1\|_2^2 \quad \text{and} \quad \sigma[X] = \frac{1}{\sqrt{n}} \|X - \mathbf{E}[X] 1\|_2.$$

- Recall the definitions of covariance and correlation coefficient:

$$\begin{aligned} \mathbf{Cov}[X, Y] &= \mathbf{E}[(X - \mathbf{E}[X] 1)(Y - \mathbf{E}[Y] 1)] \\ \mathbf{Corr}[X, Y] &= \frac{\mathbf{Cov}[X, Y]}{\sigma(X)\sigma(Y)}. \end{aligned}$$

Show that

$$\begin{aligned}\mathbf{Cov}[X, Y] &= \frac{1}{n} (X - \mathbf{E}[X] \mathbf{1}) \cdot (Y - \mathbf{E}[Y] \mathbf{1}) \\ \mathbf{Corr}[X, Y] &= \cos(\theta),\end{aligned}$$

where θ is the angles between the vectors $X - \mathbf{E}[X] \mathbf{1}$ and $Y - \mathbf{E}[Y] \mathbf{1}$.

2. (a) Prove that $\binom{n}{k} = \binom{n}{n-k}$.
- (b) Prove that if $k \leq \frac{n}{2} - 1$ then $\binom{n}{k-1} \leq \binom{n}{k}$; and if $k \geq \frac{n}{2}$ then $\binom{n}{k} \geq \binom{n}{k+1}$ (i.e., that the sequence of binomial coefficients increases until the middle one, and then it decreases).
- (c) Prove Pascal's identity

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

(Hint: use the factorial formula, and count the subsets of size k with elements in $\{1, 2, \dots, n\}$ in two ways.)

3. Prove that if the numbers in the n^{th} row of Pascal's triangle are divided by 2^n , then we obtain a probability distribution that equals the probability of flipping the corresponding number of Hs in n flips of a fair coin. Hint: prove that the numbers $\{p_i\}_{i=0}^n$, where

$$p_i = \frac{1}{2^n} \binom{n}{i}$$

satisfy the three conditions for a function to be a probability distribution.

4. Let $x \in \mathbb{R}^n$ be a random vector, in which each element $x_i \sim U[-1, 1]$.
 - (a) What is $\mathbf{E}[\|x\|_2^2]$? (Hint: consider $\mathbf{E}[x_i]$ and $\mathbf{E}[x_i^2]$ in turn.)
 - (b) What is $\mathbf{Var}[\|x\|_2^2]$?
 - (c) Let y be the projection of x onto the first standard basis vector; what is $\mathbf{E}[\|y\|_2^2]$?
 - (d) Let z be the projection of x onto the all-ones vector; what is $\mathbf{E}[\|z\|_2^2]$?
5. XXX. ANOTHER JL-TYPE QUESTION. MAYBE THAT RANDOM PROJECTIONS PRESERVE NORMS OF VECTORS AND/OR THAT IF YOU KEEP ONE COMPONENT OF A RANDOM VECTOR IT HAS AN EXPECTED AND CONCENTRATED VALUE AND/OR DITTO IF YOU PROJECT IT ON A RANDOM DIRECTION. HOW SHOULD I DEAL WITH SPHERICAL SYMMETRY.
6. XXX. MAYBE HAVE A ONE OR TWO OTHER NON-COMPUTATIONAL HWS.

8.5.2 Implementations and Applications of the Theory

1. **Sphere (L_2 ball) and cube (L_∞ ball) in higher dimensions.** Choose $N = 100$ random points in a square centered at the origin and plot the points with a scatter plot.
 - Compute the fraction of points that fall within the circle of radius equal to half the side of the square centered at the center of the square, and use this to get an estimate of π .
 - Do this $k = 10$ times to compute k such estimates. Compute the mean and variance of those k numbers.
 - Do the same with $N = 1000$, $N = 10,000$, and $N = 100,000$. What number do you approach as N gets large? What number do you expect it to approach?

Now, instead of throwing darts at a 2D square and estimating the volume of the unit circle inside that square, estimate the volume of the unit 3D sphere inside the bounding 3D box.

- Choose N and k such that when you compute the mean and variance of the estimates you get reliable estimates.

Now, do the same in 5D, 10D, and 20D, i.e., estimate the volume of the unit n-D sphere inside the bounding n-D box, for $n = 5, 10, 20$.

- The formula for the volume of the unit n-D sphere inside the bounding n-D box is known for every value of n . What is it, and how does it compare to your estimate?
2. **Properties of high-dimensional L_2 balls.** In this problem, we will illustrate computationally several of the counterintuitive properties of high-dimensional balls. XXX. WE NEED TO GIVE A ROUTINE TO GENERATE A RANDOM POINT FROM THE HIGH DIMENSIONAL BALL.
- Distribution of distances.** Select points randomly from the unit interval $[-1, 1]$ on \mathbb{R} , and compute the distances between all pairs of points, and plot a histogram of the distances. Sample enough points until the variability in the histogram is small enough that you are confident that the histogram is close to the correct answer. Do the same for points in the unit ball/sphere on \mathbb{R}^2 . Do the same for points in the unit ball/sphere on \mathbb{R}^n , for $n = 1, 2, \dots, 20$. For each n , choose parameters so that the final histogram you present is fairly smooth. What do you notice?
 - Distribution of angles.** Do the same setup as in the previous question. In fact, you can use almost all the same code. Compute the angles between all pairs of points and plot a histogram of angles. Again, vary n , as before, and choose parameters to get fairly smooth histograms. What do you notice? Compute also a histogram of the angles between the random points and e_1 . What do you notice?
 - Distance to the center of a sphere.** Select points randomly from a unit radius circle and from a unit-radius sphere in \mathbb{R}^n , for $n = 1, 2, \dots, 20$. Plot a histogram of the distance between the chosen points and the origin, for different dimensions. (If your code for “Distribution of distances” is modularized into a function, this problem should be very easy, since you can ask for the distance between each random point and the origin.) Choose enough points so that the histograms are smooth. What do you notice?
 - Mass near the equator of a sphere.** Do the same setup as in the previous question. In fact, you can use almost all the same code. Plot a histogram of the value of the first coordinate x_1 for different dimensions. Do the same for x_2 and x_3 . Q: Is there an easy way we can project onto an arbitrary vector, perhaps by giving them a routine? What do you notice?

3. XXX. NOW CUT RADIUS OF BOX IN HALF AND THROW DARTS AT THE BALL AND ASK HOW OFTEN DO YOU HIT THE BOX.

- XXX. GENERATE SAME PLOTS BUT WE NEED THE DIMENSIONS TO BE SLIGHTLY LARGER.

XXX. NOTE THIS IS A SEPARATE QUESTION, SINCE IT ISNT JUST ABOUT L2 BALLS BUT IT NEEDS THE ROUTINE TO GENERATE POINTS FROM A HIGH-DIM SPHERE. This is basically to illustrate Figure 8.1.

4. **Classification in low dimensions.** This problem will consider a toy model for classification and ask how easy/hard it is to classify points in *low* dimensions. Consider points on the plane.

- Choose point in \mathbb{R}^2 according to the following distribution: with probability 50%, choose a point from ProcessA; and otherwise, choose a point from ProcessB. Here, ProcessA a normal/Gaussian distribution $N(\mu_A, \sigma_A^2)$, centered at the $\mu_A = (0, 0) \in \mathbb{R}^2$ with variance $\sigma_A^2 = 1$; and ProcessB is a normal/Gaussian distribution $N(\mu_B, \sigma_B^2)$, centered at the $\mu_B = (x_1^*, 0) \in \mathbb{R}^2$ with variance $\sigma_B^2 = 1$. Plot a scatter plot of these points for $x_1^* = 0, 1, 2, \dots$ until the two clusters are well-separated. In these plots, color-code points into two different colors, depending on whether the points came from ProcessA or ProcessB.

- Next, draw a point from that same distribution. Call it $x_{testpoint}$. Pretend that you don't know whether it came from ProcessA or ProcessB, and use the following "nearest neighbor" rule to make a prediction: if the $\|x_{testpoint} - \mu_A\|_2 < \|x_{testpoint} - \mu_B\|_2$, i.e., if $x_{testpoint}$ is closer to the center of ProcessA, then predict that it was generated by ProcessA; otherwise, predict that it was generated by ProcessB. Plot a scatter plot of these points for the same values of $x_1^* = 0, 1, 2, \dots$ as above, and color-code the points into two different colors, depending on your prediction.
 - Discuss how do the two plots compare, e.g., how often are your predictions correct, as a function of the various parameters of the problem.
 - Define the misclassification rate to be the number of incorrect predictions divided by the total number of predictions. Compute an estimate for the misclassification rate. How does it vary as a function of the various parameters of the problem. XXX. DO FOR GAUSSIANS, AND THE DO FOR RADEMACHER.
 - Say that it is extremely important not to make an incorrect prediction for points from ProcessA. What is a trivial classification rule that will guarantee that no points from ProcessA will be incorrectly predicted? Now say that it is very important not to make an incorrect prediction for points from ProcessA, but there is still a small penalty if you make an incorrect prediction for points from ProcessB. Describe a classification rule you could use to interpolate between the trivial classification rule and the nearest neighbor rule? It's okay here just to describe the results qualitatively.
5. **Classification in high dimensions.** This problem will consider a toy model for classification and ask how easy/hard it is to classify points in *high* dimensions. (This will be related to the question: how much do two high-dimensional distributions overlap?) Consider the same setup as the previous problem, except that now we sample points from \mathbb{R}^n , with $n = 2, 3, \dots, 20$. In particular, ProcessA has mean $\mu_A = (0, \dots, 0) \in \mathbb{R}^n$ and unit variance, and ProcessB has mean $\mu_B = (x_1^*, 0, \dots, 0) \in \mathbb{R}^n$.
- Set the parameters the same as in the previous problem, and plot your best estimate of the misclassification rate as a function of x_1^* for each value of the dimension n . Describe what this means for how much the points from ProcessA and ProcessB overlap.
 - Assume that $\sigma^2 = \sigma_A^2 = \sigma_B^2$. Now, for each n , vary σ^2 so that you get a misclassification rate that is the same as the previous problem for that value of x_1^* . Plot how does σ^2 change as a function of the dimension n ? Describe what does this mean in terms of ProcessA and ProcessB, i.e., what does this mean about the assumptions you need to make on those processes to get good classification?

Part IV

The Spectral Theorem: The Central Result in Linear Algebra

Chapter 9

Eigendecompositions: Eigenvectors and Eigenvalues

9.1 Overview of the chapter

We have seen that linear algebra provides a formal way to generalize many of the intuitive geometric notions from \mathbb{R}^2 and \mathbb{R}^3 to \mathbb{R}^n , for arbitrary n (and that this might be useful for data science applications), but that \mathbb{R}^n , for large n , has many deeply-counterintuitive properties (which might make it difficult to use these generalizations for data science applications) that can be partially-understood in terms of basic probability. Thus, one might wonder whether and how one can think productively about \mathbb{R}^n and data that are modeled by vectors in \mathbb{R}^n .

There are, in fact, many algorithmic and statistical methods appropriate for such data—we will cover several of the most important in later chapters—and nearly all take advantage of a rather remarkable property of \mathbb{R}^n that is very general but that for symmetric matrices says roughly the following:

- For every $n \times n$ symmetric matrix, there exists an orthonormal basis which is determined by the matrix (and which is not typically the standard basis); and that basis is essentially the standard basis rotated (actually orthogonally-transformed) in some way; and when viewed with respect to that basis, the matrix is a diagonal matrix.

That basis consists of things called eigenvectors, and the elements of that diagonal matrix consist of things called eigenvalues. Importantly, the restriction to symmetric matrices is not too important—as we will discuss, when non-symmetric matrices arise, as is common, one typically can consider related symmetric matrices and work with them.

Eigenvectors and eigenvalues are properties of matrices that are very important in many different application areas, including many parts of data science. While they are clearly linear algebraic notions, they also have strong connections with probability theory, most notably through correlation/covariance matrices and a procedure known as PCA (Principal Component Analysis) and the related SVD (Singular Value Decomposition). During the next few classes, we will provide an introduction to them. After describing them and their basic properties, we will give several examples of how they are used in data science, including PCA/SVD, LS (least-squares), linear equation solving, PR (PageRank), high-dimensional integration and differentiation, etc.

9.2 Introduction to eigenvectors and eigenvalues

Recap/overview of applying matrices to vectors. Given an arbitrary $m \times n$ matrix A , which naïvely may be viewed as a set of mn numbers, or less naïvely may be viewed as a set of m vectors in \mathbb{R}^n or n

vectors in \mathbb{R}^m , one way to learn about the properties of that matrix is to consider the action of that matrix on vectors. This can help us understand the “shape” of the matrix in \mathbb{R}^n .

We have seen several examples of considering the action of a matrix on a vector.

- Applying an $m \times n$ matrix A to a canonical vector e_i gives the i^{th} column of A (if we post-multiply by $e_i \in \mathbb{R}^n$, i.e., if we compute Ae_i —on the other hand, it gives the i^{th} row of A if we premultiply by $e_i \in \mathbb{R}^m$ as $e_i^T A$).
- If an $m \times n$ matrix A is a term-document matrix, with the i^{th} row of A representing the i^{th} document, then for some new document $y \in \mathbb{R}^n$, the product Ay gives information on how close (e.g., in the sense of angles, if A and y are normalized properly) the document y is to each of the documents described by the rows of A .
- If an $n \times n$ matrix A is the adjacency matrix of a graph (i.e., A_{ij} equals 1 or 0 depending on whether or not there is an edge between node i and node j) and if $p \in \mathbb{R}^n$ is a probability distribution over $[n] = \{1, 2, \dots, n\}$, representing the amount of probability mass on each of the n nodes (i.e., $p_i \in [0, 1]$ and $\sum_{i=1}^n p_i = 1$), then $p' = Ap$ describes the probability distribution after one step of a random walk on the graph.

In general, if one chooses an arbitrary $n \times n$ matrix A and an arbitrary n -dimensional vector x , then the vector $y = Ax$ that equals the product of that matrix and that vector will point in some other direction than the original vector. (Of course, if the matrix is an $m \times n$ matrix, for if $m \neq n$, then the two vectors will have different dimensions and won’t even be in the same space.) That is, the matrix-vector product won’t be exactly the original vector, and it won’t be equal to a scalar multiple of the original vector, but instead it will be in some other direction. That is, in general it is not the case that $y = \lambda x$, for some $\lambda \in \mathbb{R}$. There are, however, certain vectors x for which the matrix-vector product points in the same “direction” as that vector, i.e., where there exists a $\lambda \in \mathbb{R}$ such that $y = \lambda x$. If we view a matrix as a linear transformation, i.e., as a linear function of the input vector, then this says that the direction of the input vector is unchanged by the action of the linear transformation. Here are several examples.

- Consider the matrix

$$A = \begin{pmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{pmatrix}.$$

If one applies A to the first coordinate vector, i.e., $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, then A changes the direction of it:

$$A \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.8 \\ 0.2 \end{pmatrix} \neq \lambda \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \text{for any } \lambda \in \mathbb{R}.$$

Similarly, if one applies A to $\begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, then A also changes direction of it:

$$A \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1.1 \\ 0.9 \end{pmatrix} \neq \lambda \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \text{for any } \lambda \in \mathbb{R}.$$

On the other hand, if one applies A to the vector $\begin{pmatrix} 6/10 \\ 4/10 \end{pmatrix}$, then one gets the following:

$$A \begin{pmatrix} 6/10 \\ 4/10 \end{pmatrix} = \frac{1}{10} \begin{pmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{pmatrix} \begin{pmatrix} 6 \\ 4 \end{pmatrix} = \frac{1}{10} \begin{pmatrix} 4.8 + 1.2 \\ 1.2 + 2.8 \end{pmatrix} = \frac{1}{10} \begin{pmatrix} 6 \\ 4 \end{pmatrix},$$

i.e., the direction (as well as in this case the magnitude) of the vector is unchanged. In this case, the magnitude of the vector is also unchanged, but the magnitude could also have changed and we would still be interested in that.

- Consider the matrix

$$A = \begin{pmatrix} -3 & 4 \\ 4 & 3 \end{pmatrix}.$$

If one applies A to the first coordinate vector, i.e., $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, then A changes the direction of it:

$$A \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -3 & 4 \\ 4 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -3 \\ 4 \end{pmatrix} \neq \lambda \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \text{for any } \lambda \in \mathbb{R}.$$

Similarly, if one applies A to $\begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, then A also changes direction of it:

$$A \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} -3 & 4 \\ 4 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 7 \end{pmatrix} \neq \lambda \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \text{for any } \lambda \in \mathbb{R}.$$

On the other hand, if one applies A to the vector $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$, then one gets the following:

$$A \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} -3 & 4 \\ 4 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} -3+8 \\ 4+6 \end{pmatrix} = \begin{pmatrix} 5 \\ 10 \end{pmatrix} = 5 \begin{pmatrix} 1 \\ 2 \end{pmatrix},$$

i.e., the vector is simply rescaled, and the direction of the vector (more precisely, the subspace defined by the span of vector) is unchanged.

Vectors that are unchanged by a matrix. For a given $n \times n$ matrix A , vectors x such that the action of the matrix doesn't change the direction of the vector, i.e., it leaves it unchanged (stretch factor $\lambda = 1$) or simply stretches it out (with a stretch factor < 1 or > 1) are of particular interest. The basic reason for this is the following: while matrices can seem complicated, these directions are directions in \mathbb{R}^n where the action of A (i.e., where A itself, since A represents a transformation) is particularly simple; and we will be able to use these particularly simple directions, as well as linear combinations and scalar multiples of them, to understand A more generally.

Mathematically, this "not changing direction" or "simply stretching" property can be written as

$$Ax = \lambda x, \quad \text{for some } \lambda \in \mathbb{R}.$$

Observe that there are several ways that a vector can be stretched out:

- The stretch can be an expansion or shrinkage in the same direction, i.e., if the stretch factor $\lambda > 0$.
- The stretch can shrink the vector all the way to the origin, i.e., if the stretch factor $\lambda = 0$.
- The stretch can be in the opposite direction, i.e., if the stretch factor $\lambda < 0$.

That is, stretch means multiplying by a scalar, which is more general than the colloquial use of the term.

At first, these three scenarios may seem very different. (We are calling a shrink a stretch, and we are talking about the same direction even if it is in the opposite direction.) From what we know about linear algebra, however (i.e., that a subspace is basically a line/plane/hyperplane through the origin, and since we can consider linear combinations and multiplication of a vector by a real-valued scalar, including $0 \in \mathbb{R}$, in a unified manner to define linear combinations, span, etc), it should be clear why we treat $\lambda > 0$ (including $\lambda = 1$ and $\lambda \neq 1$) and $\lambda = 0$ and $\lambda < 0$ on the same footing. More precisely, the important point here is that the action of the matrix leaves the input vector in the same subspace as it started.

In light of this discussion, in general, the question of interest is:

- Given a matrix $A \in \mathbb{R}^{n \times n}$, for what vectors $x \in \mathbb{R}^n$ and numbers $\lambda \in \mathbb{R}$ does $Ax = \lambda x$?

Eigenvectors and eigenvalues. These vectors x and numbers λ are sufficiently important that they are given a special name.

Definition 65 Given a matrix $A \in \mathbb{R}^{n \times n}$, consider vectors $x \in \mathbb{R}^n$ and numbers $\lambda \in \mathbb{R}$ such that $Ax = \lambda x$. Such vectors x are called eigenvectors, and such numbers λ are called eigenvalues.

Remark. Taken together, the eigenvectors and eigenvalues of a matrix A are sometimes called the *eigen-decomposition* of A . The reason is that we can use them to construct a decomposition of A , where by “decomposition” we mean an equivalent form for A that is simpler or more convenient in some sense. Basically, this will correspond to expressing A in terms of linear combinations of (outer products of) these special vectors and numbers. We will return to this below in Section 9.8 as well as in a later chapter.

Remark. So far, we have just provided a definition, and we gave a few examples, but in general we haven’t said anything about whether these eigenvectors/eigenvalues exist, what are their properties, how many of them there are for a general matrix or for special classes of matrices, etc. We will get to this below.

Remark. For a given matrix A , if x is an eigenvector with eigenvalue λ , then so too is $10x$ as well as $-3.5x$ as well as αx , for any $\alpha \in \mathbb{R}$. To see this, observe that:

$$\begin{aligned} Ax = \lambda x &\Rightarrow \alpha Ax = \alpha \lambda x, \quad \text{for all } \alpha \in \mathbb{R} \\ &\Rightarrow A(\alpha x) = \lambda(\alpha x). \end{aligned}$$

Thus, since $A(\alpha x) = \lambda(\alpha x)$, the vector $x' = \alpha x$ is also an eigenvector of A with eigenvalue λ . For this reason, when we work with eigenvectors, we typically assume that they are unit normalized, i.e., that they have unit Euclidean norm. Even assuming this unit normalization, it’s possible to multiply the entire vector by -1 . Thus, if we say that an eigenvector with a given eigenvalue is unique, then we mean that we ignore these trivial degeneracies. (Equivalently, we mean that we are only interested in the subspace defined by the eigenvector.) This is worth keeping in mind, both for general understanding and in particular when working with software. Two different eigenvectors that point in exactly the opposite directions are not really two different eigenvectors, but they will “appear” different, e.g., if you simply look at the output numbers or if you simply take a vector difference and ask whether it equals zero or is very small.

To summarize, when we talk about eigenvectors, we are really going to be interested in the subspaces that they define, but we would like to discuss them without being excessively formal/technical. Thus, you should keep in mind the following points.

- By the informal term “stretch,” we will mean multiplication by some scalar $\lambda \in \mathbb{R}$, whether or not $\lambda > 0$.
- By the informal terms “direction” and “same direction,” we will mean subspace and the same subspace, respectively, and so multiplying a vector by a scalar $\lambda \in \mathbb{R}$ keeps it in the same direction, regardless of whether λ is positive or zero or negative.
- By the informal term “unique,” we will mean that the subspace $\text{Span}(x)$ is unique, which means that we don’t worry about normalization or multiplication by -1 .

9.3 Some simple examples of eigenvectors and eigenvalues

Before describing eigenvectors and eigenvalues in more detail, including how to compute them, let’s go through several examples of matrices and their eigenvectors and eigenvalues. These are all relatively simple examples, but they will illustrate more general properties that we will discuss in more detail below.

Example 1: Diagonal matrix. Here is a simple example. Consider the following matrix:

$$A = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}. \tag{9.1}$$

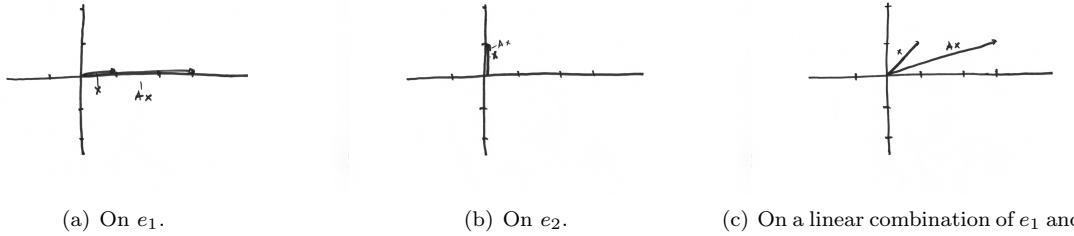


Figure 9.1: Illustration of the action of the matrix in Equation (9.1) on three vectors, e_1 , e_2 , and a linear combination of e_1 and e_2 .

Q1. What are examples of eigenvalues/eigenvectors of this matrix?

Consider the following two possibilities.

$$\begin{aligned} A \begin{pmatrix} 1 \\ 0 \end{pmatrix} &= 3 \begin{pmatrix} 1 \\ 0 \end{pmatrix} && \Rightarrow \text{ for this, } \lambda = 3 \text{ and } v_{\lambda=3} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \\ A \begin{pmatrix} 0 \\ 1 \end{pmatrix} &= 1 \begin{pmatrix} 0 \\ 1 \end{pmatrix} && \Rightarrow \text{ for this, } \lambda = 1 \text{ and } v_{\lambda=1} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \end{aligned}$$

Here, we have identified two eigenvalue-eigenvector pairs, where by “eigenvalue-eigenvector pair” we mean an eigenvalue λ and its associated eigenvector x such that $Ax = \lambda x$. Since the matrix is a square diagonal matrix with distinct entries, the eigenvalues equal the diagonal elements of the matrix, and the associated eigenvectors are the corresponding canonical coordinate axis vectors.

Q2. What can we say about the action of A on an arbitrary vector?

Let's call this vector $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. For this more general vector, we have:

$$\begin{aligned} A \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= A \left(x_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right) \\ &= x_1 A \begin{pmatrix} 1 \\ 0 \end{pmatrix} + x_2 A \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= x_1 \cdot 3 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + x_2 \cdot 1 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 3x_1 \\ x_2 \end{pmatrix}. \end{aligned}$$

Make sure you understand each of the steps of that derivation, since we used several important linear algebra results that we covered earlier. In general,

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \neq \lambda \begin{pmatrix} 3x_1 \\ x_2 \end{pmatrix}, \quad \text{for any } \lambda \in \mathbb{R}.$$

The exception to this general statement is if $x_1 = 0$ or $x_2 = 0$, which leads to the two vectors $v_{\lambda=1}$ and $v_{\lambda=3}$ above. Thus, a general vector $x \in \mathbb{R}^2$ which does not lie along one of the canonical axes is *not* an eigenvector of this A . See Figure 9.1 for an illustration.

Q3. What does the discussion for this example illustrate?

- This matrix has two eigenvalues that are distinct, and for each eigenvalue there is an associated eigenvector.

- The two eigenvectors corresponding to distinct eigenvalues are linearly-independent; and, in addition, they are orthogonal to each other.
- Each of the two directions or subspaces defined by one of the eigenvectors is stretched out by a different amount under the action of this A , and A does not cause the two directions to get “mixed up” with each other.
- A general vector pointing in an arbitrary direction of \mathbb{R}^2 is not an eigenvector of this matrix A ; and thus when A is applied to it, the output is some vector that points in some other direction.
- Since the two eigenvectors are orthogonal, they span all of \mathbb{R}^2 . Thus, a general vector $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$ can be decomposed into the eigenvectors of A (which in this case are simply coordinate basis vectors) as follows:

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Clearly, from this decomposition of an arbitrary vector into the eigenvectors of A , unless the vector lies along one of the coordinate axes, i.e., unless $x_1 = 0$ or $x_2 = 0$, this vector will under the action of A move to some other vector on the \mathbb{R}^2 plane that is not just a stretch, but instead combines both directions. The reason is that under the action of A the input vector gets stretched out a different amount along each of those two orthogonal directions.

Example 2: Identity matrix. Here is another example: $A = I$, i.e., if $A \in \mathbb{R}^{n \times n}$, then A is the the n -dimensional Identity matrix. For simplicity, let's say $n = 2$ in which case we have

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \quad (9.2)$$

This example is both simpler as well as more complicated than the previous example.

Q1. What are examples of eigenvalues/eigenvectors of this matrix?

In particular, what are vectors such that A doesn't change their direction? Well, since A is the identity matrix, A doesn't change the direction of any input vector, and so this is the case for all vectors. In particular,

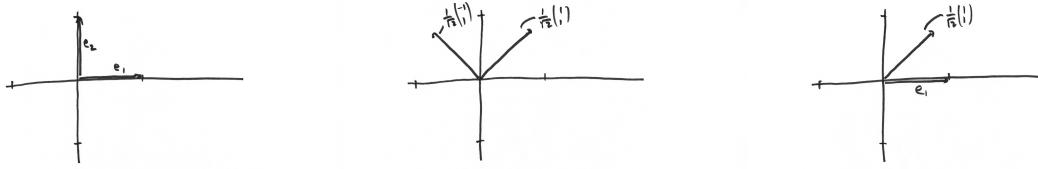
$$\begin{aligned} A \begin{pmatrix} 1 \\ 0 \end{pmatrix} &= 1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} &\Rightarrow \text{for this, } \lambda = 1 \text{ and } v_{\lambda=1} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \\ A \begin{pmatrix} 0 \\ 1 \end{pmatrix} &= 1 \begin{pmatrix} 0 \\ 1 \end{pmatrix} &\Rightarrow \text{for this, } \lambda = 1 \text{ and } v_{\lambda=1} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \end{aligned}$$

So, here are two vectors, the two vectors along the two coordinate axes, that are eigenvectors that both have the same eigenvalue 1. When there are multiple eigenvectors associated with a given eigenvalue, it is common to call that a *degenerate eigenvalue*. See Figure 9.2(a) for an illustration.

In this case, observe also that

$$\begin{aligned} A \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} &= 1 \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} &\Rightarrow \text{for this, } \lambda = 1 \text{ and } v_{\lambda=1} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \\ A \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} &= 1 \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} &\Rightarrow \text{for this, } \lambda = 1 \text{ and } v_{\lambda=1} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \end{aligned}$$

So, here are two other vectors, rotated versions of the first two, that are also eigenvectors of A ; and, in addition, this other pair of eigenvectors consists of two vectors that are orthogonal to each other. See Figure 9.2(b) for an illustration.



(a) Two orthogonal eigenvectors. (b) Two other orthogonal eigenvectors. (c) Two non-orthogonal eigenvectors.

Figure 9.2: Illustration of different orthogonal and non-orthogonal pairs of eigenvectors of the identity matrix.

Q2. What can we say about the action of A on an arbitrary vector?

Let's call this vector $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. For this more general vector, we have:

$$A \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 1 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

This should not be surprising since A in Equation (9.2) is an Identity matrix. What this says is that any vector on \mathbb{R}^2 is an eigenvector of this A , with eigenvalue 1. If we are only considering normalized unit-length vectors, then what this says is that any vector on the ℓ_2 unit ball is an eigenvector of this A , with eigenvalue 1.

Thus, in addition to the above two choices of two orthogonal eigenvectors, we could have chosen the following two linearly-independent vectors to be eigenvectors:

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}. \quad (9.3)$$

See Figure 9.2(c) for an illustration. This is correct and perfectly legitimate. The point is not that we can choose two linearly-independent eigenvectors that are not orthogonal, but instead that we can choose two eigenvectors that are orthogonal.

Q3. What does the discussion for this example illustrate?

- While the Identity is in some sense a very simple matrix (what could be simpler than a matrix that does nothing to its input), the Identity also illustrates an important subtlety regarding eigenvectors. For this matrix, there are multiple eigenvectors for a given eigenvalue λ that are not unique and are not scalar multiples of one another. One can choose those $n = 2$ eigenvectors to be linearly independent; and, in addition, one can choose those $n = 2$ eigenvectors to be orthogonal to each other.
- The span of two orthogonal eigenvectors of A is a 2-dimensional subspace (that is all of \mathbb{R}^2).
- While the actual eigenvectors are not unique, what is unique is the 2-dimensional subspace that these two vectors span. This is less interesting for 2×2 matrices, where the span is all of \mathbb{R}^2 , but it is true more generally. (This is the generalization of what we saw in Example 1 when the eigenvalues of the 2×2 matrix were distinct: that eigenvectors were not unique, but that they were unique up to scalar multiplication, i.e., up to the subspace they defined.)
- While these two vectors in Equation (9.3) are perfectly legitimate eigenvectors, and their span is all of the 2-dimensional \mathbb{R}^2 , it is common when there is a degenerate eigenvalue to choose eigenvectors that are orthonormal. It can be slightly harder to work with orthonormal vectors when you first work with them, e.g., doing pencil and paper by hand, but it is much easier to think about and much easier when you perform computations on a computer. So, if we have a degenerate eigenvalue that has multiple eigenvectors, then we will assume that they are orthogonal (whenever we can—we will describe below when we can).



(a) A matrix reflecting through the line $x_2 = x_1$.
(b) A matrix projecting onto the line $x_2 = x_1$.

Figure 9.3: Illustration of the action of a reflection matrix and a projection matrix (in Equations (9.4) and (9.5), respectively).

Example 3: Reflection matrix. Here is another example:

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \quad (9.4)$$

Q1. What are examples of eigenvalues/eigenvectors of this matrix?

For this matrix, we get the following:

$$\begin{aligned} A \begin{pmatrix} \alpha \\ \alpha \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \alpha \end{pmatrix} = 1 \begin{pmatrix} \alpha \\ \alpha \end{pmatrix} && \Rightarrow \text{ for this, } \lambda = 1 \text{ and } v_{\lambda=1} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \\ A \begin{pmatrix} \alpha \\ -\alpha \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ -\alpha \end{pmatrix} = \begin{pmatrix} -\alpha \\ \alpha \end{pmatrix} && \Rightarrow \text{ for this, } \lambda = -1 \text{ and } v_{\lambda=-1} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \end{aligned}$$

Q2. What can we say about the action of A on an arbitrary vector?

Let's call this vector $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. For this more general vector, we have:

$$A \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_1 \end{pmatrix}.$$

This matrix performs a reflection about the line $x_2 = x_1$. If one considers a vector in the subspace defined by $x_2 = x_1$, then the vector is unchanged by the matrix; and if one considers a vector in the subspace perpendicular to $x_2 = x_1$, then the vector is “reflected through” the line $x_2 = x_1$. See Figure 9.3(a).

Q3. What does the discussion for this example illustrate?

- Eigenvalues can be positive or negative; and associated with each distinct eigenvalue, there is an eigenvector.
- The two eigenvectors are perpendicular to each other.

Example 4: Projection matrix. Here is another example:

$$A = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}. \quad (9.5)$$

Q1. What are examples of eigenvalues/eigenvectors of this matrix?

For this matrix, we get the following:

$$A \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} \Rightarrow \text{for this, } \lambda = 1 \text{ and } v_{\lambda=1} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

$$A \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 0 \begin{pmatrix} 1 \\ -1 \end{pmatrix} \Rightarrow \text{for this, } \lambda = 0 \text{ and } v_{\lambda=0} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Observe that these are the same eigenvectors as for the reflection matrix given in Equation (9.4), but the eigenvalues are different. These are also the same eigenvectors as one possible choice for the Identity matrix given in Equation (9.2), but again the eigenvalues are different.

Q2. What can we say about the action of A on an arbitrary vector?

To illustrate this, consider applying the matrix A of Equation (9.5) to the vector

$$\begin{pmatrix} 5 \\ -1 \end{pmatrix} = 5 \begin{pmatrix} 1 \\ 0 \end{pmatrix} - 1 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ expressed in the basis } \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\} \quad (9.6)$$

$$= 2 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + 3 \begin{pmatrix} 1 \\ -1 \end{pmatrix} \text{ expressed in the basis } \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right\}. \quad (9.7)$$

In this case, what we get is the following:

$$A \begin{pmatrix} 5 \\ -1 \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 5 \\ -1 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

That is, A “projects” the vector $\begin{pmatrix} 5 \\ -1 \end{pmatrix}$ onto the vector $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, and it “removes” the part of the vector the vector $\begin{pmatrix} 5 \\ -1 \end{pmatrix}$ along the vector $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

More generally, consider the vector $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, in which case we get:

$$A \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} (x_1 + x_2)/2 \\ (x_1 + x_2)/2 \end{pmatrix}.$$

Here, the same projection interpretation holds. See Figure 9.3(b).

Q3. What does the discussion for this example illustrate?

- Since the two columns are linearly-dependent, A is a rank-deficient matrix. Thus, the output must lie in a lower-dimensional subspace, and that subspace is the span of the one linearly-independent column.
- This A in Eqn. (9.5) is like the matrix in Eqn. (9.1) that stretches different amounts in different directions. Here, however, it stretches by 1 along one direction, and it stretches to 0 along the other direction. That is, along one direction, the vector is unchanged, while along the other direction, the vector is shrunk to the origin.
- While the coordinate axis vectors (e_1 and e_2) may initially seem particularly nice and easy to work with, for this matrix it's actually easier to work with another set of orthogonal vectors ($v_{\lambda=1}$ and $v_{\lambda=0}$). In this latter basis, the behavior of the matrix is particularly simple to understand. For example, consider applying the matrix to the vector $\begin{pmatrix} 5 \\ -1 \end{pmatrix}$, as expressed in the two bases, as in Equation (9.6) versus Equation (9.7). The reason for this is that the basis provided by the set of eigenvectors seems particularly well-suited to this matrix A (and better-suited to A than is the canonical basis).
- This is an example of a “projection matrix,” and so it *really* is doing a projection. (The reason for this is that the non-zero eigenvalue equals unity—more generally, an $n \times n$ matrix in which all the eigenvalues equal 0 or 1 is a projection matrix onto the subspace spanned by its columns.)

9.4 Computing eigenvectors and eigenvalues

The examples of the previous subsection were sufficiently simple that we were able just to write down eigenvalue-eigenvector pairs, but in general they must be computed with some algorithm (assuming that they even exist). There are several ways to compute eigenvalues and eigenvectors.

- **With determinants.** This approach is most appropriate pedagogically for small 2×2 and 3×3 examples, and it can be understood in terms of ideas like linear dependence and independence. It is not a practically-appropriate method for larger more realistic problems.
- **With quadratic forms.** This approach illustrates the basic algebraic and geometric ideas of how to compute eigenvectors and eigenvalues for larger more realistic problems, in a way that conveys the basic ideas more generally.
- **In actual numerical practice.** The actual practical computation of eigenvectors is quite involved, and we won't discuss it. If you really want to know the details, then you can find them in a more advanced numerical linear algebra or numerical analysis class.

Here, we will consider the first two ways. In many cases, understanding the basic ideas is sufficient, basically since one typically calls existing software as a black box. Note, however, that while the actual practical computation of eigenvectors is quite involved, many of the ideas have roots in the quadratic form perspective we will discuss.

9.5 Basic properties of determinants

Determinants of 2×2 matrices. Let's start with a general 2×2 matrix:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}. \quad (9.8)$$

The so-called determinant of this matrix is given by the following:

$$\det(A) = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

To understand what this quantity captures, let's look at the action of A on the four points of the unit square in the first quadrant with one corner at the origin.

$$\begin{aligned} A \begin{pmatrix} 1 \\ 0 \end{pmatrix} &= \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ c \end{pmatrix} \\ A \begin{pmatrix} 0 \\ 1 \end{pmatrix} &= \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix} \\ A \begin{pmatrix} 1 \\ 1 \end{pmatrix} &= \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} a+b \\ c+d \end{pmatrix} \\ A \begin{pmatrix} 0 \\ 0 \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{aligned}$$

See Figure 9.4(a) for an illustration. From the figure, it is clear that the area of the parallelogram equals $|ad - bc| = \det(A)$, where the matrix A is given in Equation (9.8).

Question: When is the area = 0?

Answer: When $\begin{pmatrix} a \\ c \end{pmatrix}$ is linearly dependent on $\begin{pmatrix} b \\ d \end{pmatrix}$, i.e., when

$$\begin{pmatrix} a \\ c \end{pmatrix} = \alpha \begin{pmatrix} b \\ d \end{pmatrix} = \begin{pmatrix} \alpha b \\ \alpha d \end{pmatrix},$$

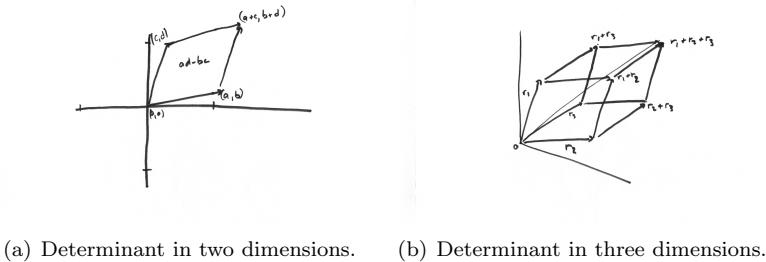


Figure 9.4: Illustration of the determinant in two dimensions and three dimensions in terms of the area of the corresponding parallelogram and the volume of the corresponding parallelepiped.

since in this case $ad - bc = (ab)d - b(ad) = 0$.

Question: Conversely, what if $\begin{pmatrix} a \\ c \end{pmatrix}$ is linearly dependent on $\begin{pmatrix} b \\ d \end{pmatrix}$?

Answer: In this case,

$$\begin{pmatrix} a \\ c \end{pmatrix} = \begin{pmatrix} \alpha b \\ \alpha d \end{pmatrix}$$

for some α , and from this it follows that the area = 0 since $A = (\text{base}) \cdot (\text{height})$. That is, the determinant equals zero.

Remark. It is true that $\det(A)$ is a number that is equal to 0 when the rows/columns of A are linearly dependent and is nonzero otherwise, and thus it can be used to characterize the linear dependence/independence of the columns/rows of A . Nevertheless, it is not a particularly good, e.g., robust or well-behaved, measure of linear dependence/independence, basically for the following reason. Recall that in many cases, we don't want to worry too much about the norms of a row or column, e.g., we may perform TFIDF normalization or we may perform a unit normalization. In these cases, $\det(A)$ does not distinguish the following two situations:

$$A_1 = \begin{pmatrix} \sqrt{\epsilon} & 0 \\ 0 & \sqrt{\epsilon} \end{pmatrix} \quad (9.9)$$

$$A_2 = \begin{pmatrix} 1 & \cos(\epsilon) \\ 0 & \sin(\epsilon) \end{pmatrix}, \quad (9.10)$$

where here ϵ is a small number such that $0 < \epsilon \ll 1$. See Figure 9.5 for an illustration.

In this case, since $\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots$ for small x , we have that

$$\begin{aligned} \det(A_1) &= \epsilon \\ \det(A_2) &= \epsilon - \frac{\epsilon^3}{3!} + \dots \approx \epsilon. \end{aligned}$$

That is, from the perspective of the determinant, the two matrices A_1 and A_2 look very similar, but (if we are not concerned about normalizing columns/rows) these two matrices are actually very different with respect to linear dependence/independence. In particular, the matrix A_1 is a scaled version of the Identity, and thus the two columns are orthogonal, i.e., the angle between them is 90° , while for the matrix A_2 the angle between the two columns is $\epsilon \approx 0$, meaning that the two columns are approximately (informally for now, but in a sense that we can make precise later) linearly dependent. In both cases, the volume of the corresponding parallelogram is very small, but it becomes very small for two quite different reasons. This is one of the reasons that the determinant is much less useful in general.

Determinants of 3×3 matrices. These ideas generalize to $n \times n$ matrices. That is, while these observations about exact/approximate linear dependence/independence in the above 2×2 example is not so

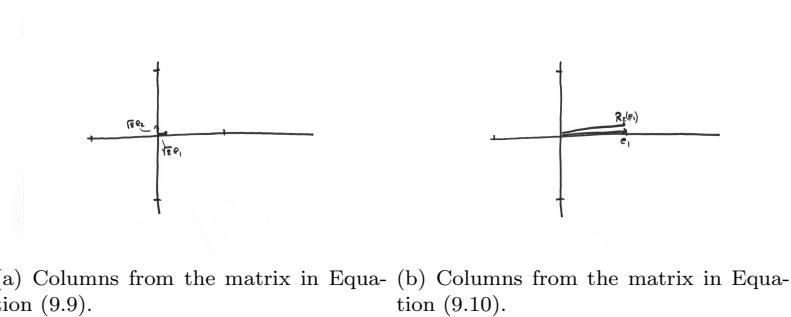


Figure 9.5: Illustration of the columns of two matrices with similar determinants.

interesting (since it corresponds to a one-dimensional subspace, i.e., linear dependence just amounts just to rescaling), in general similar observations hold for linear dependence/independence in higher dimensions. To go to larger matrices, recall that the determinant is defined recursively.

For example, let's consider a general 3×3 matrix:

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}. \quad (9.11)$$

The determinant of this matrix is given by the following:

$$\begin{aligned} \det(A) &= \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix} \\ &= a(ei - fh) - b(di - fg) + c(dh - eg) \\ &= aei + bfg + cdh - ceg - bdi - afh. \end{aligned}$$

In this 3×3 example, if we consider the unit cube in the \mathbb{R}^3 positive orthant with one corner at the origin, then the determinant measures the volume under the transformation defined by the matrix of the parallelepiped defined by the images of the 8 vectors at the corner of this unit cube. See Figure 9.4(b) for an illustration.

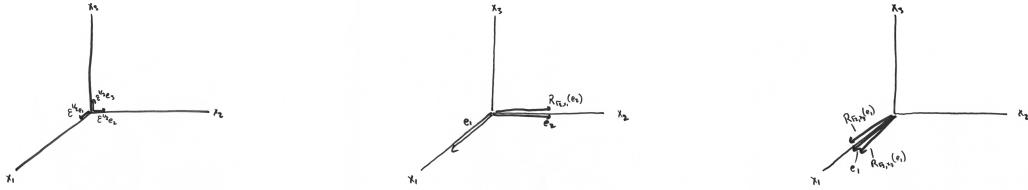
As with 2×2 matrices, for 3×3 matrices, the determinant is zero or non-zero, depending on whether or not the columns of A are linearly dependent or independent, but here too the determinant is not a particularly robust or well-behaved measure of linear dependence/independence. For example, $\det(A)$ does not distinguish the following situations:

$$A_1 = \begin{pmatrix} \epsilon^{1/3} & 0 & 0 \\ 0 & \epsilon^{1/3} & 0 \\ 0 & 0 & \epsilon^{1/3} \end{pmatrix} \quad (9.12)$$

$$A_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \cos(\epsilon) \\ 0 & 0 & \sin(\epsilon) \end{pmatrix} \quad (9.13)$$

$$A_3 = \begin{pmatrix} 1 & \cos(\sqrt{\epsilon}) & \cos(\sqrt{\epsilon}) \\ 0 & \sin(\sqrt{\epsilon}) & 0 \\ 0 & 0 & \sin(\sqrt{\epsilon}) \end{pmatrix}, \quad (9.14)$$

where here ϵ is a small number such that $0 < \epsilon \ll 1$. In all three cases, $\det(A_i) \approx \epsilon$, but the columns in each example bear a very different relationship with each other: in A_1 , since the matrix is a scaled Identity, the three columns are orthogonal; in A_2 , the second and third column are exactly or approximately orthogonal to the first, but the angle between them is very small and thus they are approximately linearly dependent



(a) Columns from the matrix in Equation (9.12). (b) Columns from the matrix in Equation (9.13). (c) Columns from the matrix in Equation (9.14).

Figure 9.6: Illustration of the columns of three matrices with similar determinants.

on each other; and in A_3 , all three columns point in approximately the same direction, and thus all three columns are approximately linearly dependent. See Figure 9.6 for an illustration. The same issues described here hold true for general $n \times n$ matrices.

9.6 Using determinants to understand eigendecompositions

Determinants and eigenvalue equations. Let's now see how to use determinants to understand eigendecompositions of matrices, i.e., to understand matrices in terms of their stretch directions and stretch amounts. To do so, recall that the basic eigenvalue equation,

$$Ax = \lambda x$$

can also be written as

$$(A - \lambda I)x = 0,$$

where $0 \in \mathbb{R}^n$ is the all-zeros vector, and where $I \in \mathbb{R}^{n \times n}$ is the Identity. Thus, in computing eigenvectors and eigenvalues, if we let

$$A' = A'(\lambda) = A - \lambda I,$$

then we are asking for vectors x and numbers λ such that

$$A'x = 0,$$

i.e., such that $A'x$ gets mapped to the all-zeros vector. (Observe that while this is a linear equation in the unknown variable x , it is not a linear equation in the variables x and λ , and we often want to find both λ and x .)

So the basic question becomes:

- When does a square $n \times n$ matrix (of the form $A - \lambda I$, where λ is a parameter, A is arbitrary and I is an Identity) acting on a non-zero vector give the all-zero vector?

From our previous discussions on linear algebra, we know that three equivalent answers to this basic question are the following:

- When the columns/rows are not linearly independent.
- When the matrix is not invertible.
- When $\det(A) \neq 0$.

Here are some important facts about determinants that are relevant for this problem.

- The determinant of an $n \times n$ matrix has n^{th} powers of the matrix entries. (This is seen in the 2×2 and 3×3 examples above.)
- Since the determinant of an $n \times n$ matrix has n^{th} powers of the matrix entries, the determinant of an $n \times n$ matrix is an n^{th} degree polynomial in λ . (This can also be seen in the 2×2 and 3×3 examples above.) Note that the terms involving λ appear on the diagonal, and of the terms that enter into the expression for the determinant, one of them is the product of all of the diagonal terms (ad in the 2×2 case, and aei in the 3×3 case).
- By the Fundamental Theorem of Algebra, if A is real, i.e., if all of the entries are real numbers, then the solution of the polynomial equation $\det(A - \lambda I) = 0$ has n (perhaps complex, perhaps repeated) roots. This n includes zero eigenvalues and counting repeated roots according to their multiplicities.

A recipe to use determinants for eigendecompositions. Given all of this, let's state a general recipe to solve the eigenvalue problem for an $n \times n$ matrix A .

1. Compute the determinant of $A - \lambda I$.
(With λ subtracted along the diagonal, the determinant starts with λ^n or $-\lambda^n$, i.e., it is a polynomial of degree n with real coefficients.)
2. Find the roots of this polynomial by solving $\det(A - \lambda I) = 0$.
(The n roots are the n eigenvalues of A . They make $A - \lambda I$ singular/noninvertible and thus have linearly dependent columns/rows.)
3. For each eigenvalue λ , solve $(A - \lambda I)x = 0$ to find eigenvector x .
(For this fixed value of λ , this is a linear equation to solve, e.g., by solving some routine.)

Remark. From this discussion, it follows that an $n \times n$ matrix has n eigenvalues. Importantly, they may not all be distinct, e.g., the expression $(x - \lambda)^{50}$ has a root at $x = \lambda$ repeated 50 times. To each *different* eigenvalue, there is an associated eigenvector. Note that this discussion doesn't say anything about whether or not there is more than one eigenvectors when there is a degeneracy in the eigenvalues (although in Example 2 above there were 2 orthogonal eigenvectors corresponding to the degenerate eigenvalue).

Remark. The general recipe works in principle for any $n \times n$ matrix, and it works in practice for 2×2 and 3×3 matrices. It is not at all practical, and it is a very bad way to even think about computing eigenvalues for larger matrices.

9.7 Using determinants to compute simple eigendecompositions

Here are several more examples of matrices and their eigenvectors and eigenvalues. Whereas the previous examples were sufficiently simple that we just stated the eigenvectors/eigenvalues, showing that they satisfied Definition 65, here we will compute them using the determinant-based procedure described in Section 9.6. Since each of these examples is a 2×2 matrix, we can use the quadratic formula to get the eigenvalues and then plug everything in to get the eigenvectors. Thus, these are all still relatively-simple examples, and this is not how one would solve larger more realistic problems, but these examples do illustrate important properties that we will discuss in more detail in later chapters.

A conclusion from these (and the previous) examples will be that eigendecompositions of symmetric matrices are particularly well-behaved and have particularly nice properties, while the eigendecompositions of general matrices may or may not have these nice properties. By nice properties, we mean that for a given $n \times n$ matrix A we can find a full set of eigenvectors that form an orthonormal basis of \mathbb{R}^n . This will be extremely useful, and we will discuss it more detail in later chapters, so keep this in mind as you go through these examples.

Several examples of “nice” matrices. Let’s start with several examples of eigendecompositions of “nice” matrices. These matrices are “nice” in the sense that they have n real-valued eigenvalues (perhaps counting multiplicity) and a full set of n orthogonal real-valued eigenvectors.

Example 5. Let’s consider

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}. \quad (9.15)$$

In this case,

$$A - \lambda I = \begin{pmatrix} 1 - \lambda & 2 \\ 2 & 4 - \lambda \end{pmatrix}.$$

In this case,

$$\begin{aligned} \det(A - \lambda I) &= (1 - \lambda)(4 - \lambda) - 4 \\ &= \lambda^2 - 5\lambda + 4 - 4 \\ &= \lambda^2 - 5\lambda = 0 \quad \text{if } \lambda = 0, 5. \end{aligned}$$

So, the eigenvalues are $\lambda = 0, 5$. Let’s compute the eigenvectors:

$$\begin{aligned} \lambda = 0 &: \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow v_{\lambda=0} = \frac{1}{\sqrt{5}} \begin{pmatrix} -2 \\ 1 \end{pmatrix}. \\ \lambda = 5 &: \begin{pmatrix} -4 & 2 \\ 2 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow v_{\lambda=5} = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ 2 \end{pmatrix}. \end{aligned}$$

This is similar to the example we saw previously, with two distinct eigenvalues and two orthogonal eigenvectors. One of the eigenvalues equals 0, and this is a consequence of the fact that the columns of A are not linearly independent.

Example 6. Let’s consider

$$A = \begin{pmatrix} -3 & 4 \\ 4 & 3 \end{pmatrix}. \quad (9.16)$$

In this case,

$$A - \lambda I = \begin{pmatrix} -3 - \lambda & 4 \\ 4 & 3 - \lambda \end{pmatrix}.$$

In this case,

$$\begin{aligned} \det(A - \lambda I) &= (3 + \lambda)(-3 + \lambda) - 16 \\ &= \lambda^2 - 9 - 16 \\ &= \lambda^2 - 25 = 0 \quad \text{if } \lambda = -5, 5. \end{aligned}$$

So, the eigenvalues are $\lambda = -5, 5$. Let’s compute the eigenvectors:

$$\begin{aligned} \lambda = -5 &: \begin{pmatrix} 2 & 4 \\ 4 & 8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow v_{\lambda=-5} = \frac{1}{\sqrt{5}} \begin{pmatrix} -2 \\ 1 \end{pmatrix}. \\ \lambda = 5 &: \begin{pmatrix} -8 & 4 \\ 4 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow v_{\lambda=5} = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ 2 \end{pmatrix}. \end{aligned}$$

Here, the eigenvectors are the same as in Example 5, but the eigenvalues are different.

Example 7. Let’s consider

$$A = \begin{pmatrix} 9 & -2 \\ -2 & 6 \end{pmatrix}. \quad (9.17)$$

In this case,

$$A - \lambda I = \begin{pmatrix} 9 - \lambda & -2 \\ -2 & 6 - \lambda \end{pmatrix}.$$

In this case,

$$\begin{aligned}\det(A - \lambda I) &= (9 - \lambda)(6 - \lambda) - 4 \\ &= \lambda^2 - 15\lambda + 54 - 4 \\ &= \lambda^2 - 15\lambda + 50 = 0 \quad \text{if } \lambda = 5, 10.\end{aligned}$$

So, the eigenvalues are $\lambda = 5, 10$. Let's compute the eigenvectors:

$$\begin{aligned}\lambda = 10 &: \begin{pmatrix} -1 & -2 \\ -2 & -4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow v_{\lambda=10} = \frac{1}{\sqrt{5}} \begin{pmatrix} -2 \\ 1 \end{pmatrix}. \\ \lambda = 5 &: \begin{pmatrix} 4 & -2 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow v_{\lambda=5} = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ 2 \end{pmatrix}.\end{aligned}$$

Here, the eigenvectors are the same as in Examples 5 and 6, but the eigenvalues are different.

These matrices are “nice” in the sense that they have n real-valued eigenvalues (perhaps counting multiplicity) and a set of n orthogonal real-valued eigenvectors. As we will see in later chapters, this is extremely important for many applications.

Several examples of “not nice” matrices. Lest you think that most matrices are “nice” in the above sense of the word, let's move onto a few examples which will illustrate several fundamental “non-nice” behaviors of eigenvalues and/or eigenvectors. These matrices are “not nice” in the sense that they do not have a full set of n orthogonal real-valued eigenvectors and associated real-valued eigenvalues.

Example 8. (An example with non-orthogonal eigenvectors.) Let's consider the matrix:

$$A = \begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix}. \quad (9.18)$$

In this case,

$$A - \lambda I = \begin{pmatrix} 3 - \lambda & 2 \\ 4 & 1 - \lambda \end{pmatrix}.$$

In this case,

$$\begin{aligned}\det(A - \lambda I) &= (3 - \lambda)(1 - \lambda) - 8 \\ &= \lambda^2 - 4\lambda + 3 - 8 \\ &= \lambda^2 - 4\lambda - 5 \\ &= (\lambda - 5)(\lambda + 1) = 0 \quad \text{if } \lambda = -1, 5.\end{aligned}$$

So, the eigenvalues are $\lambda = -1, 5$. Let's compute the eigenvectors:

$$\begin{aligned}\lambda = 5 &: \begin{pmatrix} -2 & 2 \\ 4 & -4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \rightarrow v_{\lambda=5} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ \lambda = -1 &: \begin{pmatrix} 4 & 2 \\ 4 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \rightarrow v_{\lambda=-1} = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ -2 \end{pmatrix}.\end{aligned}$$

Observe that $v_{\lambda=-1}$ and $v_{\lambda=5}$ are linearly independent, but they are *not* orthogonal to each other. Thus, the matrix in Equation (9.18) is an example of a matrix with two distinct eigenvalues, each of which has an associated eigenvector, where the two eigenvectors are linearly independent but *not* orthogonal.

Example 9. (An example with complex eigenvalues and eigenvectors.) Let's consider the matrix:

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \quad (9.19)$$

For this matrix, when applied to a vector $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, we get the following:

$$A \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_2 \\ -x_1 \end{pmatrix}.$$

In this case,

$$A - \lambda I = \begin{pmatrix} -\lambda & 1 \\ -1 & -\lambda \end{pmatrix}.$$

In this case,

$$\det(A - \lambda I) = \lambda^2 + 1 = 0 \quad \text{if } \lambda = \pm i.$$

So, the eigenvalues are $\lambda = -i, +i$. Let's compute the eigenvectors:

$$\begin{aligned} \lambda = i & : \begin{pmatrix} -i & 1 \\ -1 & -i \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \rightarrow v_{\lambda=i} = \begin{pmatrix} -i \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ i \end{pmatrix} \\ \lambda = -i & : \begin{pmatrix} -i & 1 \\ -1 & i \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \rightarrow v_{\lambda=-i} = \begin{pmatrix} 1 \\ -i \end{pmatrix} = \begin{pmatrix} i \\ 1 \end{pmatrix}. \end{aligned}$$

Thus, the matrix in Equation (9.19) is an example of a real-valued matrix, i.e., a matrix whose elements consist of only real numbers, that has eigenvalues that are imaginary/complex numbers and eigenvectors that contain imaginary/complex entries.

Example 10. (An example with fewer than two eigenvectors.) Let's consider the matrix:

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}. \tag{9.20}$$

For this matrix, when applied to a vector $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, we get the following:

$$A \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_2 \\ 0 \end{pmatrix}.$$

In this case,

$$A - \lambda I = \begin{pmatrix} -\lambda & 1 \\ 0 & -\lambda \end{pmatrix}.$$

In this case,

$$\det(A - \lambda I) = \lambda^2 - 0 = \lambda^2 = 0.$$

Thus, $\lambda = 0$ is a degenerate eigenvalue with degeneracy equal to 2. Let's compute the eigenvectors. Since

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_2 \\ 0 \end{pmatrix},$$

there is only one corresponding eigenvector, which is

$$v_{\lambda=0} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Thus, the matrix in Equation (9.20) is an example of a real-valued $n \times n$ matrix that has fewer than n eigenvectors.

Discussion of “nice” and “not nice” matrices. These last three examples show that there can be some subtleties in computing eigenvectors and eigenvalues of general matrices. In particular, for general matrices, eigenvectors and eigenvalues are not very well-behaved things. This arises for many reasons, and this is a more involved topic than we can cover in this class, but basically the reasons are variants of the three reasons in these three examples. Here are several comments about this.

- This is a mathematical fact, and so it’s okay, but—compared with what we saw in the previous examples—this represents a lack of structure that makes using such general matrices difficult in data science (and other) applications. We have illustrated this in a simple setting of 2×2 matrices, but they occur much more generally. In addition, while they may seem trivial and easy-to-diagnose for 2×2 toy matrices, they can be much less trivial and more difficult to diagnose for larger more realistic matrices. A large part of traditional linear algebra deals with these more complicated situations.
- Example 7 showed that some non-symmetric matrices are also nice in this same sense. We saw a second example of this in Section 9.2. We will consider the related example of PageRank in a later chapter, but these are the exceptions. The general topic of eigendecompositions of general matrices is more advanced than we will cover in this class.
- This should be contrasted with the first 6 examples, where one could choose a full set of n (equal to 2 in those examples) eigenvectors that are orthogonal to each other. The importance of this is that one can use the eigenvectors to form a complete orthonormal basis for \mathbb{R}^n . In general, of course, this basis is not the usual canonical basis of \mathbb{R}^n , but it is a basis that is better-suited or more well-adapted to the particular matrix.
- In data science applications, one almost never needs to worry about eigendecompositions of general matrices. Instead, if we want eigenvector information in data science applications, and in many cases we do, then we can restrict consideration to related symmetric matrices. By considering this more narrow class of matrices, then we might hope for stronger results, and we will see that this is the case.
- That claim (that we don’t need to consider eigendecompositions of non-symmetric matrices) may come as a surprise, since one of our motivating matrix examples was term-document matrices, and those are certainly not symmetric. Indeed, they are typically not even square. For non-symmetric matrices, however, e.g., term-document matrices (even if they happen to be square), we don’t use eigenvectors. Instead, we use something related known as singular vectors. Singular vectors are basically eigenvectors of related correlation/covariance matrices. Thus, for general non-symmetric matrices, in data science applications, one typically obtains insight by performing computations on related symmetric matrices.
- It is difficult to overstate the importance of the fact that for any symmetric matrix we can obtain a full set of n orthonormal eigenvectors. Basically, this says that if the matrix represents some data set, then there is a basis (which is not the standard basis) that is determined by the structure of the data and that thus can easily be used to analyze the data.

The reason that symmetric matrices are so nice is that, by restricting ourselves to symmetric matrices, we will be able to view symmetric matrices in terms of quadratic forms. This is *much* easier than viewing matrices in terms of determinants, linear independence, the Fundamental Theorem of Algebra, etc. In addition, this will lead to much more well-behaved eigenvectors and eigenvalues, e.g., where eigenvalues are real, where eigenvectors can be chosen to be orthogonal, and where we can get a full set of n orthogonal eigenvectors to serve as a basis for \mathbb{R}^n that in some sense is particularly well-adapted to the data matrix.

9.8 Expressing matrices in terms of their eigendecompositions

The reason that the set of eigenvalues and eigenvectors of a matrix is known as the eigendecomposition of that matrix is that one can use them to decompose the matrix into a simpler form. (We saw an example of a decomposition—the QR decomposition—in an earlier chapter, but this eigendecomposition is a different

decomposition.) This is a large and important topic. We will describe it in a little more detail in a few chapters. Here, we will describe it in the context of the simple 2×2 examples we have been discussing.

For the following, it may help to review what we discussed on matrix multiplication.

To do this, let's provide is a more abstract way to say what we have been discussing in this chapter. To do this, let's number the eigenvectors and eigenvalues in a way that will make it easier to generalize beyond 2×2 matrices. Let's call λ_1 the largest eigenvalue and let's call λ_i the i^{th} largest eigenvalue, where $i \in \{1, 2\}$ for 2×2 matrices. In addition, let's call v_i the eigenvector corresponding to the i^{th} largest eigenvalue. If a matrix has two (or more) identical eigenvalues, then we repeat i in λ_i according to the multiplicity, and we can number their associated orthogonal eigenvectors arbitrarily. For the 2×2 symmetric matrices we have been discussing in this chapter, we only have λ_1 and λ_2 and the corresponding v_1 and v_2 .

Given this numbering convention, recall that for the 2×2 symmetric matrices we have been discussing, we have seen that there are two eigenvalue-eigenvector pairs (λ_i, v_i) that satisfy:

$$Av_i = \lambda_i v_i, \quad (9.21)$$

where $i \in \{1, 2\}$. The LHS and RHS of this equation are both vectors, i.e., 2×1 matrices. Let's write the two equations for $i \in \{1, 2\}$ as a single matrix equation. To do so, let's define a 2×2 diagonal matrix Λ to be

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \quad (9.22)$$

and let's define a 2×2 matrix V to be

$$V = \begin{pmatrix} v_1 & v_2 \end{pmatrix}, \quad (9.23)$$

where, importantly, if the i^{th} eigenvalue is Λ_{ii} , then the corresponding i^{th} eigenvector is the i^{th} column of V . Note that, since the eigenvectors are unit-length and pair-wise orthogonal, V is an orthogonal matrix. That is, the transpose of this matrix, V^T , can be written as

$$V^T = \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix},$$

in which case

$$VV^T = V^T V = I \in \mathbb{R}^{2 \times 2}.$$

Given this, the eigenvalue equations of Equation (9.21), for $i \in \{1, 2\}$, can be written

$$AV = V\Lambda. \quad (9.24)$$

To see this, let's just plug in the expressions of Equation (9.22) and Equation (9.23) to get

$$A \begin{pmatrix} v_1 & v_2 \end{pmatrix} = \begin{pmatrix} v_1 & v_2 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \Rightarrow \begin{pmatrix} Av_1 & Av_2 \end{pmatrix} = \begin{pmatrix} \lambda_1 v_1 & \lambda_2 v_2 \end{pmatrix}.$$

Note that, in Equation (9.24), the matrix V *post*-multiplies A on the LHS, but it *pre*-multiplies Λ on the RHS. Getting the order of these matrix operations correct is *very* important, and it is a common source of confusion. Since both Λ and V are square matrices, one can multiply them in either order; but, in general, $V\Lambda \neq \Lambda V$. Students sometimes get confused and want to pre-multiply V by Λ on the RHS. To get the order correct, recall what we are trying to do (express all the n eigenvalue equations as a single matrix equation), and pre-multiply and post-multiply V by Λ to get the following:

$$\begin{aligned} \Lambda V &= \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix} = \begin{pmatrix} \lambda_1 v_{11} & \lambda_1 v_{12} \\ \lambda_2 v_{21} & \lambda_2 v_{22} \end{pmatrix} \\ V\Lambda &= \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = \begin{pmatrix} \lambda_1 v_{11} & \lambda_2 v_{12} \\ \lambda_1 v_{21} & \lambda_2 v_{22} \end{pmatrix}. \end{aligned}$$

Thus, since Λ is a diagonal matrix, the matrix ΛV has its *rows* scaled by the eigenvalues λ_i , and the matrix $V\Lambda$ has its *columns* scaled by the eigenvalues λ_i . Since the columns of V are the eigenvectors v_i , it is these that, when multiplied by the eigenvalues λ_i , should equal Av_i .

(Make sure that you understand what we did here. If you don't, then review it. Also, if the matrix multiplication was not clear, then review it.)

Spectral decomposition: expressing A as a product of three matrices. Given Equation (9.24), let's *post*-multiply the LHS and the RHS by the 2×2 matrix V^T . Recalling that $VV^T = I$, we obtain the following:

$$\begin{aligned} A &= AVV^T \\ &= V\Lambda V^T. \end{aligned} \quad (9.25)$$

This expression provides a decomposition of the matrix A into the product of three matrices, an orthogonal matrix (consisting of the eigenvectors of A), a diagonal matrix (consisting of the eigenvalues of A), and the transpose of that orthogonal matrix. For this 2×2 example, we can write Equation (9.25) in terms of individual elements, in which case we obtain:

$$A = V\Lambda V^T = \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} v_{11} & v_{21} \\ v_{12} & v_{22} \end{pmatrix}. \quad (9.26)$$

This simple eigendecomposition is the first example of *the spectral theorem*.

This decomposition of the form $A = V\Lambda V^T$ holds for general $n \times n$ symmetric matrices (we will discuss this in a later chapter), and generalizations of it hold more generally (we will discuss this in a later chapter). In particular, if we consider computing $y = Ax$, i.e., A multiplying a vector x , this is the same as

$$y = V\Lambda V^T x = V(\Lambda(V^T x)).$$

(Make sure that you understand what we did here—if you don't, then review it.)

Spectral decomposition: expressing A as a sum of outer products. Given A , as represented in Equation (9.25) as the product of three abstract matrices and in Equation (9.26) in terms of the individual elements of those matrices, let's write it in terms of the columns of V (which recall are the eigenvectors, which are also the rows of V^T) and elements of Λ (which recall are the eigenvalues):

$$\begin{aligned} A &= V\Lambda V^T \\ &= (v_1 \ v_2) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} \\ &= (v_1 \ v_2) \begin{pmatrix} \lambda_1 v_1^T \\ \lambda_2 v_2^T \end{pmatrix} \\ &= \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T. \end{aligned} \quad (9.27)$$

That is, for a given $i \in [n]$, where $n = 2$ in this case, we can view the vector $v_i \in \mathbb{R}^2$ as a 2×1 matrix $v_i \in \mathbb{R}^{2 \times 1}$. In this case, the transpose $v_i^T \in \mathbb{R}^{1 \times 2}$. In this case, the matrix-matrix product $v_i v_i^T \in \mathbb{R}^{2 \times 2}$, and if we multiply that matrix by the number λ_i , then we have a 2×2 matrix $\lambda_i v_i v_i^T$. The expression given in Equation (9.27) says that if we do this for every i and if sum up the corresponding matrices, then we get the original matrix A . Thus, the decomposition of Equation (9.27) expresses the matrix A in terms of the sum of $n = 2$ terms, each of which is the outer product of an eigenvector with its transpose, multiplied/scaled by the corresponding eigenvalue.

This decomposition holds for general $n \times n$ symmetric matrices, in which case $A = \sum_{i=1}^n \lambda_i v_i v_i^T$ (we will discuss this in a later chapter), and generalizations of it hold more generally (we will discuss this in a later chapter). In particular, if we consider computing $y = Ax$, i.e., A multiplying a vector x , this is the same as

$$y = (\lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T) x = \lambda_1 v_1 v_1^T x + \lambda_2 v_2 v_2^T x.$$

(Make sure that you understand what we did here—if you don’t, then review it.)

Problem. Show that these two expressions for the matrix A do not depend on whether we order the eigenvalue-eigenvector pairs from largest to smallest or from smallest to largest (i.e., if we called λ_1 the smallest eigenvalue and λ_i the i^{th} smallest eigenvalue, with the eigenvectors v_i being numbered in the corresponding way).

Revisiting our examples. Let’s revisit our examples in light of these two decompositions to show that these two decompositions hold for the examples we discussed previously.

Example 1 (continued): Diagonal matrix. Consider the matrix given in Equation (9.1). How do we express the matrix in two standard forms?

- **Expressing A as a sum of n outer products.** Observe that we can write the matrix A in Equation (9.1) in the following way:

$$\begin{aligned} \sum_{i=1}^2 \lambda_i v_i v_i^T &= \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T \\ &= 3 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \end{pmatrix} + 1 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 3 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} = A. \end{aligned}$$

- **Expressing A as a product of 3 matrices.** Observe also that we can write the matrix A in the following way:

$$\begin{aligned} V \Lambda V^T &= (v_1 \ v_2) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} = A. \end{aligned}$$

Remark. These two ways of expressing a matrix (as a sum of outer products consisting of eigenvalue-eigenvector pairs and as a product of three matrices containing eigenvalue-eigenvector information) are extremely important. It may seem pedantic at this point to walk through this derivation in such detail, but this is since our first example was so simple, e.g., since the eigenvectors were just the canonical axes.

Example 2 (continued): Identity matrix. Consider the matrix given in Equation (9.2). How do we express the matrix in two standard forms? Recall that this matrix was degenerate, in that it has multiple eigenvectors associated with a given eigenvalue. Here, we show that these two decompositions are valid if we work with either of the two sets of orthonormal eigenvectors discussed above, but they are not valid if we use two eigenvectors that are not orthonormal.

- **Expressing A as a sum of n outer products.**

– If we choose $v_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $v_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, then observe that we can write the matrix A in

Equation (9.2) in the following way:

$$\begin{aligned}
 \sum_{i=1}^2 \lambda_i v_i v_i^T &= \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T \\
 &= 1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \end{pmatrix} + 1 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \end{pmatrix} \\
 &= \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \\
 &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = A.
 \end{aligned}$$

– Alternatively, if we choose $v_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$ and $v_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$, then observe that we can write the matrix A in Equation (9.2) in the following way:

$$\begin{aligned}
 \sum_{i=1}^2 \lambda_i v_i v_i^T &= \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T \\
 &= 1 \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} + 1 \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \\
 &= \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} + \begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{pmatrix} \\
 &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = A.
 \end{aligned}$$

– If, on the other hand, we choose $v_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $v_2 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$, and we try to perform the same operations, then we get:

$$\begin{aligned}
 \sum_{i=1}^2 \lambda_i v_i v_i^T &= \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T \\
 &= 1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \end{pmatrix} + 1 \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} \begin{pmatrix} 1/2 & 1/2 \end{pmatrix} \\
 &= \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \\
 &= \begin{pmatrix} 3/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \neq A.
 \end{aligned}$$

• Expressing A as a product of 3 matrices.

– If we choose $v_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $v_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, then observe that we can write the matrix A in Equation (9.2) in the following way:

$$\begin{aligned}
 V \Lambda V^T &= (v_1 \ v_2) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} \\
 &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\
 &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = A.
 \end{aligned}$$

- Alternatively, if we choose $v_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$ and $v_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$, then observe that we can write the matrix A in Equation (9.2) in the following way:

$$\begin{aligned} V\Lambda V^T &= (v_1 \ v_2) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} \\ &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \\ &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = A. \end{aligned}$$

- If, on the other hand, we choose $v_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $v_2 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$, and we try to perform the same operations, then we get:

$$\begin{aligned} V\Lambda V^T &= (v_1 \ v_2) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \\ &= \begin{pmatrix} 3/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \neq A. \end{aligned}$$

Example 3 (continued): Reflection matrix. Consider the matrix given in Equation (9.4). How do we express the matrix in two standard forms?

- **Expressing A as a sum of n outer products.** Observe that we can write the matrix A in Equation (9.4) in the following way:

$$\begin{aligned} \sum_{i=1}^2 \lambda_i v_i v_i^T &= \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T \\ &= 1 \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} - 1 \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \\ &= \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} - \begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = A \end{aligned}$$

- **Expressing A as a product of 3 matrices.** Observe also that we can write the matrix A in the following way:

$$\begin{aligned} V\Lambda V^T &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \\ &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \\ &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = A \end{aligned}$$

Example 4 (continued): Projection matrix. Consider the matrix given in Equation (9.5).

- **Expressing A as a sum of n outer products.** Observe that we can write the matrix A in Equation (9.5) in the following way:

$$\begin{aligned}\sum_{i=1}^2 \lambda_i v_i v_i^T &= \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T \\ &= 1 \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} + 0 \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \\ &= \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \\ &= \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} = A.\end{aligned}$$

- **Expressing A as a product of 3 matrices.** Observe also that we can write the matrix A in the following way:

$$\begin{aligned}V \Lambda V^T &= \begin{pmatrix} v_1 & v_2 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} \\ &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \\ &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} = A.\end{aligned}$$

Example 5 (continued). Consider the matrix given in Equation (9.15).

- **Expressing A as a sum of n outer products.** Observe that we can write the matrix A in the following way:

$$\begin{aligned}\sum_{i=1}^2 \lambda_i v_i v_i^T &= \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T \\ &= 5 \begin{pmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{pmatrix} \begin{pmatrix} 1/\sqrt{5} & 2/\sqrt{5} \end{pmatrix} + 0 \begin{pmatrix} -2/\sqrt{5} \\ 1/\sqrt{5} \end{pmatrix} \begin{pmatrix} -2/\sqrt{5} & 1/\sqrt{5} \end{pmatrix} \\ &= 1 \begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 1 & 2 \end{pmatrix} + 0 \begin{pmatrix} -2 \\ 1 \end{pmatrix} \begin{pmatrix} -2 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} = A.\end{aligned}$$

- **Expressing A as a product of 3 matrices.** Observe also that we can write the matrix A in the

following way:

$$\begin{aligned}
 V\Lambda V^T &= (v_1 \ v_2) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} \\
 &= \begin{pmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{pmatrix} \begin{pmatrix} 5 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{pmatrix} \\
 &= \frac{1}{5} \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 5 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix} \\
 &= \frac{1}{5} \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 5 & 10 \\ 0 & 0 \end{pmatrix} \\
 &= \frac{1}{5} \begin{pmatrix} 5 & 10 \\ 10 & 20 \end{pmatrix} \\
 &= \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} = A.
 \end{aligned}$$

Example 6 (continued). Consider the matrix given in Equation (9.16).

- **Expressing A as a sum of n outer products.** Observe that we can write the matrix A in the following way:

$$\begin{aligned}
 \sum_{i=1}^2 \lambda_i v_i v_i^T &= \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T \\
 &= 5 \begin{pmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{pmatrix} \begin{pmatrix} 1/\sqrt{5} & 2/\sqrt{5} \end{pmatrix} - 5 \begin{pmatrix} -2/\sqrt{5} \\ 1/\sqrt{5} \end{pmatrix} \begin{pmatrix} -2/\sqrt{5} & 1/\sqrt{5} \end{pmatrix} \\
 &= \begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 1 & 2 \end{pmatrix} - \begin{pmatrix} -2 \\ 1 \end{pmatrix} \begin{pmatrix} -2 & 1 \end{pmatrix} \\
 &= \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} - \begin{pmatrix} 4 & -2 \\ -2 & 1 \end{pmatrix} \\
 &= \begin{pmatrix} -3 & 4 \\ 4 & 3 \end{pmatrix} = A.
 \end{aligned}$$

- **Expressing A as a product of 3 matrices.** Observe also that we can write the matrix A in the following way:

$$\begin{aligned}
 V\Lambda V^T &= (v_1 \ v_2) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} \\
 &= \begin{pmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{pmatrix} \begin{pmatrix} 5 & 0 \\ 0 & -5 \end{pmatrix} \begin{pmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{pmatrix} \\
 &= \frac{1}{5} \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 5 & 0 \\ 0 & -5 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix} \\
 &= \frac{1}{5} \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 5 & 10 \\ 10 & -5 \end{pmatrix} \\
 &= \frac{1}{5} \begin{pmatrix} -15 & 20 \\ 20 & 15 \end{pmatrix} \\
 &= \begin{pmatrix} -3 & 4 \\ 4 & 3 \end{pmatrix} = A.
 \end{aligned}$$

Example 7 (continued). Consider the matrix given in Equation (9.17).

- **Expressing A as a sum of n outer products.** Observe that we can write the matrix A in the following way:

$$\begin{aligned}
\sum_{i=1}^2 \lambda_i v_i v_i^T &= \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T \\
&= 5 \begin{pmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{pmatrix} \begin{pmatrix} 1/\sqrt{5} & 2/\sqrt{5} \end{pmatrix} + 10 \begin{pmatrix} -2/\sqrt{5} \\ 1/\sqrt{5} \end{pmatrix} \begin{pmatrix} -2/\sqrt{5} & 1/\sqrt{5} \end{pmatrix} \\
&= \begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 1 & 2 \end{pmatrix} + 2 \begin{pmatrix} -2 \\ 1 \end{pmatrix} \begin{pmatrix} -2 & 1 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} + \begin{pmatrix} 8 & -4 \\ -4 & 2 \end{pmatrix} \\
&= \begin{pmatrix} 9 & -2 \\ -2 & 6 \end{pmatrix} = A.
\end{aligned}$$

- **Expressing A as a product of 3 matrices.** Observe also that we can write the matrix A in the following way:

$$\begin{aligned}
V \Lambda V^T &= (v_1 \ v_2) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} \\
&= \begin{pmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{pmatrix} \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix} \begin{pmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{pmatrix} \\
&= \frac{1}{5} \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix} \\
&= \frac{1}{5} \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 5 & 10 \\ -20 & 10 \end{pmatrix} \\
&= \frac{1}{5} \begin{pmatrix} 45 & -10 \\ -10 & 30 \end{pmatrix} \\
&= \begin{pmatrix} 9 & -2 \\ -2 & 6 \end{pmatrix} = A.
\end{aligned}$$

Remark. We called these examples “nice” before since they could be expressed in terms of these two equivalent decompositions, and as we will see these two related decompositions are very nice.

9.9 A larger example

So far, we have focused on 2×2 matrices. For 2×2 matrices, the quadratic formula makes the computations very simple, and so we can illustrate the basic ideas. The same ideas hold for larger matrices. Here, we will illustrate this with a 3×3 example.

Let’s consider

$$A = \begin{pmatrix} 1 & 4 & 3 \\ 4 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix}. \quad (9.28)$$

Note that this matrix is symmetric. We will show that for this matrix, we can compute 3 eigenvalues and 3 corresponding eigenvectors that are pair-wise orthonormal and that thus form an orthonormal basis for \mathbb{R}^3 . To do so, consider

$$A - \lambda I = \begin{pmatrix} 1 - \lambda & 4 & 3 \\ 4 & 1 - \lambda & 0 \\ 3 & 0 & 1 - \lambda \end{pmatrix}.$$

In this case,

$$\begin{aligned}
 \det(A - \lambda I) &= (1 - \lambda)((1 - \lambda)^2 - 0) - 4(4(1 - \lambda) - 0) + 3(0 - 3(1 - \lambda)) \\
 &= (1 - \lambda)^3 - 25(1 - \lambda) \\
 &= (1 - \lambda)((1 - \lambda)^2 - 25) \\
 &= (1 - \lambda)(\lambda^2 - 2\lambda - 24) \\
 &= (1 - \lambda)(\lambda - 6)(\lambda + 4) \quad \text{if } \lambda = -4, 1, 6.
 \end{aligned}$$

So, the eigenvalues are $\lambda = -4, 1, 6$.

To compute the eigenspaces for each of these eigenvalues is somewhat more complex than for the 2×2 matrices we considered previously. It is, however, straightforward using a procedure (known as Gauss-Jordan reduction) that we won't cover here. Instead, we'll simply state the eigenvectors and verify that they are eigenvectors. Here they are:

$$\begin{aligned}
 \lambda_1 = 6, \quad v_1 &= \begin{pmatrix} 5 \\ 4 \\ 3 \end{pmatrix} : \begin{pmatrix} 1 & 4 & 3 \\ 4 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 5 \\ 4 \\ 3 \end{pmatrix} = \begin{pmatrix} 30 \\ 24 \\ 18 \end{pmatrix} = 6 \begin{pmatrix} 5 \\ 4 \\ 3 \end{pmatrix} \\
 \lambda_2 = 1, \quad v_2 &= \begin{pmatrix} 0 \\ -3 \\ 4 \end{pmatrix} : \begin{pmatrix} 1 & 4 & 3 \\ 4 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ -3 \\ 4 \end{pmatrix} = \begin{pmatrix} 0 \\ -3 \\ 1 \end{pmatrix} = 1 \begin{pmatrix} 0 \\ -3 \\ 4 \end{pmatrix} \\
 \lambda_3 = -4, \quad v_3 &= \begin{pmatrix} -5 \\ 4 \\ 3 \end{pmatrix} : \begin{pmatrix} 1 & 4 & 3 \\ 4 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix} \begin{pmatrix} -5 \\ 4 \\ 3 \end{pmatrix} = \begin{pmatrix} 20 \\ -16 \\ -12 \end{pmatrix} = -4 \begin{pmatrix} -5 \\ 4 \\ 3 \end{pmatrix}.
 \end{aligned}$$

While these three vectors are eigenvectors, note that we did not normalize them. We did that in order to avoid "carrying around" the normalization, and since such a normalization isn't necessary to confirm that they are in fact eigenvectors (since a scalar multiple of an eigenvector is an eigenvector, which recall is why we normalize them by convention).

Before stating the normalized eigenvectors, let's confirm that these 3 vectors are pairwise orthogonal to each other. To do so, let's define the 3×3 matrix V_{unn} to have i^{th} column equal to v_i , i.e., $V_{unn} = (v_1 \ v_2 \ v_3)$, and let's multiply it by its transpose:

$$V_{unn}^T V_{unn} = \begin{pmatrix} 5 & 4 & 3 \\ 0 & -3 & 4 \\ -5 & 4 & 3 \end{pmatrix} \begin{pmatrix} 5 & 0 & -5 \\ 4 & -3 & 4 \\ 3 & 4 & 3 \end{pmatrix} = \begin{pmatrix} 50 & 0 & 0 \\ 0 & 25 & 0 \\ 0 & 0 & 50 \end{pmatrix}.$$

So, these three vectors are eigenvectors, and they are orthogonal, and so they provide a basis for \mathbb{R}^3 . To normalize them, we divide by their norms, the square of which are the diagonal elements of $V_{unn}^T V_{unn}$. Here are the normalized eigenvectors.

$$\begin{aligned}
 \lambda_1 = 6 &: v_1 = \frac{1}{\sqrt{50}} \begin{pmatrix} 5 \\ 4 \\ 3 \end{pmatrix} \\
 \lambda_2 = 1 &: v_2 = \frac{1}{\sqrt{25}} \begin{pmatrix} 0 \\ -3 \\ 4 \end{pmatrix} \\
 \lambda_3 = -4 &: v_3 = \frac{1}{\sqrt{50}} \begin{pmatrix} -5 \\ 4 \\ 3 \end{pmatrix}.
 \end{aligned}$$

Let's now work with these normalized eigenvectors. In this case, the 3×3 matrix of normalized eigenvectors

(the columns of which form an orthonormal basis for \mathbb{R}^3) is:

$$V = \begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix} = \begin{pmatrix} \frac{5}{\sqrt{50}} & 0 & \frac{-5}{\sqrt{50}} \\ \frac{4}{\sqrt{50}} & \frac{-3}{\sqrt{25}} & \frac{4}{\sqrt{50}} \\ \frac{3}{\sqrt{50}} & \frac{4}{\sqrt{25}} & \frac{3}{\sqrt{50}} \end{pmatrix},$$

which can be shown to satisfy $V^T V = I$, and the diagonal matrix of eigenvalues is:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \begin{pmatrix} 6 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -4 \end{pmatrix},$$

in which case the eigenvalue equation can be written as a single matrix equation as follows:

$$AV = V\Lambda.$$

Next, one can show that the matrix given in Equation (9.28) can be expressed in the two standard forms we discussed. (For this, the normalization obviously is important.) Here is how do we express the matrix in two standard forms.

- **Expressing A as a sum of n outer products.** Observe that we can write the matrix A in the following way:

$$\begin{aligned} \sum_{i=1}^3 \lambda_i v_i v_i^T &= \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T + \lambda_3 v_3 v_3^T \\ &= 6 \begin{pmatrix} 5/\sqrt{50} \\ 4/\sqrt{50} \\ 3/\sqrt{50} \end{pmatrix} \begin{pmatrix} 5/\sqrt{50} & 4/\sqrt{50} & 3/\sqrt{50} \end{pmatrix} \\ &\quad + 1 \begin{pmatrix} 0/\sqrt{25} \\ -3/\sqrt{25} \\ 4/\sqrt{25} \end{pmatrix} \begin{pmatrix} 0/\sqrt{25} & -3/\sqrt{25} & 4/\sqrt{25} \end{pmatrix} \\ &\quad - 4 \begin{pmatrix} -5/\sqrt{50} \\ 4/\sqrt{50} \\ 3/\sqrt{50} \end{pmatrix} \begin{pmatrix} -5/\sqrt{50} & 4/\sqrt{50} & 3/\sqrt{50} \end{pmatrix} \\ &= \dots \\ &= A. \end{aligned}$$

- **Expressing A as a product of 3 matrices.** Observe also that we can write the matrix A in the following way:

$$\begin{aligned} V\Lambda V^T &= \begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \\ v_3^T \end{pmatrix} \\ &= \dots \\ &= A. \end{aligned}$$

You'll be asked to confirm that both of these expressions are correct, i.e., do equal A , in the homework.

9.10 Problems

9.10.1 Implementations and Applications of the Theory

1. XXX.
2. XXX.

9.10.2 Pencil-and-paper Problems

1. Consider the matrix

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

Compute the eigenvalues and eigenvectors of A , A^2 , A^{-1} , and $A + 4I$.

2. Recall that the level set of a real-valued function f of n real variables is a set where the function takes a given constant value c . More precisely,

$$L_c(f) = \{(x_1, \dots, x_n) | f(x_1, \dots, x_n) = c\}.$$

Consider each of the following three matrices:

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}, \quad A = \begin{pmatrix} -3 & 4 \\ 4 & 3 \end{pmatrix}, \quad \text{and} \quad A = \begin{pmatrix} 9 & -2 \\ -2 & 6 \end{pmatrix}.$$

For each matrix, sketch the level sets of the function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ defined as $f(x) = x^T Ax$, and describe in words what geometric structure these level sets represent.

3. Recall that for random variables X_1 and X_2 , the 2×2 matrix defined as

$$A = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix},$$

where $r \in [-1, 1]$, is known as the covariance matrix.

- (a) Compute the eigenvalues and eigenvectors of this matrix.
- (b) Verify that, given those eigenvalues and eigenvectors, you can reconstruct the matrix from $A = V\Lambda V^T$ and also from $A = \sum_{i=1}^2 \lambda_i v_i v_i^T$.
- (c) Sketch the level sets of $f(x) = x^T Ax$, and describe in words what geometric structure these level sets represent.

For each of these, pay particular attention to how the answer varies as $r \in [-1, 1]$ is varied.

4. Consider the matrix

$$A = \begin{pmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{pmatrix}.$$

- (a) Compute the eigenvectors and eigenvalues of A .
- (b) Compute A^2 and A^3 .
- (c) Compute the eigenvectors and eigenvalues of A^2 and A^3 , and compare with those of A .
- (d) Compute the eigenvectors and eigenvalues of

$$A^\infty = \begin{pmatrix} 0.6 & 0.6 \\ 0.4 & 0.4 \end{pmatrix}$$

and compare with those of A and A^2 and A^3 .

- (e) Recall the definition of a probability vector, and let p be a 2-vector that is a probability vector. What can you say about Ap ? What can you say about A^2p ? What can you say about A^3p ? What can you say about $A^\infty p$?

Chapter 10

Eigendecompositions: The Quadratic Forms Perspective

10.1 Quadratic forms and matrices

We have been viewing matrices in terms of transformations, i.e., a matrix $A \in \mathbb{R}^{m \times n}$ as a representation of a linear function from \mathbb{R}^n to \mathbb{R}^m , but we can also view matrices in terms of quadratic forms. Viewing matrices in terms of quadratic forms is a very powerful approach in machine learning and data science, and it makes many advanced high-dimensional concepts more intuitive. It also provides a very nice geometric way to think about eigenvalues and eigenvectors that complements the more algebraic approach we adopted in the last chapter and that is more common in traditional linear algebra classes. We will cover this approach in some detail.

Definition and examples. Let's start by saying what exactly we mean by a quadratic form. Here is the definition.

Definition 66 A quadratic form $Q : \mathbb{R}^n \rightarrow \mathbb{R}$ is a polynomial in the variables x_1, \dots, x_n , all of whose terms are of degree 2.

Here are several examples of quadratic forms.

$$\bullet \quad x_1^2 \quad (10.1)$$

$$\bullet \quad x_1^2 + x_2^2 \quad (10.2)$$

(In more elementary presentations, this is often written as $x^2 + y^2$, when one is not interested in generalizing to higher dimensions, but recall that we will use subscripts since we are interested in this generalization.)

$$\bullet \quad x_1^2 + x_2^2 + x_3^2 \quad (10.3)$$

$$\bullet \quad 4x_1^2 - x_2^2 \quad (10.4)$$

- $4x_1^2 + 2x_1x_2 - x_2^2 \quad (10.5)$

- $2x_1^2 - 3x_2^2 + 4x_3^2 - 5x_1x_2 + 6x_1x_3 - 7x_2x_3 \quad (10.6)$

Terms of the form $x_i x_j$, for $i \neq j$, e.g., $x_1 x_2$, $x_1 x_3$, and $x_2 x_3$ that appear in Equation (10.5) and (10.6), are sometimes called *cross terms*, since they involve the product of two different variables, rather than the product of a variable with itself. Terms of the form x_i^2 , that involve the product of a variable with itself, are sometimes called *diagonal terms*. A quadratic form can have diagonal terms or cross terms or both. Quadratic forms that do not have any cross terms, e.g., Equations (10.1) through (10.4), are sometimes called *diagonal quadratic forms*. (They are called diagonal terms since, as we will see, they are the diagonal elements of a certain matrix, in which case the cross terms are the off-diagonal elements.)

Here are few examples of things that are not quadratic forms.

- x_1
- $x_1 + x_2 + 7$
- $\exp(x_1 + x_2^2)$
- $\sin(x_1)$

As a seemingly-minor point, according to Definition 66, all the terms in a quadratic form must have degree equal to 2. In particular, the following is *not* a quadratic form:

- $4(x_1 - 3)^2 - x_2^2. \quad (10.7)$

To be more precise, Equation (10.7) is not a quadratic form in the variables x_1 and x_2 . The reason is that when we expand out $(x_1 - 3)^2$, we get a term linear in x_1 and also a constant term, as follows:

$$4x_1^2 - 24x_1 + 36 - x_2^2.$$

According to Definition 66, this is not a quadratic form. It is, however, “almost” a quadratic form, in the sense that if we define $x'_1 = x_1 - 3$, i.e., if we do a variable transformation that is a simple translation along the x_1 axis, then we get the following:

- $4x'^2_1 - x_2^2.$

According to Definition 66, this function is a quadratic form—in the variables x'_1 and x_2 , i.e., in the variables $x_1 - 3$ and x_2 —and thus Equation (10.7) is a quadratic form in the variables $x_1 - 3$ and x_2 .

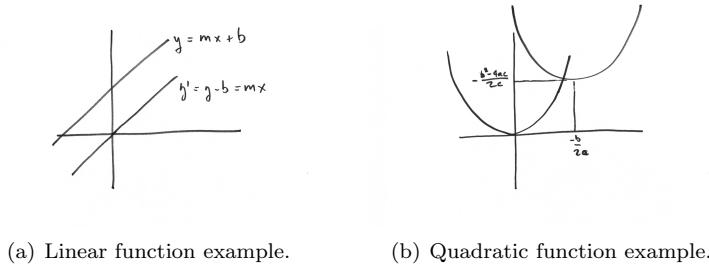


Figure 10.1: Illustration of redefining variables for linear and quadratic functions of one variable.

Importance of variable transformations. Understanding this seemingly-minor point will help to understand the geometric significance of eigenvalues and eigenvectors as well as how they arise.

A similar seemingly-minor point arose when we first discussed linear algebra. Recall that, with linear functions, we drew a distinction between a line $x_2 = ax_1 + b$ (which informally one thinks of as a linear function, but which is not technically a linear transformation) and a line through the origin $x_2 = ax_1$ (which is a linear transformation, according to the definition we presented), where the difference there was whether there was a non-zero constant or affine term. On the one hand, that simply amounted to a redefinition of the variables that involved translating the origin to a new coordinate system: $x_2 = ax_1 + b$ is simply $x'_2 = ax_1$, where $x'_2 = x_2 - b$. On the other hand, that permitted us to talk about linear transformations, linear subspaces, etc., and thereby build up the machinery of linear algebra.

Here too, with quadratic forms, we will draw a distinction between a quadratic form and something that is “almost” a quadratic form but which has lower order linear and affine terms. Redefining variables, i.e., performing a variable transformation, will permit us to remove those lower-order terms, which will help simplify/clarify the discussion. As with the linear case of removing the lower-order affine term, this redefinition involves defining new variables that are translations of original variables; and this will yield a quadratic form, but one with “cross terms” such as x_1x_2 (as opposed to x_1^2 and x_2^2). Further redefining variables will lead to a quadratic form with no cross terms, i.e., which will consist of just terms of the form x_i^2 and no terms of the form x_ix_j , and this further redefinition will involve rotations (actually, orthogonal transformations, which include rotations as well as reflections). Doing this will help us to define eigenvectors and eigenvalues more easily.

The connection with matrices. What, you may ask, is the connection between these quadratic forms and matrices?

To answer this, let’s start with the one-variable case. Consider the function $f : \mathbb{R} \rightarrow \mathbb{R}$ that is

$$f(x) = a + bx. \quad (10.8)$$

This function is the sum of an affine part (a) and a linear part (bx). Next, still working with one variable, consider the function $f : \mathbb{R} \rightarrow \mathbb{R}$ that is

$$f(x) = a + bx + cx^2. \quad (10.9)$$

This function is the sum of an affine part (a) and a linear part (bx) and a quadratic part (cx^2).

We know that we can view the equation of the line in Equation (10.8) as a linear function (in the sense that it satisfies the definition of a linear function) by “removing” the affine part and considering the function

$$g(x) = f(x) - a = bx.$$

This simply amounts to “shifting” or “translating” the function along the Y axis (if we view the function as $y = f(x)$) so that it goes through the origin. See Figure 10.1(a) for an illustration.

Similarly, we can view the equation of the quadratic function in Equation (10.9) as a quadratic form by “removing” the linear and affine parts. To do so, one can use the procedure known as *completing the square*.

Let’s remind ourselves how completing the square works in the univariate quadratic case. In this case, we are working with a function of the form of Equation (10.9), where $c \neq 0$. We first factor out the c and we then add and subtract the appropriate quantity to rewrite the quadratic term in such a way to remove the linear term:

$$\begin{aligned} f(x) &= cx^2 + bx + a \\ &= c\left(x^2 + \frac{b}{c}x + \frac{a}{c}\right) \\ &= c\left(\left(x + \frac{b}{2c}\right)^2 - \frac{b^2}{4c^2} + \frac{a}{c}\right) \\ &= c\left(\left(x + \frac{b}{2c}\right)^2 - \frac{b^2 - 4ac}{4c^2}\right) \\ &= c\left(x + \frac{b}{2c}\right)^2 - \frac{b^2 - 4ca}{4c}, \end{aligned} \tag{10.10}$$

and thus we have a quadratic with a vertex at

$$\left(-\frac{b}{2c}, -\frac{b^2 - 4ac}{4c}\right).$$

To solve a quadratic equation of the form $f(x) = 0$, we set this to zero, in which case we have that

$$\left(x + \frac{b}{2c}\right)^2 = \frac{b^2 - 4ca}{4c^2},$$

and thus that

$$x = -\frac{b}{2c} \pm \frac{\sqrt{b^2 - 4ca}}{2c},$$

which provides the usual quadratic formula. Alternatively, if we define

$$\begin{aligned} g(x) &= f(x) + \frac{b^2 - 4ca}{4c} \\ x' &= x + \frac{b}{2c}, \end{aligned}$$

i.e., if we simply shift the two coordinate axes, then we can write Equation (10.10) as

$$g(x) = cx'^2,$$

i.e., as a quadratic form in the variable x' , which is a simple transformation of the original variable x . That is, completing the square in a quadratic functions simply amounts to shifting or translating the function, here along both the X and Y axes, in order to construct a quadratic form. See Figure 10.1(b) for an illustration.

Next, let’s go to the two-variable case. Consider the function $f(x) : \mathbb{R}^2 \rightarrow \mathbb{R}$ given as

$$\begin{aligned} f(x) &= a + b_1x_1 + b_2x_2 \\ &= a + b^T x. \end{aligned} \tag{10.11}$$

In this expression, b can be viewed in one of two complementary ways: first, b is a vector, in which case $b^T x$ is a number that is the dot product of b and x ; and second, b^T is a 1×2 matrix that maps $x \in \mathbb{R}^2$ to a number in \mathbb{R} . In either case, the function f is the sum of an affine part (a) and a linear part ($b^T x$). As

before, we can convert this function in Equation (10.11) to a linear function by considering the translated function

$$g(x) = f(x) - a = b^T x.$$

Next, still working with two variables, we can add a quadratic part to consider the function $f(x) : \mathbb{R}^2 \rightarrow \mathbb{R}$ as follows:

$$\begin{aligned} f(x) &= a + b_1 x_1 + b_2 x_2 + c_1 x_1^2 + c_2 x_2^2 + c_3 x_1 x_2 \\ &= a + (b_1 \ b_2) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + (x_1 \ x_2) \begin{pmatrix} c_1 & c_3/2 \\ c_3/2 & c_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= a + b^T x + x^T C x. \end{aligned} \quad (10.12)$$

As before, we will be able to remove the linear and affine parts ($b^T x$ and a , respectively), again using the completing the square and shifting procedures. This too will lead to a quadratic form, but one with cross terms of the form $x_1 x_2$ and not just x_1^2 or x_2^2 . Removing these cross terms in order to get a much simpler quadratic form without any cross terms will be closely related to computing eigenvectors and eigenvalues.

To go to the three-variable case and beyond, observe that while Equation (10.11) and Equation (10.12) have been derived for a function $f(x) : \mathbb{R}^2 \rightarrow \mathbb{R}$, exactly the same expressions could be derived for $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$, i.e., if the input to this function were an n -dimensional vector, rather than a 2-dimensional vector. The reason is that the equations just involve dot products and matrix-vector multiplications, and there is no explicit dependence on the dimension of the input to this function. This suggests that we can write a general quadratic function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ in terms of an $n \times n$ symmetric matrix, a n -dimensional vector, and a number. This is true. As an example, for $x \in \mathbb{R}^3$, the generalization of Equation (10.12) is

$$f(x) = a + (b_1 \ b_2 \ b_3) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + (x_1 \ x_2 \ x_3) \begin{pmatrix} c_{11} & \frac{1}{2}c_{12} & \frac{1}{2}c_{13} \\ \frac{1}{2}c_{12} & c_{22} & \frac{1}{2}c_{23} \\ \frac{1}{2}c_{13} & \frac{1}{2}c_{23} & c_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad (10.13)$$

and the generalization to \mathbb{R}^n , for $n \geq 4$, is straightforward but cumbersome. Before we consider this extension to \mathbb{R}^n , however, let's consider \mathbb{R}^3 to illustrate an important point about relating quadratic functions to matrices.

A potential problem. Next, let's now turn to a *potential* problem with viewing matrices as quadratic forms. To illustrate this point, for now, to save space, let's consider the case where the function $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ has just quadratic terms, i.e., where it is a quadratic form. In this case, we are considering

$$f(x) = c_{11} x_1^2 + c_{22} x_2^2 + c_{33} x_3^2 + c_{12} x_1 x_2 + c_{13} x_1 x_3 + c_{23} x_2 x_3. \quad (10.14)$$

We can write this as a matrix—or more precisely as the product of three matrices—as follows:

$$f(x) = (x_1 \ x_2 \ x_3) \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ 0 & c_{22} & c_{23} \\ 0 & 0 & c_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

To see that this is true, just multiply out the RHS and see that we get the RHS of Equation (10.14). The point here is that we can start with an arbitrary quadratic form and construct in a very natural way a matrix.

Note however that, even if we restrict ourselves to quadratic functions that contain just terms of degree 2, this quadratic form does *not* uniquely define a matrix. For example, we could have put the c_{12} , c_{13} , and c_{23} below the diagonal, with 0s above the diagonal as follows:

$$f(x) = (x_1 \ x_2 \ x_3) \begin{pmatrix} c_{11} & 0 & 0 \\ c_{12} & c_{22} & 0 \\ c_{13} & c_{23} & c_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

Again, to see that this is true, just multiply out the RHS and see that we get the RHS of Equation (10.14). Alternatively, we can write it in a more symmetric form as follows:

$$f(x) = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} c_{11} & \frac{1}{2}c_{12} & \frac{1}{2}c_{13} \\ \frac{1}{2}c_{12} & c_{22} & \frac{1}{2}c_{23} \\ \frac{1}{2}c_{13} & \frac{1}{2}c_{23} & c_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

Again, to see that this is true, just multiply out the RHS and see that we get the RHS of Equation (10.14). In this last expression, we have split the “off diagonal” terms, i.e., c_{ij} , for $i \neq j$, into half associated with the $x_i x_j$ term and half associated with the $x_j x_i$ term. Right now, we just want to note that this is possible, i.e., we didn’t have to do this for the above expression to be true (as we saw with the previous two non-symmetric examples that also reproduce the same $f(x)$). It turns out, however, to be very convenient to do this. We’ll get back to why this is the case soon.

Let’s go the “other way,” i.e., let’s start with an arbitrary square matrix, starting with an arbitrary 3×3 matrix for simplicity,

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \in \mathbb{R}^{3 \times 3}, \quad (10.15)$$

where, in particular, we permit the possibility that $A_{12} = A_{21}$ as well as the possibility that $A_{12} \neq A_{21}$, and similarly for the other off-diagonal terms. In this case, we have that

$$Ax = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} A_{11}x_1 + A_{12}x_2 + A_{13}x_3 \\ A_{21}x_1 + A_{22}x_2 + A_{23}x_3 \\ A_{31}x_1 + A_{32}x_2 + A_{33}x_3 \end{pmatrix} \in \mathbb{R}^3,$$

and we also have that

$$\begin{aligned} x^T Ax &= \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= A_{11}x_1^2 + (A_{12} + A_{21})x_1x_2 + (A_{13} + A_{31})x_1x_3 + A_{22}x_2^2 + (A_{23} + A_{32})x_2x_3 + A_{33}x_3^2 \in \mathbb{R}. \end{aligned}$$

The point of this is to show that we can start with an arbitrary square matrix and get a quadratic form. (I.e., we don’t need to have a symmetric matrix to have a quadratic form.) Clearly, if we define the matrix A' to be

$$A' = \begin{pmatrix} A_{11} & \frac{A_{12}+A_{21}}{2} & \frac{A_{13}+A_{31}}{2} \\ \frac{A_{12}+A_{21}}{2} & A_{22} & \frac{A_{23}+A_{32}}{2} \\ \frac{A_{13}+A_{31}}{2} & \frac{A_{23}+A_{32}}{2} & A_{33} \end{pmatrix} \in \mathbb{R}^{3 \times 3}, \quad (10.16)$$

then $x^T Ax = x^T A'x$, for all $x \in \mathbb{R}^3$, illustrating again the same non-uniqueness.

Here is a specific example of two matrices that correspond to the same quadratic form.

Example. If we let $A_1 = \begin{pmatrix} 1 & 2 & 4 \\ 0 & 6 & 8 \\ 0 & 0 & 9 \end{pmatrix}$, and $A_2 = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 6 & 4 \\ 2 & 4 & 9 \end{pmatrix}$, then

$$\begin{aligned} x^T A_1 x &= \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} 1 & 2 & 4 \\ 0 & 6 & 8 \\ 0 & 0 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= x_1^2 + 6x_2^2 + 9x_3^2 + 2x_1x_2 + 4x_1x_3 + 8x_2x_3 \\ &= \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} 1 & 1 & 2 \\ 1 & 6 & 4 \\ 2 & 4 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= x^T A_2 x. \end{aligned}$$

This discussion illustrates that we can have many different square matrices that give rise to the same quadratic form. We might wonder whether there is some sort of standard or canonical form in which we can write a matrix that removes this non-uniqueness, since then we can write any quadratic form in terms of a unique matrix. The answer to this is yes. Basically, the way to do this is to do the above “average out” and write the quadratic form as a *symmetric* matrix (like we did above). That is, consider symmetric matrices. We could do it other ways, e.g., put everything above the diagonal, but we will do it this way since symmetric matrices have so many nice properties. Conversely, for any symmetric matrix, there is an associated quadratic form.

The point is the following. In general, if A is a square $n \times n$ matrix, i.e., if $A \in \mathbb{R}^{n \times n}$, then the function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ given by

$$f(x) = x^T A x$$

is a quadratic form. On the one hand, it is the product of three matrices, two of which are vectors, that for a given x and A yields a 1×1 matrix that is a number. On the other hand, it is the sum of a bunch of terms, each of which consists of elements $(A_{ij} + A_{ji})$ of the matrix A multiplied by the product $x_i x_j$ of variables from the vector x . If we consider a matrix in which the off-diagonal position A_{ij} is replaced with $\frac{1}{2}(A_{ij} + A_{ji})$, then we get the same quadratic form, and so we will work with symmetric matrices.

Here is the statement of the basic result.

Theorem 18 *Given an arbitrary quadratic form in n variables (which, recall can be written as $x^T A x$ for a square matrix $A \in \mathbb{R}^{n \times n}$), we can always find a symmetric matrix $B \in \mathbb{R}^{n \times n}$ such that*

$$x^T A x = x^T B x, \quad \forall x \in \mathbb{R}^n.$$

So, due to this result, it is typical just to assume that A is symmetric, and so we will do that.

We will be a little more precise/formal about these things in Section 10.3, but let’s give a few more examples first.

Advanced aside. *As a more advanced aside, actually more generally we get a type of quadratic form from any matrix. In this more general case, however, the variables are different dimensions. For example, if*

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{pmatrix},$$

then we can pre-multiply by a vector $x \in \mathbb{R}^2$ and post-multiply by a vector $y \in \mathbb{R}^3$ to get the following:

$$\begin{aligned} x^T A y &= (x_1 \ x_2) \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \\ &= A_{11}x_1y_1 + A_{12}x_1y_2 + A_{13}x_1y_3 + A_{21}x_2y_1 + A_{22}x_2y_2 + A_{23}x_2y_3 \in \mathbb{R} \end{aligned}$$

This is of interest in some contexts, but it is a more advanced topic, and so we won’t worry about it in this class. Instead, we will focus on square matrices, in which case the vectors on the two sides of the matrix are the same dimensionality, in which case we can represent them with the same variable.

10.2 Some simple examples

Let’s start with a few simple examples that we have already seen.

Example. If $A = I_n$, then we get the following:

$$\begin{aligned} f(x) &= x_1^2 + x_2^2 + \cdots + x_n^2 \\ &= (x_1 \ \cdots \ x_n) \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \\ &= x^T I x \\ &= x^T x \\ &= \|x\|_2^2. \end{aligned}$$

This is the usual Euclidean norm of a vector $x \in \mathbb{R}^n$.

Example. If $A = D$, a diagonal matrix, e.g.,

$$D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{pmatrix} \in \mathbb{R}^3,$$

then we get

$$\begin{aligned} f(x) &= x_1^2 + 4x_2^2 + 9x_3^2 \\ &= (x_1 \ x_2 \ x_3) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= x^T D x \\ &= (D^{1/2} x)^T D^{1/2} x \\ &= \|x\|_{2,D^{1/2}}^2. \end{aligned}$$

Observe that, if we define the vector $x' = D^{1/2}x \in \mathbb{R}^3$, then this is just the Euclidean norm of the vector x' .

Example. If $A = D$, a diagonal matrix, e.g.,

$$D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & -9 \end{pmatrix} \in \mathbb{R}^3,$$

then we get

$$\begin{aligned} f(x) &= x_1^2 + 4x_2^2 - 9x_3^2 \\ &= (x_1 \ x_2 \ x_3) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & -9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= x^T D x. \end{aligned}$$

In this case, we can't take the square root like we did in the previous example. (Below, we will get to a richer interpretation of the difference between these two examples.)

Example. If $A = D$, a diagonal matrix, e.g.,

$$D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^3,$$

then we get

$$\begin{aligned} f(x) &= x_1^2 + 4x_2^2 \\ &= \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= x^T D x. \end{aligned}$$

In this case, we still obtain a quadratic form, but one for which the coefficient of the x_3 term equals 0.

Problem. Consider the matrix

$$D = \begin{pmatrix} \alpha_1 & 0 & 0 \\ 0 & \alpha_2 & 0 \\ 0 & 0 & \alpha_3 \end{pmatrix} \in \mathbb{R}^3.$$

- Show that if $\alpha_i > 0$, for all i , then the quadratic form function $f(x) = x^T D x$ is a vector norm, in the sense that it satisfies the three conditions for a function to be a vector norm.
- Show that if $\alpha_i \geq 0$, but $\alpha_i = 0$ for at least one i , then some of the conditions for a function to be a norm are satisfied and some are violated. Determine which are violated and which are not violated. (In this case, the function is sometimes called a semi-norm.)
- If $\alpha_i > 0$ for some i and $\alpha_i < 0$ for other i , then determine which of the conditions for a function to be a norm are satisfied and which are violated.

Example. If A is a matrix that can be written as $A = B^T B$, i.e., it is a correlation/covariance matrix, then we get

$$\begin{aligned} f(x) &= x^T B^T B x \\ &= (Bx)^T B x \\ &= \|x\|_{2,B}^2. \end{aligned}$$

Observe that, if we define the vector $x' = Bx \in \mathbb{R}^3$, then this is just the Euclidean norm of the vector x' . Of course, a special case of this is the diagonal matrix with all positive entries that we saw above.

10.3 Symmetric bi-linear functions

More abstractly, in the same way that any matrix can be viewed as a representation of a *linear transformation* with respect to a basis, we can also associate to any symmetric matrix something that is known as a *symmetric bilinear transformation*. Of course, a symmetric matrix is also a matrix, so we can view it in terms of linear transformations. Since we are considering a smaller class of matrices, however, we have more structure, and this will help us to say more about them. In this case, the “more” will be intuitive geometric things having to do with eigenvalues and eigenvectors that are of widespread importance in machine learning and data science.

Let's start with the definition.

Definition 67 Let V be a vector space, e.g., \mathbb{R}^2 . Then a symmetric bilinear function on V is a mapping $B : V \times V \rightarrow \mathbb{R}$ such that

$$\begin{aligned} B(av_1 + bv_2, w) &= aB(v_1, w) + bB(v_2, w) \quad \forall v_1, v_2 \in V, a, b \in \mathbb{R} \\ B(v, w) &= B(w, v) \quad \forall v, w \in V. \end{aligned}$$

About this definition, we note the following:

1. The first condition is that B is linear in the first argument. (If we ignore the second argument, i.e., treat it as a constant, then this is the linearity condition that we have seen before.)
2. The second condition says that B is symmetric in its two arguments.
3. By combining these two results we also have that B is linear in its second argument.

Here are two special cases to compare this to the linear function perspective on matrices.

- Restricted to the case of 1×1 symmetric matrices, $A = (a)$, if we view this matrix as a linear function, then we are thinking of it as $y = Ax$, while if we think of this matrix as a symmetric bilinear function, then we are thinking of it as $y = Ax^2$.
- Restricted to the case of 2×2 symmetric matrices,

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{pmatrix}$$

(where the two off-diagonal terms are the same due to the symmetry), if we view this as a linear function, then we are thinking of this as

$$y = \begin{pmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in \mathbb{R}^2,$$

while if we are thinking of it as a symmetric bilinear function, then we are thinking of it as

$$y = (x_1 \ x_2) \begin{pmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}.$$

(Note that this is actually not a fully-general case. It is a symmetric bilinear function, but we are only considering the case where both arguments are the same.)

You may ask, which of these views should you choose? The answer is whichever you prefer. Neither one is “right” or “wrong.” Both are correct. In some cases one is more preferable than the other. That is, which one you adopt depends on which one is more useful in a given situation.

For example, viewing a symmetric matrix in terms of linear transformations leads to the notions of linear dependence, etc., that we discussed before. One can also describe eigenvalues/eigenvectors in terms of it—the determinant discussion was an example of this. On the other hand, viewing a symmetric matrix in terms of symmetric bilinear functions leads to the quadratic form perspective that has a cleaner geometric interpretation. It will also lead to a more geometric way to discuss eigenvalues/eigenvectors. Of course, there are many connections between the two perspectives, and we will discuss this.

Here is a basic result. (This is the quadratic form analogue of the theorem that said that any specific matrix can be viewed as an abstract linear transformation, and vice versa.)

Theorem 19 *We have the following.*

- If A is a symmetric $n \times n$ matrix, then $B_A(v, w) = v^T Aw$ is a symmetric bilinear form.
- If B is a symmetric bilinear function on \mathbb{R}^n , then it is of the form $B = B_A(v, w) = v^T Aw$, for some unique symmetric matrix A .

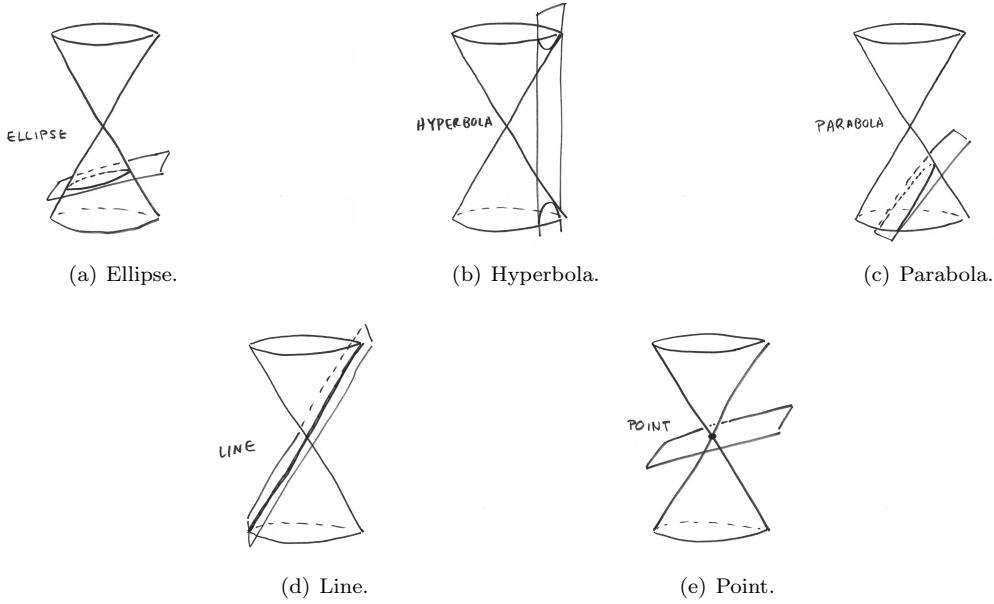


Figure 10.2: Illustration of the basic conic sections as the intersection of a cone and a plane.

For the second bullet in this theorem, note that there are also non-symmetric matrices for which the expression is true. (That has to do with the non-uniqueness we described above.) As with the example above when we symmetrized the matrix by spreading/averaging the elements above and below the diagonal, we can always symmetrize it in this manner, and Theorem 19 says that we get a unique such matrix A .

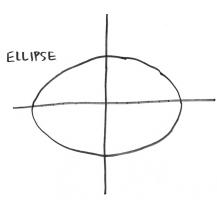
This may sound abstract or complicated, but the basic ideas are quite simple. Basically, this says that there is a clean mapping between these quadratic forms and symmetric matrices, and you should freely think of one in terms of the other.

Being pedantic about this is important for more advanced work in more advanced classes, and this is analogous to being pedantic about A as a linear function versus A as the representation of a linear function. Here, we won't be pedantic. Instead, we simply want to point out the connections, so you start to think of it in this way.

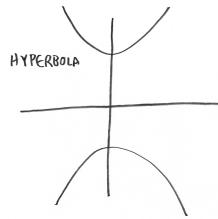
10.4 Connections with conic sections

We saw before that subspaces, linear dependence, etc. have precise definitions, and that these definitions generalized the intuitive geometric ideas that we have about lines through the origin, planes through the origin, etc. For the quadratic form perspective, Definition 66 and Theorem 19 provide precise statements, and these too generalize intuitive geometric ideas that are related to (hopefully familiar) conic sections and their generalizations.

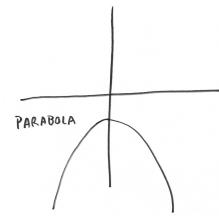
Overview of conic sections. You may recall that conic sections (ellipses, hyperbolas, and parabolas) result—geometrically—from the intersection of a cone and a plane. See Figure 10.2 for an illustration of the basic conic sections as the intersection of a cone and a plane, and see Figure 10.3 for an illustration of the basic conic sections on the two-dimensional plane. Some people like viewing them geometrically, while others like viewing them algebraically. Relevant to our discussion, when viewed algebraically, conic sections arise from considering quadratic functions of two variables.



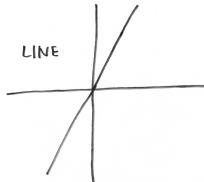
(a) Ellipse.



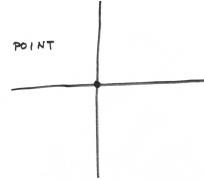
(b) Hyperbola.



(c) Parabola.



(d) Line.



(e) Point.

Figure 10.3: Illustration of the basic conic sections on the two-dimensional plane.

- **Ellipse.** An ellipse is the set of points in a plane, the sum of whose distances from two fixed points (called the foci) is a constant. Here is the equation in a standard form:

$$\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} = 1.$$

See Figures 10.2(a) and 10.3(a) for an illustration of an ellipse. Note that the major axis and minor axis of the ellipse in standard form are just the canonical axes. WLOG, let's assume that $a \geq b \geq 0$, otherwise it is longer along the other axis.

Example. Consider the graph of $9x_1^2 + 16x_2^2 = 144$. First divide by 144 to convert this into the standard form:

$$\frac{x_1^2}{16} + \frac{x_2^2}{9} = 1.$$

Remark. A circle is just an ellipse with $a = b$.

- **Hyperbola.** A hyperbola is the set of points in the plane, the difference of whose distances from two fixed points (called the foci) is a constant. Here is the equation written in a standard form:

$$\frac{x_1^2}{a^2} - \frac{x_2^2}{b^2} = 1.$$

See Figures 10.2(b) and 10.3(b) for an illustration of a hyperbola.

Example. Consider the graph of $9x_1^2 - 16x_2^2 = 144$. First divide by 144 to convert this into the standard form:

$$\frac{x_1^2}{16} - \frac{x_2^2}{9} = 1.$$

- **Degenerate cases.** In addition to the ellipse and hyperbola, there are several examples of cases where the line intersecting the cone “just barely” intersects the cone, e.g., just at the origin where the cone touches itself, or just at the edge of the cone, or just parallel to the edge of the cone. These can be viewed geometrically or algebraically.

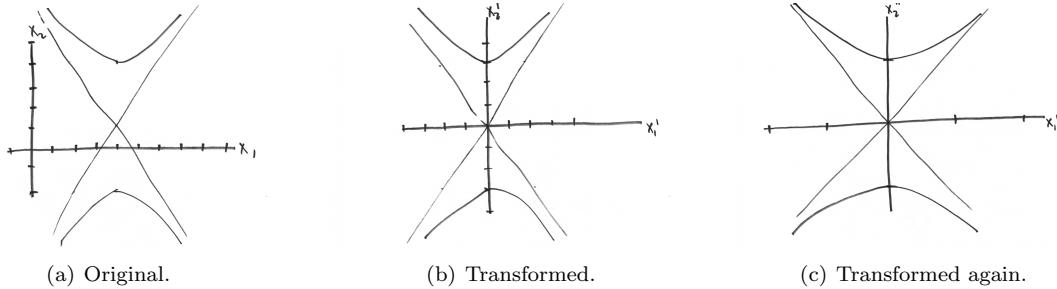


Figure 10.4: Illustration of transforming variables in a hyperbola.

- **Parabola.** This is the set of points in the plane that are equidistant from a fixed point and a fixed line. See Figures 10.2(c) and 10.3(c) for an illustration of a parabola.
- **Line.** This arises when the quadratic and constant terms are zero; we will get to below. See Figures 10.2(d) and 10.3(d) for an illustration of a line.
- **Point.** This arises when the quadratic and linear and constant terms are zero; we will get to below. See Figures 10.2(e) and 10.3(e) for an illustration of a point.

Remark. Note that the a and b above basically correspond to stretching the units of the x_1 and x_2 axes.

Examples of more complex conic sections. Let's say that, instead of being given a conic section in standard form, we are given some quadratic function expression and we have to determine the type of conic section to which it corresponds.

Question: What if we are given a slightly more general example, e.g., one that contains both linear terms and/or constant terms and/or cross-terms of the form $x_i x_j$, for $i \neq j$?

Answer: A general approach to simplify the expression is to use the familiar technique of completing the square, repeatedly if necessary.

Example. Let's sketch the graph of the conic section

$$9x_1^2 - 4x_2^2 - 72x_1 + 8x_2 + 176 = 0.$$

To do so, let's try to write it in one of the standard conic forms. To do so, complete the square. First write terms in x_1 and x_2 separately.

$$9(x_1^2 - 8x_1) - 4(x_2^2 - 2x_2) + 176 = 0.$$

Then, complete the square to get

$$9(x_1 - 4)^2 - 4(x_2 - 1)^2 + 176 - 144 + 4 = 0,$$

and then simplify this to get

$$9(x_1 - 4)^2 - 4(x_2 - 1)^2 + 36 = 0.$$

Moving the 36 to the other side of the equation and dividing by it gives

$$\frac{(x_1 - 4)^2}{4} - \frac{(x_2 - 1)^2}{9} = 1.$$

This is a hyperbola. See Figure 10.4(a).

There are two things to note about this example. First, since the negative sign is on the x_1 and not the x_2 , the hyperbola opens up-down and not left-right. Second, having those other linear and constant terms

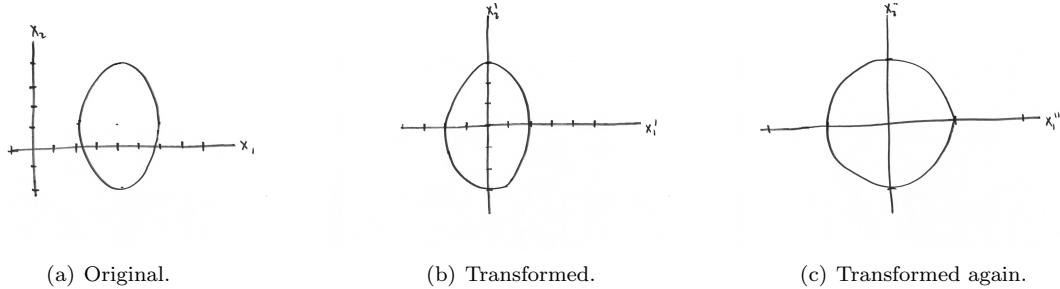


Figure 10.5: Illustration of transforming variables in an ellipse.

simply amounts to shifting the origin, i.e., defining a new set of variables. That is, if we define a new set of coordinates

$$\begin{aligned}x'_1 &= x_1 - 4 \\x'_2 &= x_2 - 1,\end{aligned}$$

which simply amounts to shifting the origin, then we get the equation

$$-\frac{(x'_1)^2}{4} + \frac{(x'_2)^2}{9} = 1.$$

See Figure 10.4(b).

BTW, if we then define yet another new set of coordinates

$$\begin{aligned}x''_1 &= \frac{x'_1}{2} \\x''_2 &= \frac{x'_2}{3},\end{aligned}$$

which simply amounts to stretching each of the axes, then we get a still simpler figure. See Figure 10.4(c).

Example. Let's sketch the graph of the conic section

$$\frac{(x_1 - 4)^2}{4} + \frac{(x_2 - 1)^2}{9} = 1.$$

This is an ellipse. See Figure 10.5(a). If we define a new set of coordinates

$$\begin{aligned}x'_1 &= x_1 - 4 \\x'_2 &= x_2 - 1,\end{aligned}$$

which simply amounts to shifting the origin, then we get the equation

$$\frac{(x'_1)^2}{4} + \frac{(x'_2)^2}{9} = 1.$$

See Figure 10.5(b). If we then define yet another new set of coordinates

$$\begin{aligned}x''_1 &= \frac{x'_1}{2} \\x''_2 &= \frac{x'_2}{3},\end{aligned}$$

which simply amounts to stretching each of the axes, then we get a still simpler figure. See Figure 10.5(c).

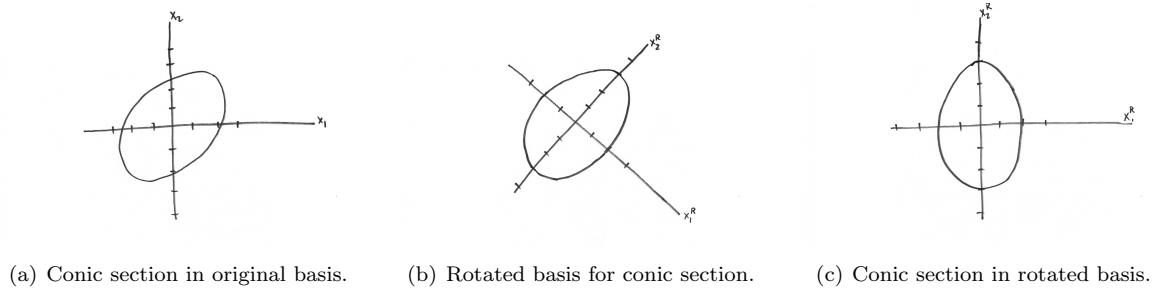


Figure 10.6: Illustration of redefining variables rotationally for conic section of two variable.

Question: What if there is an x_1x_2 term?

Answer: We still complete the square, but using the other variables. We will describe this in more detail below. In this case, we will see that A has a rotation, i.e., we don't have just a scaling and translation shift of the axes like in the previous example. Let's illustrate that.

Consider the following example.

Example. Consider the expression

$$5x_1^2 - 4x_1x_2 + 5x_2^2 = 48,$$

which is an ellipse, with major and minor axes rotated by a 45° degree angle, relative to the canonical axes. See Figure 10.6(a). (Actually, it shouldn't be immediately obvious that this is even an ellipse—but we do know that it is a symmetric bilinear equation, and so it satisfies the algebraic conditions of being a conic section, so it will be a conic section of some form.) This expression can be written as follows:

$$48 = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 5 & -2 \\ -2 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x^T Ax,$$

where the symmetric matrix associated with this quadratic form is:

$$A = \begin{pmatrix} 5 & -2 \\ -2 & 5 \end{pmatrix}.$$

Recall that a rotation in the x_1x_2 plane takes the form

$$R_\theta = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}.$$

and the coefficients of x_1x_2 term in the quadratic form is related to the angle θ . Since this is a 2×2 matrix, this is a counterclockwise rotation by θ degrees. If we choose $\theta = 45^\circ$, then we have

$$R_{\theta=45^\circ} = \begin{pmatrix} \cos(45^\circ) & -\sin(45^\circ) \\ \sin(45^\circ) & \cos(45^\circ) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Let's define a new coordinate system to be

$$x' = R_{\theta=45^\circ} x.$$

In more detail, we can write this new coordinate system as follows

$$\begin{pmatrix} x'_1 \\ x'_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} x_1 + x_2 \\ x_1 - x_2 \end{pmatrix}.$$

In the new (x'_1, x'_2) coordinate system, the associated conic section is

$$48 = 3{x'_1}^2 + 7{x'_2}^2 = 3 \left(\frac{x_1 + x_2}{\sqrt{2}} \right)^2 + 7 \left(\frac{x_1 - x_2}{\sqrt{2}} \right)^2.$$

See Figure 10.6(b) for the rotated basis and the ellipse in this rotated basis.

In this new (x'_1, x'_2) coordinate system, the original expression can be written as follows:

$$48 = \begin{pmatrix} x'_1 & x'_2 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 7 \end{pmatrix} \begin{pmatrix} x'_1 \\ x'_2 \end{pmatrix} = x'^T A' x'.$$

What this expression says is that, in the new coordinate system, the matrix A becomes a diagonal matrix A' , defined as

$$A' = \begin{pmatrix} 3 & 0 \\ 0 & 7 \end{pmatrix}.$$

See Figure 10.6(c).

Let's pause to emphasize the significance of this. What this says is that when we have cross terms such as $x_1 x_2$, we have a conic section with major and minor axes that are rotated relative to the canonical basis, and that we can view the coordinate transformation illustrated in Figure 10.6 that makes the ellipse axis-aligned in one of two complementary ways:

- As defining a new coordinate system as follows:

$$x \rightarrow x' = Rx,$$

where $R = R_{\theta=45^\circ}$.

- As defining a new matrix as follows:

$$A \rightarrow A' = R^T A R$$

To be compatible with notation we will use in the next chapter (and that we already used in the last chapter), let's do the following.

- Since the transformation R is an orthogonal transformation, let's call it V^T , i.e., $V = R^T$.
- Since the matrix A' is a diagonal matrix, let's call it D , i.e., $D = A'$.

Then, $x' \in \mathbb{R}^2$ is given by $x' = V^T x$, and if we then look at the quadratic form $x^T A x$ in the new coordinate system, then we have

$$x^T A x = x^T (V D V^T) x = x^T V D V^T x = (V^T x)^T D (V^T x) = (x')^T D x'.$$

Either way, we get

$$x^T A x = 5x_1^2 - 4x_1 x_2 + 5x_2^2 = 3 \left(\frac{x_1 + x_2}{\sqrt{2}} \right)^2 + 7 \left(\frac{x_1 - x_2}{\sqrt{2}} \right)^2 = 3x'^2 + 7x'^2 = x'^T A' x'.$$

The bottom line is that the cross term amounts to doing some sort of orthogonal transformation, and this involves defining a new coordinate system where the matrix is diagonal.

10.5 Definiteness, indefiniteness, and quadratic forms as a sum/difference of squares

The taxonomy of conic sections into ellipses, hyperbolas, and parabolas, as well as lines and points, is very useful for quadratic forms in 2 variables, but it quickly becomes awkward for quadratic forms in more than

2 variables. For quadratic functions in n variables, a related but slightly different classification is more convenient.

To understand the generalization, recall that the distinction into ellipses, hyperbolas, and parabolas depended on whether, when written in standard form, the coefficients of the variables were all the same sign (both positive, or both negative) or were different signs (one positive and one negative) or had one zero (one positive and one zero). The generalization we will consider will be a generalization of this condition.

- Definite: all positive or all negative.
- Degenerate: all non-negative including some zero, or all non-positive including some zero, i.e., all are either positive or zero, with not all positive, or all are either negative or zero, with not all negative.
- Indefinite: some positive or some negative, including potentially some that are zero.

Definition 68 Let A be an $n \times n$ symmetric matrix, and recall that $Q(x) = x^T Ax$ is the corresponding quadratic form. Then A (as well as Q) is

- negative definite if $x^T Ax < 0$, for all x
- negative semidefinite if $x^T Ax \leq 0$, for all x
- positive definite if $x^T Ax > 0$, for all x
- positive semidefinite if $x^T Ax \geq 0$, for all x
- indefinite if $x^T Ax$ is $>$ or $<$ zero, depending on x .

See Figure 10.7 for an illustration of several of these quadratic forms. In particular, we have the following.

- Figure 10.7(a) illustrates a negative definite form, in which both axes curve in a downward direction.
- Figure 10.7(b) illustrates a negative semidefinite form, in which one axis curves downward and the other axis is flat.
- Figure 10.7(c) illustrates a positive definite form, in which both axes curve in an upward direction.
- Figure 10.7(d) illustrates a positive semidefinite form, in which one axis curves upward and the other axis is flat.
- Figure 10.7(e) illustrates an indefinite form, in which the two axes curves in different directions.

Here is the basic theorem.

Theorem 20 (Quadratic forms as a sum of squares) We have the following.

- For any quadratic form Q on \mathbb{R}^n , there exists $m = k + l$ linearly independent functions, call them $\alpha_1, \dots, \alpha_m$ such that

$$Q(x) = (\alpha_1(x))^2 + \dots + (\alpha_k(x))^2 - (\alpha_{k+1}(x))^2 - \dots - (\alpha_{k+l}(x))^2$$

- The number k of positive signs and the number ℓ of minus signs depends only on Q and not on the specific linear function chosen.

Remark. Thus, if we want to consider 2-dimensional slices through the n -dimensional function, then either we have: two directions pointing up (or down); or one pointing up and one pointing down; or one pointing up (or down) and the other flat. These three possibilities mimic what we had in the 2-dimensional case, with the level sets of ellipses, hyperbolas, and parabolas.

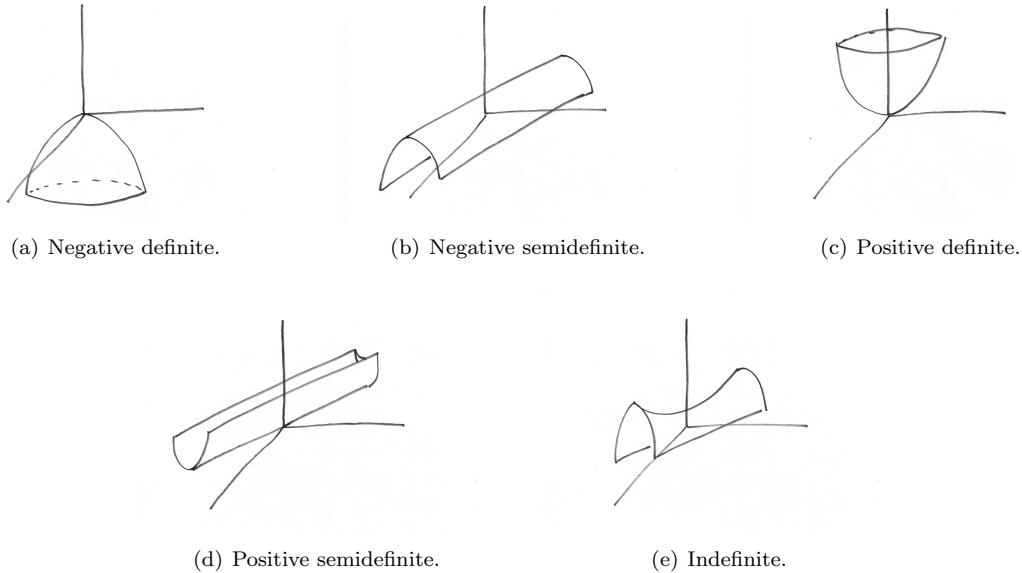


Figure 10.7: Illustration of negative definite, negative semidefinite, positive definite, positive semidefinite, and indefinite quadratic forms.

Definition 69 *The signature of a quadratic form is the pair (k, ℓ) .*

Two things to note about the signature:

- (1) it is unchanged if we use different linearly independent functions;
- (2) it does not identify uniquely the quadratic form.

The first of these isn't obvious, and we will spend some time on it, but the second of these should be obvious. For example, consider the two 1×1 matrices, (2) and (3)—both of these have the same signature, i.e., $(1, 0)$, but they are clearly different functions.

The proof of Theorem 20 is constructive, and it is sufficiently important that we will (almost) go through it. It basically amounts to finding linearly independent functions α_i by *completing the square*.

Generalizing the completing-the-square discussion from earlier in the chapter to bivariate quadratic forms, e.g.,

$$f(x_1, x_2) = ax_1^2 + bx_2^2 + cx_1x_2 + dx_1 + ex_2 + f,$$

we get the following two cases:

- If there are no cross terms, then we have a diagonal matrix and things generalize immediately.
- If there are cross terms then we get angles/rotations like we discussed before.

Here is an algorithmic idea (we wall it an algorithmic *idea* since it is a little under-specified):

1. As long as there is some single coordinate with squares, incorporate it into a perfect square. By subtracting off perfect squares in the quadratic formula, we have a quadratic formula in precisely one less variable.

Let's illustrate this algorithmic idea.

Example. Let's consider $f(x) = x_1^2 + x_1x_2$. By completing the square on x_1 , we get

$$\begin{aligned} x_1^2 + x_1x_2 &= \left(x_1 + \frac{x_2}{2}\right)^2 - \left(\frac{x_2}{2}\right)^2 \\ &= (\alpha_1(x_1, x_2))^2 - (\alpha_2(x_1, x_2))^2. \end{aligned}$$

Note that what we have essentially done by completing the square here is that we have defined the two functions, $\alpha_1 : \mathbb{R}^2 \rightarrow \mathbb{R}$ and $\alpha_2 : \mathbb{R}^2 \rightarrow \mathbb{R}$ as follows:

$$\begin{aligned} \alpha_1 &= \alpha_1 \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right) = x_1 + \frac{x_2}{2} \\ \alpha_2 &= \alpha_2 \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right) = \frac{x_2}{2}. \end{aligned}$$

Viewed as a matrix transformation, this can be written as a single function α as

$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{2} \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

i.e., as $\alpha = Ax$.

Example. Let's consider $x_1^2 + x_1x_2 - x_2^2$. One might wonder whether we should start with x_1 or with x_2 . Let's try both.

- First, try x_1 first. This gives us

$$\begin{aligned} x_1^2 + x_1x_2 - x_2^2 &= \left(x_1 + \frac{x_2}{2}\right)^2 - x_2^2 - \left(\frac{x_2}{2}\right)^2 \\ &= \left(x_1 + \frac{x_2}{2}\right)^2 - \left(\frac{\sqrt{3}}{2}x_2\right)^2, \end{aligned}$$

which implies that we are using the following transformation

$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = Ax.$$

- Second, try x_2 first. This gives us

$$\begin{aligned} x_1^2 + x_1x_2 - x_2^2 &= -(x_2^2 - x_1x_2 - x_1^2) \\ &= -\left(\left(x_2 - \frac{x_1}{2}\right)^2 - x_1^2 - \frac{x_1^2}{4}\right) \\ &= \left(\frac{\sqrt{3}}{2}x\right)^2 - \left(x_2 - \frac{x_1}{2}\right)^2, \end{aligned}$$

which implies that we are using the following transformation

$$\alpha' = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{3}}{2} & 0 \\ -\frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = A'x.$$

So, the point is that we don't get a unique answer. That's fine. We can use $\alpha = Ax$ or $\alpha' = A'x$ as our transformation. The theorem doesn't guarantee uniqueness. Moreover, if there is a subspace, then vectors aren't unique.

Example. Consider $f(x) = f(x_1, x_2, x_3) = x_1^2 + 2x_1x_2 - 4x_1x_3 + 2x_2x_3 - 4x_3^2$. Then

$$\begin{aligned} f(x) &= x_1^2 + (2x_2 - 4x_3)x_1 + 2x_2x_3 - 4x_3^2 \\ &= (x_1 + x_2 - 2x_3)^2 + 2x_2x_3 - 4x_3^2 - (x_2 - 2x_3)^2 \\ &= (x_1 + x_2 - 2x_3)^2 - x_2^2 + 6x_2x_3 - 8x_3^2 \\ &= (x_1 + x_2 - 2x_3)^2 - (x_2^2 - 6x_3x_2) - 8x_3^2 \\ &= (x_1 + x_2 - 2x_3)^2 - (x_2 - 3x_3)^2 + 9x_3^2 - 8x_3^2 \\ &= (x_1 + x_2 - 2x_3)^2 - (x_2 - 3x_3)^2 + x_3^2, \end{aligned}$$

which implies that we are using the following transformation

$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & -2 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = Ax.$$

If we had done the operations above in a different order, then we would have gotten a different matrix, $\alpha' = A'x$, but it would have had the same signature.

You can imagine that this is the sort of thing that quickly gets tedious and error-prone for a person to do, but it is very easy for a computer to do all day long. Other linear algebra classes will spend a lot of time on this, e.g., the mechanics of how to do this (although often not making the connections with quadratic forms), typically viewing this as a procedure to compute a basis for solving systems of linear equations. We are basically doing the same thing here, but we are viewing it as completing the square repeatedly, since we are viewing the matrix as a quadratic form. The reason is that this perspective is a more intuitive geometric transformation that provides some intuition for high-dimensional data science problems more generally.

So, we just repeatedly complete the square.

- If we get a sum of squares, e.g., $(u(x_1, x_2))^2 + (v(x_1, x_2))^2 = 1$, then we get a generalized ellipse, in the subspace spanned by those two vectors.
- If we get a difference of squares, e.g., $(u(x_1, x_2))^2 - (v(x_1, x_2))^2 = 1$, then we get a generalized hyperbola, in the subspace spanned by those two vectors.

More generally, if we have more than two variables, then we might have more than two terms positive, more than two terms negative, as well as some terms zero. The number of terms with each sign, as well as the subspace they span, will be the same, even if the exact terms, i.e., basis functions, are different. This is what Theorem 20 says.

Above we just followed a rule to complete the square.

Question: What if no single variable with quadratic term, e.g., $f(x_1, x_2) = x_1x_2$, then what do we do?

Answer: Just juggle things around to make it work.

Example. $Q(x) = x_1x_2 - x_1x_3 + x_2x_3$. In this case, introduce new variables, e.g., $x'_1 = x_1 - x_2$, i.e., $x_1 = x'_1 + x_2$. This amounts to a transformation

$$\begin{pmatrix} x'_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

and in this new set of variables, we have

$$\begin{aligned} Q(x) &= (x'_1 + x_2)x_2 - (x'_1 + x_2)x_3 + x_2x_3 \\ &= x_2^2 + x'_1x_2 - x_2x_3 + x'_1x_3 - x_2x_3 \\ &= x_2^2 + x'_1x_2 + x'_1x_3 \end{aligned}$$

and from this we can just proceed with the previous rule we were following.

Importantly, this doesn't depend on the specific new set of variables, e.g., we could choose $x'_1 = x_1 + x_2$, i.e., $x_1 = x'_1 - x_2$. In this case, this amounts to a transformation

$$\begin{pmatrix} x'_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

and in this new set of variables, we have

$$\begin{aligned} Q(x) &= (x'_1 - x_2)x_2 - (x'_1 - x_2)x_3 + x_2x_3 \\ &= -x_2^2 + x'_1x_2 - x_2x_3 + x_2x_3 + x_2x_3 \\ &= -x_2^2 + x'_1x_2 - x'_1x_3 + 2x_2x_3 \end{aligned}$$

and from this we can just proceed with the previous rule we were following. (Or we could have chosen many other possibilities.) Again, now that we have a term of the form x_i^2 , we can just proceed with the previous rule we were following.

Remark. The sum of squares may *not* be linearly independent. Consider the following example.

Example. Consider the following.

$$\begin{aligned} f(x) &= 2x_1^2 + 2x_2^2 + 2x_1x_2 = x_1^2 + x_2^2 + (x_1 + x_2)^2 \\ &= \left(\sqrt{2}x_1 + \frac{x_2}{\sqrt{2}}\right)^2 + \left(\frac{3}{2}x_2\right)^2 \end{aligned}$$

In this example, the first line has 3 terms in 2 variables, so they are not independent, and they can be written as a sum of two terms. Alternatively, consider the following.

$$x_1^2 + x_2^2 + 2x_1x_2 = (x_1 + x_2)^2$$

In this case, there is only 1 linearly-independent term (since, e.g., $B^2 - 4AC = 0$)

10.6 Two other topics

Here, we briefly mention two other related topics.

Connections with determinants and linear algebra. We provided several examples that illustrated that coefficients and cross terms in a symmetric matrix amount to defining new coordinate systems that are translated, rescaled, and/or rotated versions of the original coordinate system. Equivalently, they define a basis that is a translated, rescaled and/or rotated version of the original basis. How general is this for an arbitrary quadratic form? Relatedly, given an arbitrary quadratic form, is there a simple way to tell which type of conic section we have?

To answer this, recall that in general in \mathbb{R}^2 , we can have the following:

$$Ax_1^2 + Bx_1x_2 + Cx_2^2 + Dx_1 + Ex_2 + F = 0 \quad \text{where } A, B, \text{ and } C \text{ are numbers not all zero} \quad (10.17)$$

You may recall that we can classify conic sections by the discriminant: $B^2 - 4AC$. Assuming the conic section is non-degenerate, then we have the following.

- If $B^2 - 4AC < 0$, then the conic is an ellipse. (In particular, it is a circle if $A = C$ and $B = 0$.)
- If $B^2 - 4AC = 0$, then the conic is an parabola.

- If $B^2 - 4AC > 0$, then the conic is a hyperbola.

The quantity $B^2 - 4AC$ should be familiar from the quadratic formula and relatedly from the determinant discussion.

For a moment, let's ignore the D , E , and F terms, i.e., assume that we only have quadratic terms in the above equation, i.e., assume that we have taken care of the translation and rotation. In this case, we can write Equation (10.17) as follows:

$$\begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} A & B/2 \\ B/2 & C \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0.$$

Given this, if we define Ω as

$$\Omega = \begin{pmatrix} A & B/2 \\ B/2 & C \end{pmatrix},$$

then the condition we want to check is

$$\det(\Omega) = B^2 - 4AC \quad \left\{ \begin{array}{l} > \\ = \\ < \end{array} \right\} \quad 0.$$

The condition $\det(\Omega) = 0$ means that the matrix is rank-deficient, i.e., that

$$\begin{pmatrix} A \\ B/2 \end{pmatrix} = \alpha \begin{pmatrix} B/2 \\ C \end{pmatrix},$$

for some constant α . Recall that this is just the condition for linear dependence between two two-dimensional vectors. So, in particular, if $B^2 - 4AC = 0$, then we have linear dependence and a rank-deficient matrix. Alternatively, if $B^2 - 4AC \neq 0$, then we have linear independence.

If D , E , and F are non-zero, then we can use the procedures we described earlier to define a new set of coordinates in which they equal zero.

If the quadratic form involves 3 variables, then similar comments apply, except that one can obtain a zero determinant by more complex linear combinations of vectors. If the quadratic form involves 4 or more variables, then the connection with determinants tends to be less illuminating, as we discussed before.

Another way to convert degree 2 polynomials into quadratic forms. If it seems complicated to perform the procedures we described earlier to define a new set of coordinates in which the lower-order equal zero, then a “trick” that is sometimes employed is to add one extra coordinate and force it always to equal 1. In the case, of $x \in \mathbb{R}^2$, this means working with vectors $x \in \mathbb{R}^3$, where we add a third dimension to the vector x , but we always have the third component, call it x_3 , equal to 1.

If we do this, then Equation (10.17) can be viewed in one of two ways.

- As a quadratic function in two variables $x \in \mathbb{R}^2$ with lower-order terms:

$$\begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} A & B/2 \\ B/2 & C \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} D & E \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + F = 0.$$

- As a quadratic form in three variables $x \in \mathbb{R}^3$ with no lower order terms but where the last component is always equal to one:

$$\begin{pmatrix} x_1 & x_2 & 1 \end{pmatrix} \begin{pmatrix} A & B/2 & D/2 \\ B/2 & C & E/2 \\ D/2 & E/2 & F \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix} = 0.$$

Problem. If it isn't clear that both of these expressions equal Equation (10.17), then simply multiply everything out.

Remark. If we let Ω be the entire 3×3 matrix, then the condition $\det(\Omega) = 0$ says that the three vectors are linearly-dependent. For vectors in three dimensions, this isn't just that one is a scalar multiple of another, but that one of them can be written as a linear combination of the other two. In this case, we can have a point through the origin, or a line through the origin, etc.

Remark. BTW, just adding an extra dimension to all your vectors and matrices might seem strange, and it might seem like a big deal with you are working with vectors in \mathbb{R} or \mathbb{R}^2 or \mathbb{R}^3 . If your vectors are in $\mathbb{R}^{10,465}$, however, it seems like a relatively-minor change to work with vectors in $\mathbb{R}^{10,466}$, and doing this often makes things easier (which is the real justification).

10.7 Problems

10.7.1 Implementations and Applications of the Theory

1. XXX.
2. XXX.

10.7.2 Pencil-and-paper Problems

1. Recall the matrix

$$A = \begin{pmatrix} 1 & 4 & 3 \\ 4 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix}$$

that is in the notes and its eigenvalues and eigenvectors. Let Λ be the 3×3 matrix of eigenvalues, and let V be the 3×3 matrix of corresponding normalized eigenvectors. Confirm by explicit computation that $A = V\Lambda V^T$ and that $A = \sum_{i=1}^3 \lambda_i v_i v_i^T$.

2. Consider the matrix

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

- (a) Compute the eigenvalues and eigenvectors of the matrix.
- (b) Let Λ be the 3×3 matrix of eigenvalues, and let V be the 3×3 matrix of corresponding normalized eigenvectors. Confirm by explicit computation that $V^T V = I$, i.e., that the eigenvectors for an orthonormal basis for \mathbb{R}^3 .
- (c) Confirm by explicit computation that $A = V\Lambda V^T$ and that $A = \sum_{i=1}^3 \lambda_i v_i v_i^T$.

3. Consider the quadratic form

$$Q(x) = x_1^2 + 2x_1x_2 - 4x_1x_3 + 2x_2x_3 - 4x_3^2.$$

What decomposition into a sum of squares do you find if you start by eliminating terms in x_3 , then terms in x_2 , and finally terms in x_1 ? What decomposition into a sum of squares do you find if you start by eliminating terms in x_1 , then terms in x_2 , and finally terms in x_3 ?

4. Consider the quadratic form

$$Q(x) = x_1x_2 - x_1x_3 + x_2x_3.$$

- (a) Verify that the decomposition

$$(x_1/2 + x_2/2)^2 - (x_1/2 - x_2/2 + x_3)^2 + x_3^2$$

is indeed a sum of squares of linearly independent functions.

- (b) Decompose $Q(x)$ with a different choice of u , to support the statement that $u = x_1 - x_2$ is not a magical choice.
5. For each of the following equations, determine the signature of the quadratic form represented by the left-hand side. Where possible, sketch the curve or surface represented by the equation.
- (a) $x_1^2 + x_1x_2 - x_2^2 = 1$
 - (b) $x_1^2 + 2x_1x_2 - x_2^2 = 1$
 - (c) $x_1^2 + x_1x_2 + x_2x_3 = 1$
 - (d) $x_1x_2 + x_2x_3 = 1$
6. (Hubbard: Exercise 3.5.13(a,b))
 Let V be a vector space, and recall the definition of a symmetric bilinear function.
- (a) Show that if A is a symmetric $n \times n$ matrix, then the mapping $B_A(v, w) = v^T A w$ is a symmetric bilinear function.
 - (b) Show that every symmetric bilinear function on \mathbb{R}^n is of the form B_A for a unique symmetric matrix A .
7. Identify and sketch the conic sections and quadratic surfaces of equation $Q(x) = 1$, when $Q(x)$ is a quadratic form defined by one of the following matrices.
- (a) $\begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$
 - (b) $\begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$
 - (c) $\begin{pmatrix} 2 & 0 & 3 \\ 0 & 0 & 0 \\ 3 & 0 & -1 \end{pmatrix}$
 - (d) $\begin{pmatrix} 2 & 4 & -3 \\ 4 & 1 & 3 \\ -3 & 3 & -1 \end{pmatrix}$
 - (e) $\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$
8. (Hubbard: Exercise 3.7.12)
 Let A be a symmetric matrix, with p positive eigenvalues and q negative eigenvalues. Show that the quadratic form Q_A has signature (p, q) . Hint: Show that Q_A is positive definite on the span of the eigenvectors with positive eigenvalues.
9. (Hubbard: Exercise 3.7.13)
 The cone of equation $x_3^2 = x_1^2 + x_2^2$ is cut by the plane $x_3 = 1 + x_1 + x_2$ in a curve C . Describe C , and find the points of C closest and furthest from the origin.
10. Work through the jupyter notebook `eigen-nb1.ipynb`, which can be downloaded from Piazza.

Chapter 11

The Spectral Theorem: EVD and SVD

In this chapter, we will cover the so-called spectral theorem, which is one of the main results in linear algebra, and which is central to a lot of applications of linear algebra, including in data science. We have laid most of the groundwork, and we have already seen and derived various special cases of it. Here, we will discuss it more generally, in the form of the EVD of a symmetric matrix, and we will also discuss the extension of it known as the SVD.

To start, the spectral theorem is called the spectral theorem since it has to do with the eigenvalues and eigenvectors of a matrix. Recall that the set of eigenvalues of a matrix has a special name.

Definition 70 *The set of eigenvalues of a matrix is called the spectrum of a matrix.*

We said that real symmetric matrices were nice in ways that general square matrices were not. The following theorem is the reason. This theorem, stated for symmetric matrices as it is, is basically a restatement of what we covered in the last chapter. In particular, we have stated the following result in special cases, e.g., for 2×2 and 3×3 matrices, but let's discuss it more generally. It and its generalizations (e.g., to non-symmetric matrices, to infinite-dimensional operators, etc., which we will not cover, except for the SVD) are extremely important in data science and applied mathematics more generally, and so we state it explicitly.

Theorem 21 (Spectral Theorem for Symmetric Matrices) *An $n \times n$ real symmetric matrix A has the following properties:*

1. *A has n real eigenvalues, including multiplicity.*
2. *For each distinct eigenvalue, there is an associated eigenspace, and the dimension of the eigenspace is the multiplicity of the corresponding eigenvalue as the root of the characteristic equation.*
3. *Eigenspaces corresponding to distinct eigenvalues are mutually orthogonal.*
4. *A is orthogonally diagonalizable.*

This theorem might seem abstract, but—as we will discuss—it is a very practical operationalizable result.

The form of the spectral theorem given in Theorem 21, i.e., restricted to symmetric real-valued matrices, is sometimes called the EigenValue Decomposition (EVD). The reason is that it encodes information about eigenvalues and eigenvectors that we have been discussing. There are many generalizations and special cases of the spectral decomposition, as the EVD of a symmetric square matrix, perhaps the most important being the extension of Theorem 21 to arbitrary non-square matrices, known as the SVD (Singular Value Decomposition). We will cover the EVD in Chapter 11.1, and we will cover the SVD in Chapter 11.2.

11.1 The EigenValue Decomposition (EVD)

Let's discuss the EVD of a symmetric matrix A .

11.1.1 Efficiently Expressing the EVD

Writing many eigenvalue equations efficiently, i.e., as a matrix equation. We know that $Au = \lambda u$ if (λ, u) is an eigenpair. Since there are n eigenvalue-eigenvector equations of this form, let's write them as a single matrix equation. To encode all n of the equations,

$$Au_i = \lambda_i u_i, \quad \text{for } i \in [n], \quad (11.1)$$

into a single matrix equation, we need to construct the proper matrices and then perform matrix-matrix multiplication in the right way. The tricky part will be to keep track of right- and left-multiplication and of transposes and non-transposes (all in the proper order). Once we get used to this, we can drop all of the subscripts and work with a smaller number of matrix equations, and this makes everything much easier to generalize to n -dimensional spaces.

Here is how to do it.

- First, put the real eigenvalues along the diagonal of an $n \times n$ diagonal matrix Λ , and for simplicity let's order them from largest to smallest, putting in multiple entries if an eigenvalue has non-unit multiplicity.
- Second, construct an $n \times n$ orthogonal matrix U with columns consisting of the eigenvectors, in the same order. If eigenvalues are distinct, then there is a unique eigenvector associated with it. If eigenvectors have multiplicity, then there are several eigenvectors associated with it. Those eigenvectors span a subspace, and so choose any basis for that subspace.

If we follow this procedure, then we get the following equation:

$$AU = U\Lambda. \quad (11.2)$$

This expression encodes the same information as the n vector equations in Eqn. (11.1) more compactly into a single matrix equation.

Remark. The order in which the matrices on the left hand side and right hand side of Eqn. (11.2) are multiplied is very important, as this matrix equation encodes $Au_i = \lambda_i u_i$, for $i \in [n]$. If you multiply in a different order (e.g., UA or ΛU , both of which are well-defined operations, since all the matrices involved are $n \times n$ matrices), then you don't encode the eigenvector-eigenvalue information.

Most students don't have trouble with the *post*-multiplication by U . After all,

$$AU = A(u_1 \dots u_n) = (Au_1 \dots Au_n),$$

and so the i^{th} row of AU is given by Au_i . But be careful to *pre*-multiply Λ by U . (Often, you can just look at dimensions to figure out the correct order to multiply things. Here, that won't help. The matrix product ΛU is well-defined. It is the matrix consisting of rescaling the rows of U by the diagonal elements of Λ , and this doesn't have any immediate connection to $Au = \lambda u$.) If we do that, then we get

$$U\Lambda = (u_1 \dots u_n)\Lambda = (u_1\lambda_1 \dots u_n\lambda_n) = (\lambda_1 u_1 \dots \lambda_n u_n).$$

That is, for each equation, we can write $\lambda_i u_i$, and the order doesn't matter, but when we write it as a matrix equation, we need to write it in the correct order to encode the information we want.

Remark. The order, e.g., from largest to smallest, that eigenvalues (respectively, eigenvectors) are put into Λ (respectively U) is important to do properly. Observe, though, that if you change the order in Λ , and you

also change the order in U , then you do encode the same eigenvalue equation information from Eqn. (11.1). So, it is not the order per se that matters, but it is keeping the order (in the diagonal elements of Λ and the columns of U) consistent. If you change the order of one, then you must change the order of the other in exactly the same way. This should be clear from Eqn. (11.8) below, as the terms in that sum are unchanged (but the order of the terms in the sum is changed, leaving the sum unaffected) if the order of elements along the diagonal of Λ and columns in U are changed consistently. While the order is actually arbitrary, it is convenient and customary to choose them in descending (or sometimes sometimes ascending) order.

The spectral decomposition. Before proceeding with how to find these eigenvectors and eigenvalues, let's look in more detail about Equation (11.2). There are two complementary ways to view this.

- If we post-multiply on the right by U^T , then we have that

$$U\Lambda U^T = AUU^T = A. \quad (11.3)$$

The first equality should be clear; and the second equality holds since $UU^T = I$, since U is an $n \times n$ orthogonal matrix.

Equation (11.3) is a decomposition of A into three terms, $A = U\Lambda U^T$. We will describe this in more detail below.

- Alternatively, we could pre-multiply by U^T to get

$$U^T A U = U^T U \Lambda = \Lambda. \quad (11.4)$$

The first equality should be clear; and the second equality holds since $U^T U = I$, again since U is an $n \times n$ orthogonal matrix (but note that this is different than $UU^T = I$, which is also true and which we used before).

Equation (11.4) is not a decomposition of the matrix A , but instead it is expressing A in the rotated basis defined by U . We saw a particular example of this before. The point is that in the complete orthonormal basis of eigenvectors, the matrix A is a diagonal matrix, with diagonal elements equal to the eigenvalues. We will revisit this below when we discuss PCA, perhaps the most well-known example of this diagonalization in data science.

Let's look at other ways to look at Equation (11.3), i.e., the decomposition of A into eigenvectors and eigenvalues.

$$A = \begin{pmatrix} u_1 & \dots & u_n \end{pmatrix} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \begin{pmatrix} u_1^T \\ \vdots \\ u_n^T \end{pmatrix} \quad (11.5)$$

$$= \begin{pmatrix} \lambda_1 u_1 & \dots & \lambda_n u_n \end{pmatrix} \begin{pmatrix} u_1^T \\ \vdots \\ u_n^T \end{pmatrix} \quad (11.6)$$

$$= \lambda_1 u_1 u_1^T + \dots + \lambda_n u_n u_n^T \quad (11.7)$$

$$= \sum_{i=1}^n \lambda_i u_i u_i^T. \quad (11.8)$$

We saw a few examples of this last time, when we wrote a matrix as a sum of rank-1 matrices, each of which was the outer product of an eigenvector with its transpose, and each of which is scaled by its associated eigenvalue.

Remark. This is why the order of elements along the diagonal is arbitrary, as long as you put the eigenvectors as column vectors into U in the same order—when you write out A as a sum of rank-1 matrices, the eigenvalue multiplies its eigenvector (and its eigenvector transposed), and this does not depend on how the eigenvalues were originally ordered. So, it really is for convenience that we order them in descending order.

Definition 71 This last equation, equivalently either (11.5) or (11.8), is the spectral decomposition of the matrix A . Each term, $\lambda_i u_i u_i^T$ is an $n \times n$ matrix of rank 1, and $u_i u_i^T x$ orthogonal projection of the vector x onto the span of u_i .

Example. Recall our previous example of the ellipse

$$5x_1^2 - 4x_1x_2 + 5x_2^2 = 48.$$

To plot this and see the connection with the spectral decomposition, let's rewrite it as a matrix equation,

$$\begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 5 & -2 \\ -2 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 48.$$

This is a symmetric matrix, and so we know it has a full set of eigenvectors and eigenvalues. Let's compute (recompute) them. To do so, compute

$$\begin{aligned} 0 = \det(A - \lambda I) &= \begin{vmatrix} 5 - \lambda & -2 \\ -2 & 5 - \lambda \end{vmatrix} \\ &= (5 - \lambda)^2 - 4 \\ &= \lambda^2 - 10\lambda + 21 \\ &= (\lambda - 7)(\lambda - 3). \end{aligned}$$

So, the two distinct eigenvalues are $\lambda = 3$ and $\lambda = 7$. For $\lambda = 3$, we get:

$$\begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \rightarrow v_{\lambda=3} \sim \begin{pmatrix} 1 \\ 1 \end{pmatrix} \rightarrow v_{\lambda=3} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix},$$

and for $\lambda = 7$, we get:

$$\begin{pmatrix} -2 & -2 \\ -2 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \rightarrow v_{\lambda=7} \sim \begin{pmatrix} 1 \\ -1 \end{pmatrix} \rightarrow v_{\lambda=7} = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}.$$

Let's now plot this.

To plot this, let's first observe the following. If we define a matrix $V \in \mathbb{R}^{2 \times 2}$ to have columns consisting of the two eigenvectors (let's order from smallest ($\lambda = 3$) to largest ($\lambda = 7$) as

$$V = (v_{\lambda=3} \ v_{\lambda=7}) = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

and let's also define a diagonal matrix $D \in \mathbb{R}^{2 \times 2}$ to have the eigenvalues along the diagonal (in the same order as V) to be

$$D = \begin{pmatrix} 3 & 0 \\ 0 & 7 \end{pmatrix}$$

Given this definition of V and D , let's compute the product $V^T D V$ as follows.

$$\begin{aligned} V^T D V &= \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 7 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 7 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 3 & 3 \\ -7 & 7 \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} 10 & -4 \\ -4 & 10 \end{pmatrix} \\ &= \begin{pmatrix} 5 & -2 \\ -2 & 5 \end{pmatrix} \\ &= A \quad (\text{as it should!}) \end{aligned}$$

Note that here we wrote the matrix Λ in the somewhat non-traditional ascending order, but we made sure to put the eigenvectors into the eigenvector matrix in that same order.

11.1.2 Finding the EVD

For 2×2 matrices, and also for 3×3 matrices, one can compute the spectral decomposition, i.e., the EVD of a symmetric square matrix, by hand with determinants. (The reason for this is that, for $n = 2, 3$, we can use the quadratic/cubic formula to get an expression for the roots of the quadratic/cubic equation that characterizes the eigenvalues, but this approach does not generalize to larger values of n .) For larger matrices, using determinants is a very bad idea, even on a computer. Here, we will describe how to compute the spectral decomposition more generally. This approach makes strong use of what we learned about quadratic forms.

A proto-algorithm for the spectral decomposition. The basic idea is the following. To find the spectral decomposition of an $n \times n$ matrix A :

- find a/the vector such that when we evaluate it in the quadratic form defined by the matrix, i.e., $x^T Ax \in \mathbb{R}$, then we get the largest numerical value;
- then iterate this process.

This process makes sense since the n eigenvalues are real numbers and thus they can be ordered from largest to smallest, and since we have a full set of n eigenvectors.

This “basic idea” is almost an algorithm, but there are a few gotchas that we need to be careful about, in order to turn it into a real algorithm. The most prominent are the following.

- If the quadratic form points “up” then there is no largest value. For example, if you double the magnitude of x , then the numerical value of $x^T Ax$ increases by a factor of 4.
- There may not be a unique direction, i.e., eigenvector, associated with the largest numerical value, and so it’s not clear which one we should choose.
- When we iterate, a trivial answer is to have the second value/direction be the same as the first (since that too maximizes $x^T Ax$), which is presumably not what we want (since it will be the same as the first vector/direction and so won’t add any new information).

Dealing with each of these issues turns the basic idea into an algorithm. We’ll go into a bit more detail about them. To do so will require a little bit of calculus.

Eigenvalues/eigenvectors as optimization problems. To do this, let A be an $n \times n$ matrix, and let’s consider the problem

$$\max_{x \in \mathbb{R}^n} x^T Ax,$$

i.e., find the vector that gives the largest value of the quadratic form associated with the matrix. Unfortunately, this problem isn’t well-defined in the sense that it doesn’t (or, more precisely, depending on A , it may not) have a maximum: given any vector x , we can increase the value of the objective by considering the vector $x' = 2x$.

The way to avoid this issue is to “constrain” $x \in \mathbb{R}^n$. By this, we mean that we want to consider not all $x \in \mathbb{R}^n$, but only a subset of all possible x , and in particular a subset that does not permit us to consider scaling x arbitrarily by $\alpha \in \mathbb{R}^+$. There are many ways to constrain x , but the one that leads to the spectral decomposition is to constrain x to the unit ball in \mathbb{R}^n . That is, we try to solve the problem

$$\max_{x \in \mathbb{R}^n} x^T Ax \quad \text{s.t.} \quad x^T x = 1.$$

This problem is sometimes written as

$$\begin{aligned} \max_{x \in \mathbb{R}^n} \quad & x^T A x \\ \text{s.t.} \quad & x^T x = 1. \end{aligned}$$

That is, we consider all vectors on the unit ball in \mathbb{R}^n , i.e., we consider all directions but we don't worry about magnitudes, and we ask ourselves which direction has the largest value of this quadratic function. The solution of this objective is the first eigenvalue; moreover, the vector achieving it is the first eigenvector.

(That discussion assumes uniqueness of eigenvalues, and as usual we have the usual issues come up if there is multiplicity, but let's assume that for now.)

The problem just described actually gives a number as a solution. That is, it doesn't ask for the vector that achieves the maximum, but instead it asks for the numerical value of the function at the maximum. Formally, to get the vector that achieves the maximum numerical value, we need to write the "argmax" version of this problem:

$$\operatorname{argmax}_{x \in \mathbb{R}^n} x^T A x \quad \text{s.t.} \quad x^T x = 1,$$

which is sometimes also written as

$$\begin{aligned} \operatorname{argmax}_{x \in \mathbb{R}^n} \quad & x^T A x \\ \text{s.t.} \quad & x^T x = 1. \end{aligned}$$

Optimization and constrained optimization.

How does one solve a problem like this?

For a moment, assume that $x \in \mathbb{R}$, in which case

$$f(x) = ax^2,$$

for some number $a \in \mathbb{R}$. Let's also assume that $a > 0$. If we view this as a matrix,

$$A = (a) \in \mathbb{R}^{1 \times 1},$$

in which case A is a PD matrix and x is a 1-dimensional vector, $x \in \mathbb{R}^1$.

Recall, *if we want to find a minimum*, then from calculus the conditions for a function $f(x)$ to have a minimum at a point x^* are:

1. $\frac{df}{dx} = 0$: This is the first order condition. It basically says that the function is flat, but it doesn't say whether it is a minimum or a saddle point or a maximum.
2. $\frac{d^2 f}{dx^2} > 0$: This is the second order condition. It basically that the derivative of the function is upward-sloping, meaning that x^* is a minimum.

Alternatively, *if we want to find a maximum*, then the conditions for a function $f(x)$ to have a maximum at a point x^* are:

1. $\frac{df}{dx} = 0$: This is the first order condition, and it is the same as when we are trying to find a minimum.
2. $\frac{d^2 f}{dx^2} < 0$: This is the second order condition. It basically that the derivative of the function is downward-sloping, meaning that x^* is a maximum.

Let's check how these conditions apply to our function $f(x) = ax^2$. Recall that we want to find a maximum. We have:

$$\begin{aligned} \frac{df}{dx} &= 2ax \\ \frac{d^2 f}{dx^2} &= 2a. \end{aligned}$$

From the first order condition, we see that $x^* = 0$ is a candidate for a maximum (and it is also a candidate for a minimum), and also that there is no other $x \in \mathbb{R}$ that is a candidate. From the second order condition, we see that at every point, and in particular at $x^* = 0$, the second derivative is positive. In particular, this means that $x^* = 0$ is a minimum, and that there is no point that is a maximum. This is basically the same issue we saw before that if we choose a vector x and plug it into the function and scale it up then it only gets larger. In one dimension, “scale it up” means move in the direction away from the origin. Since we can always do that, there is no maximum.

That’s why we introduced the constraint that $x^T x = 1$. This is less interesting in one dimension than in higher dimensions, but let’s go through it in one-dimension. In one dimension, this constraint means that we consider the points $x = \pm 1$. Since the function is squared, either point is okay. (That is, if we view the “direction” as a “subspace,” which is what we do in linear algebra when we go to higher dimensions, then both unit length vectors in that subspace, which are scalar multiples of each other where the scalar is -1 , give the same numerical value for the quadratic form; but that is okay, since we are really interested in different directions in \mathbb{R}^n .) But this is really a degenerate case, since a line through the origin is the only possible direction.

So, let’s consider the next most simple case. In this case, we have a function $f(x_1, x_2) = x^T A x$, i.e., where $x \in \mathbb{R}^2$. In this case, the constraint means that we choose the unit Euclidean ball $x^T x = x_1^2 + x_2^2 = 1$. To do this in a way that generalizes to \mathbb{R}^3 and beyond requires some machinery.

The method of Lagrange Multipliers. The general way to deal with a constrained optimization problem is to consider the *method of Lagrange Multipliers*. This is an approach to convert a constrained optimization problem, e.g., an optimization problem with equality constraints, into an unconstrained optimization problem where the first and second order conditions can be applied. It does so by replacing the constrained optimization problem by introducing a new variable and converting it to an unconstrained optimization problem.

Example. Consider the following problem.

$$\begin{aligned} \min \quad & f(x_1, x_2) \\ \text{s.t.} \quad & g(x_1, x_2) = 0. \end{aligned}$$

Let’s introduce a parameter $\lambda \in \mathbb{R}$ and define the following function:

$$L = L(x_1, x_2, \lambda) = f(x_1, x_2) - \lambda g(x_1, x_2). \quad (11.9)$$

The function L is called the *Lagrangian*.

Fact. If (x_1^*, x_2^*) is the maximum of $f(x_1, x_2)$ such that $g(x_1, x_2) = 0$, then there exists a λ^* such that $(x_1^*, x_2^*, \lambda^*)$ is a stationary point of the Lagrangian function L .

At that stationary point, all the first derivatives of L equal zero.

Caveat. It is NOT the case that all stationary points of the Lagrangian solve the original problem. There are technical necessary and sufficient conditions for all this to go through. We won’t go through all that here (you need a more advanced class for that), but those conditions will hold in the examples to which we will apply it.

In the 2-dimensional case,

$$f(x_1, x_2) = a_{11}x_1^2 + (a_{12} + a_{21})x_1x_2 + a_{22}x_2^2,$$

in which the derivatives with respect to x_1 and x_2 are

$$\begin{aligned} \frac{\partial f}{\partial x_1} &= 2a_{11}x_1 + (a_{12} + a_{21})x_2 \\ \frac{\partial f}{\partial x_2} &= (a_{12} + a_{21})x_1 + 2a_{22}x_2, \end{aligned}$$

which can be written in more compact notation as

$$\frac{\partial f}{\partial x} = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 2a_{11} & a_{12} + a_{21} \\ a_{12} + a_{21} & 2a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 2Ax.$$

So, in the 2-dimensional case, we want to optimize

$$\begin{aligned} \min_{x \in \mathbb{R}^2} \quad & x^T Ax \\ \text{s.t.} \quad & x^T x = 1. \end{aligned}$$

Thus, in this case, the Lagrangian is

$$L = L(x_1, x_2, \lambda) = L(x, \lambda) = x^T Ax - \lambda (x^T x - 1).$$

Note here that the term multiplying the λ is $(x^T x - 1)$ and not $x^T x$. The reason for this is that the constraint is $x^T x = 1$, and so to convert it to a form $g(x) = 0$, we need to have $g(x) = x^T x - 1$. Don't forget to do this.

Now, consider the problem

$$\min_{x, \lambda} \quad x^T Ax - \lambda (x^T x - 1),$$

where this is now an unconstrained optimization problem, and thus we can apply the usual first and second order conditions to it.

We have derived this by considering $x \in \mathbb{R}^2$, but nothing in its derivation needs to be just 2-dimensional (and nothing in what follows needs that either). For example, the expression

$$\frac{\partial f}{\partial x} = 2Ax$$

generalizes to n dimensions.

In particular, the function we are interested in more generally is $f(x) = x^T Ax$, where $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$, and $f(x) \in \mathbb{R}$.

Derivatives of functions in higher dimensions. Many of the rules from one-dimensional calculus and one-dimensional probability generalize to functions that involve many variables. There are some things that are different, but the following things are good to know.

Let A be an $n \times n$ matrix (square, but not necessarily symmetric), let $a \in \mathbb{R}^n$ be a column vector (i.e., an $n \times 1$ matrix), and let $x \in \mathbb{R}^n$ be a column vector of variables. Then, we can show the following.

- If $z = a^T x$, then

$$\frac{\partial z}{\partial x} = \frac{\partial a^T x}{\partial x} = a.$$

- If $z = x^T x$, then

$$\frac{\partial z}{\partial x} = \frac{\partial x^T x}{\partial x} = 2x.$$

- If $z = a^T Ax$, then

$$\frac{\partial z}{\partial x} = \frac{\partial a^T Ax}{\partial x} = A^T a.$$

- If $z = x^T Ax$, then

$$\frac{\partial z}{\partial x} = \frac{\partial x^T Ax}{\partial x} = Ax + A^T x.$$

If, in addition, A is symmetric, then

$$\frac{\partial z}{\partial x} = \frac{\partial x^T Ax}{\partial x} = 2Ax.$$

In these expressions, $\frac{\partial}{\partial x}$ is a derivative operation that can be applied to a function and that yields slightly different results depending on the form of the function to which it is applied. In particular, it can be applied to a function

$$f = f(x) = f(x_1, \dots, x_n)$$

(i.e., a function that takes as input a vector x and) that returns as output a scalar or a vector.

- If the function f returns as output a scalar, i.e., if $f \in \mathbb{R}$, then

$$\frac{\partial f}{\partial x} = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right) \in \mathbb{R}^n,$$

i.e., it returns a vector, the i^{th} element of which is the function differentiated with respect to the i^{th} variable. That is, the derivative of a scalar-valued function of a vector input is itself a vector, the elements of which encode the derivative information of the output with respect to each element of the input. (For one-dimensional functions, derivatives are basically linear approximations of the function, as the input variable changes. So too, here, derivatives are linear approximations, but the linear approximation is in \mathbb{R}^n , since the input can vary in \mathbb{R}^n .) As with other vectors, this can be interpreted as a direction in \mathbb{R}^n and a magnitude in that direction.

- If the function f returns as output a vector, i.e., if $f = (f_1, \dots, f_m) \in \mathbb{R}^m$, then

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & & & \\ \vdots & & & \\ \frac{\partial f_m}{\partial x_1} & & & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

i.e., it returns a matrix, the $(ij)^{th}$ element of which is the i^{th} element of the vector function differentiated with respect to the j^{th} variable. That is, the derivative of a vector-valued function of a vector input is a matrix, the elements of which encode the derivative information of the i^{th} output direction with respect to the j^{th} input variable.

Observe that, by combining the above two steps, if we have a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, then

$$g = \frac{\partial f}{\partial x} \in \mathbb{R}^n \quad \text{is a vector, with elements } g_i = \frac{\partial f}{\partial x_i} \quad \text{and}$$

$$H = \frac{\partial g}{\partial x} \in \mathbb{R}^{n \times n} \quad \text{is a matrix, with elements } H_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

The vector g that contains the first derivative information of f is called the *gradient* of f , and the matrix H that contains the second derivative information of f is called the *Hessian* of f . There are some “corner cases” (basically having to do with poor continuity properties) where the so-called cross partial derivatives are not equal, but nearly always in machine learning and data science it is the case that

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i},$$

i.e., that this matrix of second derivatives is symmetric. In this case, everything we have been discussing (e.g., n eigenvalues, a full set of orthonormal eigenvectors, etc.) holds.

Solving eigendecompositions with Lagrange Multipliers. To compute the eigendecomposition of an $n \times n$ symmetric matrix A , the basic problem in which we are interested is the following. Given a symmetric matrix $A \in \mathbb{R}^{n \times n}$, solve the following constrained optimization problem.

$$\begin{array}{ll} \min_{x \in \mathbb{R}^n} & x^T A x \\ \text{s.t.} & x^T x = 1. \end{array} \tag{11.10}$$

To solve this, we will consider the Lagrangian $L : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}$, defined as follows

$$L = L(x, \lambda) = x^T Ax - \lambda (x^T x - 1), \quad (11.11)$$

and we will try to solve the following unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n, \lambda \in \mathbb{R}} x^T Ax - \lambda (x^T x - 1). \quad (11.12)$$

To do this, consider the first partial derivatives.

$$\begin{aligned} \frac{\partial L}{\partial x} &= 2Ax - 2\lambda x = 0 \\ \frac{\partial L}{\partial \lambda} &= x^T x - 1 = 0. \end{aligned}$$

From the first of these, we have

$$Ax = \lambda x, \quad (11.13)$$

i.e., we have the equation for eigenvalues and eigenvectors of a matrix; and from the second of these we have

$$x^T x = 1, \quad (11.14)$$

i.e., that that eigenvector should be normalized. These are very important results. They say that we can recover the expressions for eigenvalues and eigenvectors by considering quadratic functions, which in many ways are simpler to think about. Basically, we have unit circles as input and ellipses as output, and we want to determine the direction of maximum variance in the ellipse, since that corresponds with the vector that maximizes the quadratic form over the unit sphere.

While this figure is for 2 dimensions, similar things hold for higher dimensions. While that is true in general in linear algebra, and while we have pointed it out many times, in this case it is particularly simple, since we have one maximum direction, and we can think of the other direction as being all $n - 1$ other directions. (The partial exception to this is if the first direction isn't unique, but then the usual things we discussed hold.) So, the direction of maximum variance is given by the largest eigenvector, let's call it v_1 , and the amount by which A stretches a unit vector in that direction is given by λ_1 , the largest eigenvalue. Note that λ_1 is just the value of the Lagrangian parameter we introduced to enforce the constraint.

Note that, except in a few very special cases, we haven't said how to compute eigenvalues and eigenvectors—that is a large and involved topic for more advanced classes.

The solution to Equation (11.10), i.e., to Equation (11.13) actually gives a very specific eigenvector, namely the largest. (Recall that we can order the eigenvalues from largest to smallest since we are dealing with symmetric matrices, and thus the eigenvalues are all real numbers.) How do we get the rest of the eigenvalues?

To get the second eigenvalue, we need to solve the following problem:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & x^T Ax \\ \text{s.t.} \quad & x^T x = 1 \\ \text{and} \quad & v_1^T x = 0. \end{aligned} \quad (11.15)$$

Equation (11.15) is similar to Equation (11.10), except for the second constraint which says that we are looking for vectors x that are orthogonal to v_1 . In the 2-dimensional figure, this is harder to visualize, since there is only one dimension representing all other $n - 1$ dimensions. But if we imagine zooming-in or resolving all those dimensions, then we get another figure. In this case, we are only considering vectors perpendicular to v_1 , and we are showing the direction of maximum variance in that subspace along one axis and all other $n - 2$ dimensions aggregated along the shorter axis.

To solve this constrained optimization problem, we use again the method of Lagrange Multipliers, but now we need two multipliers to account for the two constraints. To solve this, we will consider the Lagrangian $L : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, defined as follows:

$$L = L(x, \lambda) = x^T Ax - \lambda (x^T x - 1) - \mu x^T v_1. \quad (11.16)$$

Here λ and μ are two different Lagrange parameters for two different constraints. Note here we have $\mu x^T v_1$ and not something like $\mu(x^T v_1 - 1)$ since we want to enforce the constraint $\mu v_1^T x = 0$ and thus $g(x) = v_1^T x$ in the above equation. Thus, we will try to solve the following unconstrained optimization problem:

$$\min_{x \in \mathbb{R}^n, \lambda \in \mathbb{R}, \mu \in \mathbb{R}} x^T A x - \lambda(x^T x - 1) - \mu x^T v_1. \quad (11.17)$$

To do this, consider the first partial derivatives.

$$\begin{aligned} \frac{\partial L}{\partial x} &= 2Ax - 2\lambda x - \mu v_1 = 0 \\ \frac{\partial L}{\partial \lambda} &= x^T x - 1 = 0 \\ \frac{\partial L}{\partial \mu} &= v_1^T x = 0. \end{aligned}$$

If we pre-multiply the first of these three equations by v_1^T , and then use the other two constraints, then we get the following:

$$\begin{aligned} 0 &= 2v_1^T Ax - 2\lambda v_1^T x - \mu v_1^T v_1 \\ &= 2v_1^T Ax - \mu \quad \text{due to the two constraints} \\ &= 2\lambda_1 v_1^T x - \mu \quad \text{since } Av_1 = \lambda_1 v_1 \text{ by the first eigencondition} \\ &= -\mu \quad \text{since } x^T v_1 = 0 \text{ from before.} \end{aligned}$$

and thus $\mu = 0$. This gives us

$$Ax = \lambda x$$

for the largest eigenvalue in the subspace perpendicular to v_1 . Let's call this eigenvalue/eigenvector pair (λ_2, v_2) .

To compute the third and subsequent eigenvectors, we simply iterate this process.

Question. How many eigenvectors can there be? Equivalently, for how many steps do we iterate this process?

Answer. We know that eigenvectors need to be perpendicular to each other, so in \mathbb{R}^n there can be no more than n . So, we iterate this process until we have a set of n mutually perpendicular eigenvectors, each of which has an associated eigenvalue.

All of this assumes that the eigenvalues are distinct, but if not then we get something similar, but this captures the basic idea.

Putting it all together. Let's put the n eigenvalues in a diagonal matrix with the eigenvalues along the diagonal, ordered from largest to smallest, to get

$$\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Let's also put the n eigenvectors into a matrix, with one eigenvector as each column of the matrix, to get

$$U = \begin{pmatrix} u_1 & \dots & u_n \end{pmatrix}. \quad (11.18)$$

Each of these is orthonormal, so this gives us a complete orthonormal basis for \mathbb{R}^n . In particular, $U^T U = I$, which when written out gives

$$\begin{pmatrix} u_1^T \\ \vdots \\ u_n^T \end{pmatrix} \begin{pmatrix} u_1 & \dots & u_n \end{pmatrix} = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}.$$

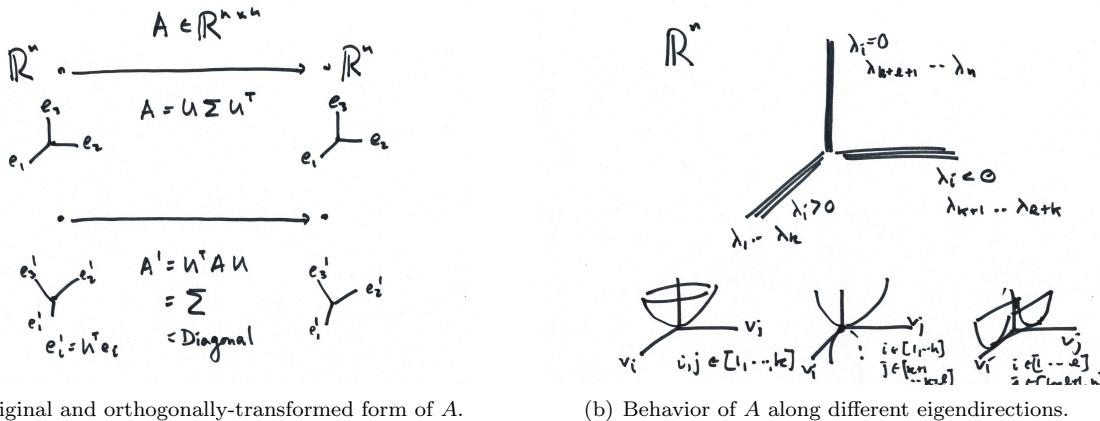


Figure 11.1: Illustrations of the eigenvalue decomposition.

The reason for this is that

$$u_i^T u_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

See Figure 11.1(a).

Question: What makes this better/worse than the basis vectors $\{e_i\}_{i=1}^n$ that populate the usual Identity matrix?

Answer: Well, they are more well-suited or adapted to the data matrix A . We will get back to this soon.

Regarding the eigenvectors, let's recall something and then write out the matrix U from Equation (11.18) in some more detail. If we write out the matrix A in full detail, we have

$$A = \begin{pmatrix} & & \\ u_1 & \dots & u_n \end{pmatrix} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \begin{pmatrix} u_1^T \\ \vdots \\ u_n^T \end{pmatrix}.$$

Since the eigenvalues are real and ordered, we have some number k that are positive, some number ℓ that are negative, and some number $n - (k + \ell)$ that equal zero.

So, we can split U up into three pieces:

$$U = \begin{pmatrix} & & & & & & & \\ u_1 & \dots & u_k & u_{k+1} & \dots & u_{n-\ell} & u_{n-\ell+1} & \dots & u_n \\ & & & & & & & \end{pmatrix}. \quad (11.19)$$

Then, we have the following.

- If we are on a subspace corresponding to positive (resp., negative) eigenvalues, then the level sets of our function are ellipsoids (if the dimension of the subspace is 2, higher dimensional generalizations otherwise), where the length of major axes are given by the eigenvalues.

- If we are on a subspace corresponding to both positive and negative eigenvalues, then the level sets of our function are hyperbolas (if the dimension of the subspace is 2, higher dimensional generalizations otherwise), where the stretch information is given by the eigenvalues.
- If some of the eigenvalues equal 0, then in those directions the function is flat.

See Figure 11.1(b).

11.1.3 Computing the EVD

There are different ways to compute the EVD of a matrix, depending on the properties of the matrix. At a high level (there are many variants of each), they boil down to two main ways.

- **Direct methods.** Direct methods run a sequence of steps, e.g., shears or rotations of two variables in a systematic way, e.g., transform the original matrix to a diagonal matrix. The diagonal elements then give the eigenvalues, and the sequence of transformation can be used to give the eigenvectors. There are many variants of these methods, and there are some subtleties, e.g., numerical issues, but they use ideas related to what we have been discussing.
- **Iterative methods.** Iterative methods start with an arbitrary vector, and they multiply that vector by the matrix iteratively, and this can be used to compute eigenvalues and eigenvectors.

We will describe some simple examples of the latter method. For simplicity, we will apply it to a symmetric positive semi-definite matrix A , but similar ideas can be applied to much more general matrices and in fact form the basis for many widely-used methods.

The power method. Consider the following algorithm. Given as input a symmetric positive semi-definite matrix A (e.g., which, given a matrix X , could be $A = X^T X$) and an arbitrary initial vector $x_0 \in \mathbb{R}^n$, do the following steps.

1. Let $y_t = Ax_t$.
2. Let $x_{t+1} = \frac{1}{\|y_t\|_2}y_t$.
3. At some step, return x_{t+1} .

It can be shown that

$$x_t \rightarrow v_1,$$

where v_1 is the eigenvector associated with the largest eigenvalue λ_1 of A ; and, given this vector v_1 , the largest eigenvalue can be computed as $\lambda_1 = \|Av_1\|_2$. That is, this algorithm gives the largest eigenvalue λ_1 and the associated eigenvector v_1 .

The reason that this algorithm works is important to understand. Recall that the matrix A can be expressed in terms of its EVD as

$$A = \sum_{i=1}^n \lambda_i v_i v_i^T = \lambda_1 v_1 v_1^T + \sum_{i=2}^n \lambda_i v_i v_i^T.$$

Recall also that since the $\{v_i\}_{i=1}^n$ provides a complete orthonormal basis, one can compute higher powers of A multiplied by itself as

$$A^t = \sum_{i=1}^n \lambda_i^t v_i v_i^T = \lambda_1^t v_1 v_1^T + \sum_{i=2}^n \lambda_i^t v_i v_i^T \rightarrow \lambda_1^t v_1 v_1^T.$$

That is, all the directions but the largest get “damped down” as the iterations proceed. Thus, if we apply A to an initial vector x_0 many times, say t times, then since that is equivalent to applying A^t to x_0 once, what we get is

$$A^t x_0 \rightarrow \lambda_1^t v_1 v_1^T x_0 = \lambda_1^t v_1^T x_0 v_1.$$

Then, to compute the second largest eigenvalue λ_2 and the associated eigenvector v_2 , one can run the following algorithm.

1. Let $y_t = Ax_t$.
2. Let $z_t = (I - P_{v_1})y_t$.
3. Let $x_{t+1} = \frac{1}{\|z_t\|_2} z_t$.
4. At some step, return x_{t+1} .

This algorithm is almost the same as the previous algorithm, except that at each iteration it removes the part of the vector that is parallel to the eigenvector associated with the largest eigenvalue. It can be shown that

$$x_t \rightarrow v_2,$$

where v_2 is the eigenvector associated with the second largest eigenvalue λ_2 of A , i.e., the largest conditioned on being orthogonal to v_1 ; and, given this vector v_2 , the associated eigenvalue can be computed as $\lambda_2 = \|Av_2\|_2$. That is, this algorithm gives the largest eigenvalue λ_2 and the associated eigenvector v_2 .

One can compute the third largest eigenvalue and associated eigenvector in the obvious way. (To get it working in a robust way in actual numerical code involves all sorts of numerical subtleties into which we will not get.)

11.2 Singular Value Decomposition (SVD)

Let's now consider the generalization of the spectral theorem to an arbitrary $m \times n$ square matrix A . (What we will say holds for non-symmetric $n \times n$ matrices, thus generalizing what we have been discussing, but since it holds for arbitrary $m \times n$ matrices, we will consider that more general case.)

11.2.1 The basic SVD

Recall that an arbitrary $m \times n$ square matrix A is a representation of a function from \mathbb{R}^n to \mathbb{R}^m , i.e., it represents with respect to the canonical basis a linear function that takes as input a vector in \mathbb{R}^n and that returns as output a vector in \mathbb{R}^m . From this, we can compute two matrices, AA^T and A^TA , both of which are symmetric, and both of which represent (different) linear functions (one from \mathbb{R}^m to \mathbb{R}^m , and the other from \mathbb{R}^n to \mathbb{R}^n).

Since both of these matrices are symmetric, each of them has an EVD, as follows.

$$\begin{aligned} AA^T &= U\Lambda U^T \in \mathbb{R}^{m \times m} \\ A^TA &= V\Lambda' V^T \in \mathbb{R}^{n \times n} \end{aligned}$$

It is a fact (that we will not prove) that $\Lambda = \Lambda'$, and that all the diagonal elements of this matrix are nonnegative. Thus, we can write $\Lambda = \Sigma^2$, for some diagonal matrix Σ .

It is also a fact that the original rectangular matrix A can be decomposed as

$$A = U\Sigma V^T. \tag{11.20}$$

This decomposition of A is known as the *Singular Value Decomposition (SVD)* of the matrix A .

Observe that the SVD of a general matrix has some similarities to the EVD of a symmetric matrix, e.g., as given in Eqn. (11.4), in that it is the product of three matrices, the first and third of which are orthogonal, and the middle of which is diagonal. There are two important differences.

- The two orthogonal matrices are not even the same size, and so they can't be the same in general.
- The singular values, being squares of numbers, are non-negative, while the eigenvalues can be negative.

Let's go into more detail on the SVD.

11.2.2 Equivalent ways to view the SVD

As with the EVD, there are two equivalent ways to view the SVD. What the SVD says is the following.

- One can represent an arbitrary matrix A as the product of three matrices, the first and third of which are orthogonal transformation in the appropriate dimension, and the second of which is diagonal with non-negative entries.
- One can fix a basis in \mathbb{R}^m as well as a basis in \mathbb{R}^n , and then the matrix A is diagonal.

Also, as with the EVD, each of these views has implications, e.g., that one can represent a general matrix A as a sum of outer products (rank-one matrices), each of which is a projection matrix onto a one-dimensional subspace, and each of which is scaled by a scaling factor.

The number of nonzero elements of Σ is the *rank* of A ; it is equal to the number of linearly-independent columns in A , and it is also equal to the number of linearly-independent rows in A .

This decomposition holds for general $m \times n$ matrices A . We will see how to prove a special case of this (restricted to square invertible matrices) in the homework.

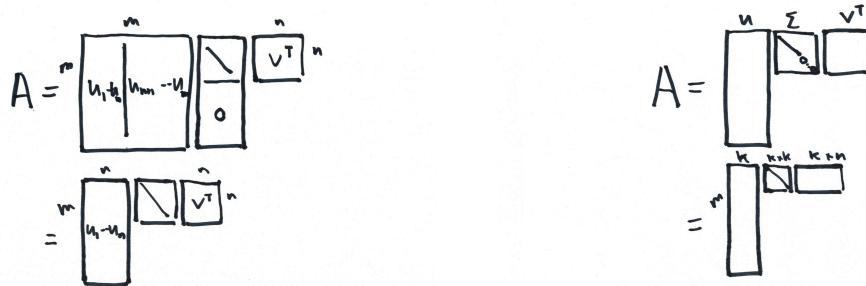
11.2.3 SVD and thin SVD and rank-deficient thin SVD

One subtlety and potential point of confusion is the following. Let's say that A is $m \times n$, and that $m > n$, and let's assume that all the singular values are non-zero. Then, the SVD, as given in Eqn. (11.20), can represent one of two things.

- In one case, there are n left singular vectors, which span an n -dimensional subspace of \mathbb{R}^m , i.e., U is an $m \times n$ matrix with orthogonal columns (so $U^T U = I$ and $U U^T \neq I$ is a projection); and there are n right singular vectors, which span all of \mathbb{R}^n , i.e., V is an $n \times n$ orthogonal matrix (so $V^T V = I$ and $V V^T = I$); in which case Σ is an $n \times n$ diagonal matrix with the singular values along the diagonal.
- In the other case, V is still an $n \times n$ orthogonal matrix; but we could choose U to be an $m \times m$ orthogonal matrix, i.e., a basis for all of \mathbb{R}^m , the first n vectors of which are singular vectors corresponding to the singular values, and the other $m - n$ of which provide an orthonormal basis for the subspace of \mathbb{R}^m perpendicular to the subspace spanned by the first n singular vectors. In this case, then we could still have an equation of the $A = U \Sigma V^T$ if the matrix Σ is $m \times n$, with an $n \times n$ diagonal matrix on the top, and an $(m - n) \times n$ all-zeros matrix on the bottom.

See Figure 11.2(a) for an illustration of these two cases. In both of these cases, it can be shown that

$$A = U \Sigma V^T = \sum_{i=1}^n \sigma_i U_i V_i^T,$$



(a) Full and thin SVD for full column-rank rectangular matrix. (b) Thin SVD for rank-deficient matrix.

Figure 11.2: Illustrations of the SVD and the thin SVD.

where this expresses A in terms of a sum of scaled outer products. (All the extra zeros on the diagonal simply “zero out” the extra columns of U .)

If the matrix is rank-deficient, meaning that some of the singular values equal to zero, then a similar issue arises. See Figure 11.2(b). When we have the SVD but include only a subset of the singular vectors then it is called the *thin SVD*. Let’s say that there are only ρ non-zero singular values. Then, from the thin SVD, it can be shown that

$$A = U\Sigma V^T = \sum_{i=1}^{\rho} \sigma_i U_i V_i^T,$$

where this expresses A in terms of a sum of scaled outer products. From this expression, it should be clear why the thin SVD for rank-deficient matrices can be written in terms of smaller matrices—it just corresponds to fewer terms in this sum. (Again, the $n - \rho$ columns of U , as well as the $n - \rho$ rows of V^T , that correspond to zero singular values are “zeroed out” when multiplying everything out.)

11.3 Additional properties of the SVD

The SVD has many properties that make it very useful in data science and beyond. Here, we will discuss several of these.

11.3.1 SVD and the structure of \mathbb{R}^m and \mathbb{R}^n

The SVD can be used to understand the structure of \mathbb{R}^m and \mathbb{R}^n , with respect to an $m \times n$ matrix A . If A is $m \times n$, then assuming that we don’t include the eigenvectors associated with zero eigenvalues in U and V , we have:

- UU^T is a projection matrix onto the span of U , equivalently onto the span of the columns of A . (It is a projection onto an n -dimensional subspace of \mathbb{R}^m , if A has full column rank, and it is a projection onto a ρ -dimensional subspace if A is rank-deficient, with rank $\rho < n$.)
- $I - UU^T$ is a projection matrix onto the orthogonal complement to span of U , equivalently onto the orthogonal complement to span of the columns of A , which is a subspace of \mathbb{R}^m . (It is a projection onto an $(m - \rho)$ -dimensional subspace of \mathbb{R}^m .)

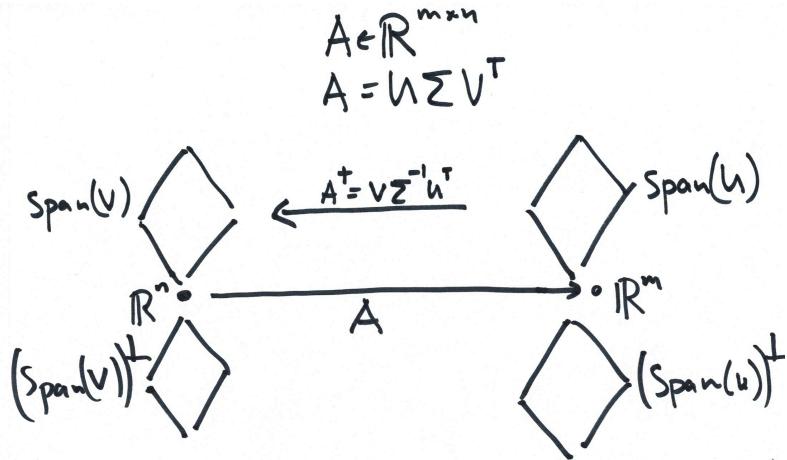


Figure 11.3: Illustration of the four fundamental subspaces associated with a matrix $A = U\Sigma V^T$.

- VV^T is a projection matrix onto the span of V , equivalently onto the span of the rows of A .
(It is a projection onto an n -dimensional subspace of \mathbb{R}^n , i.e., an identity, if A has full column rank, and it is a projection onto a ρ -dimensional subspace if A is rank-deficient, with rank $\rho < n$.)
- $I - VV^T$ is a projection matrix onto the orthogonal complement to span of V , equivalently onto the orthogonal complement to span of the rows of A , which is a subspace of \mathbb{R}^m .
(It is a projection onto an $(n - \rho)$ -dimensional subspace of \mathbb{R}^n .)

These four fundamental subspaces depend on the matrix; and, for a given matrix, they are *very* important. See Figure 11.3. Of course, any four of those could be the null vector. Also, the dimension onto which UU^T projects equals the dimension onto which VV^T projects, which is simply a manifestation of the column rank and row rank being equal.

11.3.2 SVD and the norm of a matrix

Matrix norms. As with vectors, matrices can be “large” or “small,” and it is of interest to quantify this. This can be captured by the idea of a matrix norm.

Definition 72 Given a matrix $A \in \mathbb{R}^{m \times n}$, we say that $\rho(A) \in \mathbb{R}$ is a norm or a matrix norm of A if it satisfies the following properties:

- $\rho(A) \geq 0$, and $\rho(A) = 0$ iff $A = 0$,
- $\rho(\alpha A) = |\alpha|\rho(A)$, for all numbers $\alpha \in \mathbb{R}$,
- $\rho(A + B) \leq \rho(A) + \rho(B)$.

These three conditions are simply the three conditions for a real-valued function of a vector to be a vector norm. We saw before that matrices are vectors, in that they satisfy the conditions to be a vector, and so it shouldn’t be surprising that we can use what amounts to a vector norm to define a norm for a matrix.

Recall, however, that what “really” makes a matrix a matrix is the matrix-matrix multiplication operation, i.e., that it is a linear function that can be composed or applied repeatedly. It is of interest to know how this property exhibits itself with matrix norms. For *some* matrix norms (in particular, for those of greatest interest in data science and machine learning, that we will cover), we have an additional property that captures that. In addition to satisfying the conditions of a vector norm, these norms satisfy an additional property, known as *submultiplicativity*.

Definition 73 A matrix norm, $\rho(\cdot)$ is submultiplicative or a submultiplicative matrix norm if

$$\rho(AB) \leq \rho(A)\rho(B),$$

for all A and B for which the operations are defined.

To understand this property of matrix norms, recall that the triangle inequality for the sum of two vectors (or two matrices) says that norm of the sum of two vectors is “well behaved” with respect to the norms of the original vectors. Analogously, this submultiplicativity property says that the norm of the product of two matrices is “well behaved” with respect to the norms of the original matrices. (It would be awkward if we took the norm of the sum of two small vectors and got a large vector; and, similarly, it would be awkward if we took the norm of the product of two small matrices and got a large matrix.)

As with vector norms, it is common to use the notation $\|\cdot\|$ or $\|\cdot\|_\xi$ to represent a matrix norm.

Matrix norms via the SVD. The SVD can be used to define the norm of a matrix. As with vectors, there can be many norms associated with a given $m \times n$ matrix A . Two of the most important are the following.

- **Frobenius norm.** The Frobenius norm of a matrix A can be defined as

$$\|A\|_F = \left(\sum_{i=1}^m \sum_{j=1}^n A_{ij} \right)^{1/2},$$

and it can be expressed in terms of the singular values as

$$\|A\|_F = \left(\sum_{i=1}^n \sigma_i^2 \right)^{1/2}.$$

- **Spectral norm.** The spectral norm of a matrix A can be defined as

$$\|A\|_2 = \max_{x: \|x\|=1} \|Ax\|_2,$$

and it can be expressed in terms of the singular values as

$$\|A\|_2 = \max_{i \in [n]} \sigma_i.$$

These are both matrix norms, and they are both submultiplicative matrix norms.

Observe that the Frobenius/spectral norm of a matrix A is the Euclidean/infinity norm of the vector of singular values, i.e, the diagonal elements of the singular vector matrix, viewed as a vector in \mathbb{R}^n .

Optimality properties of the SVD. One reason that matrix norms are of interest is that we often want to say that we have captured most of the information/variance in a matrix, and one way to do that is to say that the “residual” that is not captured is small. For example, we can split the thin SVD into two parts, as follows.

$$\begin{aligned} A &= U\Sigma V^T \\ &= \sum_{i=1}^{\rho} \sigma_i U_i V_i^T \\ &= \sum_{i=1}^k \sigma_i U_i V_i^T + \sum_{i=k+1}^{\rho} \sigma_i U_i V_i^T \\ &\approx \sum_{i=1}^k \sigma_i U_i V_i^T, \end{aligned}$$

where k is some integer between 1 and ρ . Since we have ordered singular values from largest to smallest, and since all are non-zero, we might hope that by keeping the first k singular values, then we get the “best” approximation in some sense. This is true, and it is extremely important in data science.

The way this is quantified is to define the matrix

$$A_k = \sum_{i=1}^k \sigma_i U_i V_i^T = U_k U_k^T A,$$

where U_k is the $m \times k$ matrix consisting of the first k singular vectors, and define the error/residual matrix

$$E = A - A_k = \sum_{i=k+1}^{\rho} \sigma_i U_i V_i^T.$$

Then, it is a fact that over all matrices of rank k , the matrix A_k that is obtained by keeping the first k singular vectors, is the best, in the sense that the error/residual E is smallest, in the sense that $\|E\|_F$ and $\|E\|_2$ is minimized.

11.3.3 SVD and inverses of non-invertible matrices: the pseudoinverse

Given an arbitrary $m \times n$ matrix A ,

- a *left inverse* is a matrix B such that $BA = I$, and
- a *right inverse* is a matrix C such that $AC = I$.

Those definitions hold even if A is not square.

Next, assume that A is square. Then, if it has a left inverse, it has a right inverse; and vice versa; and this is called the inverse, typically written A^{-1} , which satisfies

$$A^{-1}A = AA^{-1} = I.$$

The SVD can be used to define what is known as a generalization of the inverse. Given a matrix A with SVD $A = U\Sigma V^T$, then the *generalized inverse* is

$$A^\dagger = V\Sigma^{-1}U^T,$$

where Σ^{-1} is the matrix in which the non-zero elements of Σ are inverted and the zero elements of Σ are left unchanged. The generalized inverse is an inverse if you only operate on vectors that are in the row space and column space of A . See Figure 11.3. Given the generalized inverse, we can compute the following:

$$\begin{aligned} AA^\dagger &= U\Sigma V^T V \Sigma^{-1} U^T = UU^T \\ A^\dagger A &= V \Sigma^{-1} U^T U \Sigma V^T = VV^T. \end{aligned}$$

(Understand each of those steps—they are easy if you get the dimensions of the matrices correct, and they are wrong if you don't.)

Observe that one or the other or both or neither of AA^\dagger and $A^\dagger A$ are the identity matrix. More generally, they are both projection matrices (possibly an identity). The important point is that if you restrict your consideration to vectors that are in the column/row space of A , then the projection onto the column/row space is an Identity (even though it is not a matrix with ones along the diagonal and zeros off-diagonal). Thus, if you restrict your consideration to vectors that are in the column/row space of A , then the generalized inverse is simply an inverse. But since the matrix may be rank-deficient or not square, it may not be an inverse, or even a right or left inverse. See Figure 11.3.

11.4 Problems

11.4.1 Implementations and Applications of the Theory

1. XXX.
2. XXX.

11.4.2 Pencil-and-paper Problems

1. Use the method of Lagrange Multipliers to solve the following.
 - (a) Find the maximum and minimum values of $f(x_1, x_2) = 81x_1^2 + x_2^2$ subject to the constraint $4x_1^2 + x_2^2 = 9$.
 - (b) Find the maximum and minimum values of $f(x_1, x_2) = 8x_1^2 - 2x_2$ subject to the constraint $x_1^2 + x_2^2 = 1$.
 - (c) Find the maximum and minimum values of $f(x_1, x_2, x_3) = x_2^2 - 10x_3$ subject to the constraint $x_1^2 + x_2^2 + x_3^2 = 36$.
2. Let A be an $n \times n$ matrix (square, but not necessarily symmetric), let $a \in \mathbb{R}^n$ be a column vector (i.e., an $n \times 1$ matrix), and let $x \in \mathbb{R}^n$ be a column vector of variables. Then, by explicitly expanding out the vector/matrix multiplications in the definitions of z , show the following.

- (a) If $z = a^T x$, then

$$\frac{\partial z}{\partial x} = \frac{\partial a^T x}{\partial x} = a.$$

- (b) If $z = x^T x$, then

$$\frac{\partial z}{\partial x} = \frac{\partial x^T x}{\partial x} = 2x.$$

- (c) If $z = a^T A x$, then

$$\frac{\partial z}{\partial x} = \frac{\partial a^T A x}{\partial x} = A^T a.$$

- (d) If $z = x^T A x$, then

$$\frac{\partial z}{\partial x} = \frac{\partial x^T A x}{\partial x} = Ax + A^T x.$$

If, in addition, A is symmetric, then

$$\frac{\partial z}{\partial x} = \frac{\partial x^T A x}{\partial x} = 2Ax.$$

3. Let A be any $m \times n$ matrix.

- (a) Show that $A^T A$ and AA^T are both symmetric.
- (b) Show that all eigenvalues λ of $A^T A$ are nonnegative.
- (c) For an $m \times n$ matrix A , let's define the *kernel* to be the set of vectors $x \in \mathbb{R}^n$ such that $Ax = 0$. Show that all eigenvalues λ of $A^T A$ are positive if and only if the kernel of A is $\{0\}$.
- (d) For an $m \times n$ matrix A , let's define the *spectral norm* to be

$$\|A\|_2 = \max \|Ax\|_2, \text{ when } \|x\|_2 = 1,$$

where, for a vector, $\|\cdot\|_2$ is the Euclidean norm. Show that

$$\|A\|_2 = \max_{\lambda \text{ eigenvalue of } A^T A} \sqrt{\lambda}.$$

4. (SVD) This exercise shows that any invertible matrix can be viewed as a composition of rotation and/or reflection, and stretching in the direction of the axes. Recall that, for any matrix A , the matrix $A^T A$ is symmetric.

(a) Let A be an $n \times n$ matrix and set $M = A^T A$. Show that Q_M (where $Q_M(x) = x^T M x$ is the quadratic form associated with M) has signature $(m, 0)$, where $m = \text{rank}(A)$.

(b) Show that if A is invertible, then Q_M is positive definite.

The spectral theorem says that there exists an orthonormal basis v_1, \dots, v_n of \mathbb{R}^n and positive numbers $\lambda_1, \dots, \lambda_n$ such that $A^T A v_i = \lambda_i v_i$, for all i .

(c) Let $w_i = \frac{1}{\sqrt{\lambda}} A v_i$. Show that w_1, \dots, w_n is an orthonormal basis of \mathbb{R}^n .

(d) Show that

$$\begin{pmatrix} w_1 & \cdots & w_n \end{pmatrix} = A \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{1}{\sqrt{\lambda_n}} \end{pmatrix}.$$

(e) Use the previous step to show that any invertible matrix A can be written as $A = U \Sigma V^T$, where U and V are orthogonal matrices and Σ is a diagonal matrix with positive entries along the diagonal.

Hint: Recall that if a matrix Q is orthogonal, then $Q^{-1} = Q^T$.

(This factorization of A is known as the *singular value decomposition (SVD)*). This generalizes to state that any (potentially non-invertible) square matrix can be expressed as a composition of a rotation and/or reflection and a stretching, but a different proof is required than the one sketched here. This also generalizes to a similar statement for arbitrary non-square matrices, but again a different proof is required.)

Part V

**Applications: PCA, Least-Squares,
Linear Equations, PageRank,
High-Dimensional Calculus, Et Cetera**

Chapter 12

Principal Components Analysis

Principal component analysis (PCA) is a statistical technique which is central to many areas of data science. It can be viewed in one of several complementary ways. Here are several of those ways.

- **Maximizing variance.** PCA is designed to capture directions in \mathbb{R}^n of “maximum variance” in the data. As such, it is a generalization of variance/covariance matrices we discussed earlier. We have seen that for a one-dimensional function, the variance is a number, and for a two-dimensional function the variance-covariance information can be put into a 2×2 matrix. We have also seen that if two variables are strongly correlated or anti-correlated, then in some sense there is less “information” in the two variables than if they were uncorrelated. PCA amounts to performing this analysis more generally, trying to find, e.g., for a given $k \in [n]$, the k -dimensional subspace or the k linear combinations of the data that capture the maximum variance in the data, over all possible k -dimensional subspaces.
- **Minimizing reconstruction error.** Alternatively, PCA is a statistical technique which is designed to minimize the “reconstruction error” in the data. From this perspective, it generalizes the idea that the mean is the single number that best describes a one-dimensional random variable in a certain precise sense. For an n -dimensional random variable, one could ask for the one-dimensional subspace, i.e., unit vector in \mathbb{R}^n , that is best in the sense that the residual error (in the sense obtained from the Pythagorean decomposition) was minimized, when the n -dimensional random variable was projected onto it. PCA amounts to performing this analysis more generally, trying to find, e.g., for a given $k \in [n]$, the k -dimensional subspace or the k linear combinations of the data that minimize the residual error in a Pythagorean sense, over all possible k -dimensional subspaces.
- **Standardizing variables.** Alternatively, PCA is a way to construct standardized variables which generalize what we did before when we mean-centered and variance-normalized scalar random variables. Here, the variables are vectors in \mathbb{R}^n , rather than scalars in \mathbb{R} . The standardization involves mean-centering the vectors, analogously to how scalars were mean-centered; and it involves variance-normalizing, in a way that generalizes how scalars were variance-normalized. For scalars, variance-normalizing simply meant multiplying by a scalar that equaled the inverse of the standard deviation; while for vectors, variance-normalizing involves performing an orthogonal transformation (that depends on eigenvectors) as well as multiplying by a diagonal matrix of positive numbers (that depends on eigenvalues). As with many things in \mathbb{R}^n , gotchas include ensuring that the data are linearly independent, etc.

In general, if there are n random variables, the variance-covariance information can be encoded in an $n \times n$ matrix, and PCA amounts to identifying eigenvectors and eigenvalues of a suitably-normalized version of this matrix. These eigenvectors and eigenvalues provide “directions” in \mathbb{R}^n of maximum variance as well as numerical values for the variance in those directions.

Operationally, i.e., if one is interested in how to compute it and not the statistical interpretation of it, PCA can also be viewed as a “special case” of the EVD, which can be computed with the SVD.

- PCA boils down to computing the eigenvectors/eigenvalues of $A^T A$, where A is the data matrix.
- Since $A^T A$ is a symmetric matrix, all of our discussion applies.
- The only wrinkle to keep in mind is that since it is a statistical technique designed to find high-variance directions in the data, we have to mean-center the data, which basically means that we need to replace the matrix A with matrix consisting of the columns of A with their means subtracted off, and we have to be aware of variance normalization, just as we did in the ond-dimensional case.
- Computationally, since the EVD of $A^T A$ is trivially-obtainable from the SVD of A , one only needs to compute the (thin) SVD of A .

Thus, given what we have done so far, i.e., all the machinery that we have set up, understanding PCA is fairly straightforward, but it is sufficiently important that we will go through it in some detail.

12.1 The basic PCA method

PCA is a general and very common way to identify patterns in data, expressing data to highlight variability. It can be used by itself, e.g., in exploratory data analysis, and it is also the basis for many other methods. For example, since it is hard to visualize vectors in \mathbb{R}^n , PCA gives a way to visualize low-dimensional projection of the data that is optimal in a certain precise sense. These vectors can then be used for all sorts of other things, e.g., as features in other clustering algorithms.

To do PCA:

1. Get some data in the form of a numerical matrix, and preprocess it to be complete and numerical. Then we have

$$X = \begin{pmatrix} & & & \\ X_1 & X_2 & \dots & X_n \\ & & & \end{pmatrix} \in \mathbb{R}^{m \times n}.$$

Here, we are assuming that different data elements correspond to columns $X_i \in \mathbb{R}^m$ of the matrix $X \in \mathbb{R}^{m \times n}$.

Remark. Everything we say holds if the data are rows, rather than columns—we just need to mean center different things, and have lots of expressions transposed. It’s easiest to work through one way, understand it, and then go to the other way to see the differences. We’ll assume here that the data points are columns.

2. For each data point, subtract the mean. That is, $\forall i \in [n]$, do

$$X_i \rightarrow X_i - \mu_i.$$

So, in the following, we will assume that X refers to an already-mean-centered matrix.

3. Calculate the covariance matrix

$$\Sigma = X^T X \quad \text{such that} \quad \Sigma_{ij} = X_i^T X_j.$$

The reason that we mean-centered the matrix X is so that we can interpret the elements of $X^T X$ as consisting of covariances—the diagonal elements are the dot product of a mean-centered vector with itself, and the off-diagonal elements are the dot product of two different mean-centered vectors.

4. Calculate the eigenvector-eigenvalue pairs of the covariance matrix Σ , and express it as:

$$\Sigma U = U \Lambda \quad \leftrightarrow \quad \Sigma = U \Lambda U^T = \sum_{i=1}^n \lambda_i u_i u_i^T,$$

where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$.

5. Choose k to keep the top k components. This is what is known as a “model selection” rule, and there are many ways to determine it. See Figure 12.1 for an illustration. When PCA is most appropriate, there is a value for $k \ll \min\{m, n\}$ that can be used to obtain insight into the data.
6. Derive the new embedded data set. Start with

$$U = U_{\text{all cols}} = \begin{pmatrix} & & & & & & \\ u_1 & \dots & u_k & u_{k+1} & \dots & \dots & u_n \\ & & & & & & \end{pmatrix} \in \mathbb{R}^{n \times n}$$

and then if we keep only the top k then we get

$$U_k = U_{[\text{cols to keep}]} = \begin{pmatrix} & & & \\ u_1 & \dots & u_k & \\ & & & \end{pmatrix} \in \mathbb{R}^{n \times k}.$$

This matrix gives the directions in \mathbb{R}^n of maximum variance of the data. Relatedly, the subspace spanned by these columns gives the subspace that captures the maximum variance of the data. These directions are known as the PCs.

Of course, one can view an $n \times k$ matrix in one of two complementary ways: as consisting of k columns, each of which is a vector in \mathbb{R}^n ; or as consisting of n rows, each of which is a vector in \mathbb{R}^k . In the latter case, we have

$$U_k = U_{[\text{cols to keep}]} = \begin{pmatrix} u^1 \\ \vdots \\ u^n \end{pmatrix} \in \mathbb{R}^{n \times k},$$

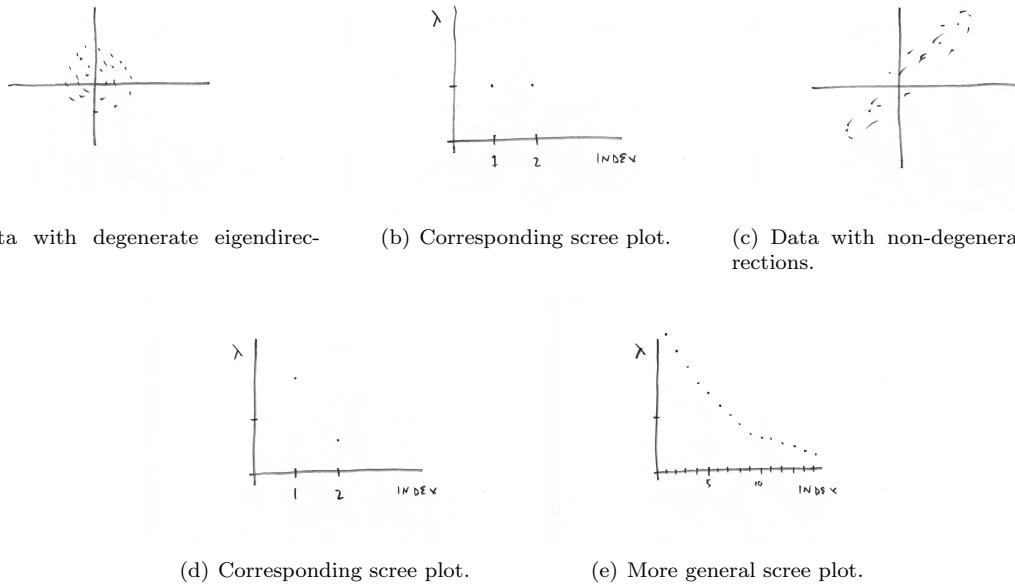


Figure 12.1: Illustration of eigenvalue decay and scree plot.

where u^i is the i^{th} row of $U_{[\text{cols to keep}]}$.

There are many things that can be done once we have the PCs. For example, we can choose $k = 2$, in which case $V_{[\text{cols to keep}]} \in \mathbb{R}^{n \times 2}$. Then, we can compute $XV_{[\text{cols to keep}]} \in \mathbb{R}^{m \times 2}$. Said another way, each row of $XV_{[\text{cols to keep}]}$ is a vector in \mathbb{R}^2 that we can plot, thereby visualizing the data.

See Figure 12.2 for an illustration. In particular, Figures 12.2(a) and 12.2(b) show data in \mathbb{R}^n that (using the top two PCs) get mapped to \mathbb{R}^2 ; and Figures 12.2(c) and 12.2(d) show data in \mathbb{R}^2 that (using the top one PC) get mapped to \mathbb{R} .

12.2 PCA and PD/PSD matrices

Let's discuss the statistical interpretation of PCA, which has to do with connections with correlations and covariances. There are strong connections between all of this and covariance/correlation matrices that we discussed in the probability part of the class.

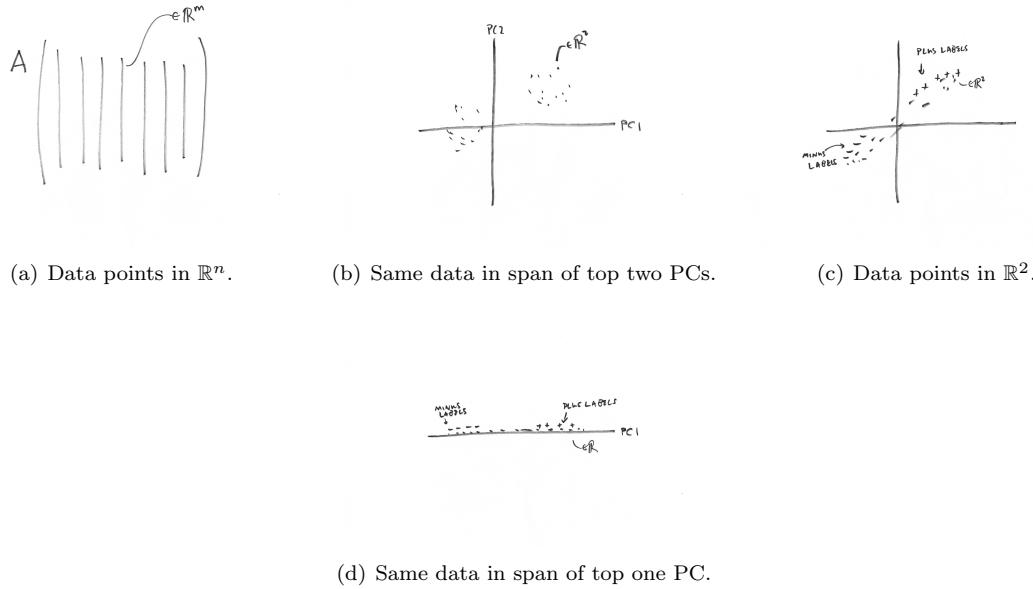
To see this, let's start with the following definition.

Definition 74 A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is a positive definite if $x^T Ax > 0$ for all $x \neq 0$; and it is positive semi-definite if $x^T Ax \geq 0$ for all $x \neq 0$

How are we to interpret this?

Since a symmetric matrix corresponds to a quadratic form, this says that the quadratic function goes up, or is flat, but does not down, in every direction. (That is, the level sets look like ellipses or parabolas but not hyperbolas.) More formally, recall that since A is symmetric, it can be decomposed in terms of its eigenvectors and eigenvalues as

$$A = U\Lambda U^T = \sum_{i=1}^n \lambda_i u_i u_i^T$$

Figure 12.2: Using PCA to map data from \mathbb{R}^n to \mathbb{R}^2 and also from \mathbb{R}^2 to \mathbb{R} .

(Here, U is orthonormal, Λ is diagonal, λ_i is the i^{th} eigenvalue, and u_i is the i^{th} eigenvector.) Since A is symmetric, all of the eigenvectors λ_i are real numbers, and if there is a degeneracy in the characteristic equation, then there are multiple eigenvectors associated with it. Thus, PD and PSD mean the following:

- If A is PD, then $\lambda_i > 0$, for all $i \in [n]$.
- If A is PSD, then $\lambda_i \geq 0$, for all $i \in [n]$.

The reason for this is that otherwise we could choose a vector x along the vector u_i (say $x = u_i$) and we would get that $x^T Ax < 0$.

Question. If $x = u_i$, then why is $x^T Ax < 0$ along this direction?

Answer. Since along that direction $Au_i = \lambda_i u_i$, with $\lambda_i < 0$, we have that $x^T Ax = x^T \lambda_i x = \lambda_i \|x\|_2^2 = \lambda_i < 0$.

So, in this case.

- If A is PD, then the corresponding quadratic function has every direction pointing up.
- If A is PSD, then the corresponding quadratic function has every direction pointing up or flat.

Let's say that $\lambda_i > 0$, for every i . (Dealing with some $\lambda_i = 0$ is a little more involved than we want to get into in this class, but the important point here is that none of the λ_i are negative.) Then, if Λ has all elements along the diagonal to be positive, then we can define a new matrix $\Lambda^{1/2}$ to be the diagonal matrix with the square root of the elements of Λ along the diagonal, i.e.,

$$\Lambda^{1/2} = \begin{pmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_n} \end{pmatrix}.$$

It is denoted $\Lambda^{1/2}$ with the superscript $1/2$ since it can be thought of as the square root of Λ , in the sense the $\Lambda^{1/2} \Lambda^{1/2} = \Lambda$. (This is a simple matrix-matrix multiplication, since the matrices are all diagonal.)

Given this, and by applying elementary operations, we see that

$$\begin{aligned}
 A &= U\Lambda U^T \\
 &= U\Lambda^{1/2}\Lambda^{1/2}U^T \quad \text{by taking the square root} \\
 &= U\Lambda^{1/2}\left(U\Lambda^{1/2}\right)^T \quad \text{by the properties of transpose} \\
 &= XX^T \quad \text{if we define } X = U\Lambda^{1/2} \\
 &= Y^TY \quad \text{if we define } Y = \left(U\Lambda^{1/2}\right)^T = \Lambda^{1/2}U^T.
 \end{aligned}$$

We took that last step just to show that the matrices XX^T and YY^T have a similar mathematical structure with respect to PD/PSD properties, their quadratic forms, and what we have been discussing. Students often think of them as very different, e.g., since one represents correlations between data points and the other represents correlations between features. In any particular application, one or the other may be more appropriate, more easily-interpretable, etc., but it should be clear that there is similar mathematics underlying them.

So, what this says is that if A is PD, then we can write A as

$$A = XX^T \text{ (or } Y^TY).$$

Conversely, if $A = Y^TY$, then

$$x^T Ax = x^T Y^T Y x = (Yx)^T Y x = z^T z = \|z\|_2^2 > 0,$$

where obviously $z = Yx$. So, if $A = Y^TY$, then A is PD.

So, let's take a step back and ask what is a PD/PSD matrix

- $A = I$: In this case,

$$x^T Ax = x^T I x = x^T x = \|x\|_2^2,$$

and A just defined the usual unit ball.

- $A = D$, where D is some diagonal matrix: In this case,

$$x^T Dx = x^T D^{1/2} D^{1/2} x = \left(D^{1/2} x\right)^T D^{1/2} x = \sum_i d_i x_i^2,$$

and thus A is just a unit ball in a stretched out norm.

- $A = Y^TY$ for general Y : in this case,

$$x^T Ax = x^T Y^T Y x = (Yx)^T Y x = \sum_i \left(\sum_j Y_{ij} x_j \right)^2.$$

See Figure 12.3 for an illustration of these three cases. In particular, the point is that—up to a rotation and scaling—they all look like circles. The rotation is determined by the eigenvectors, and the scaling is given by the eigenvalues.

12.3 When vanilla PCA is not particularly appropriate to use

It's important to note that PCA is a procedure, and it will always return an answer, whether or not it is an appropriate procedure to use in a given situation. We have been describing examples of how it performs when applied to data which it makes sense to apply it to, but it is also worth considering what happens if it is applied to other types of data. To see this, recall that here are the basic assumptions underlying the use of PCA:

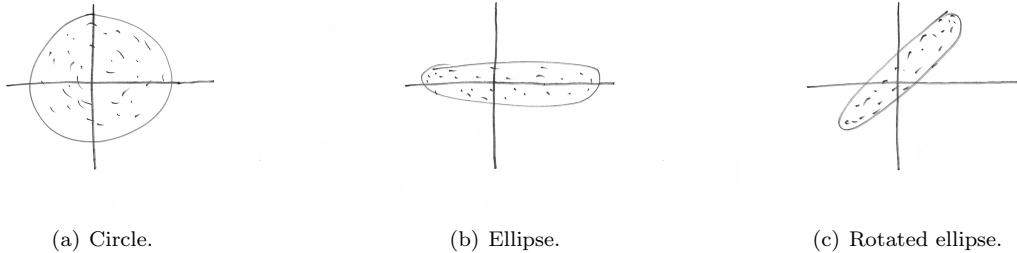


Figure 12.3: Three examples of level sets of quadratic forms associated with PD matrices.

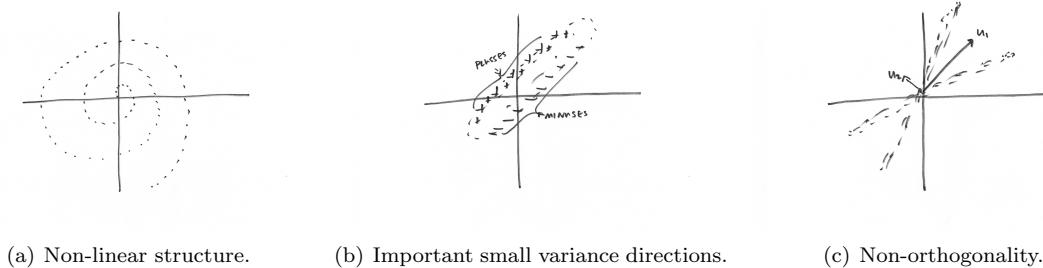


Figure 12.4: Three examples of when vanilla PCA is not particularly appropriate to use.

- **Linear structure.** That is, there is structure in the data that can be meaningfully identified by the basic operations of scaling and linear combinations, i.e., subspaces and related operations.
- **Large variance directions contain important structure.** That is, the interesting or important parts of the data are not “hidden” in the small variance directions.
- **Orthogonality.** The PCs are orthogonal, and so this property is related to structure in the data.

See Figure 12.4 for an illustration of examples where these assumptions are violated and thus where a vanilla application of PCA is not particularly appropriate. Importantly, the probabilistic and linear algebraic properties we have been discussing are so important that methods developed for these cases still have strong connections to PCS, the spectral theorem, etc.

12.4 Statistical interpretation of PCA

So far, we have been talking about symmetric matrices (when we get a full set of orthonormal eigenvectors and associated eigenvalues) and then PD/PSD matrices (when those eigenvalues are positive/non-negative). There is one other minor wrinkle/gotcha that you need to keep in mind when you apply these ideas to PCA. This will be obvious—if you think of PCA from a statistical perspective.

For us, when we are interested in connections with probability, there are two main specializations.

1. Mean Center
2. Normalized Variance Scale

We have seen both of these before, but in a different guise.

- $\mathbf{E}[X]$
- $\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2]$: the norm squared of a vector, but we took off the mean from the vector, so variability is like norm squared.
- $\sigma[X] = \sqrt{\mathbf{Var}[X]}$: square root of variance, so like the norm of a vector.
- $\mathbf{Cov}[X, Y] = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$: this is like the dot product between two vectors, but only after the mean of each vector has been removed.
- $\mathbf{Corr}[X, Y] = \frac{\mathbf{Cov}[X, Y]}{\sigma[X]\sigma[Y]}$: this is like the cosine of the angle between two mean centered vectors.

We will see that $\mathbf{Cov}[X, Y]$ and $\mathbf{Corr}[X, Y]$ are both PD/PSD matrices. When we are interested in statistical and probabilistic interpretations, it often makes sense to mean center and variance normalize these matrices, but qua PD/PSD properties, there is no need to. Thus, mathematically, these matrices should all be thought about as being nearly the same, in the sense that they are examples of general quadratic forms where none of the directions point down. It's best to figure out what question you want to ask, e.g., if you want to characterize the variability about the mean, and then perform operations to get what you want, e.g., subtract off the mean. But, many of the ideas still go through regardless of whether or not you do that.

BTW, we can view these expressions (e.g., $\mathbf{Cov}[X, Y]$ and $\mathbf{Corr}[X, Y]$) in two ways:

- Properties of two random variables and how they correlate.
- Properties of one random variable in \mathbb{R}^2 .

It's important to note that many of the operations that are performed on data have a natural linear algebraic interpretation, and this also helps generalize to higher-dimensional data.

Two-dimensional example. This discussion holds for n random variables, or equivalently for random variables that are vectors in \mathbb{R}^n , but let's restrict to the $n = 2$ case. First, mean center: $X \rightarrow X - \mathbf{E}[X]$ and $Y \rightarrow Y - \mathbf{E}[Y]$. Then, compute variances and covariances:

$$\begin{aligned}\mathbf{Var}[X] &= \mathbf{E}[(X - \mathbf{E}[X])^2] \\ \mathbf{Var}[Y] &= \mathbf{E}[(Y - \mathbf{E}[Y])^2] \\ \mathbf{Cov}[X, Y] &= \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \mathbf{Cov}[Y, X].\end{aligned}$$

We can write this as a matrix:

$$\Sigma = \begin{pmatrix} \mathbf{Var}[X] & \mathbf{Cov}[X, Y] \\ \mathbf{Cov}[Y, X] & \mathbf{Var}[Y] \end{pmatrix}.$$

Second, variance normalize: $X \rightarrow \frac{X}{\sigma(X)}$ and $Y \rightarrow \frac{Y}{\sigma(Y)}$: Then,

$$\begin{aligned}\mathbf{Var}[X] &\rightarrow \frac{1}{\sigma^2(X)} \mathbf{Var}[X] = 1 \\ \mathbf{Var}[Y] &\rightarrow \frac{1}{\sigma^2(Y)} \mathbf{Var}[Y] = 1 \\ \mathbf{Cov}[X, Y] &\rightarrow \frac{1}{\sigma(X)\sigma(Y)} \mathbf{Cov}[X, Y] = \rho \in [0, 1].\end{aligned}$$

So,

$$\Sigma \rightarrow \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = \Sigma'.$$

where $\rho = \text{Cov}[X, Y]$. Recall that if we post/pre-multiply, then we rescale the cols/rows of a matrix. Thus, we can write the previous expression as a matrix multiplication:

$$\begin{pmatrix} 1/\sigma(X) & 0 \\ 0 & 1/\sigma(Y) \end{pmatrix} \begin{pmatrix} \text{Var}[X] & \text{Cov}[X, Y] \\ \text{Cov}[Y, X] & \text{Var}[Y] \end{pmatrix} \begin{pmatrix} 1/\sigma(X) & 0 \\ 0 & 1/\sigma(Y) \end{pmatrix} = (\text{diag}(\Sigma))^{-1/2} \Sigma (\text{diag}(\Sigma))^{-1/2}.$$

Example. Let's look in more detail at the matrix

$$\Sigma' = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

In particular, let's look at the EVD of this.

First, compute it's eigenvalues. Since it is a 2×2 matrix, we have

$$\det \begin{vmatrix} 1 & \rho \\ \rho & 1 \end{vmatrix} = 0 = (1 - \lambda)^2 - \rho^2 = \lambda^2 - 2\lambda + 1 - \rho^2,$$

and from this we have

$$\begin{aligned} \lambda &= \frac{2 \pm \sqrt{4 - 4 + 4\rho^2}}{2} \\ &= 1 \pm \rho. \end{aligned}$$

If $\rho > 0$, then $\lambda_1 = 1 + \rho > 1 - \rho = \lambda_2$, and vice versa.

Second, let's compute the two eigenvectors. For $\lambda_1 = 1 + \rho$, we have

$$\begin{pmatrix} -\rho & \rho \\ \rho & -\rho \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{from which it follows that } u_{\lambda=1+\rho} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}.$$

For $\lambda_2 = 1 - \rho$, we have

$$\begin{pmatrix} \rho & \rho \\ \rho & \rho \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{from which it follows that } u_{\lambda=1-\rho} = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}.$$

Note that the eigenvectors are independent of ρ but eigenvalues depend on ρ . Let's look at this in more detail to understand why. What this says is that in this matrix there are two orthogonal directions where the action of the matrix only stretches a vector, and those directions don't depend on ρ but the magnitude of the stretch does depend on ρ .

We can write the matrix $\Sigma' = U\Lambda U^T$ as follows:

$$\begin{aligned} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} &= \frac{1}{\sqrt{2}} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\ &= \frac{1+\rho}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \end{pmatrix} + \frac{1-\rho}{2} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \begin{pmatrix} 1 & -1 \end{pmatrix}. \end{aligned}$$

The second line emphasizes that there are two directions and that the two directions stretch by different amounts.

- The first direction is given by the eigenvector $u_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, and the projection onto the span of that vector is given by

$$\Pi_{u_1} = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \end{pmatrix}.$$

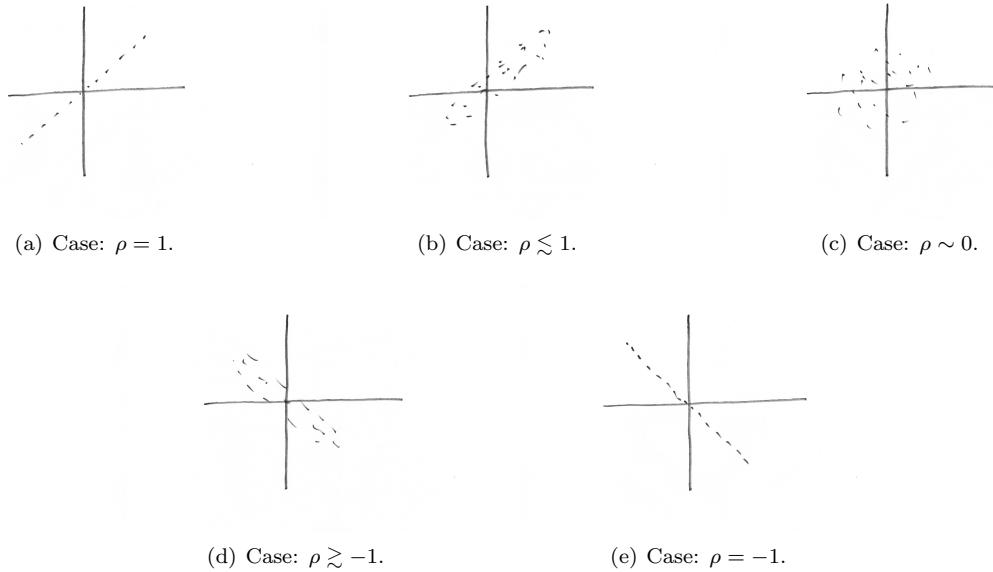


Figure 12.5: Illustration of different amounts of correlation.

- The second direction is given by the eigenvector $u_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, and the projection onto the span of that vector is given by

$$\Pi_{u_2} = \frac{1}{2} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \begin{pmatrix} 1 & -1 \end{pmatrix}.$$

Thus, we could alternatively write the matrix as

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = (1 + \rho) \Pi_{u_1} + (1 - \rho) \Pi_{u_2}.$$

It should be clear that this is a special case of Theorem 21 and Equation (11.8).

Note several cases.

- If $\rho = 1$, then the data are projected onto u_1 and no information is lost, and this is since the term $(1 - \rho) \Pi_{u_2}$ doesn't contribute anything.
- If $\rho \lesssim 1$, then the data are projected onto u_1 and very little information is lost, and this is since the term $(1 - \rho) \Pi_{u_2}$ doesn't contribute much.
- If $\rho \sim 0$, then there is little or no correlation between the two components (and we need to keep both equations in order not to lose too much information).
- If $\rho \gtrsim -1$, then the data are projected onto u_2 and very little information is lost, and this is since the term $(1 + \rho) \Pi_{u_1}$ doesn't contribute much.
- If $\rho = -1$, then the data are projected onto u_2 and no information is lost, and this is since the term $(1 + \rho) \Pi_{u_1}$ doesn't contribute anything.

See Figure 12.5 for an illustration of these cases.

Remark. Informally, it is often said that u_1 “explains” $\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1+\rho}{1+\rho+1-\rho} = \frac{1+\rho}{2}$ fraction of the variability, where in the above cases the fraction is $\sim 100\%$, $\sim 50\%$, and $\sim 0\%$ of the data; and $u + 2$ “explains”

$\frac{\lambda_2}{\lambda_1 + \lambda_2} = \frac{1-\rho}{2}$ fraction of the variability, where again in the above cases this is $\sim 0\%$, $\sim 50\%$, and $\sim 100\%$. I put explain in quotations since nothing is really being explained, e.g., in the sense of providing an explanation in terms of the processes that generated the data, but instead that much fraction of the variability is preserved when we project onto that component.

So, for 2 variables, the notion of correlation/covariance maps cleanly to what we did before in linear algebra. In particular,

- If $\rho = 0$, then we need a full 2-dimensional subspace to describe/reconstruct the data.
- If $\rho = \pm 1$, then we only need a 1-dimensional subspace to describe/reconstruct the data.
- If $\rho \in (-1, 0)$ or $\rho \in (0, 1)$, then we have a 2-dimensional subspace, but one is more important in the sense that more variability is described by one direction than the other.
- If $\rho \lesssim 1$ or $\rho \gtrsim -1$, then we have a 2-dimensional subspace, but one is much more important (in the same sense) and thus if we project the data onto the more important dimension then we don't lose too much of the variability.

Three-dimensional example. All of that discussion holds for \mathbb{R}^2 , i.e., for two random variables X and Y , and in that case the only interesting subspaces were 1-dimensional. (It also holds for one-dimensional examples, which simply reduce to standardizing via mean-centering and variance-normalizing.) But it holds more generally, and this is where we use the linear algebra in probability in more interesting ways. We could consider m random variables, each of which was a vector in \mathbb{R}^n . But, let's make more modest steps. In particular, consider the case of \mathbb{R}^3 , which is what we get when we have 3 random variables. In this case, the Σ matrix becomes

$$\Sigma = \begin{pmatrix} \text{Var}[X] & \text{Cov}[X, Y] & \text{Cov}[X, Z] \\ \text{Cov}[Y, X] & \text{Var}[Y] & \text{Cov}[Y, Z] \\ \text{Cov}[Z, X] & \text{Cov}[Z, Y] & \text{Var}[Z] \end{pmatrix}$$

Again, we can variance-normalize by pre-multiplying and post-multiplying, like we did before.

The point here is that, in this case, if there are correlations, then the matrix might be rank deficient, in which case a 2-dimensional subspace (or maybe even a 1-dimensional subspace suffices to capture most/all of the variability). In \mathbb{R}^2 , this was trivial, since for Y to be a linear combination of X meant that is is a scalar multiple, but in \mathbb{R}^3 , it is more interesting/nontrivial since A could be a linear combination of both X and Y .

Example. A simple example is

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

More generally, this holds for an arbitrary number of variables, i.e., in \mathbb{R}^n . Let

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n.$$

Given this vector, we can define a matrix $\Sigma \in \mathbb{R}^{n \times n}$ such that

$$\Sigma_{ij} = \text{Cov}[X_i, X_j] = \mathbf{E}[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j])],$$

which we can write more efficiently as

$$\Sigma = \mathbf{E}[(X - \mathbf{E}[X])(X - \mathbf{E}[X])^T],$$

from which we can define

$$\mathbf{Corr}[X,] = (\text{diag}(\Sigma))^{-1/2} \Sigma (\text{diag}(\Sigma))^{-1/2}.$$

Facts. Here are some facts that are good to know. If $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$ and if $\mu = \mathbf{E}[x]$ and $\Omega = \mathbf{E}[(x - \mathbf{E}[x])(y - \mathbf{E}[y])^T]$, then

- $\Omega = \mathbf{E}[xx^T] - \mu\mu^T$
- Ω is SPSD
- $\mathbf{Cov}[Ax + a,] = A\mathbf{Cov}[x,]A^T$
- $\mathbf{Cov}[x, y] = \mathbf{Cov}[y, x]$
- $\mathbf{Cov}[x_1 + x_2, y] = \mathbf{Cov}[x_1, y] + \mathbf{Cov}[x_2, y]$ If $p = q$, then
 - $\mathbf{Var}[x + y] = \mathbf{Var}[x] + \mathbf{Cov}[x, y] + \mathbf{Cov}[y, x] + \mathbf{Var}[y]$.
 - $\mathbf{Cov}[Ax + a, B^Ty + b] = A\mathbf{Cov}[x, y]B$
- If x and y are independent or uncorrelated, then $\mathbf{Cov}[x, y] = 0$

All of these are statements about random variables and the correlation/covariances, but as the matrix of random variables becomes larger, it becomes much easier to do this in terms of the linear algebra we have been discussing.

Facts. Here are some facts that are good to know. Let A be an $n \times n$ matrix (square, but not necessarily symmetric), let $a \in \mathbb{R}^n$ be a column vector (i.e., an $n \times 1$ matrix), and let $x \in \mathbb{R}^n$ be a random column vector of variables with mean μ and variance-covariance matrix Ω . (We can view x as a random vector, but we can also view x in terms of its elements x_i , for $i \in \{1, \dots, n\}$, in which case $\Omega_{ij} = \mathbf{Cov}[x_i, x_j]$, etc.) Then, we can show the following.

- $\mathbf{E}[a^T x] = a^T \mu$.
- $\mathbf{E}[Ax] = A\mu$.
- $\mathbf{Var}[a^T x] = a^T \Omega a$.
- $\mathbf{Var}[Ax] = A\Omega A^T$.
Note that if $\Omega = \sigma^2 I$, then $\mathbf{Var}[Ax] = \sigma^2 AA^T$.
- $\mathbf{Cov}[Ax, By] = A\mathbf{Cov}[x, y]B^T$.
- $\mathbf{E}[x^T Ax] = \mathbf{Tr}(A\Omega) + \mu^T A\mu$.
(And so $\mathbf{E}[(x - \mu)^T A(x - \mu)] = \mathbf{Tr}(A\Omega)$.)
Note that if $\Omega = \sigma^2 I$, then $\mathbf{E}[x^T Ax] = \sigma^2 \mathbf{Tr}(A) + \mu^T A\mu$.

12.5 More on variable transformations underlying PCA

We have seen that when working with one-dimensional probability functions, e.g., a normal distribution, operations such as mean-centering and variance-normalizing can be viewed as doing simple variable transformations to standardize them, i.e., to construct a new random variable that is in a standard form, and with which it is easier to work. We have also seen that when working with two-dimensional and n -dimensional functions, quadratic forms on \mathbb{R}^n , operations such as completing the square can be used to remove “cross terms” and can also be viewed as a variable transformation to put the data in a standardized diagonal form. PCA lets us view these two procedures on a common footing. Let’s discuss this, both in the special case of standardized/non-standardized normal random variables (which is sometimes convenient for the math) as well as more generally (which is typically how it is actually applied in practice).

Normal random variables. Recall that, in one dimension, a normal $N(\mu, \sigma^2)$ random variable has probability density

$$f(x) = f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

A special case of this is a standardized normal random variable, with mean $\mu = 0$ and variance $\sigma^2 = 1$, which has probability density

$$f(x) = f(x|0, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right).$$

We can go back and forth between standardized and non-standardized versions, as follows.

- **Starting with a standardized random variable.** If we are given a random variable

$$x \sim N(0, 1),$$

where “ \sim ” means that x is drawn from that distribution, then it follows that if we define a new random variable as $x' = \sigma x + \mu$, then

$$x' \sim N(\mu, \sigma^2).$$

To see this, simply note that $x = \frac{x'-\mu}{\sigma}$ and plug that into the previous expression.

We can think of starting with a standardized normal random variable and constructing an arbitrary normal random variable as “constructive,” meaning that we start with a simple random variable and construct a more complex one.

(Observe that, in this constructive approach which starts with a standardized random variable, x' is obtained from x by first scaling by σ and then translating by μ .)

- **Starting with a non-standardized random variable.** If we are given a random variable

$$x \sim N(\mu, \sigma^2),$$

where again “ \sim ” means that x is drawn from that distribution, then it follows that if we define a new random variable as $x' = \frac{x-\mu}{\sigma}$, then

$$x' \sim N(0, 1).$$

To see this, simply note that $x = \sigma x' + \mu$ and plug that into the previous expression.

We can think of starting with a non-standardized random variable and backing out the corresponding standardized variable as “operational,” meaning that we start with an arbitrary random variable and transform it into a more simple one.

(Observe that, in this operational approach which starts with an arbitrary random variable, x' is obtained from x by first translating by $-\mu$ and then scaling by $1/\sigma$.)

It is good to be able to go back and forth like that, in \mathbb{R} as well as \mathbb{R}^n .

To extend this to vectors in \mathbb{R}^n , let’s start with the constructive approach, and let’s do it in steps.

- **Simplest case.** In the simplest case, let’s consider a vector $x \in \mathbb{R}^n$, where each $x_i \sim N(0, 1)$, in which case

$$f(x) = \prod_{i=1}^n f(x_i) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2} \sum_{i=1}^n x_i^2\right) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2} x^T x\right).$$

Observe that, in this simplest case, the argument of the exponential is a quadratic form (with no cross terms and no scaling).

- **Translation.** If we consider $x' = x + b$, then $x = x' - b$ and

$$x^T x = (x' - b)^T (x' - b).$$

In this case, in the transformed variables, we have

$$f(x') = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}x'^T x'\right),$$

and thus in the original variables we have

$$f(x) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}(x - b)^T (x - b)\right).$$

Observe that the normalization is unchanged since the translation does not affect that, but that the argument to the exponential becomes $-\frac{1}{2}$ times

$$(x - b)^T (x - b) = x^T x - 2b^T x + b^T b,$$

which is a quadratic function with both unscaled quadratic terms as well as with a linear term.

- **Stretching.** If we consider $x' = \Lambda^{1/2}x$, where Λ is a diagonal matrix with positive entries, then $x = \Lambda^{-1/2}x'$ and

$$x^T x = x'^T \Lambda^{-1/2} \Lambda^{-1/2} x' = x'^T \Lambda^{-1} x' = x'^T \Sigma^{-1} x',$$

if $\Sigma = \Lambda$. In this case, in the transformed variables we have

$$f(x') = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}x'^T x'\right),$$

and thus in the original variables we have

$$f(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}x^T \Sigma^{-1} x\right).$$

Observe two things. First, there is a determinant or Jacobian factor in the normalization that must be accounted for. This is the generalization of having σ^2 appear in the denominator in the one-dimensional case. Second, in this case, Σ is a diagonal matrix, and thus the argument of the exponential becomes

$$x^T \Sigma^{-1} x = \sum_{i=1}^n \Sigma_{ii}^{-1} x_i^2,$$

which is an axis-aligned positive definite quadratic form in \mathbb{R}^n .

- **Orthogonal transformation.** If we consider $x' = Ux$, where U is an orthogonal transformation, then $x = U^T x'$ and

$$x^T x = x'^T U U^T x' = x'^T x'.$$

In this case, in the transformed variables we have

$$f(x') = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}x'^T x'\right),$$

and thus in the original variables we have

$$f(x) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}x^T x\right).$$

Observe that both are the same, i.e., simply rotating or orthogonally transforming the variables x does not affect $f(\cdot)$. This is an example of a non-uniqueness that arises often.

- **Combining all three.** If we consider

$$x' = U\Lambda^{1/2}x + b, \quad (12.1)$$

where Λ is a diagonal matrix with positive entries and U is an orthogonal matrix, then

$$x = \Lambda^{-1/2}U^T(x' - b) \quad (12.2)$$

and we have that

$$\begin{aligned} x^T x &= (x' - b)^T U \Lambda^{-1/2} \Lambda^{-1/2} U^T (x' - b) \\ &= (x' - b)^T U \Lambda^{-1} U^T (x' - b) \\ &= (x' - b)^T \Sigma^{-1} (x' - b), \end{aligned}$$

if $\Sigma = U\Lambda U^T$. In this case, in the transformed variables we have

$$f(x') = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}x'^T x'\right),$$

and thus in the original variables we have

$$f(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(x - b)^T \Sigma^{-1} (x - b)\right).$$

Again, there is a determinant or Jacobian factor in the normalization that must be accounted for; and again, the argument of the exponential is positive definite quadratic form in \mathbb{R}^n , except it is not axis-aligned, and it is translated to remove the linear terms.

We went through that in detail, since it is important to be able to “unpack” expressions like

$$f(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(x - b)^T \Sigma^{-1} (x - b)\right). \quad (12.3)$$

If, rather than *constructing* an expression of the form of Eqn. (12.3), we are given an expression of that form, then we simply observe that we can go in the reverse order. This is what *operationally* happens in practice, e.g., with PCA. That is, we consider the transformed variable

$$x' = \Lambda^{-1/2}U^T(x - b) \quad (12.4)$$

and with respect to these new variables we have

$$f(x') = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}x'^T x'\right),$$

which is simply the simplest case we started with.

How exactly does the transformation in Eqn. (12.4) work?

- First, we translate: $x \rightarrow x - b$.
- Then we rotate: $x - b \rightarrow U^T(x - b)$.
- Then, we stretch: $U^T(x - b) \rightarrow \Lambda^{-1/2}U^T(x - b)$.

The order is important. (In particular, if we do the rotation and stretch in a different order, which linear-algebraically is a legitimate operation, or we translate at some other step, then we get something different.)

Arbitrary random variables. We have been talking about this in terms of normal/Gaussian distributions, since quadratic forms are nice and since this is basically what PCA is doing under the hood, but there is really no need to discuss normality. That is, what really makes all of this work is having control on the means and variances of random variables, and transforming in a way that respects that. To see this, let's say that we start with any random variable $x \in \mathbb{R}^n$, with

$$\begin{aligned}\mathbf{E}[x] &= 0, \quad \text{and} \\ \mathbf{Var}[x] &= I.\end{aligned}$$

(Note that since x is a vector, $\mathbf{E}[x]$ is a vector, and $\mathbf{Var}[x]$ is a matrix.) Then, we can define a transformed variable

$$x' = U\Lambda^{1/2}x + b,$$

which observe is exactly the transformation we did in Eqn. (12.1) above, and we can ask what is the mean and variance of it. To compute the mean, observe that

$$\begin{aligned}\mathbf{E}[x'] &= \mathbf{E}[U\Lambda^{1/2}x + b] \\ &= \mathbf{E}[U\Lambda^{1/2}x] + \mathbf{E}[b] \\ &= U\Lambda^{1/2}\mathbf{E}[x] + b \\ &= b,\end{aligned}$$

since $\mathbf{E}[x] = 0$ and $\mathbf{E}[b] = b$. Also, observe that

$$\begin{aligned}\mathbf{Var}[x'] &= \mathbf{Var}[U\Lambda^{1/2}x + b] \\ &= U\Lambda^{1/2}\mathbf{Var}[x]\left(U\Lambda^{1/2}\right)^T \\ &= U\Lambda^{1/2}\mathbf{Var}[x]\Lambda^{1/2}U^T \\ &= U\Lambda^{1/2}I\Lambda^{1/2}U^T \\ &= U\Lambda^{1/2}\Lambda^{1/2}U^T \\ &= U\Lambda U^T \\ &= \Sigma.\end{aligned}$$

That is, this transformed random variable has expectation b and variance Σ .

As with the discussion of normal random variables, here to we can also “go the other way.” To do this, let's start with an arbitrary random variable x with

$$\begin{aligned}\mathbf{E}[x] &= b, \quad \text{and} \\ \mathbf{Var}[x] &= \Sigma.\end{aligned}$$

and let's transform it to random variable with mean 0 and variance I . To do that, consider

$$x' = \Lambda^{-1/2}U^T(x - b),$$

which observe is exactly the transformation we did in Eqn. (12.2) above, and which we obtain by first mean centering, then rotating, then rescaling. (If we didn't know b and Σ , we could estimate them, etc., but that adds an extra layer of complexity, so let's ignore that for now.) Then we can ask about the mean and variance of the x' random variable. To compute its mean, observe that

$$\begin{aligned}\mathbf{E}[x'] &= \mathbf{E}\left[\Lambda^{-1/2}U^T(x - b)\right] \\ &= \mathbf{E}\left[\Lambda^{-1/2}U^Tx\right] - \mathbf{E}\left[\Lambda^{-1/2}U^Tb\right] \\ &= \Lambda^{-1/2}U^T\mathbf{E}[x] - \Lambda^{-1/2}U^Tb \\ &= \Lambda^{-1/2}U^Tb - \Lambda^{-1/2}U^Tb \\ &= 0.\end{aligned}$$

To compute its variance, observe that

$$\begin{aligned}
 \mathbf{Var}[x'] &= \mathbf{Var}[\Lambda^{-1/2}U^T(x - b)] \\
 &= \mathbf{Var}[\Lambda^{-1/2}U^Tx - \Lambda^{-1/2}U^Tb] \\
 &= \mathbf{Var}[\Lambda^{-1/2}U^Tx] \\
 &= \Lambda^{-1/2}U^T\mathbf{Var}[x]\left(\Lambda^{-1/2}U^T\right)^T \\
 &= \Lambda^{-1/2}U^T\mathbf{Var}[x]U\Lambda^{-1/2} \\
 &= \Lambda^{-1/2}U^T\Sigma U\Lambda^{-1/2} \\
 &= \Lambda^{-1/2}U^TU\Lambda U^TU\Lambda^{-1/2} \\
 &= \Lambda^{-1/2}\Lambda\Lambda^{-1/2} \\
 &= I.
 \end{aligned}$$

So, the new variable x' has mean 0 and variance I .

At the end of the day, when you do PCA, you are finding these new variables.

12.6 Problems

12.6.1 Implementations and Applications of the Theory

1. XXX.
2. XXX.

12.6.2 Pencil-and-paper Problems

1. Let A be an $n \times n$ matrix (square, but not necessarily symmetric), let $a \in \mathbb{R}^n$ be a column vector (i.e., an $n \times 1$ matrix), and let $x \in \mathbb{R}^n$ be a random column vector of variables with mean μ and variance-covariance matrix Ω . (We can view x as a random vector, but we can also view x in terms of its elements x_i , for $i \in \{1, \dots, n\}$, in which case $\Omega_{ij} = \mathbf{Cov}[x_i, x_j]$, etc.)
 - (a) Show that $\mathbf{E}[a^T x] = a^T \mu$.
 - (b) Show that $\mathbf{E}[Ax] = A\mu$.
 - (c) Show that $\mathbf{Var}[a^T x] = a^T \Omega a$.
 - (d) Show that $\mathbf{Var}[Ax] = A\Omega A^T$.
Derive the simpler expression if $\Omega = \sigma^2 I$.
 - (e) Show that $\mathbf{E}[x^T Ax] = \mathbf{Tr}(A\Omega) + \mu^T A\mu$.
Derive the simpler expression if $\Omega = \sigma^2 I$.
2. XXX. PROBLEM INVOLVING PCA-BASED PROJECTIONS WITH DECOMPOSITIONS INTO
3. Show that if A is an $n \times n$ positive semidefinite matrix, then its EVD, let's call it $A = QDQ^T$, agrees exactly with its SVD, let's call it $A = U\Sigma V^T$ (i.e., show that $\Sigma = D$ and that $Q = U = V$).
4. Let

$$A = \begin{pmatrix} 0 & 0 & -5 \\ -9 & 12 & 0 \\ 0 & 0 & 0 \\ 8 & 6 & 0 \end{pmatrix}.$$

- (a) Compute the SVD of A . Express your answer (i) as the sum of rank-1 terms and (ii) as $A = U\Sigma V^T$ for appropriate U , Σ , and V .
- (b) Find the best rank-2 approximation A_2 of A (where “best” means the matrix that is closest in squared Frobenius norm). Express your answer (i) as the sum of rank-1 terms and (ii) as $A_2 = U\Sigma_2 V^T$ for appropriate U , Σ_2 , and V .
- (c) Compute the approximation error $\|A - A_2\|_F^2$ in terms of the singular values of A .

5. Let

$$A = \begin{pmatrix} 2 & -3 \\ 0 & 2 \end{pmatrix}.$$

- (a) Compute the SVD of A . Express your answer (i) as the sum of rank-1 terms and (ii) as $A = U\Sigma V^T$ for appropriate U , Σ , and V .
- (b) In \mathbb{R}^2 , describe the image of the unit disk under the transformation of A using the SVD. That is, draw a picture of the region $\{Ax : \|x\|_2 \leq 1\}$.
- (c) In \mathbb{R}^2 , describe the inverse image of the unit disc by drawing a picture of the region $\{x : \|Ax\|_2 \leq 1\}$.

6. (PCA) Consider the covariance matrix

$$\Sigma = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

- (a) Find the PCs of this covariance matrix, and show that they account for

$$\begin{aligned} \lambda_1 &\approx 5.83 \\ \lambda_2 &\approx 2.00 \\ \lambda_3 &\approx 0.17 \end{aligned}$$

of the total variation in Σ .

- (b) Convert the covariance matrix Σ to a correlation matrix Σ' .
- (c) Compute the PCs of the correlation matrix Σ' , and compute the proportion of the total variance explained by each component.
- (d) Are the components and proportion of variation explained the same for Σ and Σ' ? Should they be? Why or why not?

Chapter 13

Least-squares (LS) regression

Regression analysis is a set of statistical methods for estimating the relationships among variables. For example, one might hope to make a prediction for one variable, given the value of the other variables. To do that, one typically considers a model, which is some sort of mathematical function encoding input-output relationships, and which has some parameters, and one tries to find the best values of those parameters. Here, “best” often means the value of the parameters that minimizes the prediction error. There are many ways to quantify this, but perhaps the “hello world” way is with the method of least-squares. Least-squares regression is an important application of the spectral theorem ideas we have been discussing, and it is central to many other methods used throughout data science.

13.1 Least-squares (LS) regression

When we discussed PCA, we just “described” the data in terms of a small number of eigenvectors, or equivalently in terms of a ball or an ellipsoid with the “best” (in the sense of maximum variance) axes. That is, we weren’t explicitly predicting anything. This is of interest, e.g., if we want to do visualization or exploratory data analysis.

In many applications, however, one wants to try to predict something, e.g., entries in one column from the entries in the other columns. To do this, one considers different problems. For example, two of the most common are the following.

- Regression problems (roughly, if the output variable that is being predicted takes continuous values).
- Classification problems (roughly, if the output variable that is being predicted takes a discrete set of values).

This is a large topic in data science and machine learning, and we will only consider the simplest example. For this simplest example, there are strong connections with what we have been discussing, and we will highlight some of these connections. For more complicated examples, many of the ideas boil down to variants of what we are discussing.

13.1.1 Trying to model data to make predictions

Making predictions. To start, consider Figure 13.1, variants of which we have seen in different contexts before. Each subfigure of this figure illustrates a large number of data points, each of which is an element of \mathbb{R}^2 , laid out in a certain way. When we were discussing PCA, we asked what were the best axes to

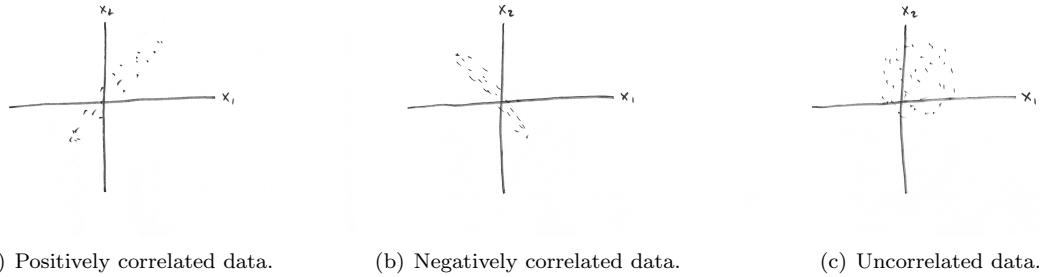


Figure 13.1: Illustration of different linear relationships in data.

describe the data in each of these subfigures, where “best” had to do with being associated with the largest eigenvector, which we showed had an interpretation in terms of variance. We saw that for figures such as Figure 13.1(a), a vector pointing up and to the right was the best, and a vector pointing orthogonal to it was next best; for Figure 13.1(b), a vector pointing down and to the right was the best, and a vector pointing orthogonal to it was the next best; and for Figure 13.1(c), there was a degeneracy, in that any two vectors in \mathbb{R}^2 are nearly equally as good in terms of capturing the variance in the data. Here, we want to ask a different question. We want to ask: if we are given a new data point $x \in \mathbb{R}^2$, but we are only told the value for x_1 , then how well can we predict the value for x_2 ? (Or vice versa?)

For example, if x_1 and x_2 are the height and weight of people, then we might expect a relationship that looks something like Figure 13.1(a), where as one variable increases, the other variable also increases. If we are given a new person, but we are only told their height (x_1), then we might wonder whether and hope that we can say something about their weight (x_2). Alternatively, if we are told their weight (x_2), we might hope that we can say something about their height (x_1). Of course, for any given person, our prediction might be wrong—any particular person might be very tall and thin or very short and fat. So, you should think of this as in a conditional probability or conditional expectation—given that we know their height (respectively, weight), what can we say about their weight (respectively, height). In particular, can we say something more informative than if we did not know their height (respectively, weight)?

In Figure 13.1(a), the two variables x_1 and x_2 are positively correlated, and so we should expect that we may be able to make such a prediction. Similarly, in Figure 13.1(b), the two variables are negatively correlated, and so we should expect that we may be able to make such a prediction. On the other hand, in Figure 13.1(c), there seems to be little relationship between the two variables, and so we should expect that it will be much more difficult to make such a prediction.

Since we are interested here in making predictions, we are going to change notation slightly. Instead of calling the variables x_1 and x_2 (which, recall, we did to emphasize that the equations we derived generalized from \mathbb{R}^2 to \mathbb{R}^n , simply by adding more subscripts), here we are going to call the variables x_1 and y . The reason is that, while in PCA both variables were “the same” in the sense that they were just variables in the data that we were trying to describe/visualize/understand, here we want to predict one variable from the other (or, more generally, from many others, which is why we are labeling the one variable x_1 rather than x , to emphasize that we can add more variables (x_2 , x_3 , x_4 , etc.) in which case the equations generalize, simply by adding more subscripts). Of course, this assumes that we want to predict x_2 given x_1 —if we want to predict x_1 from x_2 , then we would set $y \leftarrow x_1$ and $x_1 \leftarrow x_2$, and proceed as we do below.

This is a seemingly-minor point, but since with LS we are making a prediction, this introduces an asymmetry between variables that were on a similar footing when we discussed PCA. This leads to slightly different error metrics for LS versus PCA, and it highlights a difference between PCA and LS that we will get to below.



(a) Linear relationship offset from the origin. (b) Quadratic relationship. (c) Cubic relationship. (d) Still more complex relationship.

Figure 13.2: Illustration of more complex relationships in data.

More complicated relationships. Given this notation, Figure 13.2 shows several examples of somewhat more complicated relationships between y and x_1 . In Figure 13.2(a), it appears that there is a roughly linear relationship between y and x_1 , but one that potentially does not go through the origin:

$$y \approx ax_1 + b. \quad (13.1)$$

In Figure 13.2(b), it appears that there is a roughly quadratic relationship between y and x_1 :

$$y \approx ax_1^2 + b. \quad (13.2)$$

In Figure 13.2(c), it appears that there is a roughly cubic relationship between y and x_1 :

$$y \approx ax_1^3 + b. \quad (13.3)$$

Of course, there is no reason to think that real data have this simple of a relationship. We are just using this here for visual convenience. The real data might be more complicated, e.g., as illustrated in Figure 13.2(d). In this case, as a data scientist, we might hypothesize that y might depend on x_1 linearly as well as quadratically as well as cubically as well as have an affine offset:

$$y \approx b + a_1x_1 + a_2x_1^2 + a_3x_1^3. \quad (13.4)$$

Alternatively, there might be just two potentially-unrelated variables, x_1 and x_2 , in which case we might hypothesize that y might depend on x_1 linearly (if x_2 is fixed), and vice versa, in which case we obtain:

$$y \approx b + a_1x_1 + a_2x_2. \quad (13.5)$$

Alternatively, there might be just two variables x_1 and x_2 that “interact” in some way, in which case we might hypothesize that y might depend on x_1 and x_2 and x_1x_2 (to describe the interaction) as:

$$y \approx b + a_1x_1 + a_2x_2 + a_3x_1x_2. \quad (13.6)$$

More generally, we might just define some set of variables, x_1 , x_2 , and x_3 , and hypothesize that the relationship between y and $x \in \mathbb{R}^3$ is given by:

$$\begin{aligned} y &\approx b + a_1x_1 + a_2x_2 + a_3x_3 \\ &= b + \sum_{i=1}^3 a_i x_i \\ &= b + a^T x. \end{aligned} \quad (13.7)$$

In Eqn. (13.4), $x_i = x^i$, i.e., an i^{th} degree polynomial in x , but there could be other relationships. For example, there could be an exponential relationship or a sinusoidal relationship, as in:

$$y \approx b + a_1 \sin(x_1) + a_2 \exp(x_2). \quad (13.8)$$

Alternatively, the different x_i could just be different features, e.g., weight or height, income, number of phone calls made last month to a given zip code, the value of one’s most recent cholesterol test, numeric encoding of the second letter of one’s mothers’ maiden name, etc.

Linear models. Let's be clear about what is known and what is unknown in this modeling setup. The $y \in \mathbb{R}$ are known data, and all the $x_i \in \mathbb{R}^p$, where $p = 3$ here, are known data. That is, we are given many data points, each of which is of the form (y, x_i) , where $x_i \in \mathbb{R}^p$ is a vector of features and $y \in \mathbb{R}$ is a label corresponding to that feature vector. What is not known is the values of $b \in \mathbb{R}$ and $a \in \mathbb{R}^p$. Our goal will be to set up an objective function that will find the “best” such a and b . Then, if we are given a new data point, but only given its feature vector x_i and not the value of the label y for it, we will be able to predict a value of y for that new data point.

A very common way to achieve this is to work with a class of models known as *simple linear models*.

Definition 75 A simple linear model is a model of the form

$$y = b + a^T x + \varepsilon, \quad (13.9)$$

where x and ε are independent random variables, and $\varepsilon \sim N(0, \sigma^2)$, i.e., the distribution of ε has mean 0 and standard deviation σ . Here, y is called the response variable; x is called the predictor variable; and ε represents the measurement error.

Remark. Here is a very important point. This class of models is called simple linear models since there is a linear dependence on the unknowns b and a . The fact that x_2 might be quadratic or exponential or whatever in x_1 (or that there is any other connection between the elements of the feature vectors) is irrelevant. It could in fact be some completely different and unrelated feature. The goal is to predict y , and we are going to do so by taking a linear combination of feature vectors. This is as opposed to a model, e.g., such as

$$y = b + \sin(ax),$$

that is not linear in the unknowns a and b (and for which computing a and b would in general be much more difficult). Insofar as the model is concerned, y and x are fixed, and the unknowns are b and a , and those unknowns enter linearly. (Be careful about this—this is a common source of confusion.)

Remark. Here is another important point. There is a little gotcha, in that the simple linear model doesn't quite exhibit a linear relationship between y and the columns of x , since there is the constant affine offset given by b . Before, when we discussed linear algebra, we defined $y' = y - b$, and we said that this simply amounted to shifting the y axis, but we don't really want to do that here. The reason is that y is known, and b is unknown. Instead, it is more common to define a new vectors

$$\begin{aligned} x' &= \begin{pmatrix} 1 \\ x \end{pmatrix} \in \mathbb{R}^{p+1} \\ a' &= \begin{pmatrix} b \\ a \end{pmatrix} \in \mathbb{R}^{p+1}, \end{aligned}$$

in which case Eqn. (13.9) takes the form

$$y = a'^T x' + \varepsilon.$$

In this case, up to the noise ε , the relationship is linear. (Note that this is related to but a little different than adding an extra dimension, the elements of which were constructed to equal 1, when we were working with quadratic forms and we wanted to remove the linear and affine terms.) Note that this simply amounts to saying that every data point has 1 as its first entry, i.e., that the first element of the feature vector for every data point is 1, which is a perfectly legitimate modeling decision. (This too is a common source of confusion—some people obviously include it, and other people obviously exclude it.)

13.1.2 The basic LS method: simple linear regression

Suppose $n > k$ observations available, and let y_i , for $i = 1, \dots, n$ denote the i^{th} observed value (“response”), and let X_{ij} be the value of the j^{th} feature for the i^{th} data point.

Motivated by the way the problem was formulated in Definition 13.9, the general objective is the following. Given a list of paired observations,

$$\{x_i, y_i\}_{i=1}^n,$$

where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, estimate the parameters of the conditional mean formula

$$\mathbf{E}[Y|X = x] = b + a^T x.$$

This corresponds to the case of Definition 13.9.

To start, let's consider the case where $a \in \mathbb{R}^1$. Since there is only one variable (aside from the constant term), this is sometimes called *simple linear regression*.

Note that if ε has mean 0 and standard deviation σ , as is the case if $\varepsilon \sim N(0, \sigma^2)$, as in Definition 13.9, then the random variables defined as

$$Y_i = b + aX_i + \varepsilon$$

are also independent random variables, with $\mathbf{E}[Y_i] = b + ax$.

Thus, the random variables

$$Y_i - (aX_i + b),$$

which we want to be small, have mean 0. If we want $\{Y_i - (aX_i + b)\}_{i=1}^n$ to be small, then a natural next thing to consider is the variance.

So, what is the variance of this random variable?

The variance for this random variable is

$$\sigma_{(y-(ax+b))}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (ax_i + b))^2.$$

Let's define the error to be n times the variance

$$E(a, b) = \sum_{i=1}^n (y_i - (ax_i + b))^2. \quad (13.10)$$

and let's look for a and b that minimize this. (Note that multiplying by n changes the value at the solution by n , but it does not change the actual solution, i.e., the values of a and b that achieve the minimum.)

In multivariable calculus, to minimize a function such as this, we look for a and b such that

$$\frac{\partial E}{\partial a} = 0 \quad \text{and} \quad \frac{\partial E}{\partial b} = 0.$$

To do this, we get the following equations

$$\begin{aligned} 0 &= \frac{\partial E}{\partial a} = 0 = \sum_{i=1}^n 2(y_i - (ax_i + b))(-x_i) \\ 0 &= \frac{\partial E}{\partial b} = 0 = \sum_{i=1}^n 2(y_i - (ax_i + b))(-1). \end{aligned}$$

That is, we get

$$\begin{aligned} 0 &= \sum_{i=1}^n y_i x_i - a x_i^2 - b x_i \\ 0 &= \sum_{i=1}^n y_i - a x_i - b. \end{aligned}$$

Here, we have two equations for the two unknowns, a and b , and the two equations are linear in a and b . We could solve for a and b directly, but since we want to generalize our discussion to having more unknowns in the model, we want to formulate this as a matrix equation that can be generalized. So, let's write this as a matrix equation.

To do so, recall what the variables are—they are a and b , the x 's and y 's are the data that are fixed. So, we can write this as a matrix equation as

$$\begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{pmatrix}. \quad (13.11)$$

What does this mean? Well, it says that we want two parameters, and we are looking for a solution in a two dimensional space. If the matrix on the LHS is an invertible matrix, then we can invert it to find the solution a and b .

What insight can we get from the linear algebra we have been discussing? For simple 2×2 and 3×3 examples, the determinant was helpful to get some understanding, so let's look at the determinant.

$$\begin{aligned} \det \begin{vmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n 1 \end{vmatrix} &= \left(\sum_{i=1}^n x_i^2 \right) \left(\sum_{i=1}^n 1 \right) - \left(\sum_{i=1}^n x_i \right)^2 \\ &= n \left(\sum_{i=1}^n x_i^2 \right) - (n\bar{x})^2 \quad \text{where } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \\ &= n^2 \left(\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x} \right)^2. \end{aligned}$$

What this says is so long as the variance is non-zero, i.e., as long as the x_i are not all equal, then the inverse of the matrix exists and we can invert it and find the parameters a and b . In this case, a and b are given by

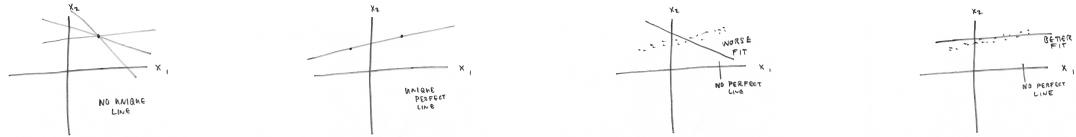
$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n 1 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{pmatrix}.$$

While it might be tedious for you to write this out by hand, you did simpler examples before, and it is straightforward for a computer to do.

Remark. As we have discussed, there are *much* better ways than computing the matrix inverse for solving an $n \times n$ system of linear equations that are larger than 2×2 . That is, while solving for a and b by computing a matrix inverse is fine for 2×2 matrices, other methods are much better in general. Those methods are the eigenvector-based methods we have discussed.

Before proceeding, let's ask when the variance is zero. The variance is zero when all the points are equal to the mean, i.e., when all the points are equal to each other.

- When variance is zero, i.e., when all the data are the same, one can't fit a *unique* line—which is what the two parameters a and b define—on the plane to a single point. See Figure 13.3(a).
- If there are two distinct points on the plane, i.e., if all of the data points lie on one of two distinct points, then one can fit a unique line to them. See Figure 13.3(b).
- If there are more than 2 unique points, which presumably is the typical case when we have more than 2 data points, then there is not any line that fits them all exactly. In this case, we try to find a line that approximates the data—and there are worse lines (see Figure 13.3(c)) and better lines (see Figure 13.3(d)) to do this. The problem we are considering is determining the “best” such line, where the best line is the line with parameters a and b determined by minimize the error in Eqn (13.10).



(a) Infinitely many solutions. (b) One unique solution. (c) No solution, and a worse approximate solution. (d) No solution, and a better approximate solution.

Figure 13.3: Illustration of fitting a line to data.

13.1.3 The basic LS method: multiple linear regression

More generally, we might assume a more complex model. For example, we might assume

$$Y = b + a_1x_1 + a_2x_2 + \cdots + a_px_p + \varepsilon.$$

This corresponds to the case of Definition 13.9, where $a \in \mathbb{R}^p$. In this case, we have

$$y = X\beta + \varepsilon$$

where $X \in \mathbb{R}^{n \times (p+1)}$ is a matrix that has an all-ones vector as its first column, $\beta \in \mathbb{R}^{p+1}$, and $y \in \mathbb{R}^n$.

As in the two-dimensional case, here we can define an error measure, the mean of which equals zero, and a variance measure, which we want to make small.

A subscript-chasing derivation. XXX. TO DO. TAKE FROM CHAPTER, PAGE 66 TO 67 AND 69 OF MONTGOMERY, PECK, VINING.

A simpler more linear algebraic derivation. XXX. TO DO. TAKE FROM CHAPTER, PAGE 68 TO 69 OF MONTGOMERY, PECK, VINING.

Least-squares objective. Regardless of which derivation one performs, one obtains the following.

XXX. TO DO. FIX BELOW IN LIGHT OF ABOVE. DO I WANT TO STATE THE OBJECTIVE OR SAY THAT THE MEAN IS FIXED AND WE WANT TO MINIMIZE THE VARIANCE.

To minimize this variance, we want to minimize

$$\sum_{i=1}^n (X\beta - y)_i^2 = \|X\beta - y\|_2^2.$$

This is the least-squares objective. It is the square of the Euclidean norm of the vector of residuals, meaning that it equals the sum of the square of the residual errors. We can take partial derivatives (as before, we won't go through the details, but we'll state the result) to get

$$X^T X \beta = X^T y.$$

In this case

$$\beta = (X^T X)^{-1} X^T y$$

and

$$\hat{y} = X\beta = X(X^T X)^{-1} X^T y.$$

Here, the $n \times n$ matrix $P_X = X(X^T X)^{-1} X^T$ is a projection matrix onto the span of the columns of X , the so-called Hat Matrix. Observe that if $X = U_X \Sigma_X V_X^T$ is the SVD of X , or if $X = Q_X R_X$ is a QR decomposition of X , then

$$\hat{y} = U_X U_X^T y = Q_X Q_X^T y.$$

While U and Q may be different matrices, they are both orthogonal matrices that span the same space (that is, the subspace of \mathbb{R}^n that is the span of the columns of X), and thus $U_X U_X^T = Q_X Q_X^T = P_X$, where P_X is a projection matrix onto the column span of X .

13.1.4 Complementary Perspectives on LS

Geometric aspects of LS. There is a very natural geometric interpretation to all of this. $X\beta$ is a vector that lies in the span ($\{X^i\}$), and recall that this span is a subspace. Then the solution \hat{y} is the “nearest” point in that subspace to the vector y . Thus, \hat{y} is simply the projection of y onto the span of the columns of X ; and the residual error is simply the projection of y onto the orthogonal complement of this; and due to the Pythagorean Theorem y can be expressed as the sum of the two vectors.

Algebraic aspects of LS. There is also a very natural algebraic interpretation to this. Assume that the columns of X are linearly independent. (What if not?) Then $\text{span}(X\beta)$ is a $(p+1)$ -dimensional subspace of \mathbb{R}^n . And $\text{span}(X\beta)^\perp$, which is the complement space, is an $n - (p+1)$ -dimensional subspace of \mathbb{R}^n . But we know that any vector $y \in \mathbb{R}^n$ can be decomposed into

$$y = y_1 + y_2$$

where $y_1 \in \text{span}(X\beta)$ and $y_2 \in \text{span}(X\beta)^\perp$. Since y_1 and y_2 are in orthogonal subspaces, we have that $y_1^T y_2 = 0$. In this case, the solution $\hat{y} = y_1$.

Statistical aspects of LS. This LS computation has a well-studied statistical interpretation. The statistical interpretation is that it is minimizing an error measure that is optimal for a certain class of hypothesized models. This perspective provides a way to construct many common generalizations of the basic LS procedure.

Mean. Here is the main result on the mean of the LS estimator.

Lemma 6 *The LS estimator has expectation $\mathbf{E}[\hat{\beta}] = \beta$.*

Proof:

$$\begin{aligned} \mathbf{E}[\hat{\beta}] &= \mathbf{E}[(X^T X)^{-1} X^T y] \\ &= \mathbf{E}[(X^T X)^{-1} X^T (X\beta + \epsilon)] \\ &= \mathbf{E}[(X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T X\epsilon] \\ &= \mathbf{E}[(X^T X)^{-1} X^T X\beta] + \mathbf{E}[(X^T X)^{-1} X^T X\epsilon] \\ &= \mathbf{E}[\beta] + (X^T X)^{-1} X^T X \mathbf{E}[\epsilon] \\ &= \beta. \end{aligned}$$

◇

What Lemma 6 says is that the LS estimator is unbiased.

Variance. Here is the main result on the mean of the LS estimator.

Lemma 7 *The LS estimator has variance $\mathbf{Var}[\hat{\beta}] = \sigma^2 (X^T X)^{-1}$.*

Proof: We know that

$$\mathbf{Var}[\hat{\beta}] = \mathbf{E} \left[(\hat{\beta} - \mathbf{E}[\hat{\beta}]) (\hat{\beta} - \mathbf{E}[\hat{\beta}])^T \right],$$

but let's not compute it directly. Instead, let's take advantage of the fact that

$$\hat{\beta} = (X^T X)^{-1} X^T y,$$

and compute it as follows.

$$\begin{aligned} \mathbf{Var}[\hat{\beta}] &= \mathbf{Var}[(X^T X)^{-1} X^T y] \\ &= (X^T X)^{-1} X^T \mathbf{Var}[y] ((X^T X)^{-1} X^T)^T \\ &= (X^T X)^{-1} X^T \mathbf{Var}[y] X^T (X^T X)^{-1} X \\ &= (X^T X)^{-1} X^T \Sigma^2 I X^T (X^T X)^{-1} X \\ &= \sigma^2 (X^T X)^{-1} X^T X^T (X^T X)^{-1} X \\ &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

◇

What Lemma 7 says is that if we let $C = (X^T X)^{-1}$, then the variance of the i^{th} element of β is

$$\mathbf{Var}[\hat{\beta}_i] = \sigma^2 C_{ii}$$

and the covariance between the i^{th} and j^{th} elements of β is

$$\mathbf{Cov}[\hat{\beta}_i, \hat{\beta}_j] = \sigma^2 C_{ij}.$$

Best linear unbiased estimator. Given that we know the expectation and variance of the LS estimator, it is natural to ask whether we can come up with a better estimator. That is, if we don't care about the LS regression problem per se, and instead if our goal is to estimate the vector β with any way possible, then is there some other procedure, perhaps using methods that go beyond basic linear algebra, that is better in some sense. It turns out that the answer is yes or no, depending on how broad a class of estimators you want to consider.

- **Yes.** If you want to consider nonlinear estimators that are biased, then the answer is yes. (This is more advanced than we can cover in this class, but it is good to know.)
- **No.** If you restrict yourself to linear estimators that are unbiased, then the answer is no. (This is the vast-majority use case, and it is the method upon which all sorts of other methods build.)

The more precise version of the latter claim is that the LS estimator is the BLUE (Best Linear Unbiased Estimator) for the vector β . XXX. FILL IN DETAILS FOR THIS, PROBABLY FROM THAT LINEAR REGRESSION BOOK I USED IN S18, PAGE 558 TO 560 OF MONTGOMERY, PECK, VINING.

Algorithmic aspects of LS. Although we have not emphasized it, the LS computation comes with well-understood algorithmic results. In particular, if there are n constraints and p variables, then the running time is something like $O(np^2)$.

- **SVD.** Given A , it takes $O(np^2)$ time to compute U , Σ , and V .

- **QR.** Given A , it takes $O(np^2)$ time to compute Q and R .
- **Normal equations.** Given A , it takes $O(np^2)$ time to compute $A^T A$.

The constants, numerical reliability, etc. differ for these methods. This is hidden from you as a user if you simply call a routine, but it is an important topic in computer science.

(Here, “something like” means for direct methods, since the running time of iterative methods depends on the number of iterations, etc., which depends on the condition number and other things.)

13.1.5 When is LS the “right thing” to do?

General considerations. XXX. TO DO. WHAT TO DO HERE, IS MLE THE ONLY SENSE IN WHICH IT IS THE RIGHT THING TO DO.

Maximum-Likelihood Estimation. XXX. TO DO. FILL IN DETAILS FOR THIS, FROM PAGE 79 OF OF MONTGOMERY, PECK, VINING.

13.2 Comparison of PCA and LS

There are strong connections between PCA and LS—in particular, some of the algorithmic techniques used to solve each of them are similar, and some of the statistical assumptions that underlie each of them are similar—but they do solve different problems. Informally, PCA tries to draw a line (more generally, a subspace) through the data that is best at capturing the variability in the data (either maximizing the projected variability or minimizing the residual reconstruction error), irrespective of any labels or prediction task; and LS tries to draw a line (the so-called regression line, or more generally, a subspace) through the data that is best at minimizing the error of predicting some set of variables from another set of variables. Somewhat more formally, PCA looks at the data and decomposes it into orthogonal components of decreasing variance, while LS is a regression problem and tries to find parameters to predict one column vector from a bunch of other column vectors. The two methods can be confused, e.g., since we draw similar pictures for each, but the pictures mean slightly different things, and it is important to understand the differences.

To illustrate this, let’s say that the data are a bunch of points on \mathbb{R}^2 , as illustrated in Figure 13.4.

- If we want to do PCA, then we want to find a subspace (a line here, through the origin) that is the “best” in the sense that when the data are orthogonally projected to that subspace the variance of the data is maximized. In this case, both x_1 and x_2 are known, and both are considered in the error measure. This is illustrated in Figure 13.4(a).
- If we want to do LS, then we may want to predict x_2 from x_1 . In this case, we project the data onto the span of x_1 , and the error we consider and want to minimize is that incurred by x_2 . This is illustrated in Figure 13.4(b).
- Alternatively, we can do LS to predict x_1 from x_2 . In this case, we project the data onto the span of x_2 , and the error we consider and want to minimize is that incurred by x_1 . This is illustrated in Figure 13.4(c).

That is, in LS regression, we are viewing this as $Y = X\beta + \varepsilon$, and then the Y axis is something we are trying to predict, and so we project down to the nearest point on the axes orthogonal to it (so, in this simple case, we have 1 dimension of data and 1 dimension of thing we are trying to predict). On the other hand, PCA just has a bunch of data (2-dimensional in this simple case), and we aren’t trying to predict anything, but instead we are trying to minimize the variance of all the coordinates for all of the data.

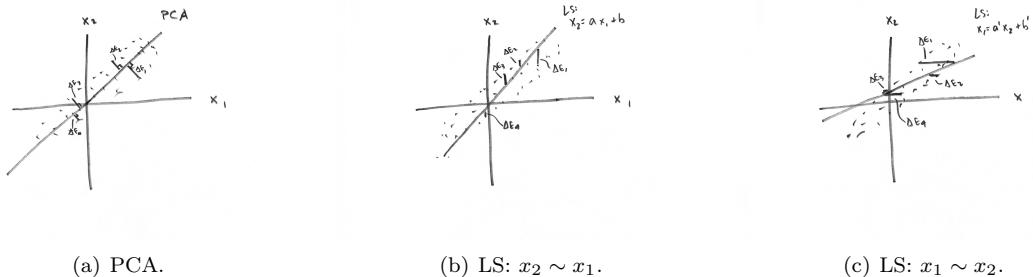


Figure 13.4: Illustration of differences between PCA and LS.

13.3 Regression Diagnostics and Related Methods

Regression diagnostics refers to things that one can compute to determine whether the regression computation makes sense. To start, note that when we compute a sample average, then each observation has the same weight in determining the outcome. In regression, this is not the case, and determining which data points exhibit a particularly large effect on the fit is a large and important topic. Here, we provide just a few examples.

- **Leverage.** Consider Figure 13.5(a). The point labeled A in this figure is far from the other points in x space, but it lies on the regression line/plane. This is an example of a *high leverage* data point. If it has an unusual x value, and if it does not affect the estimates of the regression coefficients, but it may control certain model properties as well as model summary statistics, and if it varies slightly (e.g., if it had been measured with some noise) then it could have a large effect on the fit. The diagonal elements of the projection matrix onto the span of A , i.e.,

$$(P_A)_{ii} = (A(A^T A)^\dagger A^T)_{ii} = (QQ^T)_{ii}$$

are a common measure of leverage. Recall that $H = P_A$ determined the variances and covariances of \hat{y} and e , since

$$\begin{aligned}\mathbf{Var} [\hat{y}] &= \sigma^2 H \\ \mathbf{Var} [e] &= \sigma^2 (I - H).\end{aligned}$$

The elements $H_{ii} = (P_A)_{ii}$ can be interpreted as the amount of leverage exerted by the i^{th} observation y_i on the j^{th} fitted value \hat{y}_j .

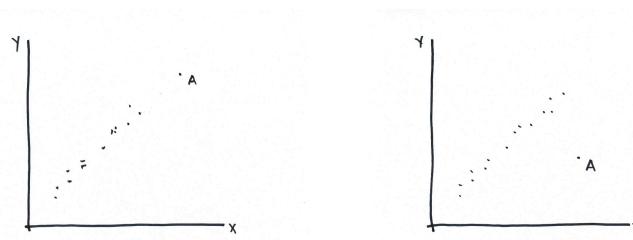
- **Influence.** Consider Figure 13.5(b). The point labeled A in this figure has somewhat unusual x value, but it has an unusual y value as well. This is an example of a *high influence* data point. It has a noticeable impact on the model coefficients in that it “pulls” the regression model in its direction. It is far from the other points in x space, but it lies on the regression line/plane. Cook’s distance and other related metrics are used to quantify this.

13.4 Regularized LS Regression

In some cases, there are more features than data points. Formally, this means that one is solving the problem

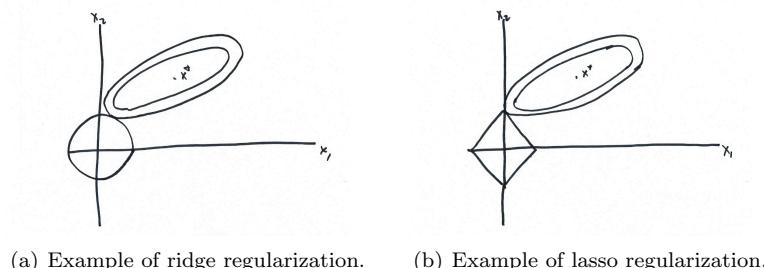
$$\min_{x \in \mathbb{R}^n} \|X\beta - y\|_2^2, \quad (13.12)$$

where $X \in \mathbb{R}^{n \times p}$, where $p > n$. We know that there is a solution, $\beta = X^+y$, but it is not unique. Regularization provides a way to enforce uniqueness. More generally, regularization provides a way to make the model more robust or less sensitive to the data. It is a large area, and here we provide just two examples.



(a) Example of a high leverage data point.
(b) Example of a high influence data point.

Figure 13.5: Examples of leverage and influence in LS regression.



(a) Example of ridge regularization.
(b) Example of lasso regularization.

Figure 13.6: Examples of regularized LS regression.

- **Ridge/Tikhonov regularization.** Here, the LS problem Eqn. (13.12) is modified to be

$$\min_{x \in \mathbb{R}^n} \|X\beta - y\|_2^2 + \lambda \|x\|_2^2,$$

for some $\lambda \in \mathbb{R}^+$. See Figure 13.6(a) for an illustration.

- **Lasso regularization.** Here, the LS problem Eqn. (13.12) is modified to be

$$\min_{x \in \mathbb{R}^n} \|X\beta - y\|_2 + \lambda \|x\|_1,$$

for some $\lambda \in \mathbb{R}^+$. See Figure 13.6(b) for an illustration.

In both of these cases, the LS objective is modified by adding a term of the form $\lambda g(x)$, where $g(x)$ is a norm (or norm squared) of the vector x . Informally, the idea is that one wants to keep the elements of x from being extremely large and “cancelling out” in ways that are not robust to noise in the data. Relatedly, one wants to consider both the “data fit” term ($\|X\beta - y\|_2$) as well as the “model complexity” term ($\lambda g(x)$) in the objective to be optimized. The form of $g(x)$ provides a bias toward solutions of different types—informally, the ridge regularization provides a bias to solutions with a lot of small entries, and lasso regularization provides a bias toward solutions that have a few large entries. Depending on the application, one or the other of these biases may be preferable.

13.5 Problems

13.5.1 Implementations and Applications of the Theory

1. XXX.
2. XXX.

13.5.2 Pencil-and-paper Problems

1. Let

$$A = \begin{pmatrix} -1 & -1 \\ -1 & 1 \\ 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} -2 \\ -2 \\ 3 \end{pmatrix}, \quad c = \begin{pmatrix} 2 \\ 3 \end{pmatrix}.$$

- (a) Find the set of all $x \in \mathbb{R}^2$ minimizing $\|Ax - b\|_2$.
- (b) Find the set of all $x \in \mathbb{R}^3$ minimizing $\|A^T x - c\|_2$.
- (c) In the second case, find the vector in this set with minimum Euclidean norm.

2. (LS) Consider the vector $b \in \mathbb{R}^m$. We would like to project this onto the line/subspace through the all-ones vector $a \in \mathbb{R}^m$, and we would like to understand this in terms of least squares. To do so, let's solve the m equations $ax = b$ in one unknown $x \in \mathbb{R}$ by least squares.

- (a) Solve $a^T a \hat{x} = a^T b$ to show that the solution \hat{x} is the mean, i.e., the average, of the elements of b .
- (b) Find $e = b - a\hat{x}$, and from this find the variance $\|e\|_2^2$ and the standard deviation $\|e\|_2$.
- 3. (LS) In our discussion of least squares, it was not essential that we considered $y = ax + b$. Instead, we could have considered

$$y = af(x) + bg(x).$$

In this case, we are trying to predict y in terms of a linear combination of two functions of x , rather than in terms of a linear combination of x itself and a constant offset.

- (a) Show that the arguments in the chapter proceed analogously for this case, and that in this case we obtain

$$\begin{pmatrix} \sum_{i=1}^n f(x_i)^2 & \sum_{i=1}^n f(x_i)g(x_i) \\ \sum_{i=1}^n f(x_i)g(x_i) & \sum_{i=1}^n g(x_i)^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n f(x_i)y_i \\ \sum_{i=1}^n g(x_i)y_i \end{pmatrix}.$$

- (b) Recall that, for $y = ax + b$, by considering the determinant condition for invertibility, we saw that the matrix equation was invertible when not all the data points were identical. In this case, for $y = af(x) + bg(x)$, what are the conditions under which the matrix is invertible?

4. XXX.

5. XXX. FROM MOORE-MCCABE-CRAIG. Use the equation for the least-squares regression line to show that this line always passes through the point (\bar{x}, \bar{y}) . XXX. MODIFY SOMEHOW.

Chapter 14

Systems of Linear Equations

The phrase “system of equations” means that we are dealing with a set or collection of linear equations all together at once. Since matrices can be used to encode information about linear transformations, solving systems of linear equations is an important application of the linear algebra methods we have been discussing.

14.1 Simple Example of System Linear Equations

The simplest non-trivial example of a system of linear equations is provided by two equations in two unknown variables. Let’s say we have

$$\begin{aligned} e &= ax + by \\ f &= cx + dy \end{aligned}$$

Depending on the values of a , b , c , d , e , and f , we have one of the three situations illustrated in Figure ???. In the notation we have been using, and with the obvious change-of-variable-names, this system of linear equations can be written as a 2×2 matrix equation:

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} e \\ f \end{pmatrix} = \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

or more compactly as

$$b = Ax.$$

Once we have it written as a matrix equation, we can think of it as a matrix equation and ask relatively-seamlessly what happens if the matrix is larger, has certain special structure, etc. Solving this problem when A is 10×10 or $10^2 \times 10^2$ or $10^{10} \times 10^{10}$, or when A has certain special types of structure, and so on, is an important problem in data science and beyond. Thinking about this problem in terms of the 3 examples that arise in the 2×2 example illustrated in Figure ?? is more limiting than thinking about it in terms of the linear algebraic ideas of linear combinations, spans, bases, and linear dependence/independence that we have been discussing. Of course, the more general ideas specialize to what we know in the 2×2 example.

14.2 Basics Ideas Underlying of Linear Equations

Here, we are interested in solving systems of linear equations, but it’s important to see how this fits into what we have been discussing, and in particular into LS regression.

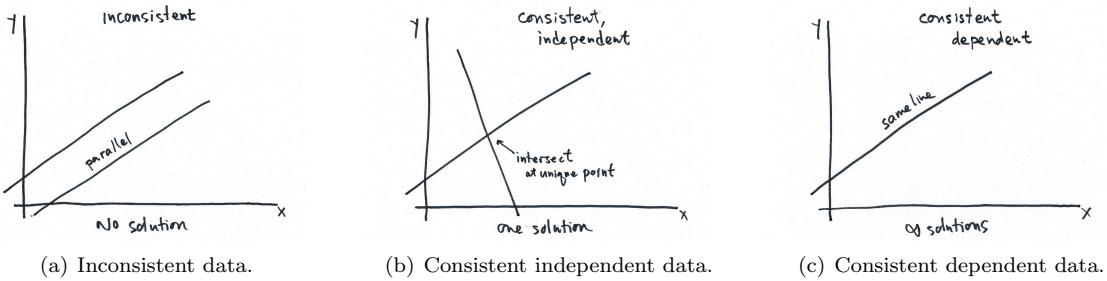


Figure 14.1: Illustration of 2×2 system of linear equations, with 0, 1, or an ∞ number of solutions, depending on the equations.

A special case of the general LS regression problem is when $m = n$ and the matrix is full rank. In that case, the LS problem

$$\min \|Ax - b\|_2^2$$

has a unique solution, and that solution has zero residual error. By this, we mean that there is a single unique vector x that can be used to construct a linear combination of the columns of A that equals the vector b . In this case, one is asking to find the vector x such that

$$Ax = b.$$

This is the basic problem of finding a solution to a system of linear equations, which is an important application of the linear algebra ideas we have been discussing. There are many ways to solve this problem, and this is a huge topic, and we will just scratch the surface.

To start, observe that one can compute the SVD of A as $A = U\Sigma V^T$, where U and V are $n \times n$ orthogonal matrices, and where Σ is a diagonal matrix with positive (i.e., not just non-negative) entries along the diagonal, in which case

$$x^* = A^\dagger b = V\Sigma^{-1}U^T b.$$

But, we know that

$$U\Sigma V^T V\Sigma^{-1}U^T = I = V\Sigma^{-1}U^T U\Sigma V^T,$$

and so in this case A^\dagger is both a left inverse and a right inverse of A and so it is *the* inverse, i.e.,

$$A^\dagger = A^{-1}$$

is the inverse of A . Thus, we can say that

$$x^* = A^{-1}b.$$

Of course, the fact that we can write down this expression does not mean that we should compute A^{-1} , either with the SVD or with some other method.

See Figure 14.2 for an illustration of this relatively-simple situation. By now it should be clear that this intuitive situation is more the exception than the rule and that a lot of subtle things can happen to mess it up.

There are many ways to compute the vector x^* , and here are just a few.

- The most obvious might be to compute A^{-1} and then compute $A^{-1}b$. It is always almost a very bad idea to do that, either since it is unnecessary (i.e., wasteful of compute time) or since algorithms to compute it do not do well in the presence of roundoff error on a computer (i.e., not just wasteful, the answer that you get may be meaningless). In case you missed that, let's say it again:
 - *It is always almost a very bad idea to try to solve $Ax = b$ by computing A^{-1} and then computing $A^{-1}b$.*

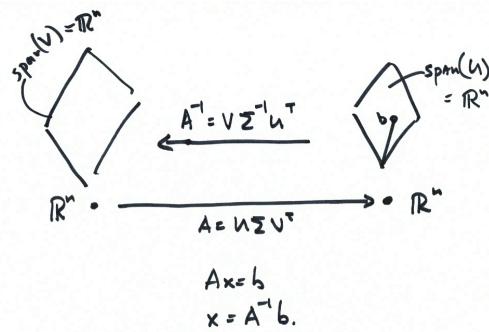


Figure 14.2: Four fundamental subspaces for an invertible $n \times n$ matrix.

- Do **not** do that. You might be tempted to do that, since python/R/etc. may have a one-line command to do it. It may work on 2×2 problems and 3×3 problems, but it will fail at some point when it matters, and you will have no idea why.
- You can compute it with the SVD, i.e., compute U , Σ , and V , and then compute $V\Sigma^{-1}U^T b$. That has better numerical properties, but it may still be overkill, i.e., wasteful in terms of compute time.
- You can compute it a QR decomposition. For example, if $A = QR$, then premultiply $Ax = b$ with Q^T to get $Rx = Q^T b = b'$, and then solve this upper triangular system.
- You can compute it via something called Reduced Row Echelon Form.
- Et cetera.

Many of these methods use ideas we have been discussing, but we won't go through them in detail.

Here is simple way to think about solving systems of linear equations: it is a special case of LS regression, where $m = n$. To solve it, note that:

- If the matrix A is diagonal and all the elements are non-zero, then $x = A^{-1}b$, where A^{-1} is easily computed as the matrix in which the non-zero diagonal elements are inverted. In this case, the solution is unique.
- If the matrix A is diagonal and some of the elements are zero, then A does not have an inverse, but it has a generalized inverse, A^\dagger , which is easily computed as the matrix in which the non-zero diagonal elements are inverted and the zero diagonal elements are left unchanged. In this case, the solution is not unique, since x is zeroed-out along the zero directions of A , and so it can take any value along those directions.
- If $A = QR$, where Q is square and R is upper triangular, with non-zeros along the diagonal, then $Ax = b$ is equivalent to $Rx = Q^T b$. The solution is easy to find by back-substitution. In this case, the solution is unique.
- If $A = QR$, where R has some zeros along the diagonal, then those directions are arbitrary, then one has to introduce parameters along those zero directions. In this case, the solution is not unique, since x is zeroed-out along those directions; but, given those parameters, the solution is easy to find by back-substitution, and the solution is given by a subspace of dimension equal to the number of parameters.

This is not the best way to compute the solution in practice, but it does solve it and give the correct solution, and it is roughly what the usual reduced row echelon form methods does.

XXX IS THIS STALE OR WHERE TO PUT

XXX. WHAT TO DO WITH THE FOLLOWING. Given an $m \times n$ vector A and a m -vector b , recall the basic least squares problem

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2.$$

If the SVD of A is $A = U\Sigma V^T$, then the solution to this is

$$x_{opt} = V\Sigma^{-1}U^T b.$$

Alternatively, if $A = QR$ is a QR decomposition of A , then one could pre-multiply by Q^T and show that this is equivalent to

$$Rx = Q^T Qx = Q^T b,$$

which can then be easily solved since R is upper-triangular. Alternatively, one could pre-multiply by A^T , in which case one gets $A^T Ax = a^T b$. If $A^T A$ is full rank one could invert it to obtain

$$x_{opt} = (A^T A)^{-1} a^T b,$$

which amounts to saying that

$$b_{opt} = A (A^T A)^{-1} a^T b = U^T U b,$$

which is simply the part of b in the span of the columns of A . If $A^T A$ is not full rank, then one can use the generalized inverse, and it still holds that

$$b_{opt} = U^T U b,$$

except here one is projecting onto a subspace of dimension less than n .

14.3 Some Mechanical Procedures to Solve Linear Equations

XXX. TO DO.

14.4 Direct Methods to Solve Linear Equations

XXX. TO DO.

14.5 Iterative Methods to Solve Linear Equations

XXX. TO DO.

14.6 Numerical Issues

XXX. TO DO.

14.7 Problems

14.7.1 Implementations and Applications of the Theory

1. XXX.
2. XXX.

14.7.2 Pencil-and-paper Problems

XXX. HW PROBLEMS.

Chapter 15

PageRank for Ranking, Clustering, and Classifying

Spectral ranking is the umbrella name for a general class of techniques that applies the theory of linear maps (in particular, eigenvalues and eigenvectors) to matrices that do not obviously represent geometric transformations, but instead that represent some other kind of relationship between entities. Since data are often represented by matrices for which there is not an *obvious* geometric structure, and since one often wants to rank things from “best” to “worst,” spectral ranking is an important topic in data science.

Spectral ranking is probably most well known via the PageRank procedure, which became well-known in the early days of the internet as a method to rank the importance of web pages, but the area has a much longer history, and many other variants exist. These methods have important connections to both probability and linear algebra, and they can be used to solve many other problems like classification and clustering that do not obviously involve ranking.

15.1 Random Walks, Diffusions, Markov Chains, and Other Approaches to PageRank

While spectral ranking methods are very general, to focus attention, let’s consider the particular case of PageRank applied to ranking web pages. The motivation for PageRank is that we want to find a way to determine important web pages that does not involve the terms on that web page and that does not simply count the number of links to/from that web page. Informally, the idea behind PageRank is that important pages are pages that are pointed to by important pages. This circular definition suggests an iterative process—and it is here that the interesting connections with linear algebra and probability arise. The iterative process can be interpreted as what is known in probability as a random walk or diffusion or Markov chain; and it can also be interpreted as an eigenvector-eigenvalue equation or as a system of linear equations such as arise in linear algebra.

The general idea behind PageRank is that the rank (or “page rank”) of a web page, call it $r(i)$ for the i^{th} page, where $i = 1, \dots, n$, where n is the number of web pages, is

$$r(i) = \sum_{j:j \text{ backlinks to } i} \frac{r(j)}{\text{outdegree}(j)}.$$

Notice that the PageRank of the linking page j , i.e., $r(j)$, is tempered/downweighted by the overall number of links that page j makes, i.e., if a page j has a large number of outlinks to other pages, then the weight of its “recommendation” as to the importance of page i is less. The problem here of course is that the values

$r(j)$ inside the sum are unknown.

A possible solution is to try to determine them via an iterative procedure. For example, at time step $t = 0$, start with an initial vector, $r_{t=0}$, e.g., the all-ones vector, and let

$$r_{t+1}(i) = \sum_{j:j \text{ backlinks to } i} \frac{r_t(j)}{\text{outdegree}(j)}.$$

This can be written more efficiently, e.g., with fewer subscripts, etc., as

$$\pi_{t+1} = W\pi_t,$$

where W is a matrix (the random walk matrix) defined below, and where $\pi = r$ is the vector of PageRank values. A few observations.

- Each iteration requires one matrix-vector multiplication.
- W is a very sparse matrix, meaning that most of the entries of W are equal to 0, because most webpages link to only a small number of other webpages.
- This iterative procedure is a simple linear stationary process that is widely-studied in numerical linear algebra. Indeed, it is the power method applied to W .
- W looks something like a stochastic transition matrix for a markov chain. We say “something like” since, in some cases, there may be so-called dangling nodes, meaning nodes with no outlinks, in which case there is an all-zeros column in the matrix. In these cases, W is called a substochastic matrix.

Let's go into more detail on the iterative process. It is a process, so it can easily be implemented, but here are some questions that arise.

- Will this iterative process continue indefinitely, or will it converge to some value, or under what properties of W is it guaranteed to converge?
- If it does converge, does it converge to something meaningful in terms of the motivating problem?
- If it does converge, does the convergence depend on the initial vector $r_{t=0}$?
- If it does converge, how long will it take, i.e., how many iterations are needed, until it converges?

The answer to several of these questions depends on results from Markov chain theory, and we will get to some of these issues below. Before that, however, let's describe a few solutions to particular problems that arose in the web application.

- **So-called sink nodes.** Some nodes may have no out-links, and for those there will be a all-zeros column in W . A solution is to replace that column by a vector with uniform entries, i.e., with $\frac{1}{n}e$, where e is the all-ones column vector. To write this in matrix notation, for all such nodes, if we let $a \in \mathbb{R}^n$, where $a_i = 1$ if page i is a dangling node and 0 otherwise, then we can replace

$$W \rightarrow W' = W + \frac{1}{n}ea^T.$$

- **So-called random surfer.** Replace W with

$$W' \rightarrow W'' = \alpha W' + (1 - \alpha)\frac{1}{n}ee^T,$$

where α is a scalar between 0 and 1.

In the lingo that we describe below, these adjustments make W a stochastic matrix that is both irreducible and aperiodic. Also, note that W'' is a dense matrix, but it is a sparse matrix plus a rank-one update. This makes many things, including doing matrix-vector multiplications, easier than if it were an arbitrary dense.

There are two ways to view this PageRank problem, which we constructed from random walks.

- As an eigenvector problem:

$$W''\pi = \pi \quad \text{s.t.} \quad e^T\pi = 1.$$

(Observe that this is of the form $Ax = \lambda x$.)

- As a system of linear equations:

$$(I - W'')\pi = 0 \quad \text{s.t.} \quad e^T\pi = 1.$$

(Observe that this is of the form $Ax = b$.)

Note that the constraint that $e^T\pi = 1$ in each of the perspectives ensures that the solution π is a probability vector. From the eigenvector problem perspective, the goal is to find the normalized dominant right-hand eigenvector of W'' corresponding to the dominant eigenvalue $\lambda_1 = 1$ (the top eigenvalue equals 1, since W'' is a stochastic matrix). From the linear systems perspective, the goal is to solve the system of linear equations with all-zeros right hand side. In this case, the two perspectives lead to the same PageRank vector.

15.2 Graphs and representing graphs as matrices

In light of what we now know about eigenvalues and eigenvectors, let's go back and revisit graphs, and let's see what this says about the PageRank motivation.

See Figure 15.1(a), which illustrates a simple graph, consisting of nodes and edges, and recall that we can define several types of matrices for this graph. In particular, we can define

- the *Adjacency Matrix* (where the (i, j) element is 1 or 0, depending on whether or not there is an edge to node i from node j in the graph, i.e., whether the two nodes, taking into account directedness, are adjacent),

$$A = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}, \quad (15.1)$$

- and the *Diagonal Degree Matrix* (which has 0 entries everywhere, except along the diagonal, where it has the total number of edges going out of the i^{th} node),

$$D_{out} = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}, \quad (15.2)$$

- and the *Random Walk Matrix*

$$W = AD_{out}^{-1} = \begin{pmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1 & 0 & 0 \end{pmatrix}. \quad (15.3)$$

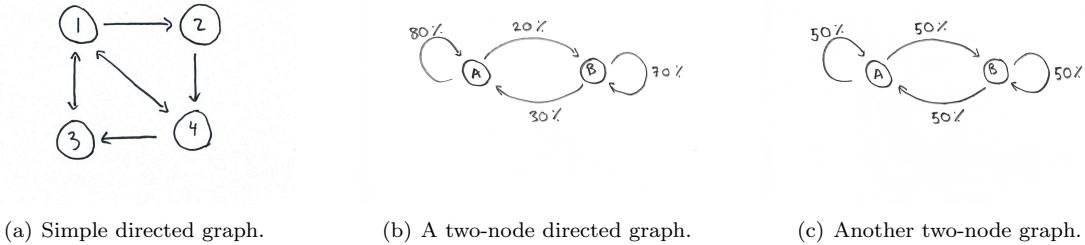


Figure 15.1: Illustration of several simple graphs.

More generally, here are the definitions of these matrices. For these definitions, by a graph $G = (V, E)$, we mean a set of *things* (called nodes/vertices, denoted by $V = \{1, \dots, n\}$) and a set of *directed pairs of things* (called edges, denoted $E = (i, j) \in (V \times V)$, for $i, j \in V$, where (ij) means to node i from node j).

Definition 76 Given a graph, $G = (V, E)$, the *Adjacency Matrix* $A \in \mathbb{R}^{n \times n}$ is

$$A_{ij} = \begin{cases} 1 & \text{if } (ij) \in E \\ 0 & \text{otherwise.} \end{cases}$$

Definition 77 Given a graph, $G = (V, E)$, the *Diagonal Degree Matrix* $D \in \mathbb{R}^{n \times n}$ is

$$D_{ii} = \begin{cases} d_i & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

Definition 78 Given a graph, $G = (V, E)$, let A be the Adjacency Matrix, and let D be the Diagonal Degree Matrix, in which case D^{-1} is the diagonal matrix with $(D^{-1})_{ii} = \frac{1}{d_i}$, for all $i \in [n]$. In this case, the Random Walk Matrix $W \in \mathbb{R}^{n \times n}$ is

$$W = AD^{-1}.$$

The Random Walk Matrix is a matrix, and thus we can perform a matrix-vector multiplication with any vector $x \in \mathbb{R}^n$, i.e., Wx is defined for all $x \in \mathbb{R}^n$, where n is the number of nodes in the graph G . It is called a Random Walk Matrix since we get particularly interesting results if we apply W to a vector x that defines a probability distribution, i.e., a vector x that has nonnegative entries that sum to one.

To see this, observe that if x is a probability distribution, then $y = Wx$ can be interpreted as applying the random walk W associated with $G = (V, E)$ for one step. In this case, the vector y is also a probability vector. The reason for this property is that the columns of W sum to one. Any matrix for which this is true is called a *stochastic matrix*.

Remark. We have said that W is a stochastic matrix if the columns of W sum to one, and we are post-multiplying W by the probability vector. The reason for this is that we are viewing column vectors as the basic thing of interest here, i.e., probability vectors are column vectors, and the Random Walk Matrix transforms one column vector into another column vector. Some books and some areas consider row vectors to be the basic thing of interest, in which case one pre-multiplies W to do one step of a random walk, in which case stochastic matrices are defined to be matrices whose rows sum to one. As with many things in this area, this is not standardized, and some people/areas think one approach or the other is obviously the right one. Yet another potential gotcha to keep in mind.

Additional examples. Before proceeding, let's consider two other stochastic matrices we have seen before.

- First, consider the matrix

$$W = \begin{pmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{pmatrix}. \quad (15.4)$$

While we could just view this W as a matrix, we can also interpret this matrix as a Random Walk Matrix for the graph which is illustrated pictorially in Figure 15.1(b). The figure illustrates that this matrix corresponds to a graph with two nodes: for one node, there is an 80% chance that the random walker stays on that node and a 20% chance that the random walker goes to the other node; and for the other node, there is a 70% chance that the random walker stays on that node and a 30% chance that the random walker goes to the other node.

In particular, if we start with all the probability mass on the first node, corresponding to starting with the vector $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, then after one step of the random walk we have

$$\begin{pmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{pmatrix} \begin{pmatrix} 1.0 \\ 0.0 \end{pmatrix} = \begin{pmatrix} 0.8 \\ 0.2 \end{pmatrix},$$

which clearly is another probability distribution. In addition, if we start with the vector $\begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix}$, i.e., where the probability mass distributed in a 60-40 ratio between the two nodes, then after one step of the random walk we have

$$\begin{pmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{pmatrix} \begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix} = \begin{pmatrix} 0.48 + 0.12 \\ 0.12 + 0.28 \end{pmatrix} = \begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix}, \quad (15.5)$$

i.e., we still have a probability distribution, and the probability distribution is unchanged by W .

Here are two complementary perspectives on this.

- From the perspective of probability, the matrix W transforms probability distributions into probability distributions; and what Eqn. (15.5) says is that if we are in the state $\begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix}$, then one step of W leaves the state unchanged.
- From the perspective of linear algebra, what Eqn. (15.5) says is that $\lambda_1 = 1$ is an eigenvalue of the stochastic matrix W in Eqn. (15.4), and $v_{\lambda_1} = \begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix}$. It is also easy to verify that $\lambda_2 = 0.5$ is also an eigenvalue of W , with associated eigenvector is $v_{\lambda_2} = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix}$, in the sense that $Wv_{\lambda_2} = \lambda_2 v_{\lambda_2}$.

(Recall that eigenvectors were defined for any square matrix. It was then for symmetric matrices that we obtained all the nice results we discussed, but some other classes of matrices, e.g., stochastic matrices, also lead to many nice properties.)

There are two things to note about this linear algebra perspective.

- This vector v_{λ_1} was not unit-normalized with respect to the more common ℓ_2 norm. We could have done that, but when working with matrices related to random walks, it is sometimes more convenient to normalize with respect to the ℓ_1 norm. That's fine, since eigenvectors are unique only up to their subspace. The reason for the ℓ_1 normalization is that it makes the probability interpretation of v_{λ_1} clearer. Either is okay—just keep track of what you do.
- The two eigenvectors of W are not orthogonal, i.e., $v_{\lambda_2}^T v_{\lambda_1} \neq 0$. This does not violate our previous results, since W is not a symmetric matrix. The matrix W is, however, related to a symmetric matrix, and we will get to this below.

- Second, consider the matrix

$$W = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \quad (15.6)$$

which is illustrated pictorially in Figure 15.1(c). The figure illustrates that this matrix corresponds to a graph with two nodes: for each node, there is a 50% chance that the random walker stays on that node and a 50% chance that the random walker goes to the other node.

Here, if we start with any probability distribution, $\begin{pmatrix} \alpha \\ 1 - \alpha \end{pmatrix}$, with $\alpha \in [0, 1]$, then

$$\begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} \begin{pmatrix} \alpha \\ 1 - \alpha \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}. \quad (15.7)$$

Thus, $\lambda_1 = 1$ is an eigenvalue, and the associated eigenvector is $v_{\lambda_1} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$. In addition, $\lambda_2 = 0$ is an eigenvalue, with associated eigenvector $v_{\lambda_2} = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix}$. Here, $v_{\lambda_2}^T v_{\lambda_1} = 0$, since Eqn (15.6) is a symmetric matrix as well as a stochastic matrix.

These are very simple examples, but they illustrate the more general phenomenon that there are two different complementary perspectives on these Random Walk Matrices.

- Linear Algebra Perspective.
- Probability Perspective.

There are many applications of these types of matrices in data science and machine learning, and we'll give just a hint of these.

15.3 Probability Perspective

Markov chains. Recall that when we discussed random variables, X , and samples from it, X_i , we typically assumed that they were presented as, or we could sample them as, a sequence of i.i.d. trials, and then we proved things about the mean, variance, tail bounds, etc.

Question. What if the trials are not independent?

Here are several answers.

- In general,

$$\Pr[X_{t+1} = x_{t+1}]$$

can depend on *all* of the previous steps. This situation is very complicated.

- The simplest dependence is that it depends on none of the previous steps, i.e., that the i.i.d. situation holds. This situation is relatively simple, and our previous result in the probability chapters basically assumed this special case of dependence.
- The next simplest dependence is that it depends on only the previous step. The situation when a random process depends only on the previous step and not on earlier steps is sufficiently important that it has a special name, a Markov chain.

Random processes that depend on only the previous step are very common and have a special name.

Definition 79 A sequence of random variables $X_0, X_1, \dots \in \Omega$ is a **Markov chain** with state space Ω , if for all possible states $x_{t+1}, x_t, \dots, x_1 \in \Omega$, we have

$$\Pr[X_{t+1} = x_{t+1} | X_t = x_t \wedge X_{t-1} = x_{t-1} \wedge \dots \wedge X_0 = x_0] = \Pr[X_{t+1} = x_{t+1} | X_t = x_t]$$

It is called **time homogeneous** if the latter probability is independent of t

Example. As an example of the situation when a random variable depends on its value only at the previous step, consider flipping a fair coin many times, and let Y_t be the number of Hs up and including the t^{th} step. Then, the number of Hs after the $(t+1)^{\text{th}}$ step clearly depends on the number of Hs at the t^{th} step, but—conditioned on that value—it doesn't depend on the value of the random variable at previous steps, i.e., on the exact order of the first t flips.

The study of Markov chains is a large and important topic, and there are many variations:

- Time homogeneous or not
- Finite versus infinite state space
- Continuous versus discrete time
- Symmetric or not
- Sinks or not
- Etc.

If you are interested (perhaps even if you are not), you will see this in more detail in later classes. Here, we are going to look at one little piece of this, but it is a piece that is very important for data science.

Transition Matrices. To start, observe that if the Markov chain is time homogeneous, then it is convenient to store all the information about the transitions in a so-called Transition Matrix.

Definition 80 *Given a time homogeneous Markov chain on a state space Ω , the transition matrix $P \in \mathbb{R}^{|\Omega| \times |\Omega|}$ has elements*

$$P_{ji} = \mathbf{Pr}[X_1 = j | X_0 = i].$$

Remark. Markov chains that are not time homogeneous also have transition matrices, but those transition matrices depend on the time variable t . When the Markov chain is time homogeneous, the Transition Matrix itself does not depend on the time variable t , and thus there is a single static matrix that basically defines the Markov chain, i.e., its state space, and its one-step transitions. In particular, for time homogeneous Markov chains, such as the one in Definition 80, it is also the case that

$$P_{ji} = \mathbf{Pr}[X_{t+1} = j | X_t = i]$$

for all $t \in \mathbb{Z}^+$.

Remark. This transition matrix P is a matrix. Given what we know about linear algebra, we can perform operations on it. Depending on the size of Ω , it may or may not be easy to write down this matrix explicitly. We get particularly interesting and useful results when we interpret the linear algebraic operations in terms of probability ideas.

Note that the elements of P are in the range $[0, 1]$, and the columns of P sum to 1. That is, P is a stochastic matrix like the Random Walk Matrix discussed above. Let's go into more detail about this connection.

In a way that complements how we can start with a graph $G = (V, E)$ and define an Adjacency Matrix, a Diagonal Degree Matrix, and a Random Walk Matrix, so too if we start with a Markov chain, then we can represent that Markov chain by a graph. To do so, consider the Markov chain with state space Ω and transition matrix P . Then, a corresponding graph representation is the weighted graph $G = (V, E, W)$, with

$$\begin{aligned} V &= \Omega \\ E &= \{(ij) \in \Omega \times \Omega \mid P(ij) > 0\} \\ W &= \{P(ij) > 0\}. \end{aligned}$$

Remark. This is a slight generalization of what we discussed before, since here we are permitting the edges to have weights.

Note. Much of Markov chain theory depends not on the values or weights of the edges but on whether the edges are present or absent, but the values or weights on the edges can have important algorithmic and statistical consequences in data science. On the other hand, since data are often preprocessed, this effect is often mitigated.

Note. Self loops are allowed here, since we can have $P(ii) > 0$. This corresponds to adding a “holding probability,” where the random walker stays at the current node with some non-zero probability.

Note. This matrix may not be symmetric, e.g., we could have

$$A = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix},$$

which clearly is a stochastic matrix.

Note. Often $P_{ij} > 0$ iff $P_{ji} > 0$. In this case, the graph is undirected (except for weighted edges which may be different). In many cases, however, this is not the case.

Reversible, irreducible, and aperiodic Markov chains. An important special case of Markov chains is when the Markov chain is reversible, irreducible, and aperiodic.

Definition 81 Let $\pi > 0$ be a probability distribution over Ω , i.e., on $\mathbb{R}^{|\Omega|}$. A Markov chain P is reversible if

$$\forall x, y \in \Omega \quad \pi_i P_{ij} = \pi_j P_{ji}.$$

Note. Any symmetric Markov chain is reversible w.r.t. the uniform distribution.

Note. A reversible Markov chain can be completely characterized by an undirected graph G with weights $Q(ij) = \pi_i P_{ij} - \pi_j P_{ji}$ on edge (ij) .

Definition 82 A Markov Chain is irreducible if $\forall i, j \exists t$ such that $P^t(ij) > 0$. Equivalently, the graph corresponding to P is strongly connected. Equivalently, if the graph is undirected, then the graph is connected.

Definition 83 A Markov chain is aperiodic if $\forall i, j$, we have that g.c.d. $\{t : P^t(i, j) > 0\}$.

Note. For a Markov chain, if the graph corresponding to P is undirected, then a periodicity is equivalent to $G(p)$ being non-bipartite.

Definition 84 A probability distribution is a stationary distribution if $P\pi = \pi$.

There are two interpretations to this:

- Stochastic process/dynamics interpretation. In this case, the stationary distribution is unchanged by one or more steps of the random walk. (Moreover, under certain assumptions, it is the limit state to which the chain converges.)
- Linear algebraic interpretation. In this case, the stationary distribution is an eigenvector with eigenvalue 1.

Fundamental Theorem of Markov Chains. When a Markov chain is irreducible and aperiodic, then we get a strong connection between the probability of Markov chains and the linear algebra of spectral ranking.

Of course, many Markov chains are not reversible, irreducible, and aperiodic. On the other hand, in data science and machine learning, one is often in a position to design the Markov chain, and one is thus in a position to ensure that these properties are satisfied.

When these conditions are satisfied, we obtain the following important result.

Theorem 22 (Fundamental Theorem of Markov Chains) *If a Markov Chain is irreducible and aperiodic, then it has a unique stationary distribution P . This is the unique right eigenvector of P with eigenvalue 1. Moreover, in this case,*

$$P^t(i, j) \xrightarrow{t \rightarrow \infty} \pi(j)$$

for all $i, j \in \Omega$. The rest of the eigenvalues satisfy $1 = \lambda_1 > \lambda_2 \geq \lambda_3 \dots > -1$.

This is actually a rather surprising connection. Let's look at it from a linear algebra perspective, and then let's give some examples.

15.4 Linear Algebra Perspective

A random walk starts at some node (or a probability distribution over nodes) and then it goes to one of the neighbors of that node, u.a.r. Instead of keeping track of where a particular random walk goes, we typically measure the probability distribution of the random walk, in order to see how the probability mass spreads out. So, let's treat the probability distribution on the vertices as a vector in $p \in \mathbb{R}^n$. Recall that $p \in \mathbb{R}^n$ is a probability distribution if $p_i \geq 0$, for all $i \in [n]$ and $\sum_{i=1}^n p_i = 1$.

Since the Markov transition matrix is a matrix, i.e., a representation of a linear transformation, we can actually do this using ideas from linear algebra without worrying too much about the initial probability distribution.

Given p_t , to derive p_{t+1} , note that the probability of being at a vertex i at time $t+1$ is the sum over neighbors j of i of the probability that the walk was at j at time t times the probability that it moved from j to i at time $t+1$. That is, the update rule is:

$$\begin{aligned} p_{t+1}(i) &= \sum_j W_{ij} p_t(j) \\ &= \sum_{j:(ij) \in E} \frac{1}{d(j)} p_t(j), \end{aligned}$$

i.e., it is just $p_{t+1} = Wp_t$. (This should remind you of the power method, as well as the motivating discussion about PageRank.)

Let's actually consider the so-called “lazy” random walk, where with probability 0.5, the random walker stays at the same node, i.e., moves according to an Identity transformation, and otherwise it follows an edge according to the original rule given by W . (This is for technical reasons that we won't discuss.) In this case, the update rule is:

$$p_{t+1}(i) = \frac{1}{2} p_t(i) + \frac{1}{2} \sum_{j:(ij) \in E} \frac{p_t(j)}{d(j)}.$$

This looks complicated, but it is very easy to represent in terms of linear algebra—the Lazy Random Walk Matrix is

$$\hat{W} = \frac{1}{2} (I + W),$$

and, in this case $p_{t+1} = \hat{W}p_t$.

Connections with symmetric matrices and the EVD. The matrix \hat{W} is not symmetric, nor is the matrix W , but it is *similar* to a symmetric matrix, in the following sense.

Observe the following. If we pre-multiply W by $D^{-1/2}$ and post-multiply W by $D^{1/2}$, then:

$$\begin{aligned} D^{-1/2}WD^{1/2} &= D^{-1/2}AD^{-1}D^{1/2} \\ &= D^{-1/2}AD^{-1/2} \\ &= M. \end{aligned}$$

From this, we have that

$$W = D^{1/2}MD^{-1/2}.$$

We also have that

$$M_{ij} = \frac{1}{\sqrt{d_i}\sqrt{d_j}}A_{ij},$$

and so M is a symmetric matrix. Then, we can also define

$$\hat{M} = \frac{1}{2}(I + M) = D^{-1/2}\hat{W}D^{1/2},$$

which is also a symmetric matrix.

Since $M \in \mathbb{R}^{n \times n}$, as well as $\hat{M} \in \mathbb{R}^{n \times n}$, is a symmetric matrix, by the spectral theorem, there is an orthonormal basis of eigenvectors and a set of eigenvalues for them. Let's work with M . Then, there are eigenvectors v_1, \dots, v_n and eigenvalues $\lambda_1, \dots, \lambda_n$ such that $\lambda_i v_i = Mv_i$, for all $i \in [n]$, which we can write more compactly as

$$MV = V\Lambda.$$

Given this, from the spectral theorem for symmetric matrices, if we have a vector $x \in \mathbb{R}^n$, e.g., a probability vector or any other vector, then we can decompose x as

$$x = \sum_{i=1}^n (v_i^T x) v_i = \sum_i \alpha_i v_i,$$

where $\alpha_i = v_i^T x$. If we have this, then multiplying any vector, and in particular a probability vector, by M is easy:

$$\begin{aligned} Mx &= M \left(\sum_{i=1}^n v_i v_i^T \right) x \\ &= \sum_{i=1}^n M v_i v_i^T x \\ &= \sum_{i=1}^n \lambda_i v_i v_i^T x \\ &= \sum_{i=1}^n (v_i^T x) \lambda_i v_i \quad (\text{since } v_i^T x \in \mathbb{R}). \end{aligned}$$

Similarly, if $t \in \mathbb{Z}^+$, then

$$M^t x = \sum_{i=1}^n (v_i^T x) \lambda_i^t v_i.$$

Relating eigen-properties of symmetric M to eigen-properties of non-symmetric W . To understand what the eigen-properties of M (a symmetric matrix, which thus has a full set of orthogonal eigenvectors, etc.) to the eigen-properties of W , observe that for each eigenvalue-eigenvector pair (λ, v) of M , we have

$$\lambda v = Mv = D^{-1/2}WD^{1/2}v.$$

Thus, by pre-multiplying each side by $D^{1/2}$, it follows that

$$\lambda D^{1/2}v = WD^{1/2}v.$$

There are several important consequences of this. First, if v is an eigenvector of M with eigenvalue λ , then $y = D^{1/2}v$ is an eigenvector of W with eigenvalue λ . Second, if we apply W to a vector, e.g., as we would do when running the random walk one step, then we have the that:

$$\begin{aligned} Wx &= D^{1/2}MD^{-1/2}x \\ &= D^{1/2}M \left(\sum_{i=1}^n v_i v_i^T \right) D^{-1/2}x \\ &= D^{1/2} \sum_{i=1}^n M v_i v_i^T D^{-1/2}x \\ &= D^{1/2} \sum_{i=1}^n \lambda_i v_i v_i^T D^{-1/2}x \\ &= \sum_{i=1}^n \lambda_i D^{1/2} v_i v_i^T D^{-1/2}x \\ &= \sum_{i=1}^n \lambda_i \left(v_i^T D^{-1/2}x \right) D^{1/2}v_i \quad (\text{since } v_i^T D^{-1/2}x \in \mathbb{R}). \end{aligned}$$

Third, if we consider W^t , where $t \in \mathbb{Z}^+$, which corresponds to applying W to a vector repeatedly t times, as we would do when running the random walk t steps, then we have that:

$$W^t x = \sum_{i=1}^n \lambda_i^t \left(v_i^T D^{-1/2}x \right) D^{1/2}v_i.$$

Observe that, in this result, t appears only in terms of the power of the eigenvalues. That is, the eigenvectors, i.e., the scaled directions, don't change direction, only the magnitudes change. (We saw this before when we considered the Power Method.) But since all the eigenvalues have magnitudes less than or equal to one, when you raise them to higher powers only the largest survives:

- $\lambda_{max} = 1$
- $\lambda_{min} \geq 0$ due to lazy random walk, and $\lambda_{min} > -1$ otherwise

So as $t \rightarrow \infty$ only terms with λ_{max} remain.

In summary, the point is the following:

- If a Random Walk Matrix is irreducible and aperiodic, then as $t \rightarrow \infty$, then any initial probability distribution approaches the stationary distribution.
- That is, in addition to being an eigenvector as well as a probability distribution that is unchanged by the Random Walk Matrix, this probability distribution is the distribution to which other distributions converge, if the random walk is run for many steps (which is why it is called the stationary distribution).

Another more complicated example. Let's look at a slightly more complicated example to illustrate an important related point.

Example. Let $n = 3$ and consider

$$A = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}.$$

In this case, we have

$$\begin{aligned} \lambda_1 &= 1 & v_1 &= \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \\ \lambda_2 &= \frac{1}{4} & v_2 &= \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \quad \text{and} \\ \lambda_3 &= \frac{1}{4} & v_3 &= \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}. \end{aligned}$$

Note that $v_i^T v_j = \delta_{ij} \leftrightarrow V^T V = VV^T = I$. So, $\{v_i\}_{i=1}^3$ defines an orthonormal basis for \mathbb{R}^3 that is particularly well-suited to the matrix A .

Let $p(0) \in \mathbb{R}^3$ be any vector such that $(p(0))_i \in [0, 1]$ and $\sum_{i=1}^3 (p(0))_i = 1$, i.e., be a probability distribution. Then

$$\begin{aligned} p(1) &= Ap(0) \\ p(2) &= Ap(1) \\ &= A^2 p(0), \end{aligned}$$

and so on.

Fact. The vector $p(t)$ is a probability distribution, for all t . For all t , we can write $p(t)$ in terms of any orthonormal basis of \mathbb{R}^3 , e.g., the canonical bases or the set of eigenvectors of A .

Fact. For all t , we can write $p(t)$ i.t.o. any orthonormal basis of \mathbb{R}^3 . The two bases that are most obvious here are the canonical basis and the basis of eigenvectors. Let's look at both. First, the canonical basis:

$$p(t) = (p(t))_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + (p(t))_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + (p(t))_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Second, the basis of eigenvectors:

$$\begin{aligned} p(t) &= c_1(t)v_1 + c_2(t)v_2 + c_3(t)v_3 \\ &= c_1(t)\frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + c_2(t)\frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + c_3(t)\frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}. \end{aligned}$$

The latter *looks* much messier, especially when it is written out explicitly, but looks can be deceiving. In particular, while it looks more complicated to you, the computer doesn't care, and it is easier to work with the latter. Here is an example of why.

Let's say that we are interested in the connection between $\{p(t)_i\}_{i=1}^3$ and $\{p(t+1)_i\}_{i=1}^3$. To do this, we need to do a matrix-vector multiplication.

In the canonical basis, it is messy.

On the other hand, in the basis of eigenvectors, we get a much simpler thing.

Question. What is the connection between $\{c(t)_i\}_{i=1}^3$ and $\{c(t+1)_i\}_{i=1}^3$?

Answer. Much simpler.

To see this, observe the following.

Lemma 8 $c_1(t) = \frac{1}{\sqrt{3}}$, when $p(t)$ is a probability distribution.

Proof:

$$\begin{aligned} c_1(t) &= v_1^T p(t) \\ &= \frac{1}{\sqrt{3}} p_1(t) + \frac{1}{\sqrt{3}} p_2(t) + \frac{1}{\sqrt{3}} p_3(t) \\ &= \frac{1}{\sqrt{3}} (p_1(t) + p_2(t) + p_3(t)) \\ &= \frac{1}{\sqrt{3}} \end{aligned}$$

◇

Next plug the expressions

$$\begin{aligned} p(t) &= c_1(t)v_1 + c_2(t)v_2 + c_3(t)v_3 \\ p(t-1) &= c_1(t-1)v_1 + c_2(t-1)v_2 + c_3(t-1)v_3 \end{aligned}$$

into the equation $p(t) = Ap(t-1)$, and the resulting equation holds iff

$$\begin{aligned} c_1(t) &= c_1(t-1) = \dots = c_1(0) \\ c_2(t) &= \frac{1}{4} c_2(t-1) = \dots = \left(\frac{1}{4}\right)^t c_2(0) \\ c_3(t) &= \frac{1}{4} c_3(t-1) = \dots = \left(\frac{1}{4}\right)^t c_3(0). \end{aligned}$$

So, plugging this in, we get

$$\begin{aligned} p(t) &= \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \left(\frac{1}{4}\right)^t \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + \left(\frac{1}{4}\right)^t \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} \\ &= \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \left(\frac{1}{4}\right)^t \frac{1}{\sqrt{2}} \begin{pmatrix} 2/3 \\ -1/3 \\ -1/3 \end{pmatrix} \\ &\xrightarrow{t \rightarrow \infty} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}. \end{aligned}$$

where for the second-to-last step we have said that if $p(0) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, then $c_1(0) = \frac{1}{\sqrt{3}}$, $c_2(0) = \frac{-1}{\sqrt{2}}$, and $c_3(0) = \frac{1}{\sqrt{6}}$.

15.5 Usefulness of these results

There are many uses of this result in data science and machine learning.

Classification, clustering, ranking. Given a graph $G = (V, E)$, which could represent a social network, a nearest neighbor graph, etc., do the following.

1. Take the graph G .
2. Construct the Adjacency matrix A .
3. Construct the random walk matrix $W = AD^{-1}$. From Markov Chain Theory or from Symmetric Matrix Theory, we know that the eigenvalues of this are

$$1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n > -1.$$

In addition, we have the matrix

$$\begin{aligned} M &= D^{-1/2}AD^{-1}D^{1/2} \\ &= D^{-1/2}WD^{1/2}, \end{aligned}$$

and since this is a symmetric matrix we have a full set of n orthogonal eigenvectors v_i and corresponding real eigenvalues λ_i , i.e., we have

$$\{\lambda_i, v_i\}_{i=1}^n.$$

Moreover, since

$$\lambda v = Mv = D^{-1/2}WD^{1/2}v,$$

and also since

$$\lambda D^{1/2}v = WD^{1/2}v,$$

we have a full set of eigenvalues/eigenvectors

$$\left\{\lambda_i, D^{1/2}v_i\right\}_{i=1}^n.$$

So, the eigenvalues are the same, and the eigenvectors are the same up to scaling.

For $\lambda = 1$, the eigenvector v_1 is the constant vector, and $D^{1/2}v_1$ is proportional to the degree. For λ_2 , we have that $v_2^T v_1 = 0$, and so v_2 has both positive and negative entries. We can use v_2 to “partition” the graph into two pieces. That is, we can cluster the nodes based on the entries in v_2 .

Here is a simple intuitive rule:

$$\begin{array}{ll} v_2^{(i)} < 0 & \text{Put in the first cluster} \\ v_2^{(i)} > 0 & \text{Put in the second cluster} \end{array}$$

This rule works in simple settings, but better rules exist. The idea is that if the data split into two meaningful clusters, e.g., see the figure, then this finds it. In particular, if that is the adjacency matrix, then we split it as you would expect; but we also split the adjacency matrix in the other subfigure (which is just the other one permuted—so eigenvectors don’t care about that, which is good since we don’t usually know that, i.e., that is what we want to find).

Here are uses of v_2 :

- Clustering
- Classification and Prediction

Here are uses of v_1 :

- Ranking

PageRank is a variant of this. Actually, the more realistic version of PageRank focuses on directed graphs. We don’t have time for this topic, but it uses a lot of the methods we have been discussing in this section.

High-dimensional integration. More generally, by running a Markov chain such as this, coupled with a filter known as the Metrolopis filter, one can construct a random walk that samples from some probability distribution, e.g., a distribution supported on a “ball inside a box” in high dimensions, thereby performing high-dimensional integrals that we couldn’t perform simply by throwing darts at the bounding box.

15.6 Problems

15.6.1 Implementations and Applications of the Theory

1. XXX.
2. XXX.

15.6.2 Pencil-and-paper Problems

1. Let λ be an eigenvalue of an $n \times n$ (not necessarily symmetric) matrix A . Show that the set of eigenvectors of A with eigenvalue λ is a subspace of \mathbb{R}^n . If λ is a root of the characteristic polynomial that is repeated k times, then what can you say about the dimension of that subspace? In general? If A is symmetric?
2. For each of the transition matrices: (i) draw the corresponding graph with edges labeled by the probabilities and (ii) state whether the resulting Markov chain is irreducible and whether it is aperiodic.
 - (a) $\begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$
 - (b) $\begin{pmatrix} 1 & 0.3 \\ 0 & 0.7 \end{pmatrix}$
 - (c) $\begin{pmatrix} 0.2 & 0 & 0 \\ 0 & 1 & 0.4 \\ 0.8 & 0 & 0.6 \end{pmatrix}$
 - (d) $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
 - (e) $\begin{pmatrix} 0.4 & 0 & 0.7 \\ 0.6 & 0.2 & 0.1 \\ 0 & 0.8 & 0.2 \end{pmatrix}$
3. Let A and B be $n \times n$ (column-)stochastic matrices, i.e., matrices in which every column has non-negative entries that sum to 1, and let I be the $n \times n$ identity matrix.
 - (a) Show that if p is a probability distribution, then Ap is a probability distribution.
 - (b) Show that $\alpha A + (1 - \alpha)B$, where $\alpha \in [0, 1]$, is a stochastic matrix.
 - (c) Let $C = \alpha A + (1 - \alpha)I$. If $1 = \lambda_1(A) > \lambda_2(A) \geq \dots \geq \lambda_n(A)$ are the eigenvalues of A , then what are the eigenvalues of C ? What are the eigenvectors of C ?
 - (d) How does replacing A with C affect the number of iterations necessary for the power method to converge? How does it affect the vector to which the power method converges? Hint: consider λ_2^t versus λ_1^t in each case.
4. XXX. OTHER PROBLEMS INVOLVING PAGERANK.

Chapter 16

High-dimensional Calculus: Integration and Differentiation

XXX. COVER THIS. MAKE THE POINT THAT THE ITERATIVE ALGORITHM FOR MCMC AND MAYBE ALSO SGD CAN THEMSELVES BE ANALYZED WITH LINEAR AGEBRA.

16.1 Integration and Differentiation in One Dimension

XXX. BRIEF REVIEW EITHER HERE OR IN HS REVIEW SECTION.

16.2 An Obvious but Not Good Way to Extend to Higher Dimension

XXX. PUT BRIEF SUMMERY HERE, BASICALLY OF TACKING ON SUBSCRIPTS, AND POINT OUT HOW TRAPEZOID RULE HITS CURSE OF DIMENSIONALITY. CAN I DO THE SAME THING WITH DERIVATIVES.

16.3 High-dimensional Integration with Markov Chain Monte Carlo

XXX. FILL IN. BASICALLY, DO ENOUGH TO ESTIMATE THE VOLUME OF THE BALL IN THE BOX IN HIGH DIM. NEED TO DO REJECTION SAMPLING AND RELATE TO GRAPH. NEED PICTURE OF CONNECTED VERSUS NOT CONNECTED CONFIGURATION SPACE.

16.4 High-dimensional Differentiation with Stochastic Gradient Descent

XXX. FILL IN. BASICALLY, MAKE THE POINT THE DERIVATIVE ISN'T A NUMBER AS MUCH AS A DIRECTION, AND YOU CAN ITERATE TO OPTIMIZE. MENTION SECOND ORDER, BUT EVEN IF I DONT DO A SECOND ORDER OPTIMIZATION ALGORITHMS IN DETAIL, DISCUSS HESSIANS AND HOW THAT GENERALIED SECOND DERIVATIVE TEST IN ONE-DIMENSION, MAYBE POINTING OUT HOW SADDLE POINTS ARE THE ISSES.

16.5 Problems

16.5.1 Implementations and Applications of the Theory

1. XXX.
2. XXX.

16.5.2 Pencil-and-paper Problems

XXX. HW PROBLEMS.

Part VI

Additional Miscellaneous Stuff To Incorporate Somewhere

Chapter 17

Additional Homework Questions to Incorporate Somewhere

17.1 From putting together s18 final and final prep

Possible Review Questions.

1. ZHEWEI:

Let A be an $n \times n$ positive definite matrix with eigenvalues $\lambda_1 > \lambda_2 > \dots > \lambda_n > 0$. Show that

$$\max_{\|x\|_2=1} x^T A x = \lambda_1, \quad \min_{\|x\|_2=1} x^T A x = \lambda_n.$$

2. ZHEWEI:

Use the method of Lagrange multipliers to find the maximum of $f(x_1, x_2) = 3x_1 + 4x_2$ subject to the constraint $x_1^2 + x_2^2 = 1$. Plot the level sets of f , the constraint equation, and the maximum of f subject to the constraint.

3. ZHEWEI:

Consider the quadratic form

$$Q(x_1, x_2, x_3) = 3x_1^2 - 2x_2^2 - 4x_3^2 + 2x_1x_2 - 7x_2x_3 + x_1x_3.$$

Decompose it into a sum/difference of squares. Do it in the order $x_1 \rightarrow x_2 \rightarrow x_3$ as well as in the order $x_3 \rightarrow x_1 \rightarrow x_2$. What is the similar thing you get from the two decompositions, and what is different between the two decompositions?

4. ZHEWEI:

If you already know the eigenvalues and eigenvectors of a symmetric matrix A , then what are the eigenvalues and eigenvectors of A^2 and A^3 and $I + A + A^2 + A^3$.

5. ZHEWEI:

Compute the eigenvalues and eigenvectors of

$$A = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 3 & 4 & 0 & 0 \\ 0 & 0 & 5 & 6 \\ 0 & 0 & 7 & 8 \end{bmatrix}.$$

6. ZHEWEI:

If the random variable $x \sim U[-1, 1]$, i.e., is uniform on the interval $[-1, 1]$, then is $x^2 \sim U[0, 1]$? Justify your answer.

7. ZHEWEI:

Suppose we flip a potentially-unfair coin, which would be heads with probability p ($0 < p < 1$) and would be tails with probability $1 - p$. Define the random variable X such that $X = 1$ if you get a head, and $X = 0$ if you get a tail.

(a) Compute the mean and variance of X .

(b) Prove $p(1 - p) \leq 1/4$.

(c) Let $S_n = X_1 + \dots + X_n$ and $n = 100$. Use the Chebychev Inequality to provide a bound for

$$P(|S_n/n - p| \geq 0.1).$$

8. OK:

Let λ be an eigenvalue of an $n \times n$ (not necessarily symmetric) matrix A . Show that the set of eigenvectors of A with eigenvalue λ is a subspace of \mathbb{R}^n . If λ is a root of the characteristic polynomial that is repeated k times, then what can you say about the dimension of that subspace? In general? If A is symmetric?

9. OK:

Let H be the subspace of \mathbb{R}^5 spanned by

$$V = \text{Span} \left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{pmatrix} \right\}.$$

Find an orthonormal basis for H .

10. EXAM:

Show that if A is an $n \times n$ positive semidefinite matrix, then its EVD, let's call it $A = QDQ^T$, agrees exactly with its SVD, let's call it $A = U\Sigma V^T$ (i.e., show that $\Sigma = D$ and that $Q = U = V$).

11. EXAM:

Let

$$A = \begin{pmatrix} 2 & -3 \\ 0 & 2 \end{pmatrix}.$$

- (a) Compute the SVD of A . Express your answer (i) as the sum of rank-1 terms and (ii) as $A = U\Sigma V^T$ for appropriate U , Σ , and V .
- (b) In \mathbb{R}^2 , describe the image of the unit disk under the transformation of A using the SVD. That is, draw a picture of the region $\{Ax : \|x\|_2 \leq 1\}$.
- (c) In \mathbb{R}^2 , describe the inverse image of the unit disc by drawing a picture of the region $\{x : \|Ax\|_2 \leq 1\}$.

12. EXAM:

Two six-sided dice are thrown sequentially, and the face values that come up are recorded.

(a) List the sample space.

(b) List the elements that make up the following events.

- i. A = the sum of the two values is at least 5;
- ii. B = the value of the first die is higher than the value of the second;
- iii. C = the first value is 4.

(c) List the elements of the following events.

- i. $A \cap C$
- ii. $B \cup C$
- iii. $A \cap (B \cup C)$

13. OK:

Draw Venn diagrams to illustrate DeMorgan's Laws:

- (a) $(A \cup B)^C = A^C \cap B^C$
- (b) $(A \cap B)^C = A^C \cup B^C$

14. TODO:

((Rice, 1.5))
XXX.

15. EXAM:

The following is an extension of the addition rule.

$$\begin{aligned} \Pr[A \cup B \cup C] &= \Pr[A] + \Pr[B] + \Pr[C] \\ &\quad - \Pr[A \cap B] - \Pr[A \cap C] - \Pr[B \cap C] \\ &\quad + \Pr[A \cap B \cap C]. \end{aligned}$$

Verify this in two ways: first, by an appropriate Venn diagram; and second, by a formal argument using the axioms of probability.

16. TODO:

((Rice, 1.32))
XXX.

17. TODO:

((Rice, 1.33))
XXX.

18. TODO:

((Rice, 1.43))
XXX.

19. EXAM:

Two dice are rolled, and the sum of the face values is six. What is the probability that at least one of the dice came up a three?

20. EXAM:

A box has three coins. One coin has two heads, one coin has two tails, and one coin is fair with one head and one tail. A coin is chosen at random, it is flipped, and it comes up heads.

- (a) What is the probability that the coin chosen is the two-headed coin?
- (b) What is the probability that if it is thrown another time it will come up heads?
- (c) Assuming that the coin is flipped again and comes up heads again, then what is the probability that the coin chosen is the two-headed coin?

21. REV:

Show that if $\Pr[A|E] \geq \Pr[B|E]$ and $\Pr[A|E^C] \geq \Pr[B|E^C]$, then $\Pr[A] \geq \Pr[B]$.

22. REV:

Suppose that the probability of living to be older than 70 is 0.6 and the probability of living to be older than 80 is 0.2. If a person reaches his or her 70th birthday, then what is the probability that he or she will reach his or her 80th birthday.

23. OK:

If B is an event with $\Pr[B] > 0$, then show that the set function $Q(A) = \Pr[A|B]$ satisfies the axioms to be a probability function. Thus, e.g,

$$\Pr[A \cup C|B] = \Pr[A|B] + \Pr[C|B] - \Pr[A \cap C|B].$$

24. REV:

Assume that A , B , and C are independent events.

(a) Show that A and B^C as well as A^C and B^C are independent.

(b) Show that

$$\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A] \Pr[B]$$

(c) Show that $A \cap B$ and C are independent and that $A \cup B$ and C are independent.

25. OK:

Show that \emptyset is independent of A , for any A .

26. OK:

Either prove the following statement or provide a counterexample to disprove it: If A is independent of B and B is independent of C , then A is independent of C .

27. OK:

Let A and B be events.

(a) If A and B are disjoint, can they be independent?

(b) If $A \subset B$, can A and B be independent?

28. TODO:

((Rice, 2.1))

XXX.

29. TODO:

((Rice, 2.2))

XXX.

30. TODO:

((Rice, 2.3))

XXX.

31. OK:

Let A and B be events, and let I_A and I_B be the associated indicator random variables. Show that

$$\begin{aligned} I_{A \cap B} &= I_A I_B \\ I_{A \cup B} &= \max\{I_A, I_B\}. \end{aligned}$$

32. OK:

Which is more likely: 9 heads in 10 tosses of a fair coin or 18 heads in 20 tosses of a fair coin.

33. TODO:

((Rice, 2.9))

XXX.

34. OK:

Three identical fair coins are thrown simultaneously until all three show the same face. What is the probability that they are thrown more than three times?

35. OK:

If X is a discrete uniform random variable, i.e., $\Pr[X = i] = 1/n$, for $i = 1, \dots, n$, then find $\mathbf{E}[X]$ and $\mathbf{Var}[X]$.

36. TODO:

((Rice, 4.3))

Find $\mathbf{E}[X]$ and $\mathbf{Var}[X]$ for Problem 3.2 XXX OF RICE.

37. TODO:

((Rice, 4.6))

XXX. HOW DOES THIS RELATE TO THE TWO WAYS I DID PROB.

38. EXAM:

Suppose that X is a random variable in \mathbb{R} with $\mathbf{E}[X] = \mu$ and $\mathbf{Var}[X] = \sigma^2$, and define the random variable $Z = (X - \mu)/\sigma$.

(a) Show that $\mathbf{E}[Z] = 0$ and $\mathbf{Var}[Z] = 1$. (The random variable Z is said to be a standardized version of the random variable X .)

(b) What is the analogue of this statement if the random variable $X \in \mathbb{R}^n$? Be careful and precise about specifying the dimensions of everything, pre-multiplying versus post-multiplying, etc.

39. OK:

Let X_1, \dots, X_n be jointly distributed discrete random variables with expectations $\mu_i = \mathbf{E}[X_i]$, and let Y be a linear (actually, an affine) function of the X_i , i.e., $Y = a + \sum_{i=1}^n b_i X_i$, then show that

$$\mathbf{E}[Y] = a + \sum_{i=1}^n b_i \mu_i.$$

40. EXAM:

State and prove Chebyshev's Inequality for discrete random variables.

41. TODO:

((Rice, 4.34))

XXX. INCLUDE ON NEXT ITERATION.

42. OK:

Show that $\mathbf{Var}[X - Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] - 2\mathbf{Cov}[X, Y]$.

43. OK:

If X and Y are independent random variables with equal variances, then find $\mathbf{Cov}[X + Y, X - Y]$.

44. OK:

Let X and Y be random variables, and define $U = a + bX$ and $V = c + dY$.

(a) Show that $|\rho_{UV}| = |\rho_{XY}|$, where $\rho_{..}$ denotes correlation.

(b) Is it true that $\rho_{UV} = \rho_{XY}$? Either prove it, or provide a counterexample.

45. OK:

Let X and Y be independent random variables, and let $Z = X - Y$. Find expressions for the covariance and the correlation of X and Z in terms of the variances of X and Y .

46. OK:

Let X and Y be independent random variables with means μ and variances σ^2 . Let $Z = \alpha X + \sqrt{1 - \alpha^2}Y$. Find $\mathbf{E}[Z]$, $\mathbf{Var}[Z]$, and ρ_{XZ} (where ρ is the correlation).

47. OK:

Two independent measurements, X and Y , are taken of a quantity, and $\mu = \mathbf{E}[X] = \mathbf{E}[Y]$, but $\sigma_X \neq \sigma_Y$. The two measurements can be combined with a weighed average to give

$$Z = \alpha X + (1 - \alpha)Y,$$

where α is a scalar $0 \leq \alpha \leq 1$.

- (a) Show that $\mathbf{E}[Z] = \mu$.
- (b) Compute $\mathbf{Var}[Z]$.
- (c) Find α in terms of σ_X and σ_Y to minimize $\mathbf{Var}[Z]$.

48. OK:

Suppose that X_i , where $i = 1, \dots, n$, are independent random variables with $\mathbf{E}[X_i] = \mu$ and $\mathbf{Var}[X_i] = \sigma^2$. Let $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. Show that $\mathbf{E}[\bar{X}] = \mu$ and that $\mathbf{Var}[\bar{X}] = \sigma^2/n$.

49. REV:

Given random variables X and Y , consider their variances and covariances.

- (a) Show that $\mathbf{Cov}[X, Y] \leq \sqrt{\mathbf{Var}[X]\mathbf{Var}[Y]}$.
- (b) If we put these variances and covariances into a 2×2 variance-covariance matrix

$$\Omega = \begin{pmatrix} \mathbf{Var}[X] & \mathbf{Cov}[X, Y] \\ \mathbf{Cov}[X, Y] & \mathbf{Var}[Y] \end{pmatrix},$$

then what does this inequality imply about the discriminant and determinant of Ω ? What does this imply about the definiteness of Ω ?

50. EXAM:

Let X , Y , and Z be uncorrelated random variables with variances σ_X^2 , σ_Y^2 , and σ_Z^2 , respectively. Let

$$\begin{aligned} U &= Z + X \\ V &= Z + Y. \end{aligned}$$

Find the covariance $\mathbf{Cov}[U, V]$ and the correlation ρ_{UV} .

51. OK:

Let X_i , for $i = 1, \dots, n$, be independent random variables with means μ and variances σ^2 . In addition, let

$$\begin{aligned} T &= \sum_{i=1}^n iX_i \\ S &= \sum_{i=1}^n X_i. \end{aligned}$$

- (a) Find $\mathbf{E}[T]$ and $\mathbf{Var}[T]$.
- (b) Find $\mathbf{E}[S]$ and $\mathbf{Var}[S]$.
- (c) Find the covariance and correlation of S and T .

52. OK:

If X and Y are independent random variables, find $\mathbf{Var}[XY]$ in terms of the means and variances of X and Y .

53. OK:

Let (X, Y) be a random point uniformly distributed on a unit disk in \mathbb{R}^2 . Show that $\mathbf{Cov}[X, Y] = 0$, but that X and Y are not independent.

54. OK:

Show that $\mathbf{E}[\mathbf{Var}[Y|X]] \leq \mathbf{Var}[Y]$.

55. OK:

If X and Y are independent, show that $\mathbf{E}[X|Y = y] = \mathbf{E}[X]$.

56. OK:

Show that, if X and Y are independent, then

$$\mathbf{Var}[aX + bY] = a^2\mathbf{Var}[X] + b^2\mathbf{Var}[Y].$$

What expression is obtained if X and Y are not independent?

57. TODO:

((Rice, 14.1))

XXX. INCLUDE IN TEXT HOW TO DO.

58. OK:

Suppose that $y_i = \mu + e_i$, where $i = 1, \dots, n$ and the e_i are independent errors with mean 0 and variance σ^2 . Show that the mean $\mathbf{E}[y]$ is the least squares estimate of μ .

59. TODO:

((Rice, 14.5))

XXX. INCLUDE IN TEXT HOW TO DO.

60. TODO:

((Rice, 14.6))

XXX. INCLUDE IN TEXT HOW TO DO.

61. TODO:

((Rice, 14.7))

XXX. INCLUDE IN TEXT HOW TO DO.

62. TODO:

((Rice, 14.8))

XXX.

63. TODO:

((Rice, 14.9))

XXX.

64. TODO:

((Rice, 14.10))

XXX.

65. TODO:

((Rice, 14.11))

XXX. TO DO NEXT TIME.

66. TODO:

((Rice, 14.14))

XXX.

67. TODO:

((Rice, 14.15))

XXX.

68. TODO:

((Rice, 14.16))

XXX. TO DO NEXT TIME.

69. TODO:

((Rice, 14.17))

XXX. TO DO NEXT TIME.

70. TODO:

((Rice, 14.18))

XXX.

71. TODO:

((Rice, 14.19))

XXX. TO DO NEXT TIME.

72. TODO:

((Rice, 14.20))

XXX. HARD EXAM.

73. TODO:

((Rice, 14.21))

XXX.

74. TODO:

((Rice, 14.22))

XXX.

75. EXAM:

Find any matrix A such that $A \neq I$ and $A = A^{-1}$.

76. OK:

Let $x = (x_1 \ x_2)^T$ and $y = (y_1 \ y_2)^T$. Find matrix A and vector c so that the two equations in each of the following cases can be written as $y = A(x - c)$.

(a) $y_1 = x_1 + 2x_2 - 1$ and $y_2 = 2x_1 + x_2 - 2$.

(b) $\sqrt{2}y_1 = x_1 - x_2 - 1$ and $\sqrt{2}y_2 = x_1 + x_2 - 1$.

(c) $y_1 = 2x_1 + 2$ and $y_2 = 4x_2 - 4$.

(d) $y_1 = 2x_2 - 4$ and $y_2 = 4x_1 - 4$.

77. EXAM:

Given $A = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$, $v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, and $v_2 = \begin{pmatrix} 0 \\ 4 \end{pmatrix}$.

(a) Compute $w_1 = Av_1$ and $w_2 = Av_2$.

(b) Compute the angle between w_1 and the horizontal axis as well as the angle between w_2 and the horizontal axis.

(c) Graph the vectors v_1 and v_2 with respect to the conventional axes.

(d) On the same graph, draw the new set of axes for which the elements of the vectors w_1 and w_2 can be regarded as the coordinates of the vectors v_1 and v_2 with respect to the new set of axes.

Do the same for $A = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$.

Do the same for $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

78. REV:

Graph the vectors

$$a = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad b = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad c = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad d = \begin{pmatrix} 2 \\ -2 \end{pmatrix},$$

as well as the vectors obtained by the following orthogonal projections.

- (a) a onto b
- (b) a onto c
- (c) a onto d
- (d) b onto d
- (e) b onto the X-axis
- (f) b onto the Y-axis

79. EXAM:

Consider the vectors

$$a = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}, \quad b = \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}, \quad c = \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \quad d = \begin{pmatrix} 2 \\ -2 \\ 2 \end{pmatrix}.$$

Compute the vector that is the projection of the vector a onto the space spanned by the following.

- (a) the vector b
- (b) the vector c
- (c) the vector d
- (d) the matrix $(b : c)$ (i.e., the matrix with b in the first column and c in the second column)
- (e) the matrix $(b : d)$
- (f) the matrix $(c : d)$
- (g) the matrix $(a : b)$
- (h) the matrix $(b : a)$
- (i) the matrix $(b : c : d)$

Hint: Not every projection needs to be computed naïvely by brute force. Consider the relationship between b, d and $(b : c), (c : b)$, etc.

80. REV:

Suppose that immediately after the transformation of a vector x to a vector $y = Ax$, e.g., using a computer, the vector x was accidentally deleted. In each of the following cases, is it possible to recover the vector x ? If yes, do it; if no, then indicate why not.

- (a) $A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ and $y = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$
- (b) $A = \begin{pmatrix} 1 & -1 \\ 1 & 2 \end{pmatrix}$ and $y = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
- (c) $A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ and $y = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$
- (d) $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $y = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$
- (e) $A = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$ and $y = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$
- (f) $A = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$ and $y = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$

81. EXAM:

Let $A = \begin{pmatrix} 2 & 0 & 4 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}$ and $b = \begin{pmatrix} 2 \\ 1 \\ 6 \end{pmatrix}$.

- (a) Write out explicitly the three equations represented by the system $Ax = b$.
 (b) Find a vector x such that $Ax = b$.

82. OK:

Let $\begin{pmatrix} 1 & s \\ 1 & 2 \end{pmatrix}$, $b = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and $c = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Find all values of s , if any, for which the system of linear equations:

- (a) $Ax = b$ has no solution.
 (b) $Ax = b$ has a unique solution.
 (c) $Ax = c$ has no solution.
 (d) $Ax = c$ has a unique solution.
 (e) $Ax = c$ has infinitely many solutions.

83. TODO:

((Hadi, 8.7))

XXX.

84. TODO:

((Hadi, 8.9))

XXX. MAYBE INCORPORATE THIS INTO TEXT.

85. REV:

Let 1 be the all-ones vector in \mathbb{R}^n , and let P be the projection matrix for the subspace of \mathbb{R}^n that is spanned by 1 .

- (a) Show that $P = \frac{1}{n}11^T$.
 (b) Let $Q = I - P$, and show that $PP = P$ and that $QQ = Q$.
 (c) Explicitly write out Q for $n = 4$.
 (d) Given an $n \times n$ matrix A , describe in words what the matrix Q does, i.e., describe what is AQ , and what is QA , in terms of A , in words.
 (e) Show that $Q1 = (I - P)1 = 0$.

86. EXAM:

For each of the following three data matrices, where data points are stored as columns,

$$X = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 1 & 0 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix}, \quad Y = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 1 & 2 \\ -1 & 0 & -1 \\ -1 & 0 & -1 \end{pmatrix},$$

compute

- (a) the mean vector, i.e., the vector with elements equal to the column-wise mean.
 (b) the mean-centered matrix.
 (c) the variance-covariance matrix.
 (d) the correlation matrix.

87. EXAM:

For each of the three data matrices in the previous problem, Problem ??, do the following.

- (a) Center the matrix by subtracting from each element the mean of its column.
 (b) Divide each element of the centered matrix by the standard deviation of its column. (The resulting matrix is called the *standardized data*.)

- (c) Compute the variance-covariance matrix of the standardized data matrix.
 (d) Verify that the covariance matrix just obtained in 87c is the same as the correlation of the original data matrix in Problem 86d.

This illustrates the fact that the correlation matrix is just the covariance matrix of the standardized data.

88. TODO:

((Hadi, 8.14))
 XXX.

89. OK:

Consider the no-intercept simple linear regression model

$$y_i = \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n,$$

or, equivalently, in matrix notation $y = \beta_1 x + \epsilon$. Show that the LS estimate of β_1 is $\hat{\beta}_1 = \frac{x^T y}{x^T x}$.

90. REV:

Consider the regression model $y = X\beta + \epsilon$, where $X \in \mathbb{R}^{n \times p}$, and where $\epsilon \in \mathbb{R}^n$ is a noise process. Let

$$\begin{aligned}\hat{\beta} &= (X^T X)^{-1} X^T y \\ \hat{y} &= X\hat{\beta} = X(X^T X)^{-1} X^T y = Py,\end{aligned}$$

where P is the projection matrix onto the span of the columns of X , be the vector of fitted values, and let

$$e = y - \hat{y} = y - Py$$

be the vector of residuals.

- (a) Show that $X^T e = 0$, i.e., that e is orthogonal to each column of X .
 (b) Show that $e^T \hat{y} = 0$, i.e., that e and \hat{y} are orthogonal to each other.
 (c) Show that $(I - P)x = 0$, i.e., that the rows of $I - P$ are orthogonal to the columns of X .

91. TODO:

((Hadi, 8.17))
 XXX.

92. TODO:

((Hadi, 8.18))
 XXX.

93. REV:

Find and compare the eigenvalues and eigenvectors for each of the three matrices below.

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}, \quad C = \begin{pmatrix} 8 & 4 \\ 4 & 8 \end{pmatrix}.$$

94. REV:

Find the eigenvalues of each of the following matrices.

$$A = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & 4 & 1 \\ 0 & 2 & 2 \\ 0 & 0 & 1 \end{pmatrix}, \quad C = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

95. EXAM:

Let $A = \begin{pmatrix} a & 6 \\ 6 & b \end{pmatrix}$. Find a and b such that the eigenvalues of A are 9 and -3.

96. TODO:

((Hadi, 9.9))
XXX.

97. EXAM:

Find the matrix A whose eigenvalues are 4 and 1 and the corresponding eigenvectors are

$$v_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad \text{and} \quad v_2 = \begin{pmatrix} -2 \\ 1 \end{pmatrix}.$$

98. REV:

Let

$$X = \begin{pmatrix} 1 & -1 \\ 0 & 0 \\ 1 & 1 \end{pmatrix}, \quad Y = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix}.$$

- (a) Compute the eigenvalues and eigenvectors of $X^T X$, $Y^T Y$, and $Z^T Z$.
- (b) Compute the singular values for each of X , Y , and Z .
- (c) Compute the SVD for each of the matrices X , Y , and Z . (Hint: Use your previous answers to compute the diagonal singular value matrix (Σ) and the right singular vector matrix (V), and then post-multiply both sides of, e.g., $X = U\Sigma V^T$ by $V\Sigma^{-1}$ to obtain U .)
- (d) What are the eigenvalues of XX^T , YY^T , and ZZ^T .
- (e) What are the eigenvectors corresponding to the non-zero eigenvalues of XX^T , YY^T , and ZZ^T .
- (f) Compute $\text{Det}(XX^T)$, $\text{Det}(YY^T)$, and $\text{Det}(ZZ^T)$.
- (g) Compute $\text{Tr}(XX^T)$, $\text{Tr}(YY^T)$, and $\text{Tr}(ZZ^T)$.

99. REV:

Let A be a symmetric matrix.

- (a) Show that the eigenvalues of A^2 are the squares of the eigenvalues of A .
 - (b) Show that A and A^2 have the same set of eigenvectors.
100. OK:
- Let $x = (x_1 \ x_2)^T$. For each of the following expressions, find the matrix $S = S^T$ so that the expression can be written as $x^T S^{-1} x$.

- (a) $4x_1^2 + 2x_2^2$
- (b) $x_1^2 + 4x_2^2$
- (c) $2x_1^2 + 5x_2^2 - 2x_1x_2$
- (d) $1.5x_1^2 + 1.5x_2^2 - x_1x_2$

101. OK:

Find the matrix $S = S^T$ such that $x^T A x = x^T S x$ in each of the following cases.

- (a) $A = \begin{pmatrix} 2 & 1 \\ 2 & 4 \end{pmatrix}$.
- (b) $A = \begin{pmatrix} 2 & -1 \\ 3 & 4 \end{pmatrix}$.
- (c) $A = \begin{pmatrix} 5 & 0 & 1 \\ 2 & 2 & 0 \\ 3 & 2 & 3 \end{pmatrix}$.

$$(d) A = \begin{pmatrix} 1 & 0 & 1 \\ 2 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

102. EXAM:

Sketch the graph of the function

$$(x - c)^T S^{-1} (x - c) = a$$

for each of the following cases.

$$(a) S^{-1} = \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}, c = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, a = 4$$

$$(b) S^{-1} = \begin{pmatrix} 4 & 0 \\ 0 & 16 \end{pmatrix}, c = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, a = 1$$

$$(c) S^{-1} = \begin{pmatrix} 5 & 1 \\ 1 & 2 \end{pmatrix}, c = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, a = 4$$

$$(d) S^{-1} = \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}, c = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, a = 1$$

103. EXAM:

For each of the ellipses defined in the previous problem.

(a) Compute the lengths of the axes for each ellipse.

(b) Compute the cosines of the angles between each of the ellipse's axes and the horizontal axis.

(c) What are the angles between each ellipse's axes and the horizontal axis.

(d) Arrange the four ellipses in an increasing order of volume.

104. TODO:

((Hadi, 10.6))

XXX.

105. TODO:

((Hadi, 11.1))

XXX.

106. TODO:

((Hadi, 11.2))

XXX.

107. TODO:

((Hadi, 11.6))

XXX.

108. REV:

Given $A = \begin{pmatrix} a & 1 \\ 1 & b \end{pmatrix}$. Find the conditions on a and b under which A is:

(a) positive definite.

(b) positive semidefinite.

(c) negative definite.

(d) indefinite.

109. OK:

Compute A^8 , when

$$\begin{pmatrix} 2 & -1 \\ 4 & 3 \end{pmatrix}.$$

110. OK:

Show that $A^3 - 9A + 10I = 0$, when

$$\begin{pmatrix} 1 & -2 & 2 \\ 0 & 2 & 0 \\ 1 & -1 & -3 \end{pmatrix}.$$

111. EXAM:

Prove that the product of lower triangular matrices is a lower triangular matrix.

112. REV:

Determine the inverse of

$$\begin{pmatrix} 2 & 1 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{pmatrix}.$$

(Hint: Solve $AX = I$ for the 3×3 matrix X .)

113. REV:

The *null space* of a matrix A is the set of all vectors x such that $Ax = 0$.

(a) Determine the null space of $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$.

(b) Determine the null space of $\begin{pmatrix} 1 & 3 \\ 2 & 6 \end{pmatrix}$.

114. REV:

Recall that the trace of a matrix, $\text{Tr}(\cdot)$, is the sum of the diagonal elements of the matrix, and thus it is a function that takes as input a matrix and returns as output a number.

(a) Show that the trace is a linear transformation, i.e., that it satisfies the linearity conditions in the definition of a linear transformation.

(b) Show that, if the sum is defined, then $\text{Tr}(A + B) = \text{Tr}(A) + \text{Tr}(B)$.

(c) Show that, if both products are defined, then $\text{Tr}(AB) = \text{Tr}(BA)$.

115. REV:

Determine the spectral and Frobenius norm of each of the following matrices.

(a) $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$

(b) $\begin{pmatrix} -1 & 7 \\ 2 & 5 \end{pmatrix}$

(c) $\begin{pmatrix} 3 & 4 \\ 4 & 5 \end{pmatrix}$

(d) $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

(e) $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$

116. OK:

Let A be a positive definite matrix.

(a) Show that A^T is a positive definite matrix.

(b) Show that A^{-1} is a positive definite matrix.

117. REV:

Let U be an $n \times n$ orthogonal matrix, and let $x, y \in \mathbb{R}^n$.

- (a) Show that the dot product between Ux and Uy equals the dot product between x and y .
 (b) Show that the angle between Ux and Uy equals the dot product between x and y .

118. REV:

Which of the following matrices are orthogonal? Justify your answer.

$$A = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & 1 \\ -2 & 1 & 0 \end{pmatrix} \quad B = \begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{3} & -1/\sqrt{2} & 0 \\ 1/\sqrt{3} & 0 & -1/\sqrt{2} \end{pmatrix} \quad C = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & -1 & 1 & 0 \\ 0 & 1 & 1 & -1 \\ 1 & 0 & -1 & -1 \end{pmatrix}$$

119. EXAM:

Determine symmetric matrix representations for the following real quadratic forms.

- (a) $3x_1^2 + 3x_2^2 + 5x_3^2 + 2x_1x_2 - 2x_1x_3 - 2x_2x_3$
 (b) $2x_1^2 + 2x_2^2 + 5x_3^2 + 4x_1x_2 - 2x_1x_3 - 2x_2x_3$
 (c) $9x_1^2 + 6x_2^2 + 9x_3^2 + 6x_4^2 - 6x_1x_2 - 6x_1x_4 + 6x_2x_3 - 6x_3x_4$
 (d) $x_1^2 + 3x_2^2 + 3x_3^2 + x_4^2 - 2x_1x_2 + 4x_1x_3 - 2x_1x_4 + 8x_2x_3 + 4x_2x_4 + 2x_3x_4$

120. EXAM:

Construct the SVD of the following matrices.

(a) $\begin{pmatrix} 1 & 1 & 3 \\ 1 & 1 & 3 \end{pmatrix}$.

(b) $\begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$.

(c) $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$.

(d) $\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

121. TODO:

((Schaum-Bronson, 21.35, HARD WITHOUT COMPUTER?))

Construct the SVD of

$$\begin{pmatrix} 7 & -9 \\ 4 & -6 \end{pmatrix}.$$

122. EXAM:

Let A^+ be the generalized inverse of A . (Recall that if $A = U\Sigma V^T$ is the thin SVD of A , where Σ is square with all diagonal elements strictly positive, then $A^+ = V\Sigma^{-1}U^T$.)

- (a) Prove that, if A is nonsingular, then $A^+ = A^{-1}$.
 (b) Prove that, if 0 is an $n \times n$ all-zeros matrix, then $0^+ = 0$.
 (c) Prove that $(A^+)^+ = A$.
 (d) Prove that, if $k \neq 0$, then $(kA)^+ = \frac{1}{k}A^+$.
 (e) Prove that, if A is a symmetric matrix, then A^+ is a symmetric matrix.

123. OK:

Find the best least squares solution \bar{x} to the equations $3x = 10$, $4x = 5$. What error ($\min \|Ax - b\|_2^2$ or $\min \|X\beta - y\|_2^2$ in the general case) is minimized in this particular case? Check that the error vector $\begin{pmatrix} 10 - 3\bar{x} \\ 5 - 4\bar{x} \end{pmatrix}$ is orthogonal to the column vector $\begin{pmatrix} 3 \\ 4 \end{pmatrix}$.

124. OK:

Write out the error $E^2 = E^2(x) = \|Ax - b\|_2^2$ and set to zero its derivatives with respect to x_1 and x_2 , if

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 3 \\ 4 \end{pmatrix}.$$

Compare the resulting equations with the normal equations $A^T A \bar{x} = A^T b$, confirming that the derivation by calculus and the derivation by geometry gives the normal equations. Find the solution \bar{x} and the projection $p = A\bar{x}$. Why is $p = b$? Do the same for

$$b = \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix}.$$

Why does $p \neq b$?

125. TODO:

((Strang, 3.3.5))

XXX.

126. TODO:

((Strang, 3.3.7))

XXX.

127. TODO:

((Strang, 3.3.10))

XXX.

128. TODO:

((Strang, 3.3.11))

XXX.

129. TODO:

((Strang, 3.3.12))

XXX.

130. TODO:

((Strang, 3.3.13))

XXX.

131. TODO:

((Strang, 3.3.14))

XXX.

132. TODO:

((Strang, 3.3.15))

XXX.

133. TODO:

((Strang, 3.3.16))

XXX.

134. TODO:

((Strang, 3.3.19))

XXX.

135. TODO:

((Strang, 5.1.1))

XXX.

136. TODO:

((Strang, 5.1.3))

XXX.

137. TODO:

((Strang, 5.1.5))

XXX.

138. TODO:

((Strang, 5.2.3))

XXX.

139. TODO:

((Strang, 5.1.7prime))

XXX.

140. EXAM:

Show that the quadratic $q(x_1, x_2) = x_1^2 + 4x_1x_2 + 2x_2^2$ has a saddle point at the origin, even though the coefficients are all positive. Hint: do a variable transformation to rewrite q as a difference of two squares.

141. REV:

Determine the definiteness (positive definite, positive semidefinite, negative definite, negative semidefinite, indefinite) of the following matrices, and write out the corresponding quadratic form $q(x) = x^T Ax$.

$$(a) \begin{pmatrix} 1 & 3 \\ 3 & 5 \end{pmatrix}$$

$$(b) \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

$$(c) \begin{pmatrix} 2 & 3 \\ 3 & 5 \end{pmatrix}$$

$$(d) \begin{pmatrix} -1 & 2 \\ 2 & -8 \end{pmatrix}$$

If the determinant of any of these matrices equals zero, then along what line is $q(x)$ zero?

142. EXAM:

Directly, i.e., without computing the EVD, show that, if $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ is positive definite, then A^{-1} is positive definite.

143. EXAM:

Under what conditions on a , b , and c is $ax_1^2 + 2bx_1x_2 + cx_2^2 \geq x_1^2 + x_2^2$, for all x_1 and x_2 .

144. EXAM:

If $A = Q\Lambda Q^T$ is symmetric positive definite, then one can define the matrix $R = Q\Lambda^{1/2}Q^T$ to be the *symmetric positive definite square root*. Why does R have real eigenvalues? Compute R and verify that $R^2 = A$ for

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} 10 & -6 \\ -6 & 10 \end{pmatrix}.$$

145. OK:

Let A be a positive definite matrix, i.e., where $A = R^T R$ for a square matrix R . Prove the generalized Cauchy-Schwarz inequality:

$$|x^T A y|^2 \leq (x^T A x) (y^T A y).$$

146. OK:

The ellipse $x_1^2 + 4x_2^2 = 1$ corresponds to the matrix $A = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$. Write down the eigenvalues and eigenvectors, and sketch the ellipse.

147. EXAM:

Transform the equation $3x_1^2 - 2\sqrt{2}x_1x_2 + 2x_2^2 = 1$ to a sum of squares by finding the eigenvalues and eigenvectors of the corresponding A , and sketch the conic section.

148. REV:

In \mathbb{R}^3 , $\lambda_1x_1^2 + \lambda_2x_2^2 + \lambda_3x_3^2 = 1$ represents an ellipsoid, when $\lambda_i > 0$, for all $i \in \{1, 2, 3\}$. Describe all the different kinds of surfaces that can appear in the positive semidefinite case when one or more of the eigenvalues is zero.

149. TODO:

((Strang, 6.3.1))

XXX.

150. TODO:

((Strang, 6.4.1))

XXX.

151. REV:

Find the SVD and generalized inverse of

$$(a) A = \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix}$$

$$(b) B = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

$$(c) C = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$$

152. EXAM:

Let Q be an $m \times n$ matrix, where $m > n$, that has orthonormal columns. What is the pseudoinverse Q^+ ?

153. OK:

Let $A = \frac{1}{\sqrt{10}} \begin{pmatrix} 10 & 6 \\ 0 & 8 \end{pmatrix}$.

(a) Compute $A^T A$, and its eigenvectors and eigenvalues, and its positive definite square root.

(b) Compute AA^T , and its eigenvectors and eigenvalues, and its positive definite square root.

(c) Compute the SVD of A .

154. OK:

Compute the minimum length least squares solution $x^* = A^+ b$, to

$$Ax = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ 2 \end{pmatrix}.$$

Solve this two ways: first, by computing A^+ and multiplying the right hand side by it; and second, by finding the solution via the normal equations and choosing the solution that is in the row space of A . Confirm that both approaches lead to the same answer.

17.2 From putting together s18 midterm

Possible Review Questions.

1. Which of the following subsets of \mathbb{R}^3 are subspaces? Justify your answer.
 - (a) The plane of vectors with first component $b_1 = 0$.
 - (b) The plane of vectors with first component $b_1 = 1$.
 - (c) The set of vectors b with $b_1 b_2 = 0$ (which is the union of two subspaces, the plane $b_1 = 0$, and the plane $b_2 = 0$).
 - (d) The single vector $b = (0 \ 0 \ 0)$.
 - (e) All linear combinations of the two vectors $x = (1 \ 1 \ 0)$ and $y = (2 \ 0 \ 1)$.
 - (f) The vectors $b = (b_1 \ b_2 \ b_3)$ that satisfy $b_3 - b_2 + 3b_1 = 0$.
2. Prove that if A is a linear transformation, say from \mathbb{R}^3 to \mathbb{R}^3 , then A^2 is a linear transformation.
3. Consider the vectors $x = (1 \ 4 \ 0 \ 2)$ and $y = (2 \ -2 \ 1 \ 3)$. Find the L_1 , L_2 , and L_∞ norms of these vectors, and compute the inner product between them.
4. From analytic geometry, two lines in the plane are orthogonal when the product of their slopes equals -1 . Apply this to the vectors $x = (x_1 \ x_2)$ and $y = (y_1 \ y_2)$, whose slopes are x_2/x_1 and y_2/y_1 , to derive the orthogonality condition $x^T y = 0$.
5. For $x, y \in \mathbb{R}^n$, show: that $x - y$ is orthogonal to $x + y$ if and only if $\|x\|_2 = \|y\|_2$. (Equivalently, show: if $x - y$ is orthogonal to $x + y$, then $\|x\|_2 = \|y\|_2$; and if $\|x\|_2 = \|y\|_2$, then $x - y$ is orthogonal to $x + y$.)
6. Find the matrix that projects every point from \mathbb{R}^2 onto the line $x_1 + 2x_2 = 0$.
7. The *trace* of a matrix is the sum of the diagonal elements of a matrix. Show that, for any non-zero vector $a \in \mathbb{R}^n$, the trace of the projection matrix $P_a = aa^T/a^T a$ equals one.
8. Compute a QR decomposition of A , and use this to compute the projection of b onto the span of the columns of A :

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

If p is the projection of b onto the span of the columns of A , verify that $b - p$ is orthogonal to the two columns of A ; and also that p can be expressed as a linear combination of the two columns of A .

9. What 2×2 matrix projects the x_1 - x_2 plane onto the -45° line $x_1 + x_2 = 0$?
10. Apply the Gram-Schmidt process to the vectors

$$a = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \quad c = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

and write the result in the form $A = QR$.

11. (a) Form an orthonormal set of vectors q_1, q_2, q_3 , for which q_1, q_2 span the same space as the columns of

$$A = \begin{pmatrix} 1 & 1 \\ 2 & -1 \\ -2 & 4 \end{pmatrix}.$$

- (b) Let $b = (1 \ 2 \ 7)^T$, and let $p_i \in \mathbb{R}^3$ be the projection of b onto q_i , for $i = 1, 2, 3$. Verify that $\|b\|_2^2 = \sum_{i=1}^3 \|p_i\|_2^2$.

12. If $A = QR$, find a simple formula for the projection matrix P_A onto the column space of A , and for the projection matrix P_{A^\perp} onto the set of vectors that are orthogonal to the column space of A .
13. For what values of a do the matrices

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix} \quad \text{satisfy} \quad AB = BA?$$

14. What is the inverse of the matrix $A = \begin{pmatrix} a & b \\ 0 & a \end{pmatrix}$, assuming $a \neq 0$?
15. The vectors $\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ form an orthogonal basis for \mathbb{R}^2 . Use this to form an orthonormal basis for \mathbb{R}^2 .

Possible Exam Questions.

1. Consider the four vectors

$$u = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad v = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \quad w = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \text{and} \quad b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}.$$

- (a) Give the values of the elements of a vector $x \in \mathbb{R}^3$ such that $x_1u + x_2v + x_3w = b$.
- (b) Show that if we define a 3×3 matrix A , where the first/second/third column is $u/v/w$, then this vector x is such that $Ax = b$.
2. If $A = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$ and $B = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$, then compute $A^T B$, $B^T A$, AB^T , and BA^T .
3. Recall the two requirements for a vector space (vector addition and scalar multiplication). Show that these two requirements are different by constructing:
- (a) a subset of \mathbb{R}^2 that is closed under vector addition and even subtraction, but not under scalar multiplication;
- (b) a subset of \mathbb{R}^2 (other than two opposite quadrants) that is closed under scalar multiplication but not under vector addition.
4. Prove that if any diagonal element of

$$A = \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix}.$$

equals zero, then the rows of A are linearly dependent.

5. For each of the following two matrices,

$$A = \begin{pmatrix} 1 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 6 \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} 2 & -2 \\ 2 & -2 \end{pmatrix},$$

find vectors u and v such that $A = uv^T$.

6. The matrix $A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ produces a *stretching* in the x_1 direction. Draw the circle $x_1^2 + x_2^2 = 1$, and sketch around it the points $(2x_1, x_2)^T$ that result from multiplication by A . What shape is that curve?

7. The matrix $A = \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix}$ produces a *shearing* transformation, which leaves the x_1 -axis unchanged. Sketch its effect on the x_2 -axis, by indicating what happens to $(1, 0)^T$ and $(2, 0)^T$ and $(-1, 0)^T$, as well as how the entire axis is transformed. XXX. THERE IS A TYPO IN THIS, I MODIFIED FROM STRANG AND DIDNT MAKE ALL THE CHANGES CORRECTLY.
8. Give an example in \mathbb{R}^2 of two linearly independent vectors that are not orthogonal. Also, give an example of two orthogonal vectors that are not linearly independent.
9. (a) Find the projection matrix P_1 onto the line through $a = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ and also the matrix P_2 that projects onto the line orthogonal to a .
- (b) Compute $P_1 + P_2$ and $P_1 P_2$, and explain in words what each of these matrices does.

10. Compute a QR decomposition of A , and use this to compute the projection of b onto the span of the columns of A :

$$A = \begin{pmatrix} 1 & 1 \\ 2 & -1 \\ -2 & 4 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 2 \\ 7 \end{pmatrix}.$$

If p is the projection of b onto the span of the columns of A , verify that $b - p$ is orthogonal to the two columns of A ; and also that p can be expressed as a linear combination of the two columns of A .

11. Let $b_1 \in \mathbb{R}^3$ and $b_2 \in \mathbb{R}^3$ be the projection of the vector $b = \begin{pmatrix} 1 & 2 \end{pmatrix}$ onto two vectors that are not orthogonal, $a_1 = \begin{pmatrix} 1 & 0 \end{pmatrix}$ and $a_2 = \begin{pmatrix} 1 & 1 \end{pmatrix}$. Show that, unlike in the case where the two vectors being projected onto are orthogonal, here $b \neq b_1 + b_2$.
12. Suppose that we are given the vectors

$$a = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad c = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.$$

Using the QR factorization applied to a , then b , then c , find three orthonormal vectors q_1, q_2, q_3 .

Some Other Questions.

1. (SOURCE: STRANG 1.4.14.)
By trial and error find examples of 2×2 matrices such that:
- (a) A matrix A with only real entries such that $A^2 = -I$;
 - (b) A matrix $B \neq 0$ such that $B^2 = 0$;
 - (c) Matrices C and D such that $CD = -DC \neq 0$;
 - (d) Matrices E and F such that $EF = 0$, even though no entries of E or F are zero.
2. (SOURCE: STRANG 1.4.16 AND HUBBARD 1.3.19)
Let A and B be two $n \times n$ matrices.

- (a) Recall the usual rule for the product of two matrices yields an $n \times n$ matrix, where an entry of the product matrix equals

$$(AB)_{ij} = \sum_k A_{ik} B_{kj}.$$

Show that this definition of matrix multiplication is associative (i.e., $(AB)C = A(BC)$) but not commutative (i.e., $AB \neq BA$).

- (b) A different notion of matrix multiplication is known as the *Jordan product* and can be defined in terms of the usual matrix product as

$$\frac{AB + BA}{2}.$$

Show that this product is commutative but not associative.

3. (SOURCE: STRANG 1.4.21.)

XXX. TO DO.

4. (SOURCE: STRANG 1.4.23.)

Consider the matrices

$$A = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \text{and} \quad C = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix}.$$

Find all powers of A^2 , A^3 (which is A^2 times A), ..., and B^2 , B^3 , ..., and C^2 , C^3 ,

5. (SOURCE: STRANG 2.1.5b.)

XXX. TO DO.

6. (SOURCE: STRANG 2.3.4.)

XXX. TO DO.

7. (SOURCE: STRANG 2.3.11.)

XXX. TO DO.

8. (SOURCE: STRANG 2.4.14.)

XXX. TO DO.

9. (SOURCE: STRANG 2.6.7.)

What 3×3 matrices represent the transformations that

(a) project every vector onto the x_1 - x_2 plane?

(b) reflect every vector through the x_1 - x_2 plane?

(c) rotate the x_1 - x_2 plane through 90° , leaving the x_3 -axis along?

(d) rotate the x_1 - x_2 plane, then the x_1 - x_3 plane, then the x_2 - x_3 plane, all through 90° ?

(e) carry out the same three rotations, but through 180° ?

10. (SOURCE: STRANG 3.1.6.)

In \mathbb{R}^3 , find all vectors that are orthogonal to $(1 \ 1 \ 1)$ and $(1 \ -1 \ 0)$. Construct from these three vectors an orthonormal basis for \mathbb{R}^3 .

11. ((Strang, 3.3.16))

XXX.

12. (SOURCE: STRANG 3.4.6.)

Find a third column so that the matrix

$$Q = \begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{14} & \cdot \\ 1/\sqrt{3} & 2/\sqrt{14} & \cdot \\ 1/\sqrt{3} & -3/\sqrt{14} & \cdot \end{pmatrix}$$

is a matrix with orthonormal columns. Since this vector must be a unit vector that is orthogonal to the other two columns, how much freedom does this leave? Verify that for this full 3×3 matrix, the rows are also orthonormal.

13. (SOURCE: MODIFIED FROM STRANG 3.4.7.)

Let $\{q_i\}_{i=1}^n$ be any set of orthonormal vectors, and let $\{x_i\}_{i=1}^n$ be any set of real numbers. Show, by forming $b^T b$ directly, that Pythagoras' law holds for any combination $b = x_1 q_1 + x_2 q_2 + \dots + x_n q_n$ of orthonormal vectors: $\|b\|_2^2 = x_1^2 + x_2^2 + \dots + x_n^2$.

14. (SOURCE: HUBBARD 1.2.5.)

Let A and B be $n \times n$ matrices, with A symmetric. Are the following true or false? Justify your answer.

- (a) $(AB)^T = B^T A$
- (b) $(A^T B)^T = B^T A^T$
- (c) $(A^T B)^T = BA$
- (d) $(AB)^T = A^T B^T$

15. (SOURCE: HUBBARD 1.2.9.)

given the two matrices $A = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$ and $B = \begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix}$:

- (a) What are their transposes?
- (b) Without computing AB , what is $(AB)^T$?
- (c) Confirm your result by computing AB and taking the transpose.
- (d) What happened if you do part (15b) using the *incorrect* formula $(AB)^T = A^T B^T$?

16. (SOURCE: HUBBARD 1.2.15.)

Show that

$$\begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix} \text{ has an inverse of the form } \begin{pmatrix} 1 & x & y \\ 0 & 1 & z \\ 0 & 0 & 1 \end{pmatrix}$$

and find it.

17. (SOURCE: HUBBARD 1.3.18.)

Rotate the unit square by 45° counterclockwise, then stretch it using the linear transformation $\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$. Sketch the result.

18. (SOURCE: ADAPTED FROM HUBBARD 1.4.2 AND HUBBARD 1.4.4.)

- (a) Consider the vectors

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} \text{ and } \begin{pmatrix} \sqrt{2} \\ \sqrt{7} \end{pmatrix}.$$

Find the L_1 , L_2 , and L_∞ norms of these vectors, and compute the inner product between them.

- (b) Do the same for the vectors

$$\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}.$$

19. (SOURCE: ADAPTED FROM HUBBARD 1.4.5.)

Compute the angle between the following pairs of vectors.

(a) $\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

(b) $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$.

$$(c) \begin{pmatrix} 1 \\ 0 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

20. (SOURCE: HUBBARD 1.4.19.)

For a given n , define the vector $v_n = e_1 + \dots + e_n \in \mathbb{R}^n$, where e_i is the i^{th} standard basis vector in \mathbb{R}^n .

(a) What is the L_1 , L_2 , and L_∞ norm of v_n ?

(b) What is the angle θ_n between v_n and e_1 ? What is $\lim_{n \rightarrow \infty} \theta_n$?

21. (SOURCE: HUBBARD 1.4.23.)

XXX. TO DO.

22. (SOURCE: HUBBARD 1.4.25.)

XXX. TO DO.

23. (SOURCE: HUBBARD 1.4.26.)

XXX. TO DO.

24. (SOURCE: HUBBARD 2.4.1.)

Show that the standard basis vectors are linearly independent.

25. (SOURCE: ADAPTED FROM HUBBARD 2.4.7.)

Let A be an $n \times n$ matrix. Show that A is orthogonal if and only if $A^T A = I$. In this case, show that both its rows as well as its columns form an orthonormal basis for \mathbb{R}^n , i.e., that $AA^T = I$ also.

17.3 XXX. MORE.

Problems not assigned in 2016:

1. Maybe parts of JL proof:

- Expectations, variances
- with constant prob, with δ high probability, with probability $1 - \frac{1}{n^2}$

2. (Maybe put this here.) Consider the following three graphs:

- A line graph consisting of n nodes, where each node is on a line and there are edges between nearest neighbors.
- A random graph consisting of n nodes, where each edge between any two possible nodes is present with probability 0.1.
- A clustered graph, consisting of n nodes, where $n/2$ of the nodes connect to each other randomly with probability 0.1, the other $n/2$ nodes connect with probability 0.1, and the nodes in the two clusters connect to each other with probability 0.01.

Do the following:

- Visualize the adjacency matrix of each of these three graphs. (Q: What is the command, in Matlab it is something like a spy plot.)
- Compute the adjacency matrix to higher powers and then apply it to an arbitrary vector; and also apply it to the same arbitrary vector recursively.
- (I'd like to do this for directed graphs.)

3. (*Probably skip. Not in 2016.*)

(Taubes 2.6.3)

Let S denote the set $\{1, 2, \dots, 10\}$.

(a) Write down three different probability functions on S by giving the probabilities that they assign to the elements of S .

(b) Write down two functions S whose values can not be those of a probability function, and explain why such is the case.

4. (*Probably skip. Not in 2016.*)

(Taubes 2.6.3)

Let S denote the set $\{1, 2, \dots, 10\}$.

(a) Write down three different probability functions on S by giving the probabilities that they assign to the elements of S .

(b) Write down two functions S whose values can not be those of a probability function, and explain why such is the case.

5. (*Probably skip. Not in 2016.*)

(Taubes 3.7.3)

Label the four bases that are used by deoxyribonucleic acid (DNA) as $\{1, 2, 3, 4\}$. (More commonly, they are called $\{A, C, G, T\}$, but we will use $\{1, 2, 3, 4\}$ here.)

(a) Given this labeling, write down the sample space for the possible bases at two given sites along the DNA molecule.

(b) Invent a probability distribution for this sample space.

(c) Let A_j , for $j = 1, 2, 3, 4$, denote the event in this two-site sample space that the first site has base j , and let B_j , for $j = 1, 2, 3, 4$, denote the analogous event for the second site. Use the definition of conditional probability to explain why, for *any* probability distribution and for any k , we have that

$$\Pr[A_1|B_k] + \Pr[A_2|B_k] + \Pr[A_3|B_k] + \Pr[A_4|B_k] = 1.$$

(d) Is there a choice for a probability function on the sample space that makes $\Pr[A_1|B_1] = \Pr[B_1|A_1]$ in the case that A_1 and B_1 are not independent? If so, give an example. If not, explain why not.

6. (*Probably skip. Not in 2016.*)

(Taubes 3.7.4)

Let's say that I have six coins, labeled $\{1, 2, 3, 4, 5, 6\}$, where the i^{th} coin has a probability of 2^{-1} of landing H when flipped and a probability $1 - 2^{-i}$ of landing T ; and the pathological six-sided dice, where the probability of the i^{th} face appearing when rolled equals $\frac{i}{21}$. Let's say that I first roll the dice, and if the i^{th} face appears then I flip the coin with label i . I don't tell you which coin was flipped, but I do tell you that H appeared.

(a) For $i = 1, 2, 3, 4, 5, 6$, give the probability that the coin with the label i was flipped.

(b) For what i , if any, is the event that the i^{th} face appears independent from the event that H appears.

7. (*Probably skip. Not in 2016.*)

(Taubes 6.9.4)

Suppose that N is a positive integer and that N selections are made from the three element set $-1, 0, +1$. Assume that these are done independently so that the probability of any one number arising on the k^{th} selection is independent of any given number arising on any other selection. Suppose, in addition, that the probability of any given number arising on any given selection is $1/3$.

(a) How many elements are in the sample space for this problem?

(b) What is the probability of any given element?

- (c) Let f denote the random variable that assigns to any given (i_1, \dots, i_N) their sum, i.e., $f = i_1 + \dots + i_N$. What are $\Pr[f = N]$ and $\Pr[f = N - 1]$? What are $\Pr[f = 0]$ and $\Pr[f = 1]$?
8. Let $A = \{1, 2, 3, 4, 5\}$, and let A_3 be the set of 3-element subsets of A .
- How many elements does A_3 have?
 - What are all possible elements of A_3 ?
 - If A instead has n distinct elements, then what is the size of the subset A_k , where $k \leq n$.

Chapter 18

Additional Material to Incorporate Somewhere

18.1 Notes of things to cover or fix in the future

While preparing Chapter 1:

- Need to do a better job being self contained and introductory.
- Show that an identity matrix doesn't change things, and that it can have different forms, depending on what it is applied to. In particular, if we always work with mean-centered matrices, then it takes a special form, and if we always work with matrices in a subspace then it takes a special form. Both of these can be written in terms of the generalized inverse of projection.

While preparing Chapter 2:

- Jim Lambers notes, including matrix norms and error analysis.
- Matrix norms are vector norms, but they also satisfy submultiplicativity. This is related to matrices being vectors but also satisfying something else related to functional properties.
- Matrix spectral norm is L_∞ norm on singular value vector, and it is vector induced with a quadratic form, so it is an eigenvalue.
- Matrix Frobenius norm is L_2 norm on singular value vector, and it is L_2 norm of the vector viewed as a matrix.
- Fix up some Ch 2 homework problems to have rotated versions, etc.
- Maybe put 3D versions of some of those problems in later chapters.

Also, incorporate or point to these.

- See <https://ontopo.wordpress.com/2009/03/03/reasoning-in-higher-dimensions-hyperspheres/>
- See <https://divisbyzero.com/2010/05/09/volumes-of-n-dimensional-balls/>

18.2 Ideas to take from other books

The book Rencher and Schaalje (Linear models in statistics) has a bunch of things.

- In Ch 2, there is a nice explicit explanation of partitioned matrices that I need to include.
- There are other things like on page 24 they make it explicit that using symmetric matrix with a quadratic form amounts to working with $\frac{1}{2}(A + A^T)$ instead of A .
- Page 49 has functions of a matrix.
- Page 56, Sec 2.14, is discussion of vector and matrix calculus. Make sure that there is nothing there that I don't already include.
- Page 77, Sec 3.3.4 had standardized/Mahalanobis distance and lots of other stuff in Chapter 3 that I need to include.
- I can take stuff from Ch 7 on multiple linear regression.

Here are some ideas from the Diaconis-Skyrms book.

- On p 12, they describe WLLN as a finite n statement (which is really a limiting statement definition in disguise). Maybe I want to present it this way.
- On p 16, they say that basic probability models include: balls from urns; flipping coins; rolling dice; and shuffling cards. I should have a section where I explain these models and why we use them.
- On p 21 there is an easy proof of the birthday paradox.
- They use Propositions. Maybe I should do that instead of Lemmas and Theorems.
- From page 215, maybe have a HW problem something like:
We know that if two events A and B are independent, then $\Pr[A|B] = \Pr[A]$, etc. Construct an example of a state space, probability, and events, Ω , $\Pr[\cdot]$, A , and B where $\Pr[A|B] = \Pr[A]$ but where A and B are not independent.
- Maybe use the urn example on page 215-216. Especially if I say that other models are similar to coins and dice.
- On p 211, they define a “standard model.” What is standard about this? Is it just that the probabilities are uniform and add up? Then on p 216 they describe random variables as a simple extension of the standard model. Why is that?
- On p 217, they have an example of what is basically a stochastic process where they can potentially use past information in different ways.
- On p 219, they define martingales. On p 219, they also give two references (Grinstead-Snell and also Lang) that they say discuss martingales within the scope of elementary probability. Maybe I should do this somewhere as a forward pointer to MCMC somewhere.

Here are some ideas from the Hirsch-Smale-Devaney book.

- Maybe emphasize how we will do into detail on up to 4×4 matrices. This would be especially good if I can compute eigenstuff and linear equation solves for that.
- Follow page 27, showing a 2×2 example of $Ax = b$, and show that when RHS is zero the there are infinitely many solutions of the matrix equation when the determinant equals zero, and use this to motivate more complices stuff for $n \times n$ matrices.

- Be explicit like on page 49 that everything boils down to a few simple types of matrices. But do a better job going beyond 2×2 matrices that show this to show that this same fact is embedded in 3×3 and 4×4 matrices.
- Like on p 62, have a section that presents the trace-determinant formula. Can we get the analogous result for 3×3 matrices.
- Also, I like how they present on page 62 the results in the trace-determinant plane, where they show that there are real distinct, real repeated, or complex roots. I should be able to relate this to the proof of the Cauchy-Schwarz result.
- Like on page 77, give those two examples of matrices not commuting; and also give that example of matrix cancellation not being permitted. Question: Is the latter simply that we can't invert the matrix A , or is there something else going on with B and C ?
- Have a separate section entitled something like “Repeated eigenvalues” or “Repeated and non-repeated eigenvalues.”
- Have a section at the end on matrix functions, e.g., the exponential function, as well as show that some algorithms such as power method and MCMC are matrix functions.

18.3 Additional Miscellaneous Stuff Maybe To Put Somewhere

18.3.1 Misc Stuff

Describe normalizing constants: https://en.wikipedia.org/wiki/Normalizing_constant E.g., for probability distributions and conditional probability distributions, as well as unit ball normalization and other things. Have as a separate section or box.

Here are some things to get back to as the semester proceeds.

- Have a HW question where we ask them to prove that the Euclidean norm is a norm.
- Have a HW question where we ask them to prove the norm equivalence properties for the Euclidean norm.
- Have a HW question like: if B and C are PD, then show that $C \geq B$ iff $B^{-1} \geq C^{-1}$.
- Have some HW question on the Mahalanobis distance.

18.3.2 Some higher-level things that will take some thought

Here are some additional things I would like to weave in at the appropriate place.

Graphs. Strand and Hubbard and presumably many others have nice intros to graphs. Can I do it simply, but also illustrate small-world graphs, spectral clustering/classification methods, and Markov chains for high-dimensional integration.

Ask a Question. As a pedagogical point: “start with a question”. Can I start different things with a questions.

Sinusoids and exponentials. Somewhere it would be nice to show that these are not orthogonal, nor are their discretized versions, and that PCA gives some funny combination of them. I can motivate this with two types of responses for bacteria, as well as IQ tests, which would be a good place to discuss reification and social policy issues.

Applications. Can I have an application for each step, e.g., bacteria; social policy of IQ; measure importance of web page with number of terms on page with or without TFIDF; number of links to a given page; and PageRank.

18.4 Additional Miscellaneous Stuff Probably To Remove

18.4.1 Some philosophy

XXX. THIS IS AN EXTRA SET OF NOTES. SOME OF THIS MATERIAL WAS NOT COVERED IN CLASS, AND MOST OF IT IS STILL WRITTEN IN TOO ADVANCED A MANNER AT THIS POINT.

The spectra of matrices—the set of eigenvalues and eigenvectors—are used in many ML and DA applications.

Given an arbitrary matrix $A \in \mathbb{R}^{n \times n}$, the eigenvectors and eigenvalues are given by solving $Ax = \lambda x$, i.e., they are the directions that are unchanged by the operation of the matrix and the amount that a vector in that direction gets scaled along that direction. Importantly,

- They may not exist.
- They may not be real numbers.
- They may not be stable.
- They may not be orthogonal to each other.

A lot of work in more advanced linear algebra classes deals with these issues.

Importantly, these issues almost never arise in data science applications. The reason is that symmetric matrices are very “nice” in several related ways. For a symmetric matrix $A \in \mathbb{R}^{n \times n}$,

- There are n eigenvalues, including multiplicity, and all n are real.
- There are n orthogonal eigenvectors.
 - Eigenvectors corresponding to different eigenvalues are orthogonal.
 - If the multiplicity of a given eigenvalue is m , then we can choose m orthogonal eigenvectors corresponding to that eigenvalue.

So, basically, if you have a symmetric matrix A , then in addition to the canonical basis, i.e., the basis given by $\{e_i\}_{i=1}^n$, we have a basis from the eigenvectors that is “adapted” to the matrix. That is, when you see a symmetric matrix (and when you have a non-symmetric matrix, you can consider the variance-covariance matrix, which is symmetric), let’s say it is an $n \times n$ matrix, then you should immediately think that the matrix defines a complete orthonormal basis for \mathbb{R}^n . That matrix may be hard for you to write down, but it is easy to think about in linear algebraic and geometric terms, and it is easy for a computer to work with.

Let’s go into this in a bit more detail. The most familiar basis is the canonical basis, which is given by

$$e_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$$

where the element 1 occurs in the i^{th} position, we also have the basis of eigenvectors. If we construct a matrix in which the i^{th} column is the vector, then we get the identity matrix:

$$I = \begin{pmatrix} & & \\ e_1 & \dots & e_n \\ & & \end{pmatrix} = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}.$$

Alternatively, we can take the n mutually-orthogonal eigenvectors and put them in a matrix

$$U = \begin{pmatrix} & & \\ u_1 & \dots & u_n \\ & & \end{pmatrix}$$

For the given matrix A , these n directions are special in that the action of A is simply to stretch the input vector. That is, if the matrix A is “written in” the basis given by it, it is a diagonal matrix, with diagonal entries given by the eigenvectors. So, if we rotate from the basis $\{e_i\}_{i=1}^n$ to the basis $\{u_i\}_{i=1}^n$, then A changes from A to $U^T A U = \Lambda$. Since that is a legitimate orthonormal basis, we can write any vector x in that basis, and that is good for many things.

For example, as follows

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix} = \sum_{i=1}^n x_i e_i = \sum_{i=1}^n \alpha_i u_i,$$

XXX WHERE α_i IS SUCH AND SUCH.

The expressions $A v_i = \lambda_i v_i$, $\forall i \in [n]$ can be written in matrix notation as $AU = U\Lambda$. Pictorially, it looks like the following.

$$\begin{aligned} \begin{pmatrix} & & \\ A & & \end{pmatrix} \begin{pmatrix} & & \\ u_1 & \dots & u_n \end{pmatrix} &= \begin{pmatrix} & & \\ u_1 & \dots & u_n \end{pmatrix} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \\ &= \begin{pmatrix} & & \\ \lambda_1 u_1 & \dots & \lambda_n u_n \end{pmatrix} \end{aligned}$$

Then, $AVV^T = V\Lambda V^T$, since $VV^T = I$. PUT THAT AS UNDERBAR.

All this holds for any symmetric matrix. But for some matrices

$$\lambda_i > 0 \quad \forall i \in [n].$$

For these matrices

$$x^T A x > 0 \quad \forall x$$

and these matrices are known as correlation/covariance matrices.

18.5 XXX REVIEWS FROM PREVIOUS WEEKS/CHAPTERS

Here are reviews from probability chapters that I have in S17, as well as the review of the chapter on quadratic forms. I can remove if I cover everything with proper notation, etc.

18.5.1 Review from last chapter

Here is a summary of what we covered last time.

- A *sample space* is the set of all possible outcomes of an experiment.
- An *event* is a subset of the sample space.
- **Definition 85** Given a sample space Ω with events A and B , a probability function is an assignment of numbers to events such that:
 1. $0 \leq \Pr[A] \leq 1$.
 2. $\Pr[\Omega] = 1$.
 3. $\Pr[A \cup B] = \Pr[A] + \Pr[B]$ if $A \cap B = \emptyset$.
- Here are some facts that are easy to prove about probability functions:
 1. $\Pr[\emptyset] = 0$
 2. $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$ (for any A and B)
 3. $\Pr[A] \leq \Pr[B]$ if $A \subset B$
 4. $\Pr[B] = \Pr[B \cap A] + \Pr[B \cap A^C]$
 5. $\Pr[A^C] = 1 - \Pr[A]$

From these basic facts, one can prove all sorts of other useful things in probability.

18.5.2 Review from last chapter

Recall the following from last time.

- Given a sample space Ω , a *random variable* X on Ω is a real-valued function on Ω , i.e., $X : \Omega \rightarrow \mathbb{R}$. A *discrete random variable* is a random variable that takes on only a finite or countably infinite number of values, i.e., the domain of Ω is discrete (and thus so too is the range).
- For a discrete random variable, the event $X = a$, sometimes denoted $(X = a)$ or $\{X = a\}$, includes all the basic/elementary events of Ω where $X = a$, i.e., $\{X = a\} = \{s \in \Omega : X(s) = a\}$. In this case, $\Pr[X = a] = \sum_{s \in \Omega : X(s)=a} \Pr[s]$.
- Two events A and B , i.e., two subsets of Ω , are *independent* if $\Pr[A \cap B] = \Pr[A]\Pr[B]$. Two random variables X and Y are *independent* if

$$\Pr[(X = x) \cap (Y = y)] = \Pr[(X = x)] \cdot \Pr[(Y = y)]$$

holds for all x and y in Ω . This is a very strong condition on X and Y , since it must hold for all $x, y \in \Omega$.

- The *mean/expectation* of a discrete random variable, denoted $\mu(X)$ or $\mathbf{E}[X]$ can be computed in one of two ways:

- $\mathbf{E}[X] = \sum_{s \in \Omega} X[s] \mathbf{Pr}[s]$, and
- $\mathbf{E}[X] = \sum_i i \cdot \mathbf{Pr}[X = i]$.

The *variance* of a random variable X is

$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2,$$

and the *standard deviation* of a random variable X is $\sigma[X] = \sqrt{\mathbf{Var}[X]}$.

18.5.3 Review from last chapter

Recall the following from last time.

- We defined the mean, conditional mean, covariance, standard deviation, and correlation of random variables, and we discussed several properties of those statistics. The mean, in particular, has properties like the linear combination properties that were satisfied by vectors in linear algebra. All of this permits us to make the following connections between *linear algebra* and *probability theory*:

length squared of a vector	\longleftrightarrow	variance
length of vector	\longleftrightarrow	standard deviation
dot product	\longleftrightarrow	covariance
cos of angle between vectors	\longleftrightarrow	correlation

- We discussed how to understand the peculiar properties of the high-dimensional Euclidean space \mathbb{R}^n , and in particular of high-dimensional balls and boxes, in terms of the measure concentration underlying the hopefully more intuitive process of flipping a fair coin n times.

18.5.4 Review of the chapter XXX ON QUADRATIC FORMS

Last time, we discussed quadratic forms and conic sections and the completing the square procedure as a way to describe symmetric matrices. As we discussed it, completing the square is essentially a not-fully-specified rule that allows us to define a new coordinate system or new set of variables that is more well-suited to describe a given symmetric matrix. Equivalently, this means that we are finding some set of linearly independent vectors that are linear combinations of our original canonical basis vectors that are more well-suited to describe a given symmetric matrix. In the same way that a diagonal matrix can be represented as the diagonal elements of that matrix (the eigenvalues) times outer products of the canonical basis vectors with themselves, in this new basis the matrix can be represented as a set of numbers (the eigenvalues) multiplied by the outer product of the new basis vectors with themselves. In that new basis, there are some terms with positive coefficients, some terms with negative coefficients, and some terms with zero coefficients.

The general point is that if we have an $n \times n$ symmetric matrix A , then we have a quadratic form, and when we consider the level sets of the quadratic form, then we have k directions that point up, ℓ directions that point down, and $n - (k + \ell)$ directions that are flat. In addition, while the different vectors we get might depend on how we complete the square to zero out elements, e.g., the order we choose, the numbers k and ℓ don't and instead depend only on the matrix.

We made this precise in a theorem that stated that for any quadratic form $Q(x)$ on \mathbb{R}^n , there exists $m = k + \ell$ linearly independent functions, call them $\alpha_1, \dots, \alpha_m$ such that

$$Q(x) = (\alpha_1(x))^2 + \dots + (\alpha_k(x))^2 - (\alpha_{k+1}(x))^2 - \dots - (\alpha_{k+\ell}(x))^2,$$

where the number k of terms with positive signs and the number ℓ of terms with minus signs depends only on Q and not on the specific linear function chosen. Given this decomposition into linearly independent

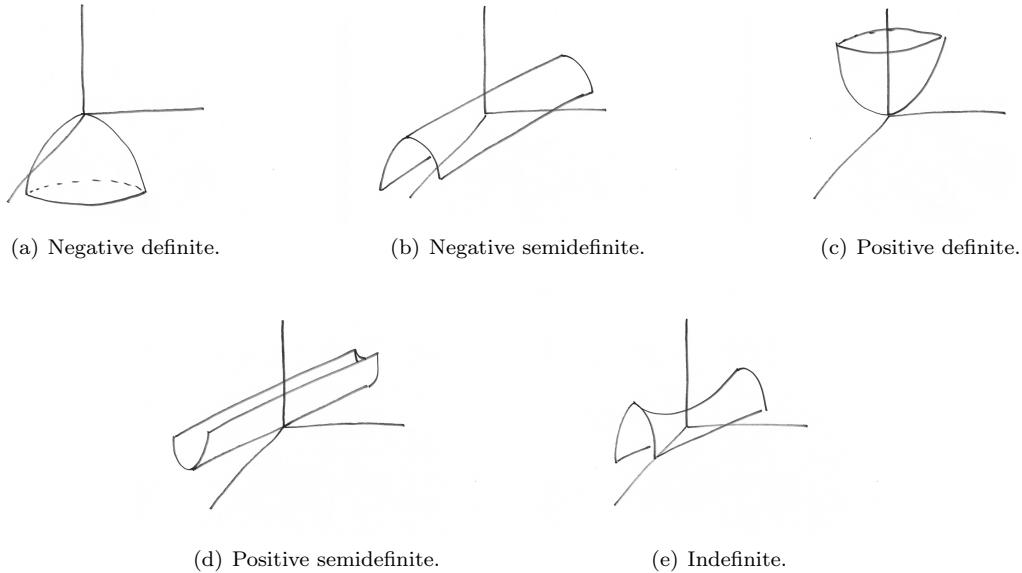


Figure 18.1: Illustration of negative definite, negative semidefinite, positive definite, positive semidefinite, and indefinite quadratic forms.

functions, we can classify symmetric matrices as follows. Let A be an $n \times n$ symmetric matrix, and recall that $Q(x) = x^T A x$ is the corresponding quadratic form. Then A (as well as Q) is

- negative definite if $x^T Ax < 0$, for all x .
(In this case, $k = 0$ directions have positive sign and point up, and $l = n$ directions have negative sign and point down.)
 - negative semidefinite if $x^T Ax \leq 0$, for all x .
(In this case, $k = 0$ directions have positive sign and point up, $l < n$ directions have negative sign and point down, and $n - l$ directions have zero coefficient and are flat.)
 - positive definite if $x^T Ax > 0$, for all x .
(In this case, $k = n$ directions have positive sign and point up, and $l = 0$ directions have negative sign and point down.)
 - positive semidefinite if $x^T Ax \geq 0$, for all x .
(In this case, $k < n$ directions have positive sign and point up, $l = 0$ directions have negative sign and point down, and $n - k$ directions have zero coefficient and are flat.)
 - indefinite if $x^T Ax$ is $>$ or $<$ zero, depending on x .
(In this case, $0 < k < n$ directions have positive sign and point up, $0 < l < n$ directions have negative sign and point down, and $n = k - l \geq 0$ directions have zero coefficient and are flat.)

Examples of each of these cases are given in Figure 18.1. When we look at a figure such as Figure 18.1, it is important to keep in mind the following: when we see a one-dimensional subspace that is flat or pointing upward or pointing down, you should really be thinking that that is potentially a higher-dimensional subspace, each direction of which behaves in qualitatively the same way. That is, rather than having just 1 dimension point up and 1 dimension point down (which recall are each a 1 dimensional subspace), in general we will have a k dimensional subspace along which the function curves up and an ℓ dimensional subspace along which the function curves down and a $n - (k + \ell)$ dimensional subspace along which the function is flat.

As an example, consider the following matrix:

$$A = \begin{pmatrix} -2 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}. \quad (18.1)$$

The eigenvalues of this matrix are $\{-2, -1, 0, 1, 2\}$, and the eigenvectors are the canonical axes, e_1, e_2, e_3, e_4, e_5 (respectively).

- If we consider $\text{span}\{e_1, e_2\}$, then the associated quadratic form is negative definite, and the corresponding picture is given by Figure 18.1(a).
- If we consider $\text{span}\{e_2, e_3\}$, then the associated quadratic form is negative semidefinite, and the corresponding picture is given by Figure 18.1(b).
- If we consider $\text{span}\{e_3, e_4\}$, then the associated quadratic form is positive semidefinite, and the corresponding picture is given by Figure 18.1(d).
- If we consider $\text{span}\{e_4, e_5\}$, then the associated quadratic form is positive definite, and the corresponding picture is given by Figure 18.1(c).
- If we consider $\text{span}\{e_5, e_1\}$, then the associated quadratic form is indefinite, and the corresponding picture is given by Figure 18.1(e).

All of those examples involve a span which is two-dimensional, but the same holds true more generally.

- If we consider $\text{span}\{e_1, e_2, e_3\}$, then the associated quadratic form is negative semidefinite, and the corresponding picture is given by Figure 18.1(b), except that the downward-pointing direction is a two-dimensional subspace rather than a one-dimensional subspace.
- If we consider $\text{span}\{e_2, e_3, e_4\}$, then the associated quadratic form is indefinite, and the corresponding picture is given by Figure 18.1(e), except that the flat direction which is perpendicular to both of the other coordinate axes is not illustrated.
- If we consider $\text{span}\{e_3, e_4, e_5\}$, then the associated quadratic form is positive semidefinite, and the corresponding picture is given by Figure 18.1(d), except that the upward-pointing direction is a two-dimensional subspace rather than a one-dimensional subspace.

The various sets of basis vectors that the procedure we described last time compute provides a basis for these different subspaces. Since a basis for a subspace is not unique (which is why the procedure, as we described it last time is not-fully-specified), so to the vectors we got were not unique.

Given all this, here we will consider several related questions.

- First, is there a basis that in some useful sense is the best basis?
- Second, if so, then how do we compute the best such bases (and this will go a long way to resolving the uniqueness issue)?
- Third, how is all of this used in machine learning and data science?

We will address each of these questions.

18.6 Advanced Aside: Discussion of other uses of the spectral decomposition

In addition to PCA, LS, and PR, there are many other uses of the spectral decomposition (in data science and machine learning, as well as more generally). We have mentioned and/or alluded to several of these. Here, we briefly describe several of them as well as several others.

18.6.1 Describing a matrix as a linear transformation

We can use the spectral decomposition to describe the action of a matrix, e.g., when viewing it as a linear transformation, in a simpler way. To see this, recall that

$$AU = U\Lambda \leftrightarrow Au_i = \lambda u_i, \text{ for } i \in 1, \dots, n, \text{ and } u_i^T u_j = \delta_{ij}.$$

Given this, consider the action of A on a vector x . If x happens to be an eigenvector of A , then $x = u_i$ for some $i \in [n]$, and we have that

$$Ax = Au_i = \lambda_i u_i = \lambda_i x,$$

i.e., the vector is just scaled in length but not changed in direction. (This should be obvious, if you recall the original motivation for eigenvectors.) If x is not an eigenvector of A , then since the eigenvectors form a complete orthonormal basis, we can decompose any arbitrary vector $x \in \mathbb{R}^n$ in that basis as

$$x = \sum_{j=1}^n \alpha_j u_j,$$

for some set of coefficients α_j . In fact, we can easily compute the coefficients α_j , for $j = 1, \dots, n$.

To compute α_j , simply compute the dot product of u_j and x . That is, the dot product of u_j and x is

$$u_j^T x = u_j^T \left(\sum_{i=1}^n \alpha_i u_i \right) = u_j^T \sum_{i=1}^n \alpha_i u_i = \sum_{i=1}^n \alpha_i u_j^T u_i = \sum_{i=1}^n \alpha_i \delta_{ij} = \alpha_j.$$

Thus,

$$\alpha_j = u_j^T x,$$

and, thus, we can write x as

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

This expression can be viewed in one of two ways:

$$\begin{aligned} x &= \sum_{i=1}^n \alpha_i u_i \\ &= \sum_{i=1}^n u_i^T x u_i \end{aligned} \tag{18.2}$$

$$\begin{aligned} &= \sum_{i=1}^n u_i u_i^T x \\ &= \left(\sum_{i=1}^n u_i u_i^T \right) x \end{aligned} \tag{18.3}$$

$$\begin{aligned} &= \left(\sum_{i=1}^n P_{u_i} \right) x \\ &= \sum_{i=1}^n (P_{u_i} x). \end{aligned} \tag{18.4}$$

Here, $P_{u_i} = u_i u_i^T$ is a projection matrix onto the span of u_i (which, when applied to x , since u_i is an eigenvector, we know equals $\alpha_i u_i$).

Make sure that you understand each step of that derivation. In particular, observe that Equation (18.3) follows from the previous step since

$$u_i^T x u_i = u_i u_i^T x,$$

which follows since $u_i^T x \in \mathbb{R}$ is a scalar. Each step is elementary, but together they highlight complementary important points.

- On the one hand, Equation (18.2) expresses x as a linear combination of the orthonormal basis provided by the eigenvectors of the matrix A , in a way exactly analogous to how one can express x as a linear combination of the canonical basis vectors.
- On the other hand, Equation (18.4) expresses x as a sum of projections onto the span of each of the n orthogonal eigenvectors, in a way that generalizes the Pythagorean Theorem from \mathbb{R}^2 to \mathbb{R}^n .

This holds for a vector x , but we can also apply similar ideas to a matrix A . Going back to Ax , we have the following:

$$\begin{aligned} Ax &= A \sum_{i=1}^n \alpha_i u_i \\ &= \sum_{i=1}^n \alpha_i A u_i \\ &= \sum_{i=1}^n \alpha_i \lambda_i u_i \\ &= \sum_{i=1}^n \alpha'_i u_i. \end{aligned} \tag{18.5}$$

So, this expression shows that the eigen-decomposition of A gives a “nice” basis, where the action of A is just to scale the length of vectors by different amounts along different eigen-basis vectors. This generalizes (to the orthonormal basis defined by the eigenvectors of the symmetric matrix A) how a diagonal matrix scales the length of the canonical basis vectors by different amounts.

That is, recall the general matrix multiplication result.

$$\begin{aligned} (Ax)_i &= \sum_j A_{ij} x_j \\ &= A_{ii} x_i \quad \text{if the matrix } A \text{ is diagonal.} \end{aligned}$$

In this case, the input vector is just scaled by the diagonal elements of A . So, the spectral decomposition shows that any symmetric matrix has a basis where this can be done.

18.6.2 Describing a matrix as a quadratic form

We can use the spectral decomposition to describe the matrix as representing a symmetric bilinear function, i.e., quadratic form, in a simpler way.

To see this, consider

$$x^T A x = \sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j,$$

and observe that there are lots of cross terms, i.e., A_{ij} , for $i \neq j$, in that expression if we write it all out. Recall that

$$AU = U\Lambda \quad \leftrightarrow \quad A = U\Lambda U^T.$$

If x is an eigenvector of A , then

$$x^T Ax = u_i^T A u_i = u_i^T \lambda_i u_i = \lambda_i \|x\|_2^2 = \lambda_i.$$

In this case, the vector is just pointing along one of the axes of the ellipsoid, and it gets scaled in length. Otherwise, let's say that $x = \alpha_1 u_1 + \alpha_2 u_2$ (which we can always do by expressing x in the orthonormal basis provided by the eigen-basis of A). Then,

$$\begin{aligned} x^T Ax &= (\alpha_1 u_1 + \alpha_2 u_2)^T A (\alpha_1 u_1 + \alpha_2 u_2) \\ &= (\alpha_1 u_1 + \alpha_2 u_2)^T (\alpha_1 \lambda_1 u_1 + \alpha_2 \lambda_2 u_2) \\ &= \alpha_1^2 \lambda_1 u_1^T u_1 + \alpha_1 \alpha_2 \lambda_1 u_1^T u_2 + \alpha_1 \alpha_2 \lambda_2 u_2^T u_1 + \alpha_2^2 \lambda_2 u_2^T u_2 \\ &= \alpha_1^2 \lambda_1 + \alpha_2^2 \lambda_2 \\ &= \sum_{i=1}^2 \alpha_i^2 \lambda_i. \end{aligned}$$

Note that this happens since the eigenvectors corresponding to different eigenvalues are orthogonal. So, we used the fact that any symmetric matrix has a full set of orthogonal eigenvectors.

So, if we calculate it this way, then the vector x is determined by α_1 and α_2 , and this determines how much of x is in the direction of u_1 and how much is in the direction of u_2 .

Remark. This equation is quadratic in α_i with no cross terms. *That is, we have completed the square to remove all the cross terms.* Thus, we are back to the generalized conic section case. In addition,

$$\text{If } \lambda_i \left\{ \begin{array}{c} > \\ = \\ < \end{array} \right\} 0, \quad \text{then} \quad \left\{ \begin{array}{c} \text{ellipse} \\ \text{hyperbola} \\ \text{flat} \end{array} \right\}.$$

18.6.3 Computing higher powers of a matrix, and iterative algorithms

If we have a symmetric matrix $A \in \mathbb{R}^{n \times n}$, represented in terms of its eigenvalue decomposition as $A = U \Lambda U^T$, then it is straightforward to compute A^2 :

$$A^2 = U \Lambda U^T U \Lambda U^T = U \Lambda^2 U^T.$$

Here, the second equality follows since $U^T U = I$, since U is a square orthogonal matrix. Clearly, it is similarly straightforward to compute A^t for and $t \in \mathbb{Z}^+$:

$$A^t = U \Lambda U^T \cdots U \Lambda U^T = U \Lambda^t U^T.$$

There are two consequences of this.

- If we have computed the eigendecomposition, we can use this to compute high powers of the matrix very easily. (Of course, if the matrix was diagonal to begin with, the same is true.)
- If we have not computed the eigendecomposition, we can use this expression to do so.

With respect to this last point, although we have described this procedure as if we know U and Λ , this observation actually forms the basis for a large body of algorithms, in particular so-called iterative algorithms, which can be used to compute the eigenvalues and eigenvectors of A when they are not known. This topic is more advanced than we want to describe in detail, but we will give a rough idea here.

Let's say that

$$Ax = \sum_{i=1}^n \alpha_i \lambda_i u_i.$$

By a similar derivation, it follows that

$$\begin{aligned} A^2 x &= A^2 \sum_{i=1}^n \alpha_i u_i \\ &= \sum_{i=1}^n \alpha_i A^2 u_i \\ &= \sum_{i=1}^n \alpha_i \lambda_i^2 u_i, \end{aligned}$$

and similarly that

$$A^t x = \sum_{i=1}^n \alpha_i \lambda_i^t u_i.$$

What this expression says is that if we apply a matrix to a vector repeatedly, then the directions along the largest eigenvector get amplified, while the directions along smaller eigenvectors get decreased. In this limit, all directions except for the largest go to zero.

As a very special case of this, consider a 2×2 diagonal matrix,

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0.9 \end{pmatrix},$$

for which the canonical basis vectors are the orthonormal eigenvectors. In this case,

$$A^t = \begin{pmatrix} 1^t & 0 \\ 0 & 0.9^t \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

as $t \rightarrow \infty$. Thus, $A^t x \rightarrow e_1$, which is the eigenvector of A corresponding to the largest eigenvalue.

Clearly, this is a larger topic than we can go into here, but more advanced classes will cover this in detail.

18.6.4 Generalize differential calculus to multiple variables

Basic calculus considers functions $f : \mathbb{R} \rightarrow \mathbb{R}$, and it then defines derivatives as rates of change along the only direction that there is (and it then defines integrals as areas as you move along the only direction there is).

Recall in 1 dimension the Taylor expansion of a function $f(x)$ about a point x^* :

$$f(x) = f(x^*) + \left(\frac{df}{dx} \right)_{x=x^*} (x - x^*) + \frac{1}{2!} \left(\frac{d^2 f}{dx^2} \right)_{x=x^*} (x - x^*)^2 + \dots$$

From this, we get derivatives, the first and second order condition for optimization, and all sorts of other things (that we are not going to describe in this class).

Consider the generalization to 2 dimensions. In this case, we have

$$f(x_1, x_2) = f(x_1^*, x_2^*) + (\nabla f)_{x=x^*}^T \begin{pmatrix} x_1 - x_1^* \\ x_2 - x_2^* \end{pmatrix} + \frac{1}{2!} \begin{pmatrix} x_1 - x_1^* & x_2 - x_2^* \end{pmatrix} (\nabla^2 f)_{x=x^*} \begin{pmatrix} x_1 - x_1^* \\ x_2 - x_2^* \end{pmatrix} + \dots$$

In this expression,

$$(\nabla f)_{x=x^*} = \left(\begin{array}{c} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{array} \right)_{(x_1, x_2)=(x_1^*, x_2^*)} \in \mathbb{R}^2$$

is known as the *gradient vector*, and it points in the direction of change of the function. (For 1-dimensional problems, the notion of direction is trivial, since there is only one direction; but the notion of direction is

non-trivial for n -dimensional problems, for $n \geq 2$, since it amounts to specifying a vector in \mathbb{R}^n .) Also, in this expression,

$$\nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} \end{pmatrix}_{(x_1, x_2) = (x_1^*, x_2^*)} \in \mathbb{R}^{2 \times 2},$$

is known as the *Hessian matrix*, and it describes the multi-dimensional curvature of the function. Importantly, the Hessian matrix is a symmetric matrix, and thus has a full set of n (in this case 2) orthonormal eigenvectors, each of which has an associated eigenvalue.

Remark. There is a caveat in this discussion. The symmetry of the Hessian relies on assuming that

$$\frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial^2 f}{\partial x_1 \partial x_2}.$$

This can fail for one of two reasons: either the second partial derivatives don't exist, in which case these derivatives aren't defined; or the function isn't sufficiently continuous, in which case the two cross second partial derivatives both exist but are unequal. The former happens quite often in data analysis, but it is usually pretty obvious when it does and when it doesn't, and when it does one can use different methods; but the latter is pretty rare in data analysis. So, we won't worry about this here, and instead we'll assume sufficient continuity for the cross second partial derivatives to equal each other.

Generalizing this discussion to n -dimensions, we have

$$f(x) = f(x^*) + (\nabla f)_{x=x^*}^T (x - x^*) + \frac{1}{2!} (x - x^*)^T (\nabla^2 f)_{x=x^*} (x - x^*) + \dots,$$

where $f(x) \in \mathbb{R}$, $x \in \mathbb{R}^n$, $(\nabla f) \in \mathbb{R}^n$ is the gradient vector, and $(\nabla^2 f) \in \mathbb{R}^{n \times n}$ is the Hessian matrix. Here, the Hessian matrix $(\nabla^2 f)$ is an $n \times n$ symmetric matrix (assuming the continuity caveat mentioned above) that has as its (ij) element:

$$(\nabla^2 f)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

Thus, in the canonical basis $\{e_i\}_{i=1}^n$, the diagonal elements $\frac{\partial^2 f}{\partial x_i^2}$ describe the curvature of the function along the i^{th} coordinate axis. The off-diagonal elements $\frac{\partial^2 f}{\partial x_i \partial x_j}$, for $i \neq j$, have to do with curvature in two different directions—they can be harder to interpret, but the easiest way to understand them is in terms of conic sections and quadratic forms. Of course, we know that we can rotate to another basis (namely, the orthonormal basis given by the eigenvectors of the Hessian) where the Hessian is diagonal, and in this basis there is no coupling between the curvature along different coordinates, and there are only the diagonal terms (which could be positive or zero or negative, as we discussed before).

18.6.5 Matrices that are non-symmetric

There are many data matrices that are not symmetric. For example,

- term-document matrices
- individual-SNP matrices
- gene-environmental condition matrices
- person-attribute matrices

What about these?

For these matrices, eigenvectors and eigenvalues are also defined (if the matrices are even square), but linear algebraic results are *much* more modest if we want to work with eigenstuff. E.g., they may or may not exist,

or they may or may not be real, or they may or may not be orthogonal, or there may or may not exist a full set of eigenvectors, and so on. So, it is harder to interpret these in terms of variance or anything like that that is meaningful in data science. So, we typically don't do this. Instead, for general non-symmetric matrices and more generally for non-square matrices—which are very common in data science—we typically work with a more robust generalization, known as *singular vectors* and *singular values*.

Given a matrix $X \in \mathbb{R}^{m \times n}$, then we can define

$$A = X^T X = V \Lambda V^T,$$

and we can also define

$$A' = X X^T = U \Lambda' U^T.$$

An important fact (that we will not prove) is that $\Lambda = \Lambda'$. Given this, we can decompose X as

$$X = U \Sigma V^T,$$

where U provides an orthonormal basis for the column space, V provides an orthonormal basis for the row space, and $\Sigma = \Lambda^2$ consists of non-negative entries. This is known as the Singular Value Decomposition (SVD). The SVD is defined for any $m \times n$ real-valued matrix A , symmetric or not, square or not.

Here is the interpretation.

1.
 - $A_{ij} = \text{dot}(X_{(i)}, X_{(j)})$, where $X_{(i)}$ is the i^{th} column, which is the i^{th} data point. So, each entry of A is the dot product between two data points. Note that we didn't do mean centering or variance scaling here. The reason is that we aren't going to interpret this as a correlation/covariance, just as a SPSD matrix.
 - $A'_{ij} = \text{dot}(X^{(i)}, X^{(j)})$, where $X^{(i)}$ is the i^{th} row, which is the i^{th} feature. So, each entry of A' is a dot product between two feature vectors. Again, this is not mean-centered or variance normalized, again for the same reason, that we aren't going to interpret this as a correlation/covariance, just as a SPSD matrix.
2. If we choose an orthonormal basis in the domain and an orthonormal basis in the range, then A is diagonal with nonnegative entries. Thus, the SVD of general matrices generalizes the EVD of symmetric matrices, and so it is an important generalization of the spectral theorem ideas that we have been discussing to general $m \times n$ matrices.

Remark. Note that we did this SVD discussion for an arbitrary $m \times n$ matrix X . In particular, that matrix need *not* be mean-centered or pre-processed in any way like that. The reason we pre-processed matrices in that way when discussing PCA was that we wanted to characterize the variability about the mean. Thus, based on the equation for the variance of a random variable, we had to subtract off the mean. But, we have seen that the quadratic form properties hold for any symmetric matrix, and not just mean-centered covariance/correlation matrices. So, those results still hold, which we know from our general PD/PSD discussion, for this SVD discussion. From this perspective, PCA is just performing the SVD on $X^T X$, after X has been mean centered; and, conversely, a common way to compute the PCs is to compute the SVD.