MATH FOR DATA SCIENCE

Samuel S. Watson

Preface

* Experience with basic calculus is necessary, and some prior experience with linear algebra and programming will be helpful.

In this book, we develop the mathematical ideas underlying the modern practice of data science. The goal is to do so accessibly and with a minimum of prerequisites.* For example, we will start by reviewing sets and functions from a data-oriented perspective. On the other hand, we will take a problem-solving approach to the material and will not shy away from challenging concepts. To get the most out of the course, you should prepare to invest a substantial amount of time on the exercises.

This text was originally written to accompany the master's course DATA 1010 at Brown University. The content of this PDF, along with hints and solutions for the exercises in this book, will be available on the edX platform starting in the fall semester of 2019, thanks to the BrownX project and the Brown University School of Professional Studies.

The author would like to acknowledge Isaac Solomon, Elvis Nunez, and Thabo Samakhoana for their contributions during the development of the DATA 1010 curriculum. They wrote some of the exercises and many of the solutions that appear in the BrownX course.

Contents

1 Sets and functions

Sets and functions are foundational to the study of mathematics and ubiquitous in quantitative disciplines, including statistics and data science. In this chapter we review the basics of sets, lists, and functions from a data perspective.

1.1 Sets

A simple grocery list is a real-life example of a *set*: the main function afforded by the grocery list is to answer the query "here's an item in the store; is it on list?" Note that for purposes of answering this question, the order of the listed items on the grocery list doesn't matter, and repeating an entry is equivalent to having a single instance of that entry. This leads us to the definition of a *set*.

Definition 1.1.1: Set

A **set** is an unordered collection of objects. The objects in a set are called *elements*.

The term *object* in this definition is deliberately vague. Sets may contain any kind of data: numbers, words, symbols, circles, squares, other sets, and many others.

If a set S contains a finite number of elements $s_1, s_2, \ldots, s_n$, we can write

$$S = \{s_1, s_2, \ldots, s_n\}.$$

The fundamental operation provided by a set is checking membership: we write $s \in S$ to indicate that s is an element of the set S. If s is not an element of S, we write $s \notin S$. If two sets have the same elements, then they are considered equal. For example, $\{1, 1, 2\} = \{1, 2\}$. For this reason, we typically list the elements of a set without duplication.

The set containing no elements is called the **empty set** and is denoted $\emptyset$ or $\{\}$.

Some sets with standard and specially typeset names include

- $\mathbb{R}$, the set of real numbers,
- $\mathbb{Q}$, the set of rational numbers,
- $\mathbb{Z}$, the set of integers, and
- $\mathbb{N}$, the set of natural numbers.

Definition 1.1.2: Subset

Suppose S and T are sets. If every element of T is also an element of S, then we say T is a subset of S, denoted $T \subset S$.

The sets S and T are equal if $S \subset T$ and $T \subset S$.

Exercise 1.1.1

Suppose that E is the set of even positive integers and that F is the set of positive integers which are one more than an odd integer.

(a) $E \subset F$

(b) $F \subset E$

(c) $E = F$

(d) all of the above

Example 1.1.1

We have $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}$, since every natural number is an integer, every integer is rational, and every rational number is real.

Definition 1.1.3: Set builder notation

If S is a set and P is a property which each element of S either satisfies or does not satisfy, then

$$\{s \in S : s \text{ satisfies } P\}$$

denotes the set of all elements in S which have the property P. This is called *set builder notation*. The colon is read as 'such that.'

Example 1.1.2

Suppose the set S denotes the set of all real numbers between 0 and 1. Then S can be expressed as

$$S = \{s \in \mathbb{R} : 0 < s < 1\}.$$

Definition 1.1.4: Cardinality

Given a set S, the cardinality of S, denoted $|S|$, denotes the number of elements in S.

Exercise 1.1.2

Let $S = \{4, 3, 4, 1\}$. Find $|S|$.

Definition 1.1.5: Countably infinite

A set is *countably infinite* if its elements can be arranged in a sequence.

Example 1.1.3

The set $\{1, 2, 3, 4, \ldots\}$ is countably infinite. The set of integers is countably infinite, since they can be arranged sequentially: $\{0, 1, -1, 2, -2, 3, -3, \ldots\}$

The set of rational numbers between 0 and 1 is countably infinite, since they all appear in the sequence $\frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \ldots$.

The set of all real numbers between 0 and 1 is not countably infinite. Any infinite sequence of real numbers will necessarily fail to include all real numbers. This may be demonstrated using an idea called *Cantor's diagonalization argument*, though we will omit the proof.

Exercise 1.1.3

Show that the set of all ordered pairs of positive integers is countably infinite.

1.1.1 Set operations

Given a set S describing a grocery list and a subset $A \subset S$ describing the set of items we've already purchased, the set we might be most interested in constructing from S and A is the set of items which are in S but not in A. This is called the **complement** of A with respect to S.

Definition 1.1.6: Complement

If A and S are sets and $A \subset S$, then the complement of A with respect to S, denoted $S \setminus A$ or A^c, is the set of all elements in S that are not in A. That is

$$A^c = \{s \in S : s \notin A\}.$$

Since S is not part of the notation A^c, we will usually only use that notation when the intended containing set S is clear from context.

Exercise 1.1.4

Suppose $S = \{1, 2, 3, 4, 5\}$ and $A = \{4, 2\}$. Find the complement A^c of A with respect to S.

Exercise 1.1.5

Suppose $A \subset S$, $|S| = 55$, and $|A| = 13$. Find $|S \setminus A|$.

If two members of your household supplied you with grocery lists as you were about to go to the store, then the first thing you might want to do is produce a combined grocery list. This set operation is called taking the *union*.

Definition 1.1.7: Union

The **union** of two sets S and T, denoted $S \cup T$, is the set containing all the elements of S and all the elements of T and no other elements. In other words, $s \in S \cup T$ if and only if either $s \in S$ or $s \in T$.

Exercise 1.1.6

Let $S = \{1, 2, 4, 5\}$ and $T = \{1, 5, 6, 7, 8\}$. Find $S \cup T$.

When we're choosing which loaf of bread to purchase, we're interested in finding one which is in *both* of two sets:

(i) the set of loaves whose price falls in our range of acceptable prices, and

(ii) the set of loaves whose taste we find satisfactory

The set of viable loaves is the *intersection* of these two sets.

Definition 1.1.8: Intersection

The **intersection** of two sets S and T, denoted $S \cap T$, is the set consisting of elements that are in both S and T. In other words, $s \in S \cap T$ if and only if $s \in S$ and $s \in T$.

Example 1.1.4

Let $S = \{1, 2, 3, 4, 5\}$ and $T = \{1, 5, 6, 7, 8\}$. Then

$$S \cap T = \{1, 5\}.$$

The union and intersection operations may be applied to any number of sets. Suppose $S_1, S_2, \ldots, S_n$ are sets—the union of these sets can be expressed as $S_1 \cup S_2 \cup \cdots \cup S_n$. More compactly,

$$\bigcup_{i=1}^{n} S_i = S_1 \cup S_2 \cdots \cup S_n = \{s : s \in S_i \text{ for some } 1 \leq i \leq n\}.$$

Similarly, we can take the intersection of an arbitrary number of sets:

$$\bigcap_{i=1}^{n} S_i = S_1 \cap S_2 \cap \cdots \cap S_n = \{s : s \in S_i \text{ for all } 1 \leq i \leq n\}.$$

Often we will want to specify whether two sets have any elements in common.

Definition 1.1.9: Disjoint

Two sets S and T are **disjoint** if they do not have any elements in common.

In other words, S and T are disjoint if $S \cap T = \emptyset$.

Definition 1.1.9 extends to an arbitrary number of sets. We say that the sets $S_1, S_2, \ldots, S_n$ are **pairwise disjoint** if $S_i \cap S_j = \emptyset$ whenever $i \neq j$.

Exercise 1.1.7

Find three sets A, B, and C which have $A \cap B \cap C = \emptyset$, but for which all of the intersections $A \cap B$, $B \cap C$, and $A \cap C$ are nonempty.

Suppose you're part of a group of n shoppers working together to purchase the items on a single grocery list. A good idea is to *partition* the set of items you want to purchase into n smaller sets so that each person can purchase only the items on their own set.

Definition 1.1.10: Partition

A **partition** of a set S is a collection of non-empty sets $S_1, S_2, \ldots, S_n$ such that

$$S = \bigcup_{i=1}^{n} S_i$$

and $S_1, S_2, \ldots, S_n$ are disjoint.

Exercise 1.1.8

Find a partition of $\{1, 2, 3, 4, 5\}$ into three sets. Is there a partition of $\{1, 2, 3, 4, 5\}$ into *six* sets?

Exercise 1.1.9

Establish the first and third of the following four identities. Use the following strategy: show that the left-hand side is a subset of the right-hand side and vice versa. To demonstrate that $A \subset B$, consider an element s of A and–assuming only that $s \in A$—apply reasoning to conclude that it must be in B as well.

$$S \cap (R \cup T) = (S \cap R) \cup (S \cap T)$$
$$S \cup (R \cap T) = (S \cup R) \cap (S \cup T)$$
$$\left(\bigcup_{i=1}^{n} S_i\right)^c = \bigcap_{i=1}^{n} S_i^c$$
$$\left(\bigcap_{i=1}^{n} S_i\right)^c = \bigcup_{i=1}^{n} S_i^c$$

Suppose we perform an experiment which consists of flipping a coin and rolling a standard six-sided die. The outcome of the coin flip is an element of the set $S_1 = \{\mathtt{H}, \mathtt{T}\}$, and the outcome of the die roll is an element of the set $S_2 = \{1, 2, 3, 4, 5, 6\}$. The set of all possible outcomes of the experiment is the set

$$S = \{(\mathtt{H}, 1), (\mathtt{H}, 2), (\mathtt{H}, 3), (\mathtt{H}, 4), (\mathtt{H}, 5), (\mathtt{H}, 6),$$
$$(\mathtt{T}, 1), (\mathtt{T}, 2), (\mathtt{T}, 3), (\mathtt{T}, 4), (\mathtt{T}, 5), (\mathtt{T}, 6)\}.$$

Definition 1.1.11: Cartesian Product

If S_1 and S_2 are sets, then the **Cartesian product** of S_1 and S_2 is defined by

$$S_1 \times S_2 = \{(s_1, s_2) : s_1 \in S_1 \text{ and } s_2 \in S_2\}.$$

Likewise, if $S_1, S_2, \ldots, S_n$ are sets, then

$$S_1 \times S_2 \times \cdots \times S_n = \{(s_1, s_2, \ldots, s_n) : s_1 \in S_1 \text{ and } s_2 \in S_2 \text{ and } \cdots \text{ and } s_n \in S_n\}.$$

Exercise 1.1.10

Find $|S \times T|$ if $|S| = 4$ and $|T| = 100$.

1.2 Lists

Sets are data containers with very little structure: you can check membership (and perform membership-checking-related operations like unions or complements), but that's all. We will define various other types of collections which provide additional structure.

For example, suppose you *do* care about the order in which the items appear on your grocery list; perhaps because you want to be able pick the items up in a certain order as you move across the store. Also, you might want to list an item multiple times as a way of reminding yourself that you should pick up more than one. *Lists* can handle both of these extra requirements:

Definition 1.2.1: List

A **list** is an ordered collection of finitely many elements.

For example, if we regard $\{1, 2, 3\}$ and $\{2, 1, 3\}$ as lists, then they are unequal because the orders in which the elements appear are different. Also, the list $\{1, 1, 2\}$ has three elements, since repeated elements are not considered redundant.

We don't distinguish sets and lists notationally, so we will rely on context to make it clear whether order matters and repetitions count.

Exercise 1.2.1

How many sets A have the property that $A \subset \{1, 2, 3\}$? How many *lists* of length 4 have all of their elements in $\{1, 2, 3\}$?

1.3 Functions

The grocery lists you make for yourself probably don't look quite like a set *or* a list, because the quickest way to indicate how many of each item to purchase is to make a separate column:

item	count
apple	3
bread	1
squash	3

We have two sets here: the set of grocery items and the set of positive integers. For each element in the former set, we want to associate with it some element of the latter set.

Note that this construction is asymmetric in the two sets: every grocery item should have exactly one number associated with it, while some positive integers may be omitted and others may be associated with multiple grocery items.

The idea of attaching a piece of data to each element of a set arises *very* often, and it deserves its own vocabulary:

Definition 1.3.1: Function, domain, and codomain

If A and B are sets, then a **function** $f : A \to B$ is an assignment to each element of A of some element of B.
The set A is called the **domain** of f and B is called the **codomain** of f.

The domain and codomain of a function should be considered part of the data of the function: to fully specify f, we must specify (i) the domain A, (ii) the codomain B, and (iii) the value of $f(x)$ for each $x \in A$. Two functions f and g are considered equal if (i) they have the same domain and codomain and (ii) $f(x) = g(x)$ for all x in the domain of f and g.

Given a subset A' of A, we define the **image** of f—denoted $f(A')$—to be the set of elements which are mapped to from some element in A':

$$f(A') = \{b \in B \ : \text{ there exists } a \in A' \text{ so that } f(a) = b\}. \tag{1.3.1}$$

The **range** of f is defined to be the image of the domain of f. Thus the range may be obtained from the codomain by removing all the elements that don't get mapped to.

Example 1.3.1

Find the range of the function from $\{\text{apple}, \text{bread}, \text{squash}\}$ to $\mathbb{N}$ represented by the following table.

item	count
apple	3
bread	1
squash	3

Solution

The range is the set of quantity counts which get mapped to from some grocery item, so the range is the two-element set $\{1,3\}$.

Exercise 1.3.1

Consider the *social-security-number function* f from the set of US citizens and permanent residents to the set of integers $\{000000000, 000000001, \ldots, 999999999\}$. For each person x, $f(x)$ is defined to be the social security number of person x.

(i) What are the largest and smallest possible values of the ratio $\frac{|f(A)|}{|A|}$ for any subset A of the domain of f?

(ii) Estimate the ratio of the cardinality of the range of f to the cardinality of the codomain of f. (You can estimate the number of social security numbers issued to be about 40% more than the current US population).

Definition 1.3.2

If $B' \subset B$, then the **preimage** $f^{-1}(B')$ of B' is defined by

$$f^{-1}(B') = \{a \in A : f(a) \in B'\}.$$

This is the subset of A consisting of every element of A that maps to some element of B'.

Exercise 1.3.2

Consider the following purported equalities.

(i) $f(A \cap B) \stackrel{?}{=} f(A) \cap f(B)$

(ii) $f(A \cup B) \stackrel{?}{=} f(A) \cup f(B)$

(iii) $f^{-1}(C \cap D) \stackrel{?}{=} f^{-1}(C) \cap f^{-1}(D)$

(iv) $f^{-1}(C \cup D) \stackrel{?}{=} f^{-1}(C) \cup f^{-1}(D)$

Which of the are true for all functions f and all subsets A and B of the domain of F and subsets C and D of the codomain of f?

(a) all of them

(b) none of them

(c) (i) and (ii) only

(d) (iii) and (iv) only

(e) (i), (iii), and (iv) only

Definition 1.3.3

A function f is **injective** if no two elements in the domain map to the same element in the codomain; in other words if $f(a) = f(a')$ implies $a = a'$.
A function f is **surjective** if the range of f is equal to the codomain of f; in other words, if $b \in B$ implies that there exists $a \in A$ with $f(a) = b$.
A function f is **bijective** if it is both injective and surjective. This means that for every $b \in B$, there is exactly one $a \in A$ such that $f(a) \in b$. If f is bijective, then the function from B to A that maps $b \in B$ to the element $a \in A$ that satisfies $f(a) = b$ is called the **inverse** of f.

Exercise 1.3.3

Identify each of the following functions as injective or not injective, surjective or not surjective, and bijective or not bijective.

1. $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$
2. $f : [0, \infty) \to \mathbb{R}$, $f(x) = x^2$
3. $f : [0, \infty) \to [0, \infty)$, $f(x) = x^2$
4. $f : \mathbb{R} \to [0, \infty)$, $f(x) = x^2$

It is frequently useful to focus on a subset of the domain of a function without changing the codomain elements that the function associates with those domain elements. For example, if we partition a grocery list with quantity counts among several shoppers, then each shopper will be interested in the *restriction* of the quantity count function to their own portion of the domain. In other words, they need to know how many of each of their items to pick up, and they don't need to know anything about the other shoppers' items.

Definition 1.3.4: Restriction

If $f : A \to B$ and $A' \subset A$, then the **restriction** of f to A' is the function $f|_{A'} : A' \to B$ defined by $f|_{A'}(x) = f(x)$ for all $x \in A'$.

Exercise 1.3.4

State a general relationship involving the terms *restriction*, *image*, and *range*.

Sometimes the elements output by a function f will themselves have associated data, and in this case we often want to connect each element in the domain of f to these data.

* the ones that have appeared on an album

For example, consider the *album* function from the set of songs* to the set of albums. Evaluated on a song, the album function returns the album on which the song appeared. Consider also the *year* function from the set of albums to the set of years (which returns the year in which each album was released). We can determine the year in which a song was released by *composing* the album function and the year function.

Definition 1.3.5: Composition

If $f : A \to B$ and $g : B \to C$, then the function $g \circ f$ which maps $x \in A$ to $g(f(x)) \in C$ is called the **composition** of g and f.

Exercise 1.3.5

Show that composition is associative: $(f \circ g) \circ h = f \circ (g \circ h)$ for all functions f, g, and h with the property that the codomain of h is equal to the domain of g and the codomain of g is equal to the domain of f.

If the rule defining a function is sufficiently simple, we can describe the function using **anonymous function notation**. For example, $x \in \mathbb{R} \mapsto x^2 \in \mathbb{R}$, or $x \mapsto x^2$ for short, is the squaring function from $\mathbb{R}$ to $\mathbb{R}$. Note that bar on the left edge of the arrow, which distinguishes the arrow in anonymous function notation from the arrow between the domain and codomain of a named function.

Exercise 1.3.6

Suppose that f is the function $(x \mapsto \sqrt{x}) \circ (y \mapsto 3y)$. Find $f\left(\frac{1}{12}\right)$.

3 Linear Algebra

Using and interpreting data requires storing and manipulating sets of numbers in conceptually and computationally helpful ways. The language of *linear algebra* provides basic vocabulary, visualizations, and mathematical results for understanding the structure of a dataset.

Exercise 3.0.1

Consider a spreadsheet of data whose rows correspond to individuals and whose three columns correspond to weight in kilograms, height in centimeters, and height in inches. Are any of the columns redundant?

In this chapter, we will develop a more general and mathematically rigorous version of the idea of *redundancy* explored in Exercise 3.0.1.

3.1 Vector spaces

3.1.1 Vectors

A **vector** in $\mathbb{R}^n$ is a column* of n real numbers. These real numbers are called the **components** or **entries** of the vector.

* For typographical convenience, we will sometimes write $\begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}$ as $[-2, 0, 1]$

Example 3.1.1

$\mathbf{v} = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}$ is a vector in $\mathbb{R}^3$. We say that the first component of $\mathbf{v}$ is equal to -2, the second component is equal to 0, and the third component is equal to 1.

We draw a vector in $\mathbb{R}^2$ as an arrow from one point to another so that the horizontal separation between the points is equal to the first component of the vector and the vertical separation between the points is equal to the second component. We define the **norm** $|\mathbf{v}|$ of a vector $\mathbf{v} \in \mathbb{R}^n$ to be the length of the associated arrow, which may be calculated as the square root of the sum of the squares of $\mathbf{v}$'s components. A vector whose norm is 1 is called a **unit vector**.

The fundamental vector operations are

1. **Vector addition** (addition of two vectors), and
2. **Scalar multiplication** (multiplication of a real number and a vector).

These operations are defined componentwise, and they have natural geometric interpretations (see Figures 3.1 and 3.2)

* Multiplying by -1 reverses the direction of the vector and preserves its norm

1. Summing vectors concatenates them tail-to-head, and
2. Multiplying a vector by a positive* real number k preserves its direction and multiplies its norm by k.

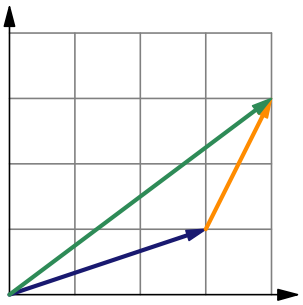


Figure 3.1 Vector addition: $[3,1]+[1,2]=[4,3]$

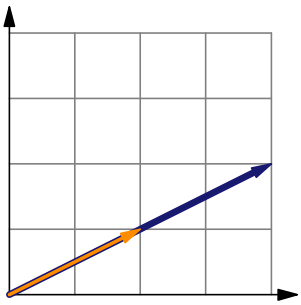


Figure 3.2 Scalar multiplication: $2[2,1]=[4,2]$

Scalar multiplication is denoted by placing the scalar adjacent to the vector, and vector addition is denoted with "+" between two vectors.

Exercise 3.1.1

Simplify $3\begin{bmatrix}-2\\11\end{bmatrix}-\begin{bmatrix}4\\0\end{bmatrix}$.

Exercise 3.1.2

Determine whether there exists a real number r satisfying the vector equation

$$r\begin{bmatrix}-3\\2\end{bmatrix}-\begin{bmatrix}2\\1\end{bmatrix}=\begin{bmatrix}4\\2\end{bmatrix}.$$

Exercise 3.1.3

Show that every nonzero vector $\mathbf{v}$ can be written as the product of a nonnegative real number c and a unit vector $\mathbf{u}$.

Exercise 3.1.4

Find a formula in terms of $\mathbf{u}$ and $\mathbf{v}$ which represents the vector from the head of $\mathbf{v}$ to the head of $\mathbf{u}$ when $\mathbf{u}$ and $\mathbf{v}$ are situated so that their tails coincide.

Exercise 3.1.5

Solve for $\mathbf{u}$ in terms of c and $\mathbf{v}$ in the equation $c\,\mathbf{u}+\mathbf{v}=\mathbf{0}$, assuming that $\mathbf{u}$ and $\mathbf{v}$ are vectors in $\mathbb{R}^n$ and c is a nonzero real number.

3.1.2 Linear independence, span, and basis

A **linear combination** of a list of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ is an expression of the form

$$c_1\mathbf{v_1} + c_2\mathbf{v_2} + \cdots + c_k\mathbf{v_k},$$

where $c_1, \ldots, c_k$ are real numbers. The c's are called the **weights** of the linear combination.

Exercise 3.1.6

Suppose that $\mathbf{u} = [2, 0]$ and $\mathbf{v} = [1, 2]$. Draw the set of all points (a, b) in $\mathbb{R}^2$ for which the vector $[a, b]$ can be written as an *integer* linear combination* of $\mathbf{u}$ and $\mathbf{v}$.

* An integer linear combination is a linear combination in which all weights are integers

The **span** of a list of vectors is the set of all vectors which can be written as a linear combination of the vectors in the list.

Exercise 3.1.7

Is $\mathbf{w} = \begin{bmatrix} 1 \\ 4 \\ 0 \end{bmatrix}$ in the span of $\mathbf{u} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$? If so, find values α and β such that $\mathbf{w} = \alpha\mathbf{u} + \beta\mathbf{v}$.

We visualize a set S of vectors in $\mathbb{R}^n$ by associating the vector $[v_1, v_2, \ldots, v_n]$ with the point $(v_1, \ldots, v_n)$—in other words, we associate each vector with the location of its head when its tail is drawn at the origin.

Exercise 3.1.8

The span of two vectors in $\mathbb{R}^2$

(a) can be any shape

(b) must be either a circle or a line

(c) can be all of $\mathbb{R}^2$

(d) must be either a line or a point

(e) must be either a line or a point or all of $\mathbb{R}^2$

The span of three vectors in $\mathbb{R}^3$

(a) can be any shape

(b) must be a sphere or a line

(c) must be a plane

(d) must be a point, a plane, a line, or all of $\mathbb{R}^3$

(e) must be a plane, a line, or a point

A list of vectors is **linearly independent** if none of the vectors in the list can be written as a linear combination of the others.

Example 3.1.2

The list of vectors $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ where $\mathbf{u}_1 = \begin{bmatrix}1\\1\\2\end{bmatrix}$, $\mathbf{u}_2 = \begin{bmatrix}0\\1\\0\end{bmatrix}$, $\mathbf{u}_3 = \begin{bmatrix}4\\7\\8\end{bmatrix}$ is not linearly independent, since $\mathbf{u}_3 = 4\mathbf{u}_1 + 3\mathbf{u}_2$.

The list of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ where $\mathbf{v}_1 = \begin{bmatrix}1\\0\\0\end{bmatrix}$, $\mathbf{v}_2 = \begin{bmatrix}0\\1\\0\end{bmatrix}$, $\mathbf{v}_3 = \begin{bmatrix}0\\0\\1\end{bmatrix}$ is linearly independent, since any linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$ is unequal to $\mathbf{v}_3$, and similarly for $\mathbf{v}_1$ and $\mathbf{v}_2$.

Theorem 3.1.1

A list of vectors is linearly independent if and only if there is no vector in the list which is in the span of the *preceding* vectors.

For example, to check that $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$, it suffices to check that $\mathbf{v}_1 \neq \mathbf{0}$, that $\mathbf{v}_2$ is not a scalar multiple of $\mathbf{v}_1$ and that $\mathbf{v}_3$ is not in the span of $\{\mathbf{v}_1, \mathbf{v}_2\}$.

Proof

If a list is linearly independent, then no vector in the list can be represented as a linear combination of others by defintion, so no vector can be in the span of the previous ones. Now suppose a list of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is such that no vector in the list is in the span of the preceding vectors. Note that such a list necessarily does not contain $\mathbf{0}$. If this list were linearly dependent, then one of the vectors could be written as linear combination of the others. Lets assume, without loss of generality, that $\mathbf{v}_1$ is such a vector, then

$$\mathbf{v}_1 = c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n$$

for some $c_2, \ldots, c_n$ which are not all zero. If we define k so that c_k is the *last* of the nonzero c's, then we can rearrange the above to get

$$\mathbf{v}_k = \frac{\mathbf{v}_1 - (c_2\mathbf{v}_2 + \cdots + c_{k-1}\mathbf{v}_{k-1})}{c_k}$$

which is a contradiction. Therefore the list must be linearly independent.

Exercise 3.1.9

Let's say that a linear combination of a list of vectors is **trivial** if all of the weights are zero.

Show that a list of vectors is **linearly independent** if and only if every nontrivial linear combination of the vectors is not equal to the zero vector.

Spans of lists of vectors are so important that we give them a special name: a **vector space** is a nonempty set of vectors which is closed under the vector space operations. If V and W are vector spaces and $V \subset W$, then V is called a **subspace** of W.

Example 3.1.3

Lines and planes through the origin are vector subspaces of $\mathbb{R}^3$. More generally, the span of any list of vectors in $\mathbb{R}^n$ is a vector subspace of $\mathbb{R}^n$.

A **spanning list** of a vector space V is a list of vectors in V whose span is equal to V.

Example 3.1.4

The list $\left\{\begin{bmatrix}2\\1\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}7\\11\end{bmatrix}\right\}$ is a spanning list for $\mathbb{R}^2$ because any vector $\mathbf{v} = \begin{bmatrix}x\\y\end{bmatrix} \in \mathbb{R}^2$ can be represented as

$$\mathbf{v} = (x-y)\begin{bmatrix}2\\1\end{bmatrix} + (2y-x)\begin{bmatrix}1\\1\end{bmatrix} + 0\begin{bmatrix}7\\11\end{bmatrix}$$

A linearly independent spanning list for a vector space V is called a **basis** for V.

Example 3.1.5

The list $\left\{\begin{bmatrix}2\\1\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}\right\}$ is a basis for $\mathbb{R}^2$ and the list $\left\{\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}\right\}$ is also a basis for $\mathbb{R}^2$.

Theorem 3.1.2

If V is a vector space, then any spanning list of V is at least as long* as any linearly independent list of vectors in V.

* The length of a list of vectors is the number of vectors in the list

In other words, Theorem 3.1.2 says that if L_1 is a linearly independent list of vectors in V and L_2 is a list of vectors which spans V, then the length of L_1 is less than or equal to the length of L_2.

Exercise 3.1.10

Use Theorem 3.1.2 to show that all bases of a vector space V have the same length. In other words, if B_1 is a basis for V, and B_2 is a basis for V, then the lengths of B_1 and B_2 are equal.

The **dimension** of a vector space V is the length of any basis of V.

Given a basis of V, we can represent each vector in V uniquely as a linear combination of the vectors in the basis. In other words, if a vector space V has a basis $\mathcal{B} = \{\mathbf{b}_1, \ldots \mathbf{b}_n\}$ and $\mathbf{v} \in V$, then there exists a unique n-tuple of real numbers $(v_1, \ldots, v_n)$ such that

$$\mathbf{v} = v_1\mathbf{b}_1 + \cdots + v_n\mathbf{b}_n.$$

We call $(v_1, \ldots, v_n)$ the **coordinates** of $\mathbf{v}$ with respect to $\mathcal{B}$.

Example 3.1.6

For $1 \leq i \leq n$, let $\mathbf{e}_i \in \mathbb{R}^n$ be a vector with 1 in the ith position and zeros elsewhere. Then $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is called the **standard basis** for $\mathbb{R}^n$. The components of a vector in $\mathbb{R}^n$ coincide with its coordinates with respect to this basis.

Exercise 3.1.11

Show that any linearly independent list of vectors in a vector space $V \subset \mathbb{R}^n$ can be extended to form a basis of V, and show that any spanning list of V can be trimmed to form a basis of V.

Exercise 3.1.12

Suppose that U and V are vector spaces in $\mathbb{R}^n$. Suppose that $\{\mathbf{u}_1, \ldots, \mathbf{u}_j\}$ is a basis for $U \cap V$, that $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is a basis for U, and that $\{\mathbf{u}_1, \ldots, \mathbf{u}_j, \mathbf{v}_1, \ldots, \mathbf{v}_\ell\}$ is a basis for V. Show that

$$\{\mathbf{u}_1, \ldots, \mathbf{u}_k, \mathbf{v}_1, \ldots, \mathbf{v}_\ell\}$$

is a linearly independent list.

Exercise 3.1.13

Suppose that V and W are subspaces of $\mathbb{R}^{10}$ and that V has dimension 4 and W has dimension 8. Which of the following could possibly be equal to the dimension of $V \cap W$? Select all that apply.

1. 0
2. 1
3. 2
4. 3
5. 4
6. 5
7. 8
8. 9

Hint: consider two two-dimensional spaces in $\mathbb{R}^3$: what are the possible dimensions for the intersection of two planes through the origin in $\mathbb{R}^3$?

Exercise 3.1.14

In Julia, a set of 5 column vectors in $\mathbb{R}^7$ with entries selected uniformly at random from $[0,1]$ may be generated using `rand(7,5)`. The dimension of the span of the columns of a matrix may then by computed using the function `rank`.

(a) Calculate the dimension of many such spans of random lists of five vectors in $\mathbb{R}^7$. What sorts of values do you get?

(i) All fives

(ii) Mostly fives, some numbers fewer than five

(iii) Mostly threes, some twos and fours, occasional ones and fives

(b) Repeat with random vectors whose entries are 0 or 1 with probability $\frac{1}{2}$.

(i) All fives

(ii) Mostly fives, some numbers fewer than five

(iii) Mostly threes, some twos and fours, occasional zeros, ones and fives

Hint: for part (b), `[rand(0:1) for i=1:7,j=1:5]` generates the desired random vectors.

3.1.3 Linear transformations

A **linear transformation** L is a function from one vector space to another which satisfies $L(\alpha\mathbf{v}+\beta\mathbf{w}) = \alpha L(\mathbf{v}) + \beta L(\mathbf{w})$. Geometrically, these are "flat maps": a function is linear if and only if it maps equally spaced lines to equally spaced lines or points.

Example 3.1.7

In $\mathbb{R}^2$, reflection along the line $y=x$, defined by $L\left(\begin{bmatrix}x\\y\end{bmatrix}\right) = \begin{bmatrix}y\\x\end{bmatrix}$, is linear because

$$\begin{aligned} L\left(\alpha\begin{bmatrix}x_1\\y_1\end{bmatrix} + \beta\begin{bmatrix}x_2\\y_2\end{bmatrix}\right) &= \begin{bmatrix}\alpha y_1 + \beta y_2\\ \alpha x_1 + \beta x_2\end{bmatrix} \\ &= \alpha\begin{bmatrix}y_1\\x_1\end{bmatrix} + \beta\begin{bmatrix}y_2\\x_2\end{bmatrix} \\ &= \alpha L\left(\begin{bmatrix}x_1\\y_1\end{bmatrix}\right) + \beta L\left(\begin{bmatrix}x_2\\y_2\end{bmatrix}\right). \end{aligned}$$

The **rank** of a linear transformation from one vector space to another is the dimension of its range.

Example 3.1.8

If $L\left(\begin{bmatrix}x\\y\\z\end{bmatrix}\right) = \begin{bmatrix}z+y\\z-y\\0\end{bmatrix}$, then the rank of L is 2, since its range is the xy-plane in $\mathbb{R}^3$.

The **null space** of a linear transformation is the set of vectors which are mapped to the zero vector by the

linear transformation.

Example 3.1.9

If $L\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix}\right) = \begin{bmatrix} z+y \\ z-y \\ 0 \end{bmatrix}$, then the null space of L is span $\left(\left\{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right\}\right)$, since $L(\mathbf{v}) = 0$ implies that $\mathbf{v} = \begin{bmatrix} x \\ 0 \\ 0 \end{bmatrix}$ for some $x \in \mathbb{R}$.

Exercise 3.1.15

Suppose that V and W are vector spaces and that L_1 and L_2 are linear transformations from V to W. Suppose that $\mathcal{B}$ is a basis of V and that $L_1(\mathbf{b}) = L_2(\mathbf{b})$ for all $\mathbf{b} \in \mathcal{B}$. Show that $L_1(\mathbf{v}) = L_2(\mathbf{v})$ for all $\mathbf{v} \in V$.

Exercise 3.1.16

What is the dimension of the null space of the linear transformation $L([x, y, z]) = [y, z, 0]$? What is the rank of L?

Exercise 3.1.17

(a) For $k \leq n$, let $P_k : \mathbb{R}^n \to \mathbb{R}^k$ be the linear transformation that projects a vector on to its first k components, i.e.

$$P_k(a_1, a_2, \cdots, a_k, \cdots, a_n) = (a_1, a_2, \cdots, a_k)$$

What is the rank of P_k? What is the nullity of P_k? What is the sum of the rank and the nullity of P_k?

(b) In this part, we will show that for any transformation T from $\mathbb{R}^n$ to $\mathbb{R}^m$, the sum of the rank of T and the nullity of T is equal to the value you found above for P_k.

(i) Consider a basis $\{\mathbf{v}_1, \ldots \mathbf{v}_k\}$ of the null space of T, and extend it to a basis

$$\{\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{v}_{k+1}, \ldots, \mathbf{v}_n\}$$

of $\mathbb{R}^n$.

(ii) Show that $\{T(\mathbf{v}_{k+1}), \ldots, T(\mathbf{v}_n)\}$ is linearly independent. Begin by assuming that a linear combination of these vectors is equal to the zero vector and do some work to conclude that all the weights must have been zero.

(iii) Show that $\{T(\mathbf{v}_{k+1}), \ldots, T(\mathbf{v}_n)\}$ spans the range of T. To do this, consider an arbitrary vector $\mathbf{w}$ in the range of T and show how it can be written as a linear combination of vectors in this list.

Exercise 3.1.18

Suppose you're designing an app that recommends cars. For every person in your database, you have collected twenty variables of data: age, height, gender, income, credit score, etc. In your warehouse are ten types of cars. You envision your recommendation system as a linear transformation $T : \mathbb{R}^{20} \to \mathbb{R}^{10}$ that takes in a person's data and then returns a number for each car, reflecting how well that car fits their needs. The rank of T can be as high as ten, which we might summarize by saying that your recommendation system can have ten degrees of complexity.

After some time, you find that storing all twenty variables takes up too much space in your database. Instead, you decide to take those twenty variables and apply a linear aggregate score function $S : \mathbb{R}^{20} \to \mathbb{R}^3$, with the three output components corresponding to health, personality, and finances. You also compute a linear map $R : \mathbb{R}^3 \to \mathbb{R}^{10}$ that takes in these three aggregate scores and returns a vector of recommendation values. The total recommendation system is the composition $R \circ S : \mathbb{R}^{20} \to \mathbb{R}^{10}$. What is the maximum possible rank of $R \circ S$? What does this mean for the complexity of this recommendation system?

3.2 Matrix algebra

A **matrix** is a rectangular array of numbers. We report the size of a matrix as *number of rows by number of columns*. In other words, a matrix with m rows and n columns is said to be an $m \times n$ matrix. We refer to the entry in the ith row and jth column of a matrix A as A's (i, j)th entry, and we denote it by $A_{i,j}$. In Julia or Python, the (i, j)th entry may be referenced as `A[i,j]`.

Matrices are versatile structures with a variety of problem-solving uses. For example,

1. A matrix can be thought of as a list of column vectors, so we can use a matrix to package many column vectors into a single mathematical object.
2. An $m \times n$ matrix can be thought of as a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$.

In this section, we will develop both of these perspectives and define some operations which facilitate common manipulations that arise when handling matrices.

3.2.1 Matrix operations

Definition 3.2.1: Matrix addition and scalar multiplication

We define **matrix addition** for two $m \times n$ matrices A and B entrywise: the sum $A + B$ is $m \times n$, and each entry is defined to be the sum of the corresponding entries in A and B.
Likewise, the product of a number c and an $m \times n$ matrix A is defined to be the $m \times n$ matrix each of whose entries is c times the corresponding entry of A.

Exercise 3.2.1

Find the value of c such that

$$\begin{bmatrix} 6 & 7 & -1 \\ 1 & 3 & 5 \end{bmatrix} + (1-c)\begin{bmatrix} 4 & -4 & 2 \\ -2 & 0 & 1 \end{bmatrix} = \begin{bmatrix} -2 & 15 & -5 \\ 5 & 3 & 3 \end{bmatrix}$$

Definition 3.2.2: Matrix-vector multiplication

If A is an $m \times n$ matrix and $\mathbf{x}$ is a column vector in $\mathbb{R}^n$, then $A\mathbf{x}$ is defined to be the linear combination of the columns of A with weights given by the entries of $\mathbf{x}$.

Example 3.2.1

If $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and $\mathbf{x} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$, then $A\mathbf{x} = 2\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 3\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$.

Exercise 3.2.2

Suppose that A is an $m \times n$ matrix. Show that $\mathbf{x} \mapsto A\mathbf{x}$ is a linear transformation.

In fact, *every* linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$ can be represented as $\mathbf{x} \mapsto A\mathbf{x}$ for some matrix A. The entries of the matrix A may be obtained from L by placing the components of $L(\mathbf{e}_1)$ in the first column of A, the components of $L(\mathbf{e}_2)$ in the second column, and so on. Then $L(\mathbf{x}) = A\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^n$, by Exercise 3.1.15.

Exercise 3.2.3

Find the matrix corresponding to the linear transformation $T([x, y, z]) = [z, x, y]$.

Exercise 3.2.4

Suppose that A is an $m \times n$ matrix and $\mathbf{b}$ is a vector in $\mathbb{R}^m$ with the property that the equation $A\mathbf{x} = \mathbf{b}$ has at least one solution $\mathbf{x} \in \mathbb{R}^n$. Show that the solution is unique if and only the columns of A are linearly independent.

We define matrix multiplication so that it corresponds to composition of the corresponding linear transformations.

Definition 3.2.3

If A is an $m \times n$ matrix and B is an $n \times p$ matrix, then AB is defined to be the matrix for which $(AB)(\mathbf{x}) = A(B\mathbf{x})$ for all $\mathbf{x}$.

Exercise 3.2.5: (Matrix Product)

Suppose that $A = \begin{bmatrix} 3 & -1 & 2 \\ 4 & 2 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 4 & -5 & 0 & 1 \\ 2 & 8 & 0 & 0 \\ -1 & 5 & 3 & 2 \end{bmatrix}$. Consider the matrix C defined so that, for all $1 \leq k \leq 4$, the kth column of C is defined to be the product of A and the kth column of B. Show that $C = AB$ according to Definition 3.2.3.

The principle you worked out in Exercise 3.2.5 is universal: the kth column of AB is the product of A and the kth column of B, for each column index k.

3.2.2 The inverse of a matrix

The range or null space of a matrix A is defined to be the range or null space of the corresponding linear transformation $\mathbf{x} \mapsto A\mathbf{x}$. The rank of A is defined to be the dimension of its range.

Example 3.2.2

The matrix $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ has rank 2, because the span of its columns is all of $\mathbb{R}^2$. The null space has dimension 1, since the solution of $A\mathbf{x} = \mathbf{0}$ is the span of $\{[1, 0, 0]\}$.

If the range of an $n \times n$ matrix is $\mathbb{R}^n$, then the corresponding linear transformation is an invertible function from $\mathbb{R}^n$ to $\mathbb{R}^n$:

Theorem 3.2.1: Invertible Matrix Theorem

Suppose that A is an $n \times n$ matrix. Then the following are equivalent (that is, for a given matrix they are either all true or all false).

(i) The transformation $\mathbf{x} \mapsto A\mathbf{x}$ from $\mathbb{R}^n$ to $\mathbb{R}^n$ is bijective.

(ii) The range of A is $\mathbb{R}^n$.

(iii) The null space of A is $\{\mathbf{0}\}$

In other words, for a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$, bijectivity, surjectivity, and injectivity are equivalent.

Proof

We begin by showing that (ii) and (iii) are equivalent. If the columns of A are linearly dependent, then the range of A is spanned by fewer than n vectors. Therefore, if the rank of A is equal to n, then the columns of A are linearly independent. This implies that a linear combination of the columns is equal to the zero vector only if the weights are all zero. In other words, the only solution of the equation $A\mathbf{x} = \mathbf{0}$ is the zero vector. In other words, the null space of A is $\{\mathbf{0}\}$.

Conversely, if the null space of A is $\{\mathbf{0}\}$, then the columns of A are linearly independent, and the rank of A is therefore equal to n.

By definition of bijectivity, (ii) and (iii) together imply (i), and (i) implies (ii) and (iii). Therefore, the three statements are equivalent.

If A is invertible, then the inverse function is also a linear transformation:

Exercise 3.2.6

Show that if T is a bijective linear transformation, then the inverse function T^{-1} is also linear.

Its matrix is called the **inverse** of A and is denoted A^{-1}. The matrices A and A^{-1} satisfy the equations $AA^{-1} = A^{-1}A = I$, where I denotes the $n \times n$ *identity* matrix, which has ones along the diagonal starting at the top left entry and zeros elsewhere.

Example 3.2.3

If $A = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$, then

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

Therefore $B(A\mathbf{x}) = (BA)\mathbf{x} = \mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^2$. So $B = A^{-1}$.

Exercise 3.2.7

Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be a linear transformation defined to be a reflection across the y-axis followed by a 15-degree clockwise rotation about the origin. Which of the following maps $T\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right)$ back to $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$?

(a) Reflection across the y-axis followed by a 15-degree counterclockwise rotation about the origin.

(b) A 15-degree counterclockwise rotation about the origin followed by a reflection across the y-axis.

Use the above example to write $(AB)^{-1}$ in terms of A and B when A and B are invertible matrices.

Exercise 3.2.8

Let A be a non-zero $n \times n$ matrix whose rank is k.

1. If $k = n$ and $\mathbf{b} \in \mathbb{R}^n$, explain why there exists only one vector $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{b}$.
2. Suppose $k < n$ and show that there are vectors in $\mathbb{R}^n$ for which the equation $A\mathbf{x} = \mathbf{b}$ has no solution.
3. If $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$ both satisfy $A\mathbf{x} = \mathbf{b}$ and $A\mathbf{y} = \mathbf{b}$ for some fixed vector $\mathbf{b} \in \mathbb{R}^n$, describe geometrically the set of points $(c_1, c_2) \in \mathbb{R}^2$ such that $A(c_1\mathbf{x} + c_2\mathbf{y}) = \mathbf{b}$.

Based on the above observations, can the equation $A\mathbf{x} = \mathbf{b}$ have exactly 10 solutions?

3.3 Dot products and orthogonality

3.3.1 The dot product

Consider a shop inventory which lists unit prices and quantities for each of the products they carry. For example, if the store has 32 small storage boxes at \$4.99 each, 18 medium-sized boxes at \$7.99 each, and 14 large boxes at \$9.99 each, then the inventory's price vector $\mathbf{p}$ and quantity vector $\mathbf{q}$ are

$$\mathbf{p} = \begin{bmatrix} 4.99 \\ 7.99 \\ 9.99 \end{bmatrix}, \quad \mathbf{q} = \begin{bmatrix} 32 \\ 18 \\ 14 \end{bmatrix}.$$

The total value of the boxes in stock is

$$(32)(\$4.99) + (18)(\$7.99) + (14)(\$9.99) = \$443.36.$$

This operation—multiplying two vectors' entries in pairs and summing—arises often in applications of linear algebra and is also foundational in basic linear algebra theory.

Definition 3.3.1

The **dot product** of two vectors in $\mathbb{R}^n$ is defined by

$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n.$$

Example 3.3.1

If $\mathbf{x} = \begin{bmatrix} 1 \\ 3 \\ 5 \\ 7 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} 2 \\ 4 \\ 6 \\ 8 \end{bmatrix}$, then $\mathbf{x} \cdot \mathbf{y} = 1 \cdot 2 + 3 \cdot 4 + 5 \cdot 6 + 7 \cdot 8 = 100$.

One of the most algebraically useful features of the dot product is its linearity: $\mathbf{x} \cdot (c\mathbf{y} + \mathbf{z}) = c\mathbf{x} \cdot \mathbf{y} + \mathbf{x} \cdot \mathbf{z}$.

Exercise 3.3.1

Show that $(\mathbf{a} + \mathbf{b}) \cdot (\mathbf{a} + \mathbf{b}) = |\mathbf{a}|^2 + 2\mathbf{a} \cdot \mathbf{b} + |\mathbf{b}|^2$ for all vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^n$.

The dot product $\mathbf{x} \cdot \mathbf{y}$ has a geometric connection with the angle θ between two vectors $\mathbf{x}$ and $\mathbf{y}$, via

$$\mathbf{x} \cdot \mathbf{y} = |\mathbf{x}||\mathbf{y}| \cos\theta. \tag{3.3.1}$$

$\mathbf{x} \cdot \mathbf{y} = 0$ if and only if $\mathbf{x}$ and $\mathbf{y}$ are orthogonal.

Exercise 3.3.2

In natural language processing, one basic way to compare a finite number of text documents is to use *vectorized word counts*. Suppose the documents have a combined total of n distinct words, which are arranged in some order. Each document is then associated with a vector of length n whose ith entry indicates the number of times the ith word occurs in the associated document.

One way to measure similarity between two documents is to take the dot product of the associated unit vectors: If two documents A and B have associated vectors $\mathbf{a}$ and $\mathbf{b}$ respectively, their similarity is defined by

$$S(A, B) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}.$$

By (3.3.1), we have $0 \leq S(A, B) \leq 1$ for any two documents A and B. Documents with no words in common are associated with orthogonal vectors and thus have 0 similarity. If two documents have similarity 1, their associated vectors are scalar multiples of each other, meaning that they have the same words and that the words appear in the same proportions.

Find the vectorized word count similarity between the following sentences:

"The rain in Spain falls mainly in the plain"

"The plain lane in Spain is mainly a pain"

Exercise 3.3.3

Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be a list of orthogonal non-zero vectors, that is, for all $i, j \in \{1, \ldots, n\}$, suppose that $\mathbf{v}_i \cdot \mathbf{v}_j = 0$ whenever $i \neq j$. Show that this list is linearly independent.

3.3.2 The transpose

The dot product gives us a compact way to express the formula for an entry of a matrix product: to obtain the (i, j)th entry of a matrix product AB, we dot the ith row of A and the jth column of B.

However, the matrix product by itself is not quite flexible enough to handle a common use case: suppose we have two matrices A and B which contain tabular data stored in the same format. For example, suppose that the columns of A store the vectorized word counts for a series of emails sent from Alice, while B stores vectorized word counts for a series of emails sent from Bob. If we want to calculate the similarity of each of Alice's email to each of Bob's emails, then we want to dot the *columns* of A—not its rows—with the columns of B.

So we define the **transpose** A' of a matrix A to be the matrix resulting from switching A's rows and columns.

Definition 3.3.2

If A is an $m \times n$ matrix, then its **transpose** A' is defined to be the matrix with n rows whose ith row is equal to the ith column of A, for each i from 1 to n.

Example 3.3.2

If $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$, then $A' = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$.

With this definition in hand, we can write the matrix whose entries are the dot products of columns of A and B as $A'B$.

Let's develop a few properties of the transpose so that we can manipulte matrix expressions involving the transpose. First, we note that the transpose is a *linear* operator, meaning that $(cA + B)' = cA' + B'$ whenever c is a constant and A and B are matrices.

Taking the transpose also interacts nicely with matrix multiplication:

Exercise 3.3.4

Suppose that A is an $m \times n$ matrix and that B is an $n \times p$ matrix. Exactly one of the following expressions is equal to $(AB)'$ in general—identify the correct answer choice by checking the dimensions of each matrix in each expression.

1. $A'B'$
2. $B'A'$
3. ABA'

Confirm your conjecture numerically in Julia and paste your code in the answer box. You can generate a random $m \times n$ matrix using `rand(m,n)`, the transpose of `A` is computed as `A'`, and the product of `A` and `B` is `A * B`.

In some applications, a matrix will have the property that its (i, j)th entry is necessarily equal to its (j, i)th entry. For example, suppose we have an ordered list of 100 cell phone towers, and we define the 100×100 matrix whose (i, j)th entry is equal to the distance from tower i to tower j. Such a matrix is said to be *symmetric*.

Definition 3.3.3

If A is an $n \times n$ matrix satisfying the equation $A = A'$, we say that A is **symmetric**.

Exercise 3.3.5

Suppose that A is a symmetric matrix, B is a matrix, and $c \in \mathbb{R}$. Which of the following is necessarily equal to $(c^2(A + B)' + A)'$?

1. $c^2A' + B$
2. $(c^2 - 1)A' + B'$
3. $(c^2 + 1)A + c^2B$
4. $(c^2 - 1)A + B'$
5. $(c^2 + 1)A + c^2B'$

In the case where A is a $n \times 1$ matrix and B is an $n \times 1$ for some n, then $A'B$ is a 1×1 matrix, which we may think of as just a number. This means that if $\mathbf{x}$ and $\mathbf{y}$ are vectors in $\mathbb{R}^n$ then the dot product $\mathbf{x} \cdot \mathbf{y}$ may be written as $\mathbf{x}'\mathbf{y}$. This representation can be useful for manipulating expressions involving dot products.

Exercise 3.3.6

Show that

$$\mathbf{u} \cdot (A\mathbf{v}) = (A'\mathbf{u}) \cdot \mathbf{v}$$

for all $m \times n$ matrices A and all vectors $\mathbf{u} \in \mathbb{R}^m$ and $\mathbf{v} \in \mathbb{R}^n$.

In other words, we may move a matrix which is pre-multiplying one of the vectors in a dot product to the other vector, at the cost of taking its transpose. Let's look at one important application of this property.

Exercise 3.3.7

Show that $\mathbf{x} \cdot (A'A\mathbf{x}) \geq 0$ for all $m \times n$ matrices A and all $\mathbf{x} \in \mathbb{R}^n$.

The **orthogonal complement** $V^\perp$ of a vector space $V \subset \mathbb{R}^n$ is the set of vectors in $\mathbb{R}^n$ which are orthogonal to every vector in V. For example, the orthogonal complement a two-dimensional subspace V of $\mathbb{R}^3$ is the line through the origin perpendicular to the plane of vectors in V.

Exercise 3.3.8

The orthogonal complement of the span of the columns of a matrix A is equal to which of the following?

(a) The range of A

(b) The null space of A

(c) The range of A'

(d) the null space of A'

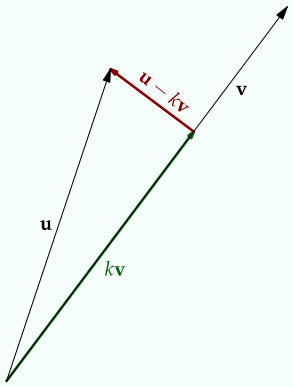
For any vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$, it is always possible to write $\mathbf{u}$ as the sum of a multiple of $\mathbf{v}$ and a vector which is perpendicular to $\mathbf{v}$:

Exercise 3.3.9: Orthogonal decomposition

Suppose that $\mathbf{u}$ and $\mathbf{v}$ are nonzero vectors in $\mathbb{R}^n$. Solve the equation

$$k\mathbf{v} \cdot (\mathbf{u} - k\mathbf{v}) = 0$$

for k to find the multiple of $\mathbf{v}$ which is perpendicular to its difference with $\mathbf{u}$.



If $\mathbf{u}$ is written as $k\mathbf{v} + \mathbf{w}$ where $\mathbf{w}$ is perpendicular to $\mathbf{v}$, then we call $k\mathbf{v}$ the **projection** of $\mathbf{u}$ onto $\mathbf{v}$.

Theorem 3.3.1: Gram-Schmidt

Every vector space $V \subset \mathbb{R}^n$ has an orthogonal basis.

Proof

Suppose that $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is a basis of $\mathbb{R}^n$. We will build our orthogonal basis by orthogonalizing $\mathcal{V}$ one vector at a time.

Define $\mathbf{b}_1 = \mathbf{v}_1$, and then define $\mathbf{b}_2$ to be the part of $\mathbf{v}_2$ which is orthogonal to $\mathbf{b}_1$:

$$\mathbf{b}_2 = \mathbf{v}_2 - \frac{\mathbf{b}_1 \cdot \mathbf{v}_2}{|\mathbf{b}_1|^2}\mathbf{b}_1.$$

Similarly, we project $\mathbf{v}_3$ onto $\mathbf{b}_1$ and onto $\mathbf{b}_2$ and subtract off both of these projections:

$$\mathbf{b}_3 = \mathbf{v}_3 - \frac{\mathbf{b}_2 \cdot \mathbf{v}_3}{|\mathbf{b}_2|^2}\mathbf{b}_2 - \frac{\mathbf{b}_1 \cdot \mathbf{v}_3}{|\mathbf{b}_1|^2}\mathbf{b}_1.$$

Then $\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$ has the same span as $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ and is pairwise orthogonal. We may continue this procedure (project each new $\mathbf{v}_i$ onto each of the preceding $\mathbf{b}$'s and subtract off all of these projections) until we reach the end of the list, thereby obtaining an orthogonal basis of V.

Theorem 3.3.2

Suppose $V \subset \mathbb{R}^n$ is a vector space. Then every vector $\mathbf{u} \in \mathbb{R}^n$ can be written as the sum of a vector in V and a vector in $V^\perp$.

Proof

Consider an orthogonal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ of V, and define

$$\mathbf{v} = \frac{\mathbf{v}_1 \cdot \mathbf{u}}{|\mathbf{v}_1|^2}\mathbf{v}_1 + \cdots + \frac{\mathbf{v}_k \cdot \mathbf{u}}{|\mathbf{v}_k|^2}\mathbf{v}_k$$

Then $\mathbf{v}$ is in V and $\mathbf{u} - \mathbf{v}$ is perpendicular to all of the $\mathbf{v}_i$'s and therefore to every vector in V.

Exercise 3.3.10

Suppose that V is a d-dimensional vector space in $\mathbb{R}^n$. Show that there is a basis of $\mathbb{R}^n$ whose first d vectors form a basis for V and whose last $n - d$ vectors form a basis for $V^\perp$.

3.3.3 Matrices with orthonormal columns

Suppose we can write a given transformation T as a composition involving (i) a single transformation Λ which scales space along the coordinate axes, and (ii) some other transformations which preserve distances and angles—like rotations and reflections in $\mathbb{R}^2$ or $\mathbb{R}^3$. Such a decomposition of T would be useful because it isolates the space-distorting behavior of T in the simple transformation Λ. In preparation for developing such a decomposition, let's study transformations which are distance-preserving and angle-preserving.

A transformation $x \mapsto U\mathbf{x}$ from $\mathbb{R}^n$ to $\mathbb{R}^n$ is distance-preserving if the norm of $\mathbf{x}$ is the same as the norm of $U\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^n$. Using dot products, we can write the distance-preserving condition as

$$\mathbf{x} \cdot \mathbf{x} = (U\mathbf{x}) \cdot (U\mathbf{x})$$

If the transformation preserves angles as well as distances, then $(U\mathbf{x}) \cdot (U\mathbf{y})$ must also be equal to $\mathbf{x} \cdot \mathbf{y}$ for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$. Rewriting this equation using transposes, we see that we want

$$\mathbf{x}'\mathbf{y} = \mathbf{x}'U'U\mathbf{y}$$

for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$. This identity only holds if $U'U$ is equal to the identity matrix. This leads us to the following definition.

Definition 3.3.4: Orthogonal matrix

A square matrix U is **orthogonal** if $U'U$ is equal to the identity matrix.

Equivalently, we can say that a square matrix is orthogonal if and only if its columns are *orthonormal*, which means that they are orthogonal and have unit norm. If a non-square matrix U satisfies $U'U = I$, then we refer to U as a *matrix with orthonormal columns*.

Exercise 3.3.11

(i) Explain why, for an $m \times n$ matrix U with orthonormal columns, we must have $m \geq n$. (ii) Explain why the rank of U is n.

If U is an $m \times n$ matrix with orthonormal columns and if $n < m$, then UU' is an $m \times m$ matrix of rank n and therefore cannot be the identity matrix. In fact, UU' is a projection matrix:

Exercise 3.3.12

Show that if U is an $m \times n$ matrix with orthonormal columns, then UU' is the matrix of the transformation which projects each vector in $\mathbb{R}^m$ onto the n-dimensional subspace of $\mathbb{R}^m$ spanned by the columns of U.

Exercise 3.3.13

Let $\mathbf{v}$ be a vector in $\mathbb{R}^n$, and consider the linear transformation $T : \mathbb{R}^n \to \mathbb{R}$ defined as $T(\mathbf{x}) = \mathbf{v} \cdot \mathbf{x}$. What is the rank of T? Geometrically describe the null space of T.

3.4 Eigenvalues and matrix diagonalization

3.4.1 Eigenpairs

In this section we will see how we can better understand a linear transformation by breaking it down into simpler linear transformations.

Let T be a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$. Suppose that $\mathcal{B}$ is a basis of $\mathbb{R}^n$, that V is the span of some of the vectors in $\mathcal{B}$, and that W is the span of the remaining vectors in $\mathcal{B}$. Then any vector in $\mathbb{R}^n$ can be written as the sum of a vector $\mathbf{v}$ in V and a vector $\mathbf{w}$ in W. Since $T(\mathbf{v} + \mathbf{w}) = T(\mathbf{v}) + T(\mathbf{w})$, we can see how T behaves on all of $\mathbb{R}^n$ if we know how it behaves on V and on W. This decomposition is particularly helpful if V and W are chosen so that T behaves in a simple way on V and on W.

Given such a decomposition of $\mathbb{R}^n$ into the vector spaces V and W, we can apply the same idea to split V and W into lower-dimensional vector spaces and repeat until no more splits are possible. The most optimistic outcome of this procedure would be that we get all the way down to n one-dimensional subspaces and that T acts on each of these subspaces by simply scaling each vector in that subspace by some factor. In other words, we would like to find n vectors $\mathbf{v}$ for which $T(\mathbf{v})$ is a scalar multiple of $\mathbf{v}$. This leads us to the following definition.

Definition 3.4.1

An eigenvector $\mathbf{v}$ of an $n \times n$ matrix A is a *nonzero* vector with the property that $A\mathbf{v} = \lambda\mathbf{v}$ for some $\lambda \in \mathbb{R}$ (in other words, A maps $\mathbf{v}$ to a vector which is either zero or parallel to $\mathbf{v}$). We call λ an **eigenvalue** of A, and we call the eigenvector together with its eigenvalue an **eigenpair**.

Example 3.4.1

Every nonzero vector is an eigenvector (with eigenvalue 1) of the identity matrix.

Exercise 3.4.1

Find a matrix with eigenpairs $([1,0],2)$ and $([1,1],3)$. Sketch the images of some gridlines under multiplication by this matrix to show how it scales space along the lines through its eigenvectors.

Exercise 3.4.2

In general, if $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are eigenvectors of A with the same eigenvalue λ and $\mathbf{v} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$ for some weights $c_1, \ldots, c_n$ such that $c_i \neq 0$ for at least one $i \in \{1, \ldots, n\}$, then $\mathbf{v}$ is also an eigenvector of A with eigenvalue λ because

$$\begin{aligned} A\mathbf{v} &= A(c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n) \\ &= c_1A\mathbf{v}_1 + \cdots + c_nA\mathbf{v}_n \\ &= c_1\lambda\mathbf{v}_1 + \cdots c_n\lambda\mathbf{v}_n \\ &= \lambda(c_1\mathbf{v}_1 + \cdots c_n\mathbf{v}_n) \\ &= \lambda\mathbf{v}. \end{aligned}$$

Let A be a 4×4 matrix, with eigenvectors $\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 0 \\ 2 \\ -3 \end{bmatrix}$, both with eigenvalue 3. Find $A\left(\begin{bmatrix} 5 \\ 5 \\ 8 \\ -12 \end{bmatrix}\right)$.

Exercise 3.4.3

Let $V \subset \mathbb{R}^n$ be a subspace spanned by the eigenvectors of a matrix A. If $\mathbf{v} \in V$, which of the following are necessarily true?

1. $A\mathbf{v} \in V$.
2. $A\mathbf{v}$ is orthogonal to every vector in V.
3. $A\mathbf{v}$ and $\mathbf{v}$ are always linearly dependent.

Exercise 3.4.4

Suppose A is a matrix with a 3-eigenvector $\mathbf{v}$ and a 2-eigenvector $\mathbf{w}$. Let $\mathbf{u} = \mathbf{v} + \mathbf{w}$. Explain why

$$\lim_{n\to\infty} \frac{|A^n\mathbf{u}|}{|A^n\mathbf{v}|} = 1$$

If an $n \times n$ matrix A has n linearly independent eigenvectors, then we can think of the one-dimensional subspaces spanned by each of these vectors as (not necessarily orthogonal) axes along which A acts by scaling.

In matrix terms, we can define V to be the matrix with the eigenvectors of A as columns. Then from the definition of an eigenpair, we have

$$AV = V\Lambda,$$

where Λ is a matrix whose diagonal entries are the eigenvalues (in order corresponding to the columns of V) and whose other entries are zero. We conclude that $A = V\Lambda V^{-1}$, where Λ is a diagonal matrix, and we say that A is **diagonalizable**.

Exercise 3.4.5

Some matrices are not diagonalizable, because they correspond to geometric transformations that cannot be viewed as scaling along any set of axes. Use this geometric intuition to come up with a 2×2 matrix which is not diagonalizable.

Exercise 3.4.6

Suppose that we have diagonalized A as $A = VDV^{-1}$. Using matrix multiplication, determine which of the following is equal to A^3.

1. $V^3D^3V^{-3}$.
2. VD^3V^{-1}.
3. V^3DV^{-3}.

Let B be another matrix, with 3-eigenvector $\mathbf{v}_1$ and (-2)-eigenvector $\mathbf{v}_2$. Let $\mathbf{u} = 2\mathbf{v}_1 + \mathbf{v}_2$. Which of the following is equal to $B^n(\mathbf{u})$?

1. $2(3)^n\mathbf{v}_1 + (-2)^n\mathbf{v}_2$.
2. $(2(3) - 1)^n\mathbf{u}$.
3. $(2(3)^n - 1)\mathbf{u}$.
4. None of the above.

3.4.2 Positive definite matrices

A **positive definite** matrix A is a symmetric matrix whose eigenvalues are all positive. A **positive semidefinite** matrix A is a symmetric matrix whose eigenvalues are all nonnegative. Equivalently, a matrix A is positive semidefinite if $\mathbf{x}'A\mathbf{x} \geq 0$ for all $\mathbf{x}$.

Negative definite and *negative semidefinite* matrices are defined analogously.

Exercise 3.4.7

(i) Is the sum of two positive definite matrices necessarily positive definite?

(ii) Is the product of two positive definite matrices necessarily positive definite?

If A is an $m \times n$ matrix, then $A'A$ is its *Gram matrix*. The Gram matrix of A is always positive semidefinite:

Exercise 3.4.8

Let $X = A'A$ be a Gram matrix, and let $\mathbf{v}$ be a vector. Which of the following is equal to $\mathbf{v}'X\mathbf{v}$?

1. $|A\mathbf{v}|^2$.
2. $A^2\mathbf{v}$.
3. $\mathbf{v}'A^2\mathbf{v}$.

Using your answer above, explain why a Gram matrix is always positive semidefinite, but not necessarily positive definite.

Exercise 3.4.9

Explain why the rank of A is equal to the rank of $A'A$. (Hint: consider the null spaces of A and $A'A$)

The eigenspace decomposition is even easier to understand if the eigenvectors happen to be orthogonal. It turns out that this happens exactly when the matrix is *symmetric*:

!!!

Theorem 3.4.1: Spectral Theorem

If A is an $n \times n$ symmetric matrix, then A is *orthogonally* diagonalizable, meaning that A has n eigenvectors which are pairwise orthogonal.
Conversely, every orthogonally diagonalizable matrix is symmetric.

In other words, if A is symmetric, then the one-dimensional subspaces along which A is decomposed form a set of axes for $\mathbb{R}^n$ which are orthogonal. In matrix terms, we have

$$A = V\Lambda V',$$

for some orthogonal matrix V.

Exercise 3.4.10

Given an invertible matrix A, we are often interested in solving a system of the form $A\mathbf{x} = \mathbf{b}$. Our knowledge of $\mathbf{b}$ is seldom perfect however, so it is important to consider what happens to the solution if we replace $\mathbf{b}$ with a slightly different vector $\widehat{\mathbf{b}}$.

It is possible that a small change in $\mathbf{b}$ leads to a substantial change in the vector $\mathbf{x} = A^{-1}\mathbf{b}$.

(i) Find an invertible 2×2 matrix A all of whose entries are between -2 and 2 and a vector $\mathbf{b}$ with entries between -2 and 2 and another vector $\widehat{\mathbf{b}}$ whose components are nearly equal to those of $\mathbf{b}$ for which $A^{-1}\mathbf{b}$ and $A^{-1}\widehat{\mathbf{b}}$ are not very close.

To be concrete, let's say "nearly equal" means "having ratio between 0.99 and 1.01", and let's say that "not very close" means "having a difference whose norm is greater than the norm of either".

(ii) Find the eigenvalues of your matrix A.

3.4.3 Polar decomposition

The Gram matrix of a square matrix A is a useful tool for understanding the behavior of A. Let's define the matrix $\sqrt{A'A}$ to be $V\Lambda^{1/2}V'$, where $V\Lambda V'$ is the diagonalization of $A'A$ and $\Lambda^{1/2}$ is the matrix obtained by taking the square root of the diagonal entries of Λ. Then $\sqrt{A'A}$ is symmetric and satisfies

$$\sqrt{A'A}\sqrt{A'A} = V\Lambda^{1/2}V'V\Lambda^{1/2}V' = A'A.$$

The matrix $\sqrt{A'A}$ is simpler to understand than A because it is symmetric and positive definite, yet it transforms space in nearly the same way as A: if $\mathbf{x} \in \mathbb{R}^n$, then

$$|A\mathbf{x}|^2 = \mathbf{x}'A'A\mathbf{x} = \mathbf{x}'\sqrt{A'A}\sqrt{A'A}\mathbf{x} = |\sqrt{A'A}\,\mathbf{x}|^2.$$

In other words, for all $\mathbf{x}$, the images of $\mathbf{x}$ under A and under $\sqrt{A'A}$ have equal norm. This means that for each $\mathbf{x} \in \mathbb{R}^n$, there is an orthogonal transformation from the range of $\sqrt{A'A}$ to the range of A which sends $A\mathbf{x}$ to $\sqrt{A'A}\mathbf{x}$. It turns out that this orthogonal transformation is the same for all $\mathbf{x}$.

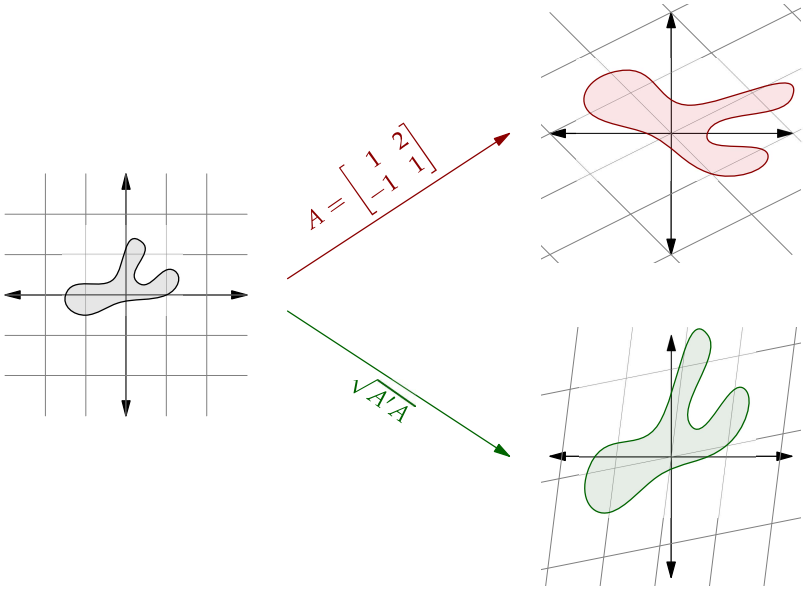


Figure 3.3 The grid-line images under A and $\sqrt{A'A}$ have the same shape; they are related by an orthogonal transformation.

Even if the range of $\sqrt{A'A}$ is not all of $\mathbb{R}^n$, we can extend this orthogonal transformation to an orthogonal transformation on $\mathbb{R}^n$. Thus we arrive at the *polar decomposition*:

Theorem 3.4.2: Polar Decomposition

For any $n \times n$ matrix A, there exists an orthogonal matrix R such that

$$A = R\sqrt{A'A}.$$

This representation is useful because it represents an arbitrary square matrix as a product of matrices whose properties are easier to understand (the orthogonal matrix because it is distance- and angle-preserving, and the positive-definite matrix $\sqrt{A'A}$ because it is orthogonally diagonalizable, by the Spectral Theorem [Theorem 3.4.1]).

Exercise 3.4.11

Let's explore a fast method of computing a polar decomposition $A = R\sqrt{A'A}$. This method actually works by calculating R and then recovering $\sqrt{A'A}$ as $R^{-1}A$ (since this is computationally faster than calculating the matrix square root). We call R the *orthogonal part* of A and $\sqrt{A'A}$ the *symmetric part* of A.

We set $R_0 = A$ and define the iteration

$$R_{k+1} = \frac{R_k + (R_k')^{-1}}{2}$$

Let's see why this converges to R.

1. Defining $P = \sqrt{A'A}$ and using the equation $A = RP$, show that

$$R_1 = \frac{A + (A')^{-1}}{2} = R\left(\frac{P + P^{-1}}{2}\right).$$

2. Use the prior step to explain why the R_k's all have the same orthogonal parts and have symmetric parts converging to the identity matrix.

   Hint: consider the eigendecompositions of the symmetric parts. You may assume that the sequence defined by $x_{k+1} = \frac{1}{2}(x_k + 1/x_k)$ converges to 1 regardless of the starting value $x_0 \in \mathbb{R}$.

3. Write some code to apply this algorithm to the matrix

```
A = [1 3 4; 7 -2 5; -3 4 11]
```

   and confirm that the resulting matrices R and P satisfy $R'R = I$ and $P^2 = A'A$.

Exercise 3.4.12

Show that the product of two matrices with orthonormal columns has orthonormal columns.

3.5 Singular value decomposition

In this section we will combine the polar decomposition and the spectral theorem to obtain one of the most powerful ideas in linear algebra: the **singular value decomposition**.

The polar decomposition tells us that any square matrix A is almost the same as some symmetric matrix, and the spectral theorem tells us that a symmetric matrix is almost the same as a simple scaling along the coordinate axes. (In both cases, the phrase "almost the same" disguises a composition with an orthogonal transformation.) We should be able to combine these ideas to conclude that *any* square matrix is basically the same as a simple scaling along the coordinate axes!

Let's be more precise. Suppose that A is a square matrix. The polar decomposition tells us that

$$A = R\sqrt{A'A}$$

for some orthogonal matrix R. The spectral theorem tells us that $\sqrt{A'A} = V\Sigma V'$ for some orthogonal matrix V and a diagonal matrix Σ with nonnegative diagonal entries. Combining these equations, we get

$$A = RV\Sigma V'.$$

Since a product of orthogonal matrices is orthogonal, we can define $U = RV$ and obtain the **singular value decomposition** (SVD) of A:

$$A = U\Sigma V' \tag{3.5.1}$$

where U and V are orthogonal matrices.

We can visualize (3.5.1) geometrically making a figure like the one shown here, which illustrates the successive effects of each map in the composition $U\Sigma V'$. If we draw grid lines on the *second* plot (just before Σ is applied) and propagate those grid lines to the other plots, then we endow the domain and range of A with orthogonal sets of gridlines with A mapping one to the other.

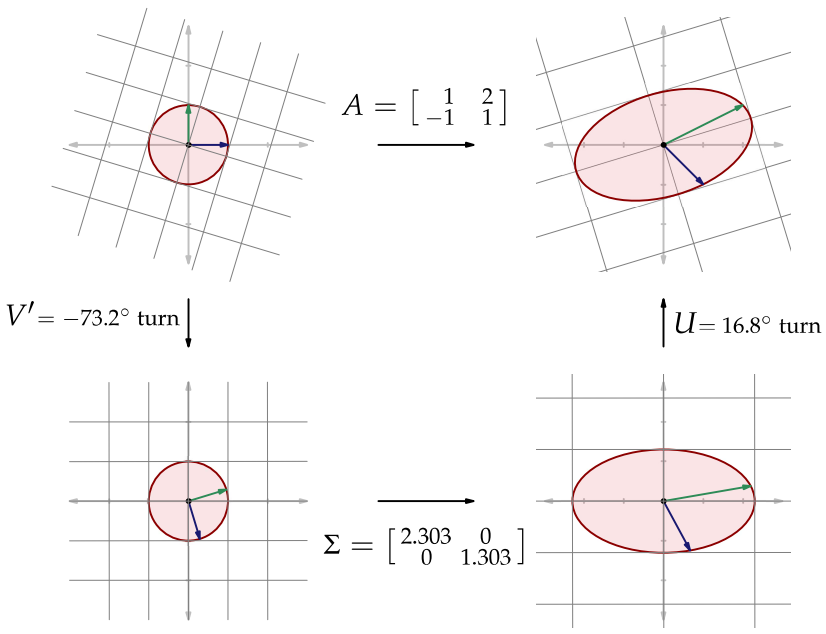


Figure 3.4 The matrix A maps one set of orthogonal grid lines to another

We can extend the singular value decomposition to rectangular matrices A (that is, matrices which are not necessarily square) by adding rows or columns of zeros to a rectangular matrix to get a square matrix, applying the SVD to that square matrix, and then

We can trim the resulting Σ matrix as well as either U or V' (depending on which dimension of A is smaller) and get a decomposition of the form $A = U\Sigma V'$ where U is an $m \times m$ orthogonal matrix, V' is an $n \times n$ orthogonal matrix, and Σ is a rectangular $m \times n$ diagonal matrix. This version of the SVD, called the *full* SVD, can be reduced further to obtain the following *thin* SVD:

Theorem 3.5.1: Singular value decomposition (thin)

Suppose that A is an $m \times n$ matrix with $m \leq n$. Then there exist matrices U and V with orthonormal columns and a diagonal matrix Σ such that

$$A = \underbrace{U}_{m\times m}\underbrace{\Sigma}_{m\times m}\underbrace{V'}_{m\times n},$$

Similarly, if $n \leq m$, there exist U, V, and Σ satisfying the above properties such that

$$A = \underbrace{U}_{m\times n}\underbrace{\Sigma}_{n\times n}\underbrace{V'}_{n\times n}.$$

We call $A = U\Sigma V'$ the a **singular value decomposition** (or SVD) of A.The diagonal entries of Σ are called the **singular values** of A.

The diagonal entries of Σ, which are the square roots of the eigenvalues of $A'A$, are called the **singular values** of A. The columns of U are called *left* singular vectors, and the columns of V are called *right* singular vectors.

Looking at the bottom half of Figure 3.4, we see that the singular values of A are the lengths of the semi-axes of the ellipsoid in $\mathbb{R}^m$ obtained as the image under A of the unit ball in $\mathbb{R}^n$. Moreover, the directions of these axes are the columns of U, since they are the images under U of the standard basis vectors. We will see an important application of this feature of the SVD in the probability chapter when we discuss *principal component analysis*.

As an example of how the singular value decomposition can be used to understand the structure of a linear transformation, suppose that A is an $m \times n$ matrix. The **Moore-Penrose pseudoinverse** A^+ of A is defined to be $V\Sigma^+U'$, where Σ^+ is the matrix obtained by inverting each nonzero element of Σ. The pseudoinverse is a swiss-army knife for solving the linear system $A\mathbf{x} = \mathbf{b}$:

1. If A is square and invertible, then $A^+ = A^{-1}$
2. If $A\mathbf{x} = \mathbf{b}$ has no solution, then $A^+\mathbf{b}$ is the value of $\mathbf{x}$ which minimizes $|A\mathbf{x} - \mathbf{b}|^2$.
3. If $A\mathbf{x} = \mathbf{b}$ has multiple solutions, then $A^+\mathbf{b}$ is the solution with minimal norm.

Exercise 3.5.1

Show that $\begin{bmatrix} -160 & -120 \\ -12 & -134 \\ 141 & 12 \end{bmatrix}$ has SVD $\begin{bmatrix} -\frac{4}{5} & 0 \\ -\frac{9}{25} & -\frac{4}{5} \\ \frac{12}{25} & -\frac{3}{5} \end{bmatrix} \begin{bmatrix} 250 & 0 \\ 0 & 125 \end{bmatrix} \begin{bmatrix} \frac{4}{5} & \frac{3}{5} \\ -\frac{3}{5} & \frac{4}{5} \end{bmatrix}$. Find its Moore-Penrose pseudoinverse.

We close this section with a computational exercise illustrating another widely applicable feature of the singular value decomposition.

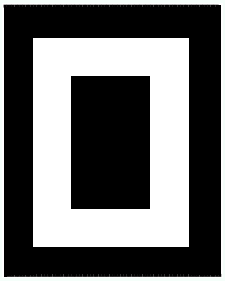
Exercise 3.5.2

(i) Show that if $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are the columns of U, $\mathbf{v}_1, \ldots \mathbf{v}_n$ are the columns of V, and $\sigma_1, \ldots, \sigma_n$ are the diagonal entries of Σ, then

$$A = \sigma_1 \mathbf{u}_1 \mathbf{v}_1' + \sigma_2 \mathbf{u}_2 \mathbf{v}_2' + \cdots + \sigma_n \mathbf{u}_n \mathbf{v}_n'. \tag{3.5.2}$$

(ii) The equation (3.5.2) is useful for *compression*, because terms with sufficiently small singular value factors can be dropped and the remaining vectors and singular values can be stored using less space. Suppose that A is a 256×128 matrix—how many entries does A have, and how many entries do $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$, $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ have in total?

(iii) The Julia code below creates a matrix A with pixel values for the image shown. How many nonzero singular values does A have? Explain how you can tell just from looking at the picture.



Note: You can type `\div«tab»` to get a division symbol for integer division and `\Sigma«tab»` to get Σ

```
using LinearAlgebra, Plots
m = 80
n = 100
a = m ÷ 8
b = m ÷ 4
A = ones(m,n)

function pixel(i,j)
    if (a ≤ i ≤ b || m-b ≤ i ≤ m-a) && a ≤ j ≤ n - a
        0
    elseif (a ≤ j ≤ b || n-b ≤ j ≤ n-a) && a ≤ i ≤ m - a
        0
    else
        1
    end
end

A = [pixel(i,j) for i=1:m,j=1:n]

U, Σ, V = svd(A)
heatmap(A)
```

(iv) Now add some noise to the image: `B = A + 0.05*randn(m,n)`.

Display this new matrix B, and also find the matrix obtained by keeping only the first three terms of (3.5.2) for this matrix B. Which looks more like the original image A: (i) B or (ii) the three-term approximation of B?

Hint: you can achieve this computationally either by setting some singular values to 0 or by indexing the matrices U, Σ, and V' appropriately. Also, you will need the function `diagm` to generate a diagonal matrix from the vector of Σ values returned by `svd`.

3.6 Determinants

The *determinant* of a square matrix A is a single number which captures some important information about how the transformation $\mathbf{x} \mapsto A\mathbf{x}$ behaves. In this section, we will develop a geometrically-motivated definition of the determinant.

Exercise 3.6.1

Suppose that R is a region in $\mathbb{R}^n$ and that A is an $n \times n$ matrix. Consider the singular value decomposition $A = U\Sigma V'$.

1. Let $L_1(\mathbf{x}) = V'\mathbf{x}$. By what factor does L_1 transform volumes?
2. Let $L_2(\mathbf{x}) = \Sigma\mathbf{x}$. In terms of the entries of Σ, by what factor does L_1 transform volumes?
3. Let $L_3(\mathbf{x}) = U\mathbf{x}$. By what factor does L_3 transform volumes?

From Exercise 3.6.1, we see that a linear transformation T from $\mathbb{R}^n$ to $\mathbb{R}^n$ scales the volume of any n-dimensional region by the same factor: the *volume scale factor* of T.

Exercise 3.6.2

Find the volume scale factor of the matrix $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & k & 0 \end{bmatrix}$ by describing how the matrix transforms a region in $\mathbb{R}^3$.

Another geometrically relevant piece of information about T is whether it preserves or reverses orientations. For example, rotations in $\mathbb{R}^2$ are orientation preserving, while reflections are orientation reversing. Let's define the *orientation factor* of T to be $+1$ if T is orientation preserving and -1 if T is orientation reversing.

Definition 3.6.1

We define the **determinant** of a transformation T to be the product of its orientation factor and its volume scale factor.
We define the determinant of a matrix A to be the determinant of the corresponding linear transformation $\mathbf{x} \mapsto A\mathbf{x}$.

Exercise 3.6.3

Interpret $A = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$ geometrically and use this interpretation to find $\det A$, the determinant of A.

There is relatively simple formula for $\det A$ in terms of the entries of A. For example,

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

is the determinant of a 2×2 matrix. However this formula is terribly inefficient if A has many entries (it has

$n!$ terms for an $n \times n$ matrix), and all scientific computing environments have a `det` function which uses much faster methods.

Exercise 3.6.4

For various values of n, use the Julia expression `det(rand(-9:9,n,n))` to find the determinant of an $n \times n$ matrix filled with random single-digit numbers. How large does n have to be for the determinant to be large enough to consistently overflow?

Exercise 3.6.5

Suppose that A and B are 3×3 matrices, with determinant 5 and $\frac{1}{2}$ respectively. Suppose that $R \subset \mathbb{R}^3$ is a 3D region modeling a fish whose volume is 14. What is the volume of the transformed fish $BA(R)$?

1. 19.5
2. 35
3. 12
4. 16.5

Exercise 3.6.6

Let $R \subset \mathbb{R}^3$ be 3D region modeling a fish, and suppose A an invertible 3×3 matrix. If R has volume 15 and $A^{-1}(R)$ has volume 5, what is the determinant of A?

1. 3
2. 5
3. 10

Determinants can be used to check whether a matrix is invertible, since A is noninvertible if and only if it maps $\mathbb{R}^n$ to a lower-dimensional subspace of $\mathbb{R}^n$, and in that case A squishes positive-volume regions down to zero-volume regions.

Exercise 3.6.7

Let $A = \begin{bmatrix} 2 & -2 \\ -4 & 0 \end{bmatrix}$. Find the values of $\lambda \in \mathbb{R}$ for which the equation $A\mathbf{v} = \lambda\mathbf{v}$ has nonzero solutions for $\mathbf{v}$.

Exercise 3.6.8

For an $n \times n$ square matrix, which of the following is the relationship between $\det A$ and $\det(3A)$?

1. $\det(3A) = 3n + \det(A)$.
2. $\det(3A) = 3n \det(A)$.
3. $\det(3A) = n^3 \det(A)$.
4. $\det(3A) = 3^n \det(A)$.

Exercise 3.6.9

Is every matrix with positive determinant positive definite?

3.7 Matrix Norms

The **operator norm** $\|A\|$ of an $m \times n$ matrix A is defined to be the largest value of $|A\mathbf{v}|$ for any unit vector $\mathbf{v} \in \mathbb{R}^n$.

Example 3.7.1

The operator norm of $A = \left[\begin{smallmatrix} 0 & 2 \\ 3 & 0 \end{smallmatrix}\right]$ is the maximum value of $|A \left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]| = \sqrt{(2y)^2 + (3x)^2}$ subject to the constraint $x^2 + y^2 = 1$. This optimization problem may be solved by solving the second equation for x^2, substituting into the first equation, and differentiating (or with Lagrange multipliers, which is introduced in the section on multivariable calculus) to find that the maximum occurs for $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] = \left[\begin{smallmatrix} \pm 1 \\ 0 \end{smallmatrix}\right]$. So $\|A\| = \boxed{3}$.

Exercise 3.7.1

Explain why $\|AB\| \leq \|A\|\|B\|$. Give an example when equality holds and when the left-hand side is strictly smaller than the right-hand side.

Exercise 3.7.2

Explain why the operator norm of a matrix is equal to its largest singular value.
Confirm that the largest singular value of $A = \left[\begin{smallmatrix} 0 & 2 \\ 3 & 0 \end{smallmatrix}\right]$ is 3.

Exercise 3.7.3

If A has SVD $A = U\Sigma V'$, find the SVD of the Gram matrix $A'A$, and use it to prove that $\|A'A\| = \|A\|^2$.

4 Multivariable Calculus

Calculus is the study of continuously varying functions. Specifically, we examine instantaneous rates of change and learn how to average (or total) the values of a function over a region. In multivariable calculus, we generalize differentiation and integration ideas developed for functions defined on $\mathbb{R}^1$ to the setting where our functions are defined on $\mathbb{R}^d$ for some $d > 1$.

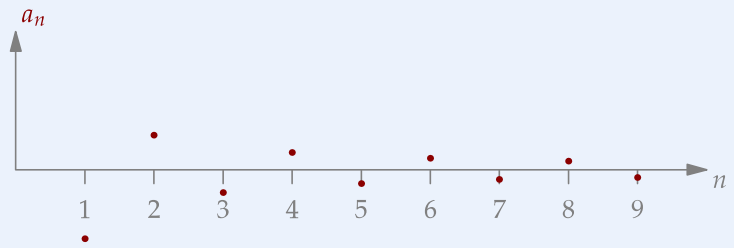
The ideas of multivariable calculus are useful for data science in at least a couple of ways: (i) the functions we use to gauge the goodness of a model typically depend on many model variables. To optimize these functions, we need to think about how they increase or decrease under small perturbations of the variables. And (ii) we will mathematically represent the idea of *probability* using functions on $\mathbb{R}^d$, and in that context probabilities will be recovered by integrating these functions.

4.1 Sequences and series

A **sequence** of real numbers $(x_n)_{n=1}^{\infty} = x_1, x_2, \ldots$ converges to a number $x \in \mathbb{R}$ if the distance from x_n to x on the number line can be made as small as desired by choosing n sufficiently large. In that case, we say that $x_n \to x$ as $n \to \infty$, or $\lim_{n\to\infty} x_n = x$.

Example 4.1.1

The sequence $(-1)^n/n$ converges to 0 as $n \to \infty$, since the distance on the number line from 0 to $(-1)^n/n$ is $1/n$, and that distance may be made as small as desired by choosing n large enough. For example, if you want $1/n$ to be less than 0.001, all the values of n larger than 1000 will work.



Theorem 4.1.1: Squeeze theorem

If $a_n \leq b_n \leq c_n$ for all $n \geq 1$ and if $\lim_{n\to\infty} a_n = \lim_{n\to\infty} c_n$, then the sequence $(b_n)_{n=1}^{\infty}$ converges, and its limiting value is equal to the common limiting value of $(a_n)_{n=1}^{\infty}$ and $(c_n)_{n=1}^{\infty}$.

Exercise 4.1.1

Suppose that $|x_n| \leq n^{-1/2}$ for all $n \geq 1$. Show that $x_n \to 0$ as $n \to \infty$.

A **series** $\sum_{n=1}^{\infty} x_n = x_1 + x_2 + x_3 + \cdots$ converges if the sequence $(S_n)_{n=1}^{\infty}$ converges, where

$$S_n = x_1 + x_2 + \cdots + x_n$$

ster is the nth *partial sum*. Roughly speaking, a series converges if its terms converge to 0 fast enough. In particular, the terms must converge to zero:

Theorem 4.1.2: Term test

If a_n does not converge to zero, then the series $\sum_{n=1}^{\infty} a_n$ does not converge.

Another valid statement suggested by the "terms go to 0 fast enough" intuition is that convergence of one series implies convergence of any other series whose terms go to 0 at least as fast:

Theorem 4.1.3: Comparison test

If $\sum_{n=1}^{\infty} b_n$ converges and if $|a_n| \leq b_n$ for all n, then $\sum_{n=1}^{\infty} a_n$ converges.
Conversely, if $\sum_{n=1}^{\infty} b_n$ does not converge and $0 \leq b_n < a_n$, then $\Sigma_{n=1}^{\infty} a_n$ also does not converge.

The comparison test works well in conjunction with a list of basic series which are known to converge or not.

Theorem 4.1.4

(i) The series $\sum_{n=1}^{\infty} n^p$ converges if and only if $p < -1$.

(ii) The series $\sum_{n=1}^{\infty} a^n$ converges if and only if $-1 < a < 1$.

Example 4.1.2

Show that the series $\sum_{n=1}^{\infty} \frac{1}{n^2+n}$ converges.

Solution

We know that $\frac{1}{n^2+n} < \frac{1}{n^2}$ and that $\sum_{n=1}^{\infty} \frac{1}{n^2}$ converges. Therefore, the comparison test implies that $\frac{1}{n^2+n}$ converges.

Exercise 4.1.2

Numerically examine the statement that $\sum_{n=1}^{\infty} \frac{1}{n^2}$ converges to $\frac{\pi^2}{6}$.

4.2 Taylor series

We can define a polynomial which approximates a smooth function in the vicinity of a point with the following idea: *match as many derivatives as possible*.

First, a bit of review on the exponential function $x \mapsto \exp(x)$: we define exp to be the function which maps 0 to 1 and which is everywhere equal to its own derivative. It follows (nontrivially) from this definition that $\exp(x) = \exp(1)^x$, so may define $e = \exp(1)$ and write the exponential function as $x \mapsto e^x$. The value of e is approximately 2.718.

Example 4.2.1

Find the quadratic polynomial P_2 whose zeroth, first, and second derivatives at the origin match those of the exponential function.

Solution

Since P_2 is quadratic, we must have

$$P_2(x) = a_0 + a_1 x + a_2 x^2$$

for some a_0, a_1, and a_2. To match the zeroth derivative, we check that $P_2(0) = a_0$ and $f(0) = 1$. So we must have $a_0 = 1$. Similarly, $P_2'(0) = a_1$, so if we want $P_2'(0) = f'(0) = 1$, have to choose $a_1 = 1$ as well.

For a_2, we calculate $P_2''(x) = (a_1 + 2a_2 x)' = 2a_2$, so to get $P_2''(0) = f''(0) = 1$, we have to let $a_2 = \frac{1}{2}$. So

$$P_2(x) = 1 + x + \tfrac{1}{2}x^2$$

is the best we can do. Looking at the figure, we set that P_2 does indeed do a better job of 'hugging' the graph of f near $x = 0$ than the best linear approximation ($L(x) = 1 + x$) does.

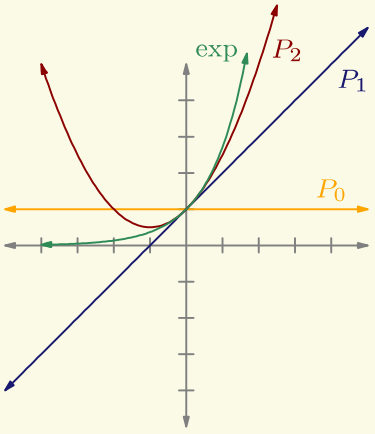


Figure 4.1 The best constant, linear, and quadratic approximations of $\exp(x) = e^x$ near the origin

We can extend this idea to higher order polynomials, and we can even include terms for *all* powers of x, thereby obtaining an infinite series:

Definition 4.2.1: Taylor Series

The Taylor series, centered at c, of an infinitely differentiable function f is defined to be

$$f(c) + f'(c)(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \frac{f'''(c)}{3!}(x-c)^3 + \cdots$$

Example 4.2.2

Find the Taylor series centered at the origin for the exponential function.

If the Taylor series for a function converges, then it does so in an interval centered around c. Furthermore, inside the interval of convergence, it is valid to perform term-by-term operations with the Taylor series as though it were a polynomial:

1. We can multiply or add Taylor series term-by-term.
2. We can integrate or differentiate a Taylor series term-by-term.
3. We can substitute one Taylor series into another to obtain a Taylor series for the composition.

Theorem 4.2.1

All the operations described above may be applied wherever all the series in question are convergent. In other words, f and g have Taylor series P and Q converging to f and g in some open interval, then the Taylor series for fg, $f+g$, f', and $\int f$ converge in that interval and are given by PQ, $P+Q$, P', and $\int P$, respectively. If P has an infinite radius of convergence, then the Taylor series for $f \circ g$ is given by $P \circ Q$.

Example 4.2.3

Find the Taylor series for $f(x) = \cos x + xe^{x^2}$ centered at $c = 0$.

Solution

Taking many derivatives is going to be no fun, especially with that second term. What we can do, however, is just substitute x^2 into the Taylor series for the exponential function, multiply that by x, and add the Taylor series for cosine:

$$\left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots\right) + x\left(1 + x^2 + \frac{(x^2)^2}{2!} + \frac{(x^2)^3}{3!} + \cdots\right) = 1 + x - \frac{x^2}{2!} + x^3 + \frac{x^4}{4!} + \frac{x^5}{2!} + \cdots.$$

In summation notation, we could write this series as $\sum_{n=0}^{\infty} a_n x^n$ where a_n is equal to $(-1)^{n/2}/n!$ if n is even and $1/((n-1)/2)!$ if n is odd.

Exercise 4.2.1

Find the Taylor series for $1/(1-x)$ centered at the origin, and show that it converges to $1/(1-x)$ for all $-1 < x < 1$.

Use your result to find $x + 2x^2 + 3x^3 + 4x^4 + \cdots$. Hint: think about differentiation.

Exercise 4.2.2

Show that $\lim_{n\to\infty}(1 + x/n)^n$ is equal to e^x by showing that $\lim_{n\to\infty} \log(1 + x/n)^n = x$.

4.3 Partial differentiation

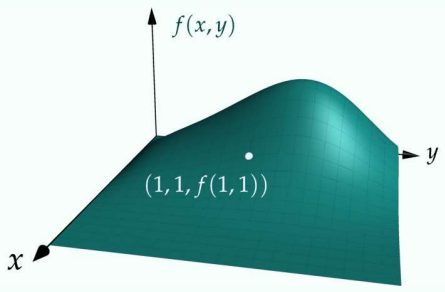
Differentiating a single-variable function involves answering the question *near a given point, how much does the value of the function change per unit change in the input*? In the higher-dimensional setting, the question must be made more specific, since the change in output depends not only on the change in input but also the direction in which the input is changed.

Consider, for example, the function $f(x,y)$ which returns the altitude of the point on earth with latitude x and longitude y. If the point (x,y) identifies a point on a sloping hillside, then there are some directions in which f increases, others in which f decreases, and two directions in which f neither increases nor decreases (these are the directions along the hill's contour lines).

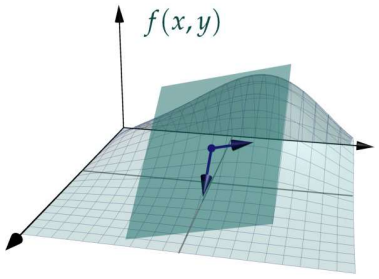
The simplest directions for inquiring about the instantaneous rate of change of f are those along the axes: The **partial derivative** $\frac{\partial f}{\partial x}(x_0,y_0)$ of a function $f(x,y)$ at a point (x_0,y_0) is the slo $\frac{\partial f}{\partial x}(x_0,y_0)$ of a function $f(x,y)$ at a point (x_0,y_0) is the slope of the graph of f in the x-direction at the point (x_0,y_0) pe of the graph of f in the x-direction at the point (x_0,y_0). In other words, it's the slope of the intersection of the graph of f with the plane $y=y_0$. The partial derivative $\frac{\partial f}{\partial x}(x_0,y_0)$ may also be denoted $f_x(x_0,y_0)$.

Exercise 4.3.1

Consider the function f whose graph is shown. Determine the sign of $f_x(1,1)$ and the sign of $f_y(1,1)$.



A single-variable function is differentiable at a point if and only if its graph looks increasingly like that of a particular non-vertical line when zoomed increasingly far in. In other words, f is differentiable if and only if there's a linear function L such that $\frac{f(x)-L(x)}{x-a}$ goes to 0 as $x \to a$.



Likewise, a function of two variables is **differentiable** at a point if its graph looks like a plane when you zoom in sufficiently around the point; that is, f is differentiable at (a,b) if

$$\lim_{(x,y)\to(a,b)} \frac{f(x,y)-c_0-c_1(x-a)-c_2(y-b)}{||[x,y]-[a,b]||} = 0$$

for some real numbers c_0, c_1, and c_2. If such a linear function $c_0+c_1(x-a)+c_2(y-b)$ exists, then its coefficients are necessarily $c_0=f(a,b)$, $c_1=f_x(a,b)$, and $c_2=f_y(a,b)$.

So, the equation of the plane tangent to the graph of a differentiable function f at the point $(a,b,f(a,b))$ is given by

$$z = f(a,b)+f_x(a,b)(x-a)+f_y(a,b)(y-b) \tag{4.3.1}$$

This equation says how f behaves for values of (x,y) very close to (a,b): the output changes by the x-change

$x - a$ times f's sensitivity to changes in x (namely $f_x(a,b)$) *plus* the y-change times f's sensitivity to changes in y (namely $f_y(a,b)$).

Once we know how a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ changes in the coordinate-axis directions, we can use (4.3.1) to succinctly express how it changes in any direction: we form the **gradient** ∇f of f by putting all of the partial derivatives of a function f together into a vector. Then, for any unit vector $\mathbf{u}$, the rate of change of f in the $\mathbf{u}$ direction is equal to $\nabla f \cdot \mathbf{u}$.

Since $\nabla f \cdot \mathbf{u} = |\nabla f| \cos\theta$, the direction of the gradient is the direction in which f increases most rapidly. The direction opposite to the gradient is the direction of maximum decrease, and the directions orthogonal to these are the ones in which f is constant.

Exercise 4.3.2

Suppose that $f : \mathbb{R}^2 \to \mathbb{R}$ is a differentiable function at the point $(a,b) \in \mathbb{R}^2$ and that its instantaneous rates of change in the directions $\mathbf{u}$ and $\mathbf{v}$ are known. Show that if $\mathbf{u}$ and $\mathbf{v}$ are not parallel, then it is always possible to infer f's rates of change in the coordinate-axis directions.

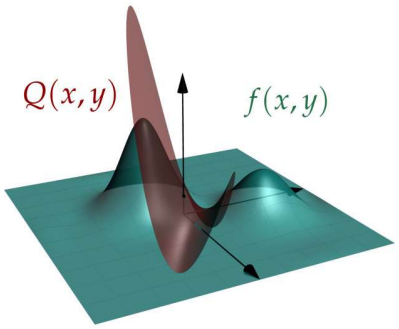
We can take the notion of a gradient, which measures the *linear* change of a function, up a degree. The **Hessian** of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined to be the matrix

$$\mathbf{H}(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

The best quadratic approximation to the graph of a twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ at the origin is

$$Q(\mathbf{x}) = f(\mathbf{0}) + (\nabla f(\mathbf{0}))'\mathbf{x} + \frac{1}{2}\mathbf{x}'\mathbf{H}(\mathbf{0})\mathbf{x}.$$

The same is true at points $\mathbf{a}$ other than the origin if we evaluate the gradient and Hessian at $\mathbf{a}$ instead of $\mathbf{0}$ and if we replace $\mathbf{x}$ with $\mathbf{x} - \mathbf{a}$.



Exercise 4.3.3

Suppose that a, b, c, d, e and f are real numbers and that $f(x,y) = a + bx + cy + dx^2 + exy + fy^2$. Show that the quadratic approximation of f at the origin is equal to f.

4.4 Optimization

To find the largest or smallest value of a differentiable function defined on a subset D of $\mathbb{R}^d$, we may make an observation regarding the instantaneous rates of change of f: if ∇f is nonzero at a point away from the boundary of D, then there are directions in which f decreases away from that point and other directions where it increases. Therefore, any minima or maxima of f away from the boundary must occur at points where the gradient of f is zero. We term such points—where the gradient is zero or where the function is non-differentiable—**critical points**. If a function has a minimum or a maximum at a point, then either that point is a critical point, or it is on the boundary of the domain of the function.

It's sometimes possible to check using derivatives whether a function has a minimum or maximum in the immediate vicinity of a critical point $\mathbf{a}$: f has

1. a local maximum if the Hessian at $\mathbf{a}$ is negative definite
2. a local minimum if the Hessian at $\mathbf{a}$ is positive definite
3. neither a local min nor max if the Hessian at $\mathbf{a}$ has at least one positive and one negative eigenvalue

Since a function's maximum or minimum may also occur on its boundary, we must also identify candidate points on the boundary where maxima or minima may occur. This may be done in a couple ways: we parametrize the boundary and solve an optimization problem in a lower dimension. For example, if we want to optimize a function on the unit disk, we can identify boundary critical points by finding critical points of the single-variable function $\theta \mapsto f(\cos\theta, \sin\theta)$.

Parametrizing the boundary is only possible in simple cases, so we rely more commonly on the method of *Lagrange multipliers*. The idea is that the function does *not* have a maximum at a boundary point if it's possible to move along the boundary in a direction whose angle with ∇f is less than a right angle. Therefore, ∇f must perpendicular to the boundary of D at a point if f is to have an extremum there. If the boundary of D is specified as a level set of a function g, we arrive at the equation

$$\nabla f = \lambda \nabla g,$$

for some $\lambda \in \mathbb{R}$.

Theorem 4.4.1: Extreme value theorem and Lagrange multipliers

Suppose that f is a continuous function defined on a closed and bounded subset D of $\mathbb{R}^n$. Then

1. f realizes an absolute maximum and absolute minimum on D (the extreme value theorem)
2. any point where f realizes an extremum is either a critical point—meaning that $\nabla f = 0$ or f is non-differentiable at that point—or at a point on the boundary.
3. if f realizes an extremum at a point on a portion of the boundary which is the level set of a differentiable function g with non-vanishing gradient ∇g, then either f is non-differentiable at that point or the equation
$$\nabla f = \lambda \nabla g$$
is satisfied at that point, for some $\lambda \in \mathbb{R}$.

Exercise 4.4.1

Find the point closest to the origin in the region $3x + 2y + z \geq 6$.

4.5 Matrix differentiation

Just as elementary differentiation rules are helpful for optimizing single-variable functions, *matrix* differentiation rules are helpful for optimizing expressions written in matrix form. This technique is used often in statistics.

* This is called the *numerator layout* convention. Sometimes you'll see the transpose of this matrix instead. See the Wikipedia article on matrix calculus for details.

Suppose $\mathbf{f}$ is a function from $\mathbb{R}^n$ to $\mathbb{R}^m$. Writing $\mathbf{f}(\mathbf{x}) = \mathbf{f}(x_1, \ldots, x_n)$, we define the Jacobian matrix to be*

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Note that if $m = 1$, then differentiating f with respect to $\mathbf{x}$ is the same as taking the gradient of f. With this definition, we obtain the following analogues to some basic single-variable differentiation results: if A is a constant matrix, then

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}}(A\mathbf{x}) &= A \\ \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'A) &= A' \\ \frac{\partial}{\partial \mathbf{x}}(\mathbf{u}'\mathbf{v}) &= \mathbf{u}'\frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}'\frac{\partial \mathbf{u}}{\partial \mathbf{x}} \end{aligned}$$

The Hessian of a function $f : \mathbb{R}^n \to \mathbb{R}$ may be written in terms of the matrix differentiation operator as follows:

$$\mathbf{H}(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} \left(\frac{\partial f}{\partial \mathbf{x}} \right)'.$$

Some authors define $\frac{\partial f}{\partial \mathbf{x}'}$ to be $\left(\frac{\partial f}{\partial \mathbf{x}} \right)'$, in which case the Hessian operator can be written as $\frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}'}$.

Exercise 4.5.1

Let $f : \mathbb{R}^n \to \mathbb{R}$ be defined by $f(\mathbf{x}) = \mathbf{x}'A\mathbf{x}$ where A is a symmetric matrix. Find $\frac{\partial f}{\partial \mathbf{x}}$.

Exercise 4.5.2

Suppose A is an $m \times n$ matrix and $\mathbf{b} \in \mathbb{R}^m$. Use matrix differentiation to find the vector $\mathbf{x}$ which minimizes $|A\mathbf{x} - \mathbf{b}|^2$. Hint: begin by writing $|A\mathbf{x} - \mathbf{b}|^2$ as $(A\mathbf{x} - \mathbf{b})'(A\mathbf{x} - \mathbf{b})$. You may assume that the rank of A is n.

4.6 Multivariable integration

Integrating a function is a way of totaling up its values. For example, if f is a function from a region D in $\mathbb{R}^n$ to $\mathbb{R}$ which represents the mass density of a solid occupying the region D, we can find the total mass of the solid as follows: (i) split the region D into many tiny pieces, (ii) multiply the volume of each piece by the value of the function at some point on that piece, (iii) and add up the results. If we take the number of pieces to ∞ and the piece size to zero, then this sum converges to the total mass of the solid.

We may apply this procedure to any function f defined on D, and we call the result the integral of f over D, denoted $\int_D f$.

To find the integral of a function f defined on a 2D region D, we set up a double iterated integral over D: the bounds for the outer integral are the smallest and largest possible values of y for point in D, and the bounds for the inner integral are the smallest and largest values of x for any point in a given *each "y = constant" slice* of the region (assuming that each slice intersects the region in a line segment).

Exercise 4.6.1

Find the integral over the triangle T with vertices $(0,0)$, $(2,0)$, and $(0,3)$ of the function $f(x,y) = x^2y$.

To set up an integral of a function over a 3D region (for the order $\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$): the bounds for the outer integral are the smallest and largest values of z for any point in the region of integration, then the bounds for the middle integral are the smallest and largest values of y for any point in the region in each "z = constant" plane, and the inner bounds are the smallest and largest values of x for any point in the region in each "(y,z) = constant" line.

Exercise 4.6.2

Integrate the function $f(x,y,z) = 1$ over the tetrahedron with vertices $(0,0,0)$, $(2,0,0)$, $(0,3,0)$, and $(0,0,4)$.

4.7 The chain rule

If we compose a differentiable function $\mathbf{r} : \mathbb{R}^1 \to \mathbb{R}^2$ with a differentiable function $f : \mathbb{R}^2 \to \mathbb{R}^1$, we get a function whose derivative is

$$(f \circ \mathbf{r})'(t) = (\nabla f)(\mathbf{r}(t)) \cdot \mathbf{r}'(t).$$

This follows from linearizing f: the change that results from making a small move from $\mathbf{r}(t)$ to $\mathbf{r}(t) + \mathbf{r}'(t)\Delta t$ is the dot product of the gradient of f and the small step $\mathbf{r}'(t)\Delta t$.

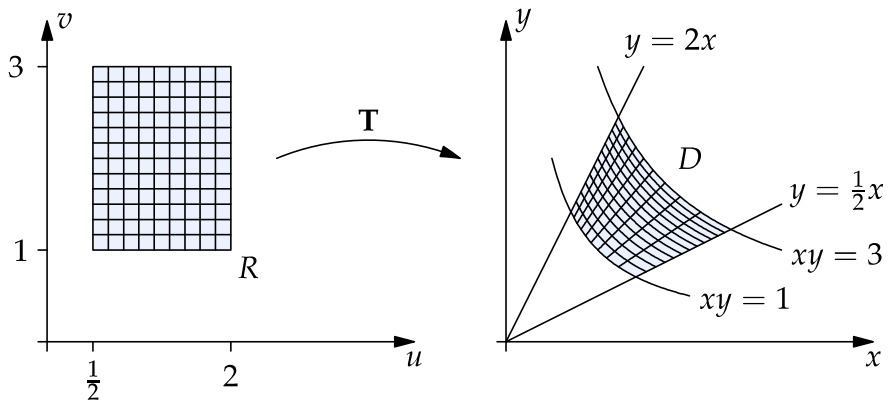
Exercise 4.7.1

Suppose that $f_x(3,2) = 4$, that $f_y(3,2) = -2$, and that $x(t) = 1 + 2t$ and $y(t) = 4 - 2t^2$. Find the derivative of the function $f(x(t), y(t))$ at the point $t = 1$.

4.8 The Jacobian determinant of a transformation

If we want to integrate over a region which doesn't split nicely along lines parallel to the coordinate axes, we can split the region up along other lines.

For example, consider the region bounded by the hyperbolas $xy = 1$ and $xy = 3$ and the lines $y = \frac{1}{2}x$ and $y = 2x$. This region naturally splits along hyperbols of the form $xy = v$ where v ranges from 1 to 3 and lines of the form $y/x = u$ where u ranges from $\frac{1}{2}$ to 2. We can therefore write the region as the image of the rectangle $[\frac{1}{2}, 2] \times [1, 3]$ under the inverse $\mathbf{T}$ of the transformation $(x, y) \mapsto (y/x, xy)$.



To find the area of each small piece of D in this subdivision, we may multiply the area of the corresponding piece of the rectangle by the area distortion factor of the transformation $\mathbf{T}$. This local area distortion factor, or *Jacobian determinant* is the absolute value of the determinant of the Jacobian matrix $\frac{\partial \mathbf{T}(\mathbf{x})}{\partial \mathbf{x}}$. Thus we arrive at the formula

$$\iint_D f(x, y)\, dx\, dy = \iint_R f(\mathbf{T}(u, v)) \left| \frac{\partial \mathbf{T}(u, v)}{\partial (u, v)} \right| du\, dv.$$

Exercise 4.8.1

Show that the map $(u, v) \mapsto (u \cos v, u \sin v)$ maps the rectangle $[0, 1] \times [0, 2\pi]$ onto the unit disk, and calculate the Jacobian for this transformation. Use your result to integrate 1 over the unit disk and confirm that the result is equal to the area of the unit disk.

5 Numerical Computation

5.1 Machine arithmetic

Computers store all information, including numerical values, as sequences of bits. The **type** of a numeric value specifies how to interpret the underlying sequence of bits as a number.

You can access the bit representation of a numerical value in Julia using the function `bitstring`. In this section we will introduce several of the most important numeric types:*

* These representations are not specific to Julia; they are universal computing standards. See the Wikipedia article on IEEE 754

Exercise 5.1.1

We interpret a string of digits as an integer using *place value*: the units digit is worth 10^0, the next digit to the left is worth 10^1, and so on. Then $709 = 7 \cdot 10^2 + 0 \cdot 10^1 + 9 \cdot 10^0$, for example. This is called the decimal representation of a number.

We can do the same thing with 2 in place of 10: the rightmost digit is worth 2^0, the next digit is worth 2^1, and so on. Instead of 10 digits we only have two *bits*: 0 and 1. This is called the binary representation of a number. The binary representation of 13, for example, is 1101, since $13 = 1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0$.

Find the binary representations of each of the following numbers: 2, 16, 20, and 100.

5.1.1 64-bit integers

There are 2^{64} length-64 strings of zeros and ones, so with 64 bits we can represent 2^{64} integers. For example, we can represent the integers from 0 to $2^{64} - 1$ by interpreting each string of 0's and 1's as a binary number. In order to make room for negative numbers, however, we will only use half these bitstrings to represent positive numbers (from 0 to $2^{63} - 1$), and we will allocate the other half to negative integers (from -2^{63} to -1).

More precisely, for $0 \leq n \leq 2^{63} - 1$, we represent n using its binary representation, with leading zeros as necessary to get 64 total bits. For $1 \leq n \leq 2^{63}$, we represent $-n$ using the binary representation of $2^{64} - n$.

Example 5.1.1

The expression `bitstring(+34)` evaluates to

```
0000000000000000000000000000000000000000000000000000000000100010;
```

This checks out: $34 = 1 \cdot 2^5 + 0 \cdot 2^4 + 0 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0$.

The expression `bitstring(-34)` evaluates to

```
1111111111111111111111111111111111111111111111111111111011110.
```

We could check that this is the binary representation of $2^{64} - 34$, but in the following exercise we will learn a trick for doing that without having to deal with all those 1's.

Exercise 5.1.2

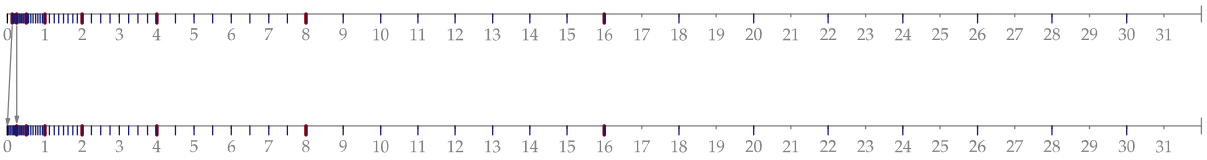
Show that if $1 \leq n \leq 2^{63} - 1$, then you can find `bitstring(-n)` from `bitstring(n)` by (i) flipping every bit, and (ii) adding 1 to the resulting number (interpreted as an integer represented in binary).

5.1.2 64-bit floating point numbers

Integer types are appropriate for calculations which only involve integer operations (multiplication, addition, and negation), but most integers do not have integer reciprocals. So performing calculations involving division requires a new number type.

Let's visualize our number system by placing a tick mark on the number line for each number we're representing. If we want to represent very large numbers and very small numbers accurately (relative to the size of the number), we need the tick marks to be much denser around 0.

One way to achieve this is to put equally spaced tick marks between 1 and 2, and then scale that interval up repeatedly into $[2,4)$, then $[4,8)$, then $[8,16)$, and so on, and also scale it down to $[1/2,1)$, $[1/4,1/2)$, and so on. Here's an example of such a scheme: we place 8 tick marks between 1 and 2, and then we scale that interval's worth of tick marks four times by a factor of 2, and also 3 times by a factor of $\frac{1}{2}$.



We can see on the left edge of the picture that we didn't cover zero! There is also quite a large gap between 0 and the smallest representable number in this system. So we replace the leftmost interval between successive powers of 2 with ticks twice as far apart so that they reach down to 0. The locations of these new ticks are called *subnormal numbers*.

If we have 64 bits, we can do the same thing but on a grander scale. We define σ to be the first bit of the sequence, e to be the next 11 bits interpreted as a binary integer, and f to be the remaining 52 bits interpreted as a binary integer. If $0 < e < 2047$, then the sequence represents the number

$$(-1)^{\sigma} \left(1 + \left(\frac{1}{2}\right)^{52} f\right) \cdot 2^{e-1023}.$$

Note that the exponent $e - 1023$ ranges from -1022 up to 1023 as e ranges from 1 to 2046. If $e = 0$, then the sequence represents

$$(-1)^{\sigma} \left(0 + \left(\frac{1}{2}\right)^{52} f\right) \cdot 2^{-1022},$$

These are the subnormal numbers.

We also appropriate the *last* value of e for a special meaning: if $e = 2047$, then the sequence represents one of the special values `Inf` or `NaN`, depending on the value of f.

Example 5.1.2

`bitstring(-0.75)` returns `1011111111101000000000000000000000000000000000000000000000000000` The leading `1` indicates that the number is negative, the next eleven digits `01111111110` give $e = 1022$, and interpreting the remaining 52 digits as a base-2 fraction $(0.1000\ldots)_2$ gives $f = \frac{1}{2}$. So the value represented is

$$(-1)\left(1+\frac{1}{2}\right)2^{1022-1023} = -\frac{3}{4}.$$

Thus -0.75 can be represented exactly as a `Float64`.

The nonnegative representable numbers are laid out as shown (figure not drawn to scale!):

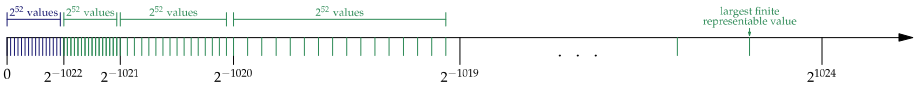


Figure 5.1 The tick marks indicate the positive values which are representable as 64-bit floating point numbers. There are 2^{52} of them between any two successive powers of 2 from 2^{-1022} up to 2^{1024}, and the interval from 0 to 2^{-1022} also contains 2^{52} representable values (these are the subnormal numbers, indicated in blue).

Exercise 5.1.3

Show that 0.1 cannot be represented exactly as a `Float64`.

Exercise 5.1.4

Show that there are 2^{52} representable numbers in each interval of the form $[2^k, 2^{k+1})$, where $-1022 \leq k < 1024$.

5.1.3 32-BIT FLOATING POINT NUMBERS

Each 32-bit sequence represents the value

$$(-1)^\sigma \left(1+\left(\frac{1}{2}\right)^{23} f\right)\cdot 2^{e-127},$$

where σ is the first bit of the sequence, e is the next 8 bits interpreted as a binary integer, and f is the remaining 23 bits are interpreted as a binary integer.

Example 5.1.3

`bitstring(Float32(-0.75))` returns `10111111010000000000000000000000`

Exercise 5.1.5

Find the positive difference between 1 and the first number greater than 1 which is representable as a `Float32`.

5.1.4 Arbitrary-precision numbers

The number of bits in the representation of an arbitrary precision number is not fixed. `BigInt` values are useful for when dealing with very large integers, and `BigFloat` values are useful when very high precision is required.

Exercise 5.1.6

(i) Arbitrary-precision arithmetic is helpful for inspecting the behavior of lower precision formats. Find the exact value of the difference between 1.1 the `Float64` value nearest 1.1. (Hint: `big(1.1)` interprets 1.1 as a Float64 value—look at that value and mentally subtract 1.1).

(ii) Confirm that calculating the decimal representation of $2^{100,000}$ is no problem with big number arithmetic. Convert the resulting sequence of decimal digits to a string and find its length.

5.1.5 General comments

Choice of numerical representation depends on the application at hand, since each has its advantages and disadvantages.

`Int64` arithmetic is actually *modular* arithmetic* with a modulus of 2^{64}. This means that Int64 arithmetic is exact unless our calculation takes us outside the window of representable values. Basic `Float64` operations return the same number as computing the mathematical result of the operation and rounding to the nearest `Float64`-representable value (or one of the two nearest ones if it's halfway between)

* An expression evaluates to the integer in $[-2^{63}, 2^{63})$ which leaves the same remainder when divided by 2^{64} as the mathematical value of the expression

Exercise 5.1.7

Without using a computer, perform the operations in the expression $(1.0 + 0.4/2^{52}) + 0.4/2^{52}$ using `Float64` arithmetic. Repeat with the expression $1.0 + (0.4/2^{52} + 0.4/2^{52})$. Then check your findings by evaluating these expressions in Julia.

A *numeric literal* is a sequence of characters to be parsed and interpreted as a numerical value.

1. Numeric literals with a decimal point are *real literals*.
2. Numeric literals without a decimal point are *integer literals*.

Example 5.1.4

In the expression `2.718^50+1` , `2.718` is a real literal, and `50` and `1` are both integer literals.

Integer literals are interpreted as `Int64` values, and real literals are interpreted as `Float64` values.

Example 5.1.5

`2^100` returns `0` , since `2` and `100` are interpreted as `Int64` values, and 2^{100} is equivalent to 0 modulo 2^{64}.

To obtain numerical values of other types in Julia, use **`parse`** or **`big`**.

```
julia> parse(BigInt,"1267650600228229401496703205376")^(big(1)/100)
2.000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
```

`Float64` and `Int64` operations are performed *in hardware*, meaning that they use instructions programmed directly into the computer's microprocessor. They are much faster and more memory efficient than arbitrary precision arithmetic, which has to be done in software*:

* `@elapsed` returns the number of seconds an expression takes to evaluate

```
julia> A = [i^2 for i=1:1_000_000]
julia> B = [BigInt(a) for a in A]
julia> @elapsed(sum(B))/@elapsed(sum(A))
25.681853565096898
```

Exercise 5.1.8

Explain why it is never necessary to use a `BigInt` for a loop counter (that is, a variable which starts at 0 or 1 and is incremented by 1 each time the body of the loop runs).

5.2 Error

Error is the discrepancy between a quantity and the value used to represent it in the program. A result is **accurate** if its error is small. If $\widehat{A}$ is an approximation for A, then

- the **absolute error** is $\widehat{A} - A$, and
- the **relative error** is $\frac{\widehat{A}-A}{A}$.

We are usually more interested in relative error, since the relevance of an error is usually in proportion to the quantity being represented. For example, misreporting the weight of an animal by one kilogram would be much more significant if the animal were a squirrel than if it were a blue whale.

Example 5.2.1

The expression `sqrt(200.0)`, which returns the `Float64`-square root of `Float64`, yields

$$14.142135623730951010657008737325668334960937\overline{50}.$$

The actual decimal representation of $\sqrt{200}$ is

$$14.142135623730950488016887242096980785696718\ldots$$

The difference between these values, 5.23×10^{-16}, is the absolute error, and $\frac{5.23\times10^{-16}}{\sqrt{200}} = 3.7 \times 10^{-17}$ is the relative error.

5.2.1 Sources of numerical error

There are a few categories of numerical error.

Roundoff error comes from rounding numbers to fit them into a floating point representation.

Example 5.2.2

`0.2 + 0.1` is equal to $3.000000000000000444089209850062616169452667236328125\overline{0}$ in `Float64` arithmetic. The discrepancy between 3 and this value is roundoff error.

Truncation error comes from using approximate mathematical formulas or algorithms.

Example 5.2.3

The Maclaurin series of $\sin x$ is $x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$, so approximating $\sin(0.1)$ as $0.1 - \frac{0.1^3}{6}$ yields a truncation error equal to $\frac{0.1^5}{5!} - \frac{0.1^7}{7!} + \cdots$.

Example 5.2.4

Newton's method approximates a zero of a function f by starting with a value x_0 near the desired zero and defining $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ for all $n \geq 0$.

Under certain conditions, x_n converges to a zero of f as $n \to \infty$. The discrepancy between x_n and $\lim_{n\to\infty} x_n$ is the truncation error associated with stopping Newton's method at the nth iteration.

Example 5.2.5

We may approximate $\int_0^1 \sin(x^2)\,dx$ using the sum $\sum_{k=1}^{100} \sin\left(\left(\frac{k}{100}\right)^2\right)\frac{1}{100}$. The error associated this approximation is a type of truncation error.

Statistical error arises from using randomness in an approximation.

Example 5.2.6

We can approximate the average height of a population of 100,000 people by selecting 100 people uniformly at random and averaging their measured heights. The error associated with this approximation is an example of statistical error.

Exercise 5.2.1

Discuss the error in each of the following scenarios using the terms *roundoff error, truncation error, or statistical error*.

(i) We use the trapezoid rule with 1000 trapezoids to approximate $\int_0^{10} \frac{1}{4+x^4}\,\mathrm{d}x$.

(ii) We are trying to approximate $f'(5)$ for some function `f` that we can compute, and we attempt to do so by running `(f(5 + 0.5^100) - f(5))/0.5^100`. We fail to get a reasonable answer.

(iii) To approximate the minimum of a function $f : [0,1] \to \mathbb{R}$, we evaluate f at 100 randomly selected points in $[0,1]$ and return the smallest value obtained.

5.2.2 Condition number

The derivative of a function measures how it stretches or compresses absolute error. The **condition number** of a function measures how it stretches or compresses *relative* error. Just as the derivative helps us understand how small changes in input transform to small changes in output, the condition number tells us how a small relative error in the initial data of a problem affects the relative error of the solution. We will use the variable a to denote a problem's initial data and $S(a)$ to denote the solution of the problem with initial data a.

The condition number of a function is defined to be the absolute value of the ratio of the relative change in output of the function to a very small* relative change in the input. The condition number of a *problem* is the condition number of the function which maps the problem's initial data to its solution.

* "very small" means that we define the condition number to be the *limit* of the stated ratio as the relative change in input goes to 0

Definition 5.2.1

If S is the map from the initial data $a \in \mathbb{R}$ of a problem to its solution $\mathbf{S}(a) \in \mathbb{R}^n$, then the condition number κ of the problem is

$$\kappa(a) = \frac{|a|\left|\frac{\mathrm{d}}{\mathrm{d}a}\mathbf{S}(a)\right|}{|\mathbf{S}(a)|}. \tag{5.2.1}$$

Example 5.2.7

Show that the condition number of $a \mapsto a^n$ is constant, for any $n \in \mathbb{R}$.

Solution

We have

$$\kappa(a) = \frac{ana^{n-1}}{a^n} = n,$$

for all $a \in \mathbb{R}$.

Example 5.2.8

Show that the condition number of the function $a \mapsto a - 1$ is very large for values of a near 1.

Solution

We substitute into (6.4.1) and get

$$\kappa(a) = \frac{a}{|a-1|}$$

for values of a near 1. This expression goes to infinity as $a \to 1$, so the condition number is very large. Subtracting 1 from two numbers near 1 preserves their difference, but the *relative* size of this difference is increased because the numbers themselves are much smaller.

Example 5.2.9

If $a \neq 0$, then the solution of the equation $ax + 1 = 0$ is $x = -1/a$. If we change the initial data a to $a(1+r)$, then the solution changes to $-\frac{1}{a(1+r)}$, which represents a relative change of

$$\frac{-\frac{1}{a(1+r)} - \left(-\frac{1}{a}\right)}{-1/a} = -\frac{r}{1+r}$$

in the solution. The relative change in input is $(a(1+r) - a)/a) = r$, so taking the absolute value of the ratio of $-\frac{1}{1+r}$ to r and sending $r \to 0$, we see that condition number of this problem is $\boxed{1}$.

Exercise 5.2.2

Consider a function $S : \mathbb{R} \to \mathbb{R}$. If the input changes from a to $a + \Delta a$ for some small value Δa, then the output changes to approximately $S(a) + \frac{\mathrm{d}}{\mathrm{d}a}S(a)\,\Delta a$. Calculate the ratio of the *relative change* in the output to the relative change in the input, and show that you get

$$\frac{a\frac{\mathrm{d}}{\mathrm{d}a}S(a)}{S(a)}.$$

More generally, if the initial data is in $\mathbb{R}^n$ and the solution is in $\mathbb{R}^m$, then the condition number is defined to be

$$\kappa(\mathbf{a}) = \frac{|\mathbf{a}|\|J(\mathbf{a})\|}{|\mathbf{S}(\mathbf{a})|}, \tag{5.2.2}$$

where $J(\mathbf{a})$ is the Jacobian matrix of $\mathbf{S}$ and $\|J(\mathbf{a})\|$ is its operator norm. The operator norm of the Jacobian is the appropriate generalization of the norm of the derivative of $\mathbf{S}$ since it measures how $\mathbf{S}$ stretches space near $\mathbf{a}$.

If the condition number of a problem is very large, then small errors in the problem data lead to large changes in the result. A problem with large condition number is said to be **ill-conditioned**.* Unless the initial data can be specified with correspondingly high precision, it will not be possible to solve the problem meaningfully.*

* This would be true even if we could compute with infinite precision arithmetic. Condition number is a property of a *problem*, not of the method used to solve it.

Example 5.2.10

Consider the following matrix equation for x and y.

$$\begin{bmatrix} a & 3 \\ 6 & 9 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$$

Find the values of a for which this matrix is ill-conditioned.

Solution

If $a \neq 2$, then the solution of this equation is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{7}{3(a-2)} \\ \frac{5a-24}{9(a-2)} \end{bmatrix}$$

Using (6.4.1), we can work out that

$$\kappa(a) = \frac{7|a|\sqrt{13}}{|a-2|\sqrt{(5a-24)^2+441}}.$$

If a is very close to 2, then $\kappa(a)$ is very large, and the matrix is ill-conditioned:*

* The expression `A\b` computes $A^{-1}\mathbf{b}$

```
julia> [2.01 3; 6 9] \ [4; 5]
2-element Array{Float64,1}:
233.333
-155.0

julia> [2.02 3; 6 9] \ [4; 5]
2-element Array{Float64,1}:
116.667
-77.2222
```

Machine epsilon, denoted ϵ_{mach}, is the maximum relative error associated with rounding a real number to the nearest value representable as a given floating point type. For `Float64`, this value is $\epsilon_{\text{mach}} = 2^{-53} \approx 1.11 \times 10^{-16}$. A competing convention—more widely used outside academia—defines ϵ_{mach} to be the difference between 1 and the next representable number, which for `Float64` is 2^{-52}. This is the value returned by `eps()` in Julia.

Since we typically introduce a relative error on the order of ϵ_{mach} to encode the initial data of a problem, the relative error of the computed solution should be expected to be no smaller than $\kappa\epsilon_{\text{mach}}$, regardless of the algorithm used.

An algorithm used to solve a problem is **stable** if it is approximately as accurate as the condition number of the problem allows. In other words, an algorithm is *un*stable if the answers it produces have relative error

* To recap, a *problem* is well-conditioned or ill-conditioned, and a particular *algorithm* for solving a problem is stable or unstable

many times larger than $\kappa\epsilon_{\text{mach}}$.*

Example 5.2.11

Consider the problem of evaluating $f(x) = \sqrt{1+x} - 1$ near for values of x near 0. Show that the problem is well-conditioned, but algorithm of substituting directly into the function is unstable.

Comment on whether there are stable algorithms for evaluating $f(x)$ near $x = 0$.

Solution

Substituting this function into the condition number formula, we find that

$$\kappa(x) = \frac{\sqrt{1+x}+1}{2\sqrt{1+x}}.$$

Therefore, $\kappa(0) = 1$, which means that this problem is well-conditioned at 0. However, the algorithm of substituting directly includes an ill-conditioned step: subtracting 1.

What's happening is that a roundoff error of approximately ϵ_{mach} is introduced when $1+x$ is rounded to the nearest `Float64`. When 1 is subtracted, we still have an error of around ϵ_{mach}. Since $\sqrt{1+x} \approx 1 + x/2$, we will have $f(x) \approx x/2$, and that means that the relative error in the value we find for $f(x)$ will be approximately $2\epsilon_{\text{mach}}/x$. If x is small, this will be many times larger than ϵ_{mach}.

There are stable algorithms for approximating $f(x)$ near $x = 0$. For example, we could use the Taylor series

$$\sqrt{1+x} = 1 + \frac{x}{2} - \frac{x^2}{8} + \frac{x^3}{16} - \frac{5x^4}{128} + O\left(x^5\right)$$

and approximate $f(x)$ as a sum of the first several terms on the right-hand side. Since power functions are well-conditioned (and performing the subtractions is also well-conditioned as long as x is small enough that each term is much smaller than the preceding one), this algorithm is stable. Alternatively, we can use the identity

$$\sqrt{1+x} - 1 = \frac{1}{\sqrt{1+x}+1},$$

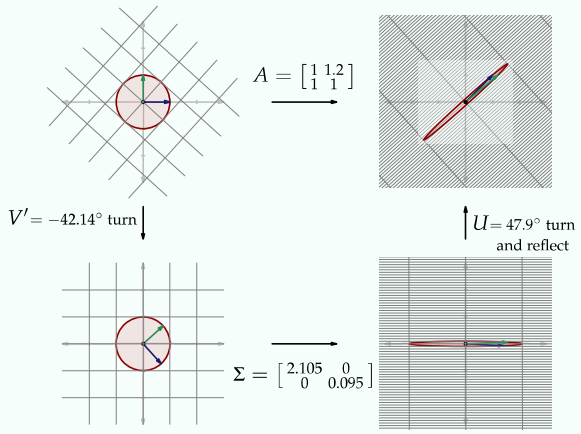
which can be obtained by multiplying by $\frac{\sqrt{1+x}+1}{\sqrt{1+x}+1}$ and simplifying the numerator. Substituting into this expression is stable, because adding 1, square rooting, and reciprocating are well-conditioned.

The condition number of an $m \times n$ matrix A is defined to be the maximum condition number of the function $\mathbf{x} \mapsto A\mathbf{x}$ as $\mathbf{x}$ ranges over $\mathbb{R}^n$. The condition number of A can be computed using its singular value decomposition:

Exercise 5.2.3

Show that the condition number of a matrix A is equal to the ratio of its largest and smallest singular values.

Interpret your resulting by explaining how to choose two vectors with small relative difference which are mapped to two vectors with large relative difference by A, assuming that A has a singular value which is many times larger than another. Use the figure below to help with the intuition.



5.2.3 Hazards

Integer or floating point arithmetic can **overflow**, and may do so without warning.

```
julia> 2^63
-9223372036854775808
julia> 10.0^309
Inf
```

Example 5.2.12

In September 2013, NASA lost touch with the *Deep Impact* space probe because* systems on board tracked time as a 32-bit-signed-integer number of tenth-second increments from January 1, 2000. The number of such increments reached the maximum size of a 32-bit signed integer in August of 2013.

*(it is believed)

Errors resulting from performing ill-conditioned subtractions are called *catastrophic cancellation*.

Example 5.2.13

Approximating $\sqrt{10^6+1} - \sqrt{10^6}$ with the result of `sqrt(10^6 + 1) - sqrt(10^6)`, we get a relative error of approximately 10^{-13}, while using `1/(sqrt(10^6 + 1) + sqrt(10^6))` gives a relative error of 5×10^{-17} (more than a thousand times smaller).

If you rely on exact comparisons for floating point numbers, be alert to the differences between `Float64` arithmetic and real number arithmetic:

```
julia> function increment(n)
           a = 1.0
               for i = 1:n
                   a = a + 0.01
               end
           a
       end
julia> increment(100) > 2
true
julia> (increment(100) - 2) / eps(2.0)
2.0
```

Each time we add 0.01, we have to round off the result to represent it as a `Float64`. These roundoff errors accumulate and lead to a result which is two ticks to the right of 2.0.

It is often more appropriate to compare real numbers using ≈ (`\approx «tab»`), which checks that two numbers x and y differ by at most $\sqrt{\epsilon_{\text{mach}}}\max(x, y)$.

Exercise 5.2.4

Guess what value the following code block returns. Run it and see what happens. Discuss why your initial guess was correct or incorrect, and suggest a value near 0.1 that you could use in place of 0.1 to get the expected behavior.

```
function increment_till(t,step=0.1)
    x = 0.0
    while x < t
        x += step
    end
    x
end
increment_till(1.0)
```

5.3 Pseudorandom number generation

When random numbers are needed in a scientific computing application, we generally use deterministic processes which mimic the behavior of random processes. These are called **pseudo-random number generators** (PRNG).

A PRNG takes an initial value, called the **seed**, and uses it to produce a sequence of numbers which are supposed to "look random". The seed determines the sequence, so you can make the random number generation in a program reproducible by providing an explicit seed. If no seed is provided, a different one will be used each time the program runs.

A simple PRNG is the **linear congruential generator**: fix positive integers M, a, and c, and consider a seed

$X_0 \in \{0, 1, \ldots, M-1\}$. We return the sequence $X_0, X_1, X_2, \ldots$, where $X_n = \text{mod}(aX_{n-1} + c, M)$ for $n \geq 1$. *

* where $\text{mod}(n, d)$ denotes the remainder when n is divided by d.

Exercise 5.3.1

A sequence of numbers $X_0, X_1, \ldots$ is periodic with period $p > 0$ if p is the smallest number such that $X_k = X_{k+p}$ for all $k \geq 0$. We say that a linear congruential generator (LCG) with $c = 0$ is *full-cyle* if the generated sequence has a period $p = M - 1$. For what values of a is the LCG with $c = 0$, $M = 5$, and $X_0 = a$ full-cycle?

Since the sequence of numbers produced by a PRNG is determined by the initial seed, we cannot say that the sequence of numbers is random. The *pseudo* part of the term *pseudorandom* is meant to emphasize this distinction. However, we can subject the sequence of numbers to a battery of **statistical tests** to check whether it can be readily distinguished from a random sequence.

For example, we can check that each number appears with approximately equal frequency. For example, a sequence purporting to sample uniformly from $\{1, 2, 3, 4, 5\}$ which begins

$$1, 1, 5, 5, 5, 1, 1, 5, 1, 1, 5, 5, 5, 1, 1, 1, 2, 1, 5, 5, 1, 2, 1, 5, 1, 1, 1, \ldots$$

is probably not a good pseudorandom number generator. However, some clearly non-random sequences pass the basic frequency test:

$$\{a_n\}_{n=1}^{\infty} = \{1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, \ldots\}$$

To detect such failures, we can split up the sequence into pairs of consecutive terms and ensure that these pairs are also approximately equally represented. Since $\{1, 2\}$ appears often and $\{2, 1\}$ never appears in $\{a_n\}_{n=1}^{\infty}$, the pair frequency test is sufficient to distinguish this sequence from a random one. We can apply the same idea with blocks of length 3 or 4 or more.

Exercise 5.3.2

Consider the sequence $\{\text{mod}(3 \cdot 2^n, 11)\}_{n=1}^{100}$. Use Julia to show that each number from 1 to 10 appears exactly 10 times in this sequence. Also, use Julia to show that a_{2k} is smaller than a_{2k-1} for far more than half the values of k from 1 to 50. Hint: `countmap(a)` tells you how many times each element in the collection `a` appears. To use this function, do `using StatsBase` first.

Repeat these tests on the sequence whose kth term is the kth digit in the decimal representation of π: `reverse(digits(floor(BigInt,big(10)^99*π)))`.

A PRNG is *cryptographically secure* if an agent who knows the algorithm used to generate the numbers (but who does not know the value of the seed) cannot feasibly infer the kth number generated based on observation of the first $k-1$ numbers generated.

Most PRNGs are not cryptographically secure. In other words, they hold up well under the scrutiny of *general* statistical tests, but not to tests which exploit knowledge of the specific algorithm used.

As a simple example of a PRNG that is not cryptographically secure, consider the digits of π starting from some unknown position. This sequence does behave *statistically* as though it were random (as far as we know), but an adversary who knew you were using successive digits of π would be able to use several values output by your PRNG to find your position and start anticipating subsequent values.

5.4 Automatic differentiation

* Symbolic differentiation means applying differentiation rules to a symbolic representation of f, like $\left(\frac{\mathrm{d}}{\mathrm{d}x}\sqrt{x} = \frac{1}{2\sqrt{x}}\right)$

Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a function whose definition is too complicated for us to feasibly differentiate symbolically.* Perhaps the most straightforward way to approximate the derivative of f is to calculate the difference quotient

$$\frac{f(x+\epsilon) - f(x)}{\epsilon}$$

a small value of ϵ. However, this approach is very inaccurate because the subtraction step is ill-conditioned.

Exercise 5.4.1

Use difference quotients to approximate the derivative of $f(x) = x^2$ at $x = \frac{2}{3}$, with $\epsilon = 2^k$ as k ranges from -60 to -20. What is the least error over these values of k? How does that error compare to machine epsilon?

The problem with difference quotients is that the accuracy of $f(x+\epsilon) - f(x)$ degrades as $\epsilon \to 0$ (due to catastrophic cancellation—see Example 5.2.13). Even at the optimal value of ϵ, the precision is still very poor.

On the other hand, the problem of calculating the derivative of f is well-conditioned as long as the condition number $|xf''(x)/f'(x)|$ of $f'(x)$ isn't too large. So the difference quotient algorithm is unstable, and we may hope for a stable alternative.

Indeed, there is an approach to derivative computation which is precise, fast, and scalable: **automatic differentiation**. The idea is to substitute the *matrix*

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

in place of x in the program that computes $f(x)$. This requires that any internal calculations performed by f are able to handle 2×2 matrices as well as plain numbers. The matrix resulting from this calculation will be equal to

$$\begin{bmatrix} f(x) & f'(x) \\ 0 & f(x) \end{bmatrix},$$

allowing us to read off the derivative as the top-right entry.

Exercise 5.4.2

In this exercise, we will explain why

$$f\left(\begin{bmatrix} x & 1 \\ 0 & x \end{bmatrix}\right) = \begin{bmatrix} f(x) & f'(x) \\ 0 & f(x) \end{bmatrix}, \tag{5.4.1}$$

for any polynomial f.

(i) Check that (5.4.1) holds for the identity function (the function which returns its input) and for the function which returns the multiplicative identity.

(ii) Check that if (5.4.1) holds for two differentiable functions f and g, then it holds for the sum $f+g$ and the product fg.

(iii) Explain why (5.4.1) holds for any polynomial function $f(x)$.

Exercise 5.4.3

Use automatic differentiation to find the derivative of $f(x) = (x^4 - 2x^3 - x^2 + 3x - 1)e^{-x^4/4}$ at the point $x = 2$. Compare your answer to the true value of $f'(2)$.

Hint: You'll want to define f using

```
using LinearAlgebra
f(t) = exp(-t^2/4)*(t^4 - 2t^3 - t^2 + 3t - I)
```

where `I` is an object which is defined to behave like multiplicative identity (note that subtracting the identity matrix is the appropriate matrix analogue of subtracting 1 from a real number).

Also, to help check your answer, here's the symbolic derivative of f:

```
df(t) = (-t^5 + 2*t^4 + 9*t^3 - 15*t^2 - 3*t + 6)*exp(-t^2/4)/2
```

In practice, you will usually want to use a library to perform automatic differentiation, because ensuring suitable matrix-awareness of all of the functions called by f can be a daunting task. Julia has the package `ForwardDiff` for this purpose, and in Python you can use `autograd` (which works for all of the NumPy functions).

5.5 Optimization

5.5.1 Gradient descent

Gradient descent is an approach to finding the minimum of a function f from $\mathbb{R}^n$ to $\mathbb{R}$. The basic idea is to repeatedly step in the direction of $-\nabla f$, since that is f's direction of maximum decrease from a given point, beginning with some initial guess $\mathbf{x}_0 \in \mathbb{R}^n$.

We can choose how large each step should be and when to stop. A common way to determine step size is to fix a **learning rate** ϵ and set $\mathbf{x}_{n+1} = \mathbf{x}_n - \epsilon\nabla f(\mathbf{x}_{n-1})$. Note that the size of the step naturally gets smaller as we get closer to a local minimum, since the norm of the gradient decreases. One way to choose when to terminate the algorithm is to set a threshold for the norm of the gradient of f.

Gradient descent is fundamentally local: it is not guaranteed to find the global minimum since the search can get stuck in a local minimum.

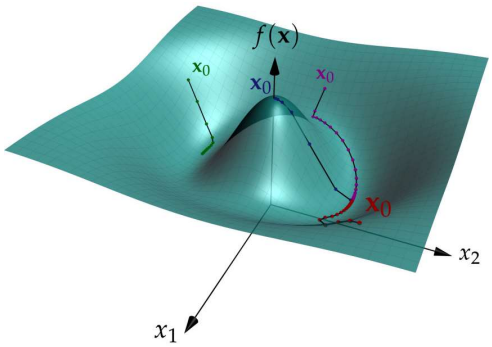


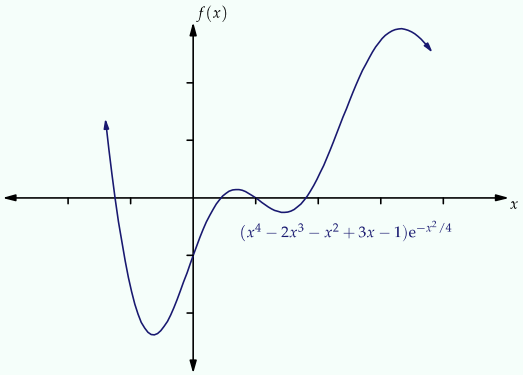
Figure 5.2 To find a local minimum of a function, we repeatedly take steps in the direction of maximum decrease. Results are shown for several starting points $\mathbf{x}_0$.

Exercise 5.5.1

Consider the function $f(x) = (x^4 - 2x^3 - x^2 + 3x - 1)e^{-x^2/4}$. Implement the gradient descent algorithm for finding the minimum of this function.

(i) If the learning rate is $\epsilon = 0.1$, which values of x_0 have the property that $f(x_n)$ is close to the global minimum of f when n is large?

(ii) Is there a starting value x_0 between -2 and -1 and a learning rate ϵ such that the gradient descent algorithm does not reach the global minimum of f? Use the graph for intuition.



A subset of $\mathbb{R}^n$ is **convex** if it contains every line segment connecting any two points in the set (sets like ▲ and ⬣ are convex, while ⬓ isn't convex).

A function from $\mathbb{R}^n$ to $\mathbb{R}$ is **convex** if a line segment connecting any two points on its graph lies on or above the graph. For example, a function like $f(x, y) = x^2 + y^2$ with a bowl-shaped graph ( ) is convex, while a function like $g(x, y) = x^2 - y^2$ with a saddle-shaped graph ( ) is not convex. A convex function is *strictly convex* if a line segment connecting any two points on its graph touches the graph only at the endpoints. You can check whether a smooth function is convex by checking whether its Hessian is positive semidefinite everywhere. If the Hessian is positive definite everywhere, then the function is also strictly convex.

A **convex optimization problem** is a problem of the form *find the minimum value of* $f : A \to \mathbb{R}$, where f is convex and $A \subset \mathbb{R}^n$ is convex. Compared to general optimization, convex optimization is particularly well-behaved:

Theorem 5.5.1

If $f : A \to \mathbb{R}$ is convex and $A \subset \mathbb{R}^n$ is convex, then any local minimum of f is also a global minimum of f. Furthermore, if f is strictly convex, the f has at most one local minimum.

Convex optimization problems play an important role in applied math and data science, because (1) many optimization problems of interest can be expressed in the form of a convex optimization problem, and (2) specialized, fast numerical methods are available for such problems.

Example 5.5.1

Use the Julia* package `JuMP` with the `Ipopt` solver to find the minimum of the function $f(x, y) = x^2 + 2y^2$ on the half-plane $x + y \geq 1$.

* The Python package `cvxopt` and the R package `cvxr` have a similar interface.

Exercise 5.5.2

Use JuMP to find the line of best fit for the points $(1,2), (2,5), (4,4)$. In other words, find the values m and b such that the sum of squared vertical distances from these three points to the line $y = mx + b$ is minimized.

5.6 Parallel Computing

Parallel computing involves decomposing a computational task into subtasks which may be performed concurrently.

Example 5.6.1

The problem of adding the numbers in an array is readily *parallelizable*, since we can subdivide the array, sum the values in each smaller array, and add up the resulting sums at the end.

You can start a Julia session with n worker processes via `julia -p n` and loading the distributed computing tools with `using Distributed`.

1. `pmap(f,A)` applies the function `f` to each element of the collection `A`, taking advantage of the available worker processes. For example, to parallelly check the primality of the positive integers up to 100,000:

```
using Primes
pmap(isprime,2:100_000)
```

2. If `(op)` is an operator, then `@parallel (op) for ... end` assigns a subrange of the given `for` loop to each worker. The values returned by the body of the loop are combined using the operator `op`. For example, to sum a million random Gaussians in parallel fashion:

```
@parallel (+) for i=1:1_000_000
  randn()
end
```

6 Probability

When we do data science, we begin with a data set and work to gain insights about the process that generated the data. Crucial to this endeavor is a robust vocabulary for discussing the behavior of data-generating processes.

It is helpful to initially consider data-generating processes whose randomness properties are specified completely and precisely. The study of such processes is called **probability**. For example, "What's the probability that I get at least 7 heads in 10 independent flips of a fair coin?" is a probability question, because the setup is fully specified: the coins have exactly 50% probability of heads, and the different flips do not affect one another.

The question of whether the coins are really fair or whether the flips are really independent will be deferred to our study of *statistics*. In statistics, we will have the *outcome* of a random experiment in hand and will be looking to draw inferences about the unknown *setup*. Once we are able to answer questions in the "setup $\rightarrow$ outcome" direction, we will be well positioned to approach the "outcome $\rightarrow$ setup" direction.

Exercise 6.0.1

Label each of the following questions as a probability question or a statistics question.

(i) On days when the weather forecast says that the chance of rain is 10%, it actually rains only about 5% of the time. What is the probability of rain on a day when the weather forecast says "10% chance of rain"?

(ii) If it will rain today with probability 40%, what is the probability that it will *not* rain today?

(iii) If you roll two fair dice, what is the average total number of pips showing on the top faces?

(iv) Your friend rolled 12 on each of the first three rolls of the board game they're playing with you. What is the probability that the dice they're using are weighted in favor of the 6's?

6.1 Counting

We begin our study of probability with a closely related skill: *counting*.

Theorem 6.1.1: Fundamental principle of counting

If one experiment has m possible outcomes, and if a second experiment has n possible outcomes for each of the outcomes in the first experiment, then there are mn possible outcomes for the pair of experiments.

Example 6.1.1

If you flip a coin and roll a die, there are $2 \times 6 = 12$ possible flip-roll pairs.

One simple way to prove the fundamental theorem of counting is to observe that the possible outcomes for the pair of experiments can be arranged to form an $m \times n$ rectangle:

	1	2	3	4	5	6
H	(H,1)	(H,2)	(H,3)	(H,4)	(H,5)	(H,6)
T	(T,1)	(T,2)	(T,3)	(T,4)	(T,5)	(T,6)

The fundamental principle of counting may be used to determine the number of ordered r-tuples of distinct elements of $\{1, 2, \ldots n\}$: we begin forming an r-tuple by selecting any one of the n possibilities for the first entry. Given any of the choices for the first entry, there are $n-1$ choices for the second entry. By the fundamental principle of counting, there are $n(n-1)$ choices for the first two entries. Continuing in this way, we find that there are

$$n(n-1)(n-2)\cdots(n-r+1)$$

choices for filling in all r entries.

Example 6.1.2

How many three-digit positive integers have distinct digits?

Note: a positive integer must be between 100 and 999 (inclusive) to count as a three-digit integer.

The number of r-element subsets of an n-element set is denoted $\binom{n}{r}$. Expressions of the form $\binom{n}{r}$ are called **binomial coefficients***.

*for reasons that we will explore in Exercise 6.1.4

Example 6.1.3

We have $\binom{4}{3} = 4$, since there are four ways to choose a 3-element subset of a 4-element set. The sets

$$\{1,2,3\}, \{1,2,4\}, \{1,3,4\}, \{2,3,4\}$$

are all of the 3-element subsets of $\{1,2,3,4\}$.

To work out a general procedure for evaluating $\binom{n}{r}$, we may first count the number of r-tuples and then account for all of the repeats. For example, if $r = 3$, then the tuples

$$(1,2,3), (1,3,2), (2,1,3), (2,3,1), (3,1,2), (3,2,1)$$

should collectively contribute 1, rather than 6, to the total count. Since every set of r elements corresponds to $r(r-1)(r-2)\cdots(2)(1)$ r-tuples of distinct elements, we divide the number of r-tuples by this number to obtain an expression for $\binom{n}{r}$:

$$\binom{n}{r} = \frac{n(n-1)(n-2)\cdots(n-r+1)}{r(r-1)(r-2)\cdots(2)(1)}.$$

We often abbreviate the product $r(r-1)(r-2)\cdots(2)(1)$ as $r!$.

Exercise 6.1.1

Of the 1024 total length-10 strings composed of the symbols H and T, how many of them have exactly 6 T's and 4 H's? (HHTHTTTHTT is such a string).

Exercise 6.1.2: Principle of Inclusion-Exclusion

Let $\Omega = \{0, 1, 2, \cdots, 100\}$ be the set of natural numbers up to and including 100. Let $A \subset \Omega$ the subset of integers divisible by 3, and $B \subset \Omega$ the subset of integers divisible by 5.

- Compute $|A|$.
- Compute $|B|$.
- Compute $|A \cap B|$.
- Explain why $|A \cup B| = |A| + |B| - |A \cap B|$.
- Use the prior steps to find $|A \cup B|$.

Exercise 6.1.3

How many subsets does the English alphabet have? For example, $\{a, r, w\}$ and $\{d, g, m, x, y, z\}$ are two such subsets.

Exercise 6.1.4

Expand the algebraic expression $(x + y)^3$. Show that the coefficients of this expansion are given by the binomial coefficients of the form $\binom{3}{r}$ where r ranges from 0 to 3:

$$(x+y)^3 = \binom{3}{0}x^3y^0 + \binom{3}{1}x^2y^1 + \binom{3}{2}x^1y^2 + \binom{3}{3}x^0y^3$$

Write a corresponding expansion for $(x + y)^4$.

6.2 Probability models

In this section we will learn how to mathematically represent and reason about randomness. The benefit of having an explicit mathematical model is that the intuitive approach to probability has serious limitations when analyzing tricky or sophisticated phenomena. Consider the following example.

Example 6.2.1: Exchange paradox

Two envelopes are placed on the table in front of you, containing X and $2X$ dollars for some unknown positive number X (you don't know which envelope is which). You choose one of the envelopes and discover \$10 inside. You have a choice to switch envelopes; should you?

On one hand, your chance of getting the better envelope was 50% to begin with, and opening the envelope did not provide any information on whether you succeeded. From this perspective, you should be indifferent to switching.

On the other hand, you might reason that the unopened envelope contains either \$20 or \$5, with a 50% chance of each. So on average the other envelope contains \$12.50. from this perspective, you should switch.

How can we adjudicate between these contradictory analyses? We need a **model** for the situation—that is, a mathematical object together with a way to translate questions about the situation to unambiguous questions about the object.

6.2.1 Discrete probability models

Let's develop a model for the following simple experiment: **two flips of a fair coin**. The first thing to observe about this experiment is that we can write down all of the possible outcomes:

$$\{(\mathsf{H},\mathsf{H}),(\mathsf{H},\mathsf{T}),(\mathsf{T},\mathsf{H}),(\mathsf{T},\mathsf{T})\}.$$

This set clearly bears an important relationship to the experiment; let's call it the **sample space** and denote it as Ω.

Furthermore, we need a way to specify how likely each outcome is to occur. It seems reasonable in this scenario to believe that each of the four outcomes is equally likely, in which case we should assign a probability value of $\frac{1}{4}$ to each outcome. The mathematical object which assigns a particular value to each element in a set is a *function*, so we will call this assignment of probability values the **probability mass function*** and denote it as m. So all together, we have

* *Mass* here is a metaphor for probability: we think of outcomes that are more likely as more massive

- the sample space Ω, which contains the possible outcomes of the experiment, and
- the probability mass function m from Ω to $[0,1]$ which indicates the probability of each outcome in Ω.

The pair (Ω, m) is already enough to specify the experiment, but we need a few more translations for the model to be useful: in the context of the experiment, an *event* is a predicate whose occurrence can be determined based on the outcome. For example, "the first flip turns up heads" is an event.

Exercise 6.2.1

Identify a mathematical object in our model (Ω, m) which can be said to correspond to the phrase "the first flip turns up heads". Which of the following is true of this object?

(i) It is one of the values of the function m

(ii) It is the set Ω

(iii) It is a subset of Ω

(iv) It is one of the elements of Ω

Exercise 6.2.2

Explain how to obtain the probability of an event from the probability mass function.

For concreteness, consider $\Omega = \{(\mathsf{H},\mathsf{H}),(\mathsf{H},\mathsf{T}),(\mathsf{T},\mathsf{H}),(\mathsf{T},\mathsf{T})\}$, a probability mass function which assigns mass $\frac{1}{4}$ to each outcome, and the event $\{(\mathsf{H},\mathsf{H}),(\mathsf{H},\mathsf{T})\}$.

Some common terms for combining and modifying predicates include **and**, **or**, and **not**. For example, we might be interested in the event "the first flip comes up heads and the second does not come up heads, or the first flip comes tails". Each of these corresponds to one of the set-theoretic operations we have learned:

Exercise 6.2.3

Match each term to its corresponding set-theoretic operation. Assume that E and F are events.

For concreteness, you can think about the events "first flip comes up heads" and "second flip comes up heads" for the two-flip probability space we've been considering.

(a) the event that E and F both occur	(i) the intersection $E \cap F$
(b) the event that E does not occur	(ii) the union $E \cup F$
(c) the event that either E occurs or F occurs	(iii) the complement E^c

Exercise 6.2.4

Suppose a group of n friends enter the lottery. For $i \in \{1, \ldots, n\}$ let E_i be the event that the ith friend wins. Express the following events using set notation.

1. At least one friend loses.
2. All friends win.
3. At least one friend wins.

Since events play a more prominent role than individual outcomes in discussions of probability, we will demote the probability mass function to auxiliary status and instead focus on the function $\mathbb{P}$ from the set of *events* to $[0, 1]$ which assigns to each event the total probability mass therein. For example, for our two-flip experiment, the function $\mathbb{P}$ satisfies

$$\begin{aligned} \mathbb{P}(\{(\mathsf{H},\mathsf{T})\}) &= \tfrac{1}{4} \\ \mathbb{P}(\{\}) &= 0 \\ \mathbb{P}(\{(\mathsf{H},\mathsf{H}),(\mathsf{H},\mathsf{T}),(\mathsf{T},\mathsf{T})\}) &= \tfrac{3}{4} \\ \mathbb{P}(\Omega) &= 1, \end{aligned}$$

and so on.

Exercise 6.2.5

What is the cardinality of the domain of the function $\mathbb{P}$ if

$$\Omega = \{(\mathsf{H},\mathsf{H}),(\mathsf{H},\mathsf{T}),(\mathsf{T},\mathsf{H}),(\mathsf{T},\mathsf{T})\}?$$

This notation means that we sum the function m over all outcomes ω in the event E.

We call* $\mathbb{P}(E) = \sum_{\omega \in E} m(\omega)$ the **probability measure** associated with the probability mass function m. The pair $(\Omega, \mathbb{P})$ is called a **probability space**. Probability measures satisfy the following properties.

Theorem 6.2.1: Properties of a probability measure

If $(\Omega, \mathbb{P})$ is a probability space, then

1. $\mathbb{P}(\Omega) = 1$ — "something has to happen"
2. $\mathbb{P}(E) \geq 0$ for all $E \subset \Omega$ — "probabilities are non-negative"
3. $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$ **if** E and F are mutually exclusive events — "probability is additive"

These are the fundamental properties of a probability measure on a finite sample space Ω, in the sense that

functions from the set of events to $[0,1]$ satisfying the above properties are in one-to-one correspondence with probability mass functions.

One further important property is a consequence of the properties in Theorem 6.2.1. It says that if B's occurrence implies A's occurrence, then $\mathbb{P}(B) \leq \mathbb{P}(A)$.

Exercise 6.2.6: Monotonicity

Use the additivity property and the fact that $A = (A \cap B) \cup (A \cap B^c)$ to show that if $B \subset A \subset \Omega$, then $\mathbb{P}(B) \leq \mathbb{P}(A)$.

Exercise 6.2.7: Subadditivity

Show that $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ for all events A and B.

Use this property to show that if A occurs with probability zero and B occurs with probability zero, then the probability that A *or* B occurs is also zero.

If Ω is countably infinite, then the additivity property extends to *countable additivity*: If $E_1, E_2, \ldots$ is a pairwise disjoint sequence of events, then $\mathbb{P}(E_1 \cup E_2 \cup \cdots) = \mathbb{P}(E_1) + \mathbb{P}(E_2) + \cdots$.

Example 6.2.2: Countable additivity

Suppose that Ω is the set of ordered pairs of positive integers, with probability mass $m((i,j)) = 2^{-i-j}$ at each pair (i,j). Show that the probability of the event $\{(i,j) \in \Omega : i > 2\}$ is equal to the sum of the probabilities of the events $\{(i,j) \in \Omega : i = t\}$ as t ranges over $\{3,4,5,\ldots\}$

Exercise 6.2.8

Show that the function $m((i,j)) = 2^{-i-j}$ sums to 1 as (i,j) ranges over the set of ordered pairs of positive integers.

6.3 Random variables

An event is a binary function of the outcome of an experiment: based on the outcome, we can say that the event occurred or didn't occur. Sometimes, however, we are interested in specifying *real*-valued information based on the outcome of the experiment.

For example, suppose that you will receive a dollar for each head flipped in our two-fair-flips experiment. Then your payout X might be 0 dollars, 1 dollar, or 2 dollars. Because X is a variable whose value is random (that is, dependent on the outcome of a random experiment), it is called a **random variable**. A random variable which takes values in some finite or countably infinite set (such as $\{0,1,2\}$, in this case) is called a **discrete** random variable.

Since a random variable associates a real number to each outcome of the experiment, in mathematical terms

a random variable is a *function* from the sample space to $\mathbb{R}$. Using function notation, the dollar-per-head payout random variable X satisfies

$$\begin{aligned} X((\mathsf{T},\mathsf{T})) &= 0, \\ X((\mathsf{H},\mathsf{T})) &= 1, \\ X((\mathsf{T},\mathsf{H})) &= 1, \text{ and} \\ X((\mathsf{H},\mathsf{H})) &= 2. \end{aligned}$$

We can combine random variables using any operations or functions we can use to combine numbers. For example, suppose X_1 is defined to be the number of heads in the first flip—that is,

$$\begin{aligned} X_1((\mathsf{T},\mathsf{T})) &= 0 \\ X_1((\mathsf{H},\mathsf{T})) &= 1 \\ X_1((\mathsf{T},\mathsf{H})) &= 0 \\ X_1((\mathsf{H},\mathsf{H})) &= 1, \end{aligned}$$

and X_2 is defined to be the number of heads in the second flip. Then the random variable $X_1 + X_2$ maps each $\omega \in \Omega$ to $X_1(\omega) + X_2(\omega)$. Note that this random variable is equal to X, since $X(\omega) = X_1(\omega) + X_2(\omega)$ for every $\omega \in \Omega$.

Exercise 6.3.1

Suppose that the random variable X represents a fair die roll and Y is defined to be the remainder when X is divided by 4.

Define a six-element probability space Ω on which X and Y may be defined, and find* $\mathbb{P}(X - Y = k)$ for every integer value of k.

* This notation is shorthand for $\mathbb{P}(\{\omega : X(\omega) - Y(\omega) = k\})$

Exercise 6.3.2

Consider a sample space Ω and an event $E \subset \Omega$. We define the random variable $\mathbf{1}_E : \Omega \to \{0, 1\}$ by

$$\mathbf{1}_E(\omega) = \begin{cases} 1 & \text{if } \omega \in E \\ 0 & \text{otherwise.} \end{cases}$$

The random variable $\mathbf{1}_E$ is called the indicator random variable for E. If F is another event, which of the following random variables are necessarily equal?

(i) $\mathbf{1}_{E \cap F}$ and $\mathbf{1}_E \cdot \mathbf{1}_F$

(ii) $\mathbf{1}_{E \cup F}$ and $\mathbf{1}_E + \mathbf{1}_F$

(iii) $\mathbf{1}_E$ and $1 - \mathbf{1}_{E^c}$

6.3.1 Marginal distributions

Given a probability space $(\Omega, \mathbb{P})$ and a random variable X, the **distribution** of X tells us how X *distributes* probability mass on the real number line. Loosely speaking, the distribution tells us where we can expect to find X and with what probabilities.

Definition 6.3.1: Distribution of a random variable

The distribution (or *law*) of a random variable X is the probability measure on $\mathbb{R}$ which maps a set $A \subset \mathbb{R}$ to $\mathbb{P}(X \in A)$.

We can think of X as pushing forward the probability mass from Ω to $\mathbb{R}$ by sending the probability mass at ω to $X(\omega)$ for each $\omega \in \Omega$. As you can see in Figure 6.1, the probability masses at multiple ω's can stack up at the same point on the real line if X maps the ω's to the same value.

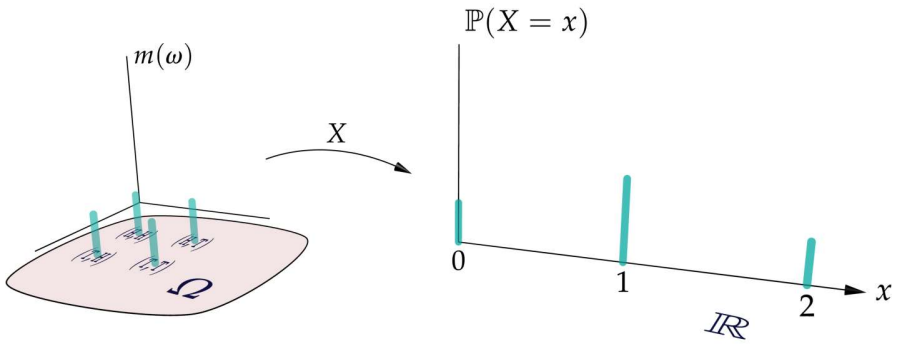


Figure 6.1 The distribution of a discrete random variable is the measure on $\mathbb{R}$ obtained by pushing forward the probability masses at elements of the sample space to their locations on the real line.

Exercise 6.3.3

A problem on a test requires students to match molecule diagrams to their appropriate labels. Suppose there are three labels and three diagrams and that a student guesses a matching uniformly* at random. Let X denote the number of diagrams the student correctly labels. What is the probability mass function of the distribution of X?

* A *uniform* probability measure spreads the probability mass out evenly over Ω

6.3.2 Cumulative distribution function

The distribution of a random variable X may be specified by its probability mass function or by its **cumulative distribution function*** F_X:

* CDF, for short

Definition 6.3.2: Cumulative distribution function

If X is a random variable, then its cumulative distribution function F_X is the function from $\mathbb{R}$ to $[0, 1]$ defined by

$$F_X(x) = \mathbb{P}(X \leq x).$$

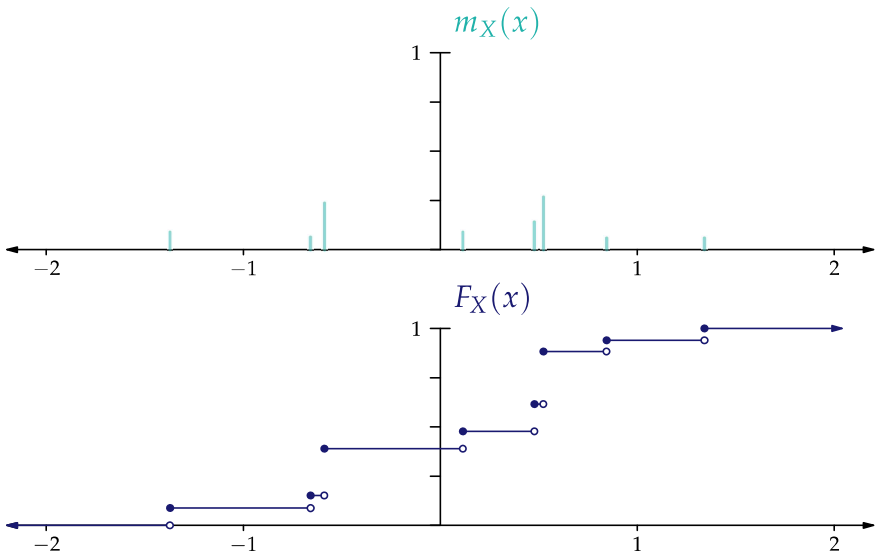


Figure 6.2 A probability mass function m_X and its corresponding CDF F_X

Exercise 6.3.4

Consider a random variable X whose distribution is as shown in Figure 6.2. Identify each of the following statements as true or false.

(a) $\mathbb{P}(-1 < X < 1)$ is greater than $\frac{3}{5}$

(b) $\mathbb{P}(X \geq 2) = 0$

(c) $\mathbb{P}\left(-\frac{1}{2} < X < 0\right)$ is greater than $\frac{1}{100}$

(d) $\mathbb{P}(100X < 1)$ is greater than $\frac{1}{2}$

Exercise 6.3.5

Suppose that X is a random variable with CDF F_X and that $Y = X^2$. Express $\mathbb{P}(Y > 9)$ in terms of the function F_X. For simplicity, assume that $\mathbb{P}(X = -3) = 0$.

Exercise 6.3.6

Random variables with the same cumulative distribution function are not necessarily equal as random variables, because the probability mass sitting at each point on the real line can come from different ω's.

For example, consider the two-fair-coin-flip experiment and let X be the number of heads. Find another random variable Y which is not equal to X but which has the same distribution as X.

6.3.3 Joint distributions

The distribution of a random variable is sometimes its called its **marginal** distribution, with the term *marginal* emphasizing that distribution includes information only about a single random variable. If we are interested in two random variables X and Y, it is often important to consider their *joint* distribution, which captures probabilistic information about where the pair (X, Y) falls in $\mathbb{R}^2$.

Definition 6.3.3

If X and Y are two random variables defined on the same probability space, then the **joint distribution** of X and Y is the measure on $\mathbb{R}^2$ which assigns to each set $A \subset \mathbb{R}^2$ the value $\mathbb{P}((X, Y) \in A)$.

We can find the probability mass function of (X, Y) by (i) finding all of the pairs $(x, y) \in \mathbb{R}^2$ with the property that the event $\{X = x\} \cap \{Y = y\}$ has positive probability, and (ii) finding the probability of each such event.

Example 6.3.1

Consider the two-fair-coin-flip experiment, and let X_1 be the number of heads in the first flip and X_2 the number of heads in the second flip. Let Y_1 be the number of tails in the first flip.

Show that X_1, X_2, and Y_1 all have the same marginal distributions and but that (X_1, X_2) and (X_1, Y_1) have different joint distributions.

The marginal distributions of two random variables may be recovered from their joint distribution.

Exercise 6.3.7

Consider a computer program which rolls two virtual dice and returns roll results with probabilities shown in the table.

What is the probability that Die 1 shows 4?

Die 2 \ Die 1	1	2	3	4	5	6
1	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{1}{36}$
2	$\frac{1}{72}$	$\frac{1}{36}$	$\frac{1}{72}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{1}{36}$	$\frac{1}{72}$	$\frac{1}{72}$	$\frac{1}{72}$	$\frac{1}{36}$	$\frac{2}{36}$
4	$\frac{1}{72}$	$\frac{1}{72}$	$\frac{1}{72}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{72}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{72}$

6.4 Conditional probability

6.4.1 Conditional probability measures

One of the most important goals of modeling random phenomena is to account for *partial information*. We often discover something about the outcome of an experiment before we know the outcome exactly. For example, when we flip a fair coin twice, we see the result of the first flip before we see the result of the second flip, and we would like to define a new probability measure which reflects this intermediate knowledge. We call this a **conditional probability measure**.

Suppose we observe that the first of two flips is a tail. Then all of the ω's which are incompatible with this observation should receive a probability of zero under our conditional probability measure. Since we have no new information about the remaining ω's, it makes sense to keep their probabilities in the same proportions as in the original probability measure.

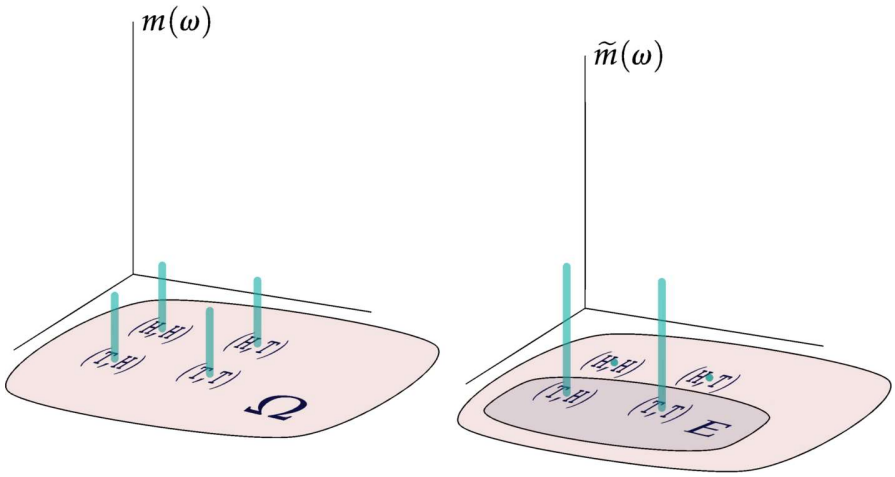


Figure 6.3 Consider the event E that the first flip is a tail. The conditional probability mass function $\widetilde{m}$ given E assigns probability mass $\frac{1}{2}$ to each of the ω's in E.

These two observations are sufficient to determine the conditional probability measure. In other words, to condition on an event E, we set the masses at elements of E^c to 0 and multiply the amount of mass at each point in E by $1/\mathbb{P}(E)$ to get the total mass up to 1 without changing the proportions:

Definition 6.4.1

Given a probability space $(\Omega, \mathbb{P})$ and an event $E \subset \Omega$ whose probability is positive, the *conditional probability mass function* given E, written as $\omega \mapsto m(\omega \mid E)$ is defined by

$$m(\omega \mid E) = \begin{cases} \frac{m(\omega)}{P(E)} & \text{if } \omega \in E \\ 0 & \text{otherwise.} \end{cases}$$

The conditional probability measure given E is the measure associated to $\omega \mapsto m(\omega \mid E)$: for all events F, we have

$$\mathbb{P}(F \mid E) = \frac{\mathbb{P}(F \cap E)}{\mathbb{P}(E)}. \tag{6.4.1}$$

Exercise 6.4.1

Two objects are submerged in a deep and murky body of water. The objects are chosen to be both positively buoyant* with probability $\frac{1}{4}$, both are negatively buoyant with probability $\frac{1}{4}$, and with probability $\frac{1}{2}$ the objects have opposite buoyancy. The objects, if they float, rise in the water at different rates, but they are visually indistinguishable.

After the objects are released, an observer sees one of them emerge at the water's surface. What is the conditional probability, given the observed information, that the second object will emerge?

* An object with positive buoyancy floats and an object with negative buoyancy sinks

One reason that conditional probabilities play such an important role in the study of probability is that in many scenarios they are more fundamental than the probability measure on Ω.

Example 6.4.1

Consider the following experiment: we roll a die, and if it shows 2 or less we select Urn A, and otherwise we select Urn B. Next, we draw a ball uniformly at random from the selected urn. Urn A contains one red and one blue ball, while urn B contains 3 blue balls and one red ball.

Find a probability space Ω which models this experiment, find a pair of events E and F such that $\mathbb{P}(E \mid F) = \frac{3}{4}$.

Exercise 6.4.2

Consider three random variables X_1, X_2, and X_3, each of which is equal to 1 with probability 0.6 and to 0 with probability 0.4. These random variables are not necessarily independent.

(i) Find the greatest possible value of the event $X_1 + X_2 + X_3 = 0$.

(ii) Find the least possible value of the event $X_1 + X_2 + X_3 = 0$.

6.4.2 Bayes' theorem

Bayes' theorem tells us how to update beliefs in light of new evidence. It relates the conditional probabilities $\mathbb{P}(A \mid E)$ and $\mathbb{P}(E \mid A)$:

$$\mathbb{P}(A \mid E) = \frac{\mathbb{P}(E \mid A)\mathbb{P}(A)}{\mathbb{P}(E)} = \frac{\mathbb{P}(E \mid A)\mathbb{P}(A)}{\mathbb{P}(E \mid A)\mathbb{P}(A) + \mathbb{P}(E \mid A^c)\mathbb{P}(A^c)}.$$

The last step follows from writing out $\mathbb{P}(E)$ as $\mathbb{P}(E \cap A) + \mathbb{P}(E \cap A^c)$.

Bayes' theorem has many applications to everyday life, some intuitive and others counterintuitive.

Example 6.4.2

Suppose you're 90% sure that your package was delivered today and 75% sure that if it was delivered it would be on your door step rather than tucked away in your mailbox. When you arrive at home and do not see your package right away, what is the conditional probability—given the observed information—that you'll find it in your mailbox?

Exercise 6.4.3

Suppose a disease has 0.1% prevalence in the population and has a test with 90% reliability. A random selected person is tested for the disease and tests positive. What is the conditional probability that the person has the disease, given the positive test result?

6.4.3 Independence

In the context of a random experiment, two positive-probability events E and F are **independent** if knowledge of the occurrence of one of the events gives no information about the occurrence of the other event. In other words, E and F are independent if the probability of E is the same as the conditional probability of E given F, and vice versa. In other words, E and F are independent if

$$\mathbb{P}(E) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)} \quad \text{and} \quad \mathbb{P}(F) = \frac{\mathbb{P}(F \cap E)}{\mathbb{P}(E)}.$$

Both of these equations rearrange to

$$\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F).$$

This equation is clearly symmetric in E and F, and it does not require that E and F have positive probability, so we take it as our fundamental independence equation for two events:

Definition 6.4.2: Independence

If $(\Omega, \mathbb{P})$ is a probability space, then two events E and F are said to be independent if

$$\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F).$$

If we want to check whether two positive-probability events are independent, we may check any one of the equations $\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F)$ or $\mathbb{P}(E) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)}$ or $\mathbb{P}(F) = \frac{\mathbb{P}(F \cap E)}{\mathbb{P}(E)}$, since they are all equivalent.

Exercise 6.4.4

Let X be the result of a six-sided die roll. Consider the following events.

$$A = \{X \text{ is even}\}$$
$$B = \{X \text{ is odd}\}$$
$$C = \{X \leq 4\}$$

Are events A and B independent? Are events A and C independent?

We say that two random variables X and Y are independent if the every pair of events of the form $\{X \in A\}$ and $\{Y \in B\}$ are independent, where $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$.

Exercise 6.4.5

Suppose that $\Omega = \{(\mathsf{H},\mathsf{H}), (\mathsf{H},\mathsf{T}), (\mathsf{T},\mathsf{H}), (\mathsf{T},\mathsf{T})\}$ and $\mathbb{P}$ is the uniform probability measure on Ω. Let X_1 be the number of heads in the first flip and let X_2 be the number of heads in the second flip. Show that X_1 and X_2 are independent.

Directly showing that random variables are independent can be tedious, because there are many events to check. However, there is a general way to construct Ω to get independent random variables. The idea is to build Ω as a rectangle:

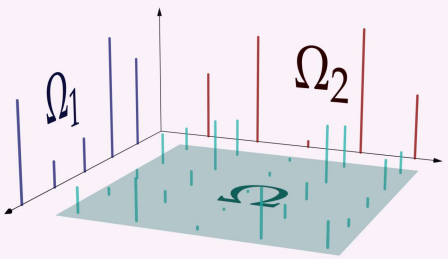
Theorem 6.4.1: Product measure

Suppose that $(\Omega_1, \mathbb{P}_1)$ and $(\Omega_2, \mathbb{P}_2)$ are probability spaces with associated probability mass functions m_1 and m_2. Define a probability space Ω by defining

$$\Omega = \Omega_1 \times \Omega_2$$

and

$$m((\omega_1, \omega_2)) = m_1(\omega_1)m_2(\omega_2)$$



for every $(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2$. Let $\mathbb{P}$ be the probability measure with probability mass function m. Then the random variables $X_1((\omega_1, \omega_2)) = \omega_1$ and $X_2((\omega_1, \omega_2)) = \omega_2$ are independent.
We call $\mathbb{P}$ a **product measure** and $(\Omega, \mathbb{P})$ a **product space**.

We say that a collection of random variables $(X_1, X_2, \ldots, X_n)$ is independent if

$$\mathbb{P}(\{X_1 \in A_1\} \cap \{X_2 \in A_2\} \cap \cdots \cap \{X_n \in A_n\}) = \mathbb{P}(X_1 \in A_1)\mathbb{P}(X_2 \in A_2) \cdots \mathbb{P}(X_n \in A_n)$$

for any events $A_1, A_2, \ldots, A_n$.

We may extend the product measure construction to achieve as many independent random variables as desired: for three random variables we let Ω be cube-shaped (that is, $\Omega = \Omega_1 \times \Omega_2 \times \Omega_3$), and so on.*

* These are not possible to directly visualize in dimensions higher than 3, but there is no problem with making longer tuples

Exercise 6.4.6

Define a probability space Ω and 10 independent random variables which are uniformly distributed on $\{1,2,3,4,5,6\}$.

The product measure construction can be extended further still to give a supply of *infinitely many* independent random variables. The idea is use a space of the form $\Omega = \Omega_1 \times \Omega_2 \times \Omega_3 \cdots$ (whose elements are infinite tuples $\omega = (\omega_1, \omega_2, \omega_3, \ldots)$) and define a measure which makes the random variables $X_n(\omega) = \omega_n$ independent. We will not need the details of this construction, although we will use it indirectly when we discuss infinite sequences of independent random variables.

We say that a collection of events is independent if the corresponding indicator random variables are independent. Independence for three or more events is more subtle than independence for two events:

Exercise 6.4.7

Three events can be *pairwise* independent without being independent: Suppose that ω is selected uniformly at random from the set

$$\Omega = \{(0,0,0),(0,1,1),(1,0,1),(1,1,0)\}$$

and define A to be the event that the first entry is 1, B to be the event that the second entry is 1, and C to be the event that the third entry is 1. For example, if $\omega = (0,1,1)$, then B and C occurred but A did not.

Show that A and B are independent, that A and C are independent, and that B and C are independent.

Show that the equation $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$ does **not** hold and that the triple of events is therefore not independent.

Independence satisfies many basic relationships suggested by the intuition that random variables are independent if they are computed from separate streams of randomness. For example, if X_1, X_2, X_3, X_4, X_5 are independent random variables, then $X_1 + X_2 + X_3$ and $X_4^2 + X_5^2$ are independent of each other.

Exercise 6.4.8

Consider as sequence of 8 independent coin flips. Show that the probability of getting at least one pair of consecutive heads is at least $1 - (3/4)^4$.

6.5 Expectation and Variance

6.5.1 Expectation

Sometimes we want to distill a random variable's distribution down to a single (non-random) number. For example, consider the height of an individual selected uniformly at random from a given population. This is a random variable, and communicating its distribution would involve communicating the heights of every person in the population. However, we may summarize the distribution by reporting an *average* height: we

add up the heights of the people in the population and divide by the number of people.

If the random individual were selected according to some non-uniform probability distribution on the population, then it would make sense to calculate a *weighted* average rather than a straight average. The probability-weighted average of the values of a random variable is called its **expectation**.

Definition 6.5.1

The **expectation** $\mathbb{E}[X]$ (or **mean** μ_X) of a random variable X is the *probability-weighted average of* X:

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) m(\omega)$$

For example, the expected number of heads in two fair coin flips is

$$\mathbb{E}[X] = \tfrac{1}{4} \cdot 2 + \tfrac{1}{4} \cdot 1 + \tfrac{1}{4} \cdot 1 + \tfrac{1}{4} \cdot 0 = 1.$$

There are two common ways of interpreting expected value.

1. The expectation $\mathbb{E}[X]$ may be thought of as the value of a random game with payout X. According to this interpretation, you should be willing to pay anything less than \$1 to play the game where you get a dollar for each head in two fair coin flips. For more than \$1 you should be unwilling to play the game, and at \$1 you should be indifferent.
2. The second way of thinking about expected value is as a *long-run average*. If you play the dollar-per-head two-coin-flip game a very large number of times, then your average payout per play is very likely to be close to \$1.

We can test this second interpretation out:

Exercise 6.5.1

Use the expression* `mean(rand(0:1) + rand(0:1) for k=1:10^6)` to play the dollar-per-head two-coin-flip game a million times and calculate the average payout in those milion runs.

How close to 1 is the result typically? Choose the best answer.

(a) Around 0.1

(b) Around 0.01

(c) Around 0.0001

(d) Around 0.0000001

In Julia you can drop the brackets in an array comprehension if you are substituting the array directly into a function. This saves time and memory because it skips generating the array.

We will see that this second interpretation is actually a *theorem* in probability, called the **law of large numbers**. In the meantime, however, this interpretation gives us a useful tool for investigation: if a random variable is easy to simulate, then we can sample from it many times and calculate the average of the resulting samples. This will not give us the expected value exactly, but we can get as close as desired by using sufficiently many samples. This is called the **Monte Carlo** method of approximating the expectation of a random variable.

Exercise 6.5.2

Use a Monte Carlo simulation to estimate the expectation of X/Y, where X and Y are independent die rolls.

Exercise 6.5.3

Explain why $\mathbb{E}[X] \leq \mathbb{E}[Y]$ if $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$.

Although the definition $\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)m(\omega)$ involves the probability space Ω, we can also write a formula for expectation in terms of the probability mass function of the *distribution* of X:

Theorem 6.5.1

The expectation of a discrete random variable X is equal to

$$\mathbb{E}[X] = \sum_{x \in \mathbb{R}} x\mathbb{P}(X = x).$$

Proof

Let's consider an example first. Suppose $\Omega = \{1, 2, 3\}$ with probability mass function m satisfying $m(1) = 1/6$, $m(2) = 2/6$, and $m(3) = 3/6$. Suppose $X(1) = 5, X(2) = 5$ and $X(3) = 7$. Then

$$\mathbb{E}[X] = \frac{1}{6} \cdot 5 + \frac{2}{6} \cdot 5 + \frac{3}{6} \cdot 7.$$

We can group the first two terms together to get

$$\mathbb{E}[X] = \left(\frac{1}{6} + \frac{2}{6}\right) \cdot 5 + \frac{3}{6} \cdot 7.$$

This expression is the one we would get if we wrote out

$$\sum_{x \in \mathbb{R}} x\mathbb{P}(X = x).$$

Therefore, we can see that the two sides are the same.

Let's write this idea down in general form. We group terms on the right-hand side in the formula $\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)m(\omega)$ according to the value of $X(\omega)$:

$$\mathbb{E}[X] = \sum_{x \in \mathbb{R}} \sum_{\omega \in \Omega \,:\, X(\omega) = x} X(\omega)m(\omega).$$

Then we can replace $X(\omega)$ with x and pull it out of the inside sum to get

$$\mathbb{E}[X] = \sum_{x \in \mathbb{R}} x \sum_{\omega \in \Omega \,:\, X(\omega) = x} m(\omega).$$

Since $\sum_{\omega \in \Omega \,:\, X(\omega) = x} m(\omega)$ is equal to $\mathbb{P}(X = x)$, we get

$$\mathbb{E}[X] = \sum_{x \in \mathbb{R}} x\mathbb{P}(X = x),$$

as desired.

Exercise 6.5.4

The expectation of a random variable need not be finite or even well-defined. Show that the expectation of the random variable which assigns a probability mass of 2^{-n} to the point 2^n (for all $n \geq 1$) is not finite.

Consider a random variable X whose distribution assigns a probability mass of $2^{-|n|-1}$ to each point 2^n for $n \geq 1$ and a probability mass of $2^{-|n|-1}$ to -2^n for each $n \leq -1$. Show that $\mathbb{E}[X]$ is not well-defined. (Note: a sum $\sum_{x\in\mathbb{R}} f(x)$ is not defined if $\sum_{x\in\mathbb{R}\,:\,f(x)>0} f(x)$ and $\sum_{x\in\mathbb{R}\,:\,f(x)<0} f(x)$ are equal to ∞ and $-\infty$, respectively.)

Theorem 6.5.2

If $f : \mathbb{R}^2 \to \mathbb{R}$, and X and Y are discrete random variables defined on the same probability space, then

$$\mathbb{E}[f(X,Y)] = \sum_{(x,y)\in\mathbb{R}^2} f(x,y)\mathbb{P}(X = x \text{ and } Y = y).$$

Proof

We use the same idea we used in the proof of Theorem 6.5.1: group terms in the definition of expectation according the value of the pair $(X(\omega), Y(\omega))$. We get

$$\begin{aligned}
\mathbb{E}[f(X,Y)] &= \sum_{\omega\in\Omega} f(X(\omega),Y(\omega))m(\omega) \\
&= \sum_{(x,y)\in\mathbb{R}^2} \sum_{\substack{\omega\in\Omega\,:\\ X(\omega)=x \text{ and } Y(\omega)=y}} f(X(\omega),Y(\omega))m(\omega) \\
&= \sum_{(x,y)\in\mathbb{R}^2} f(x,y)\mathbb{P}(X = x \text{ and } Y = y).
\end{aligned}$$

6.5.2 Linearity of expectation

In this section, we will develop a powerful tool for computing expected values and manipulating expressions involving expectations.

Example 6.5.1

Each of your n coworkers left their new (one-size-fits-all) company t-shirt on a table in the break room. Each person grabs a shirt at random on their way home. What is the expected number of employees who grab the same shirt they left on the table?

(i) Solve this problem in the special cases $n = 2$ and $n = 3$.

(ii) Note that if we define X_k to be the indicator of the event that the kth person collects the same shirt they left, then the total number of coworkers who collected the same shirt is $X_1 + X_2 + \cdots + X_n$. Show that the answer from (i) is the same as $\mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]$.

(iii) Assuming that the principle you discovered in (ii) holds in general, write down a formula in terms of n for the answer to the general question.

The surprising aspect of Example 6.5.1 is that the distribution of $X_1 + X_2 + \cdots + X_n$ depends on the joint distribution of the random variables $X_1, \ldots, X_n$, while the value of $\mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]$ depends only on the marginal distributions of these random variables. Nevertheless, it is indeed the case that expectation distributes across addition.

Theorem 6.5.3: Linearity of expectation

Let $c \in \mathbb{R}$. For all random variables X and Y, we have

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y] \quad \text{and} \quad \mathbb{E}[cX] = c\mathbb{E}[X].$$

Example 6.5.2

Shuffle a standard 52-card deck, and let X be the number of consecutive pairs of cards in the deck which are both red. Find $E[X]$.

Write some code to simulate this experiment and confirm that your answer is correct. Hint: store the deck of undrawn cards as a `Set`, and `pop!` cards from it as you draw. You can draw a random element from a set `S` using `rand(S)`.

Example 6.5.3

Show that if X and Y are independent discrete random variables, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Show that this equation does *not* hold in general if X and Y are not assumed to be independent.

Solution

We have

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{\omega \in \Omega} X(\omega)Y(\omega)m(\omega) \\
&= \sum_{(x,y)\in\mathbb{R}^2} \sum_{\substack{\omega\in\Omega: \\ X(\omega)=x \text{ and } Y(\omega)=y}} xym(\omega) \\
&= \left(\sum_{x\in\mathbb{R}} x\mathbb{P}(X=x)\right)\left(\sum_{y\in\mathbb{R}} y\mathbb{P}(Y=y)\right) \\
&= \mathbb{E}[X]\mathbb{E}[Y].
\end{aligned}$$

If X and Y are not independent, then $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$ is the typical situation. For example, let X be any mean-zero random variable (other than the random variable which is always zero) and let $Y = X$. Then $\mathbb{E}[XY] = \mathbb{E}[X^2] > 0$, and $\mathbb{E}[X]\mathbb{E}[Y] = 0 \cdot 0 = 0$.

6.5.3 Variance

The expectation of a random variable gives us some coarse information about where on the number line the random variable's probability mass is located. The **variance** gives us some information about how widely the probability mass is spread around its mean. A random variable whose distribution is highly concentrated

about its mean will have a small variance, and a random variable which is likely to be very far from its mean will have a large variance. We define the variance of a random variable X to be the average squared* distance from X to its mean:

* Why square? To ensure, in a smooth way, that both negative and positive deviations contribute positively.

Definition 6.5.2: Variance

The variance of a random variable X is defined to be

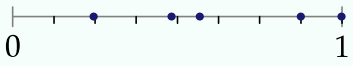
$$\operatorname{Var} X = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

The standard deviation σ_X of X is the square root of the variance:

$$\sigma(X) = \sqrt{\operatorname{Var} X}.$$

Exercise 6.5.5

Consider a random variable which is obtained by making a selection from the list



$$[0.245, 0.874, 0.998, 0.567, 0.482]$$

uniformly at random. Make a rough estimate of the mean and variance of this random variable just from looking at the number line. Then use Julia to calculate the mean and variance exactly to see how close your estimates were.

Note: *don't* use Julia's built-in `var` function; that will give you the correct answer to a different question, as we will see when we study statistics.

Exercise 6.5.6

Consider the following game. We begin by picking a number in $\{0, \frac{1}{1000}, \frac{2}{1000}, \ldots, \frac{1000}{1000}\}$ with uniform probability. If that number is less than 1, we pick another number from the same distribution and add it to the first. We repeat this procedure until the running sum exceeds 1. Let X be the random variable whose value is the number of draws needed to end the game. Use a simulation to approximate the expected value and variance of X. Include your code in your answer as well as some discussion of your results.

Tips: `rand(0:1000)/1000` returns a sample from the desired distribution. Also, it's a good idea to wrap a single run of the game into a zero-argument function.

We can use linearity of expectation to rewrite the formula for variance in a simpler form:

$$\begin{aligned}
\operatorname{Var} X &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2] \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X\mathbb{E}[X]] + \mathbb{E}[X]^2 \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2.
\end{aligned}$$

We can use this formula to show how variance interacts with linear operations:

Exercise 6.5.7

Show that variance satisfies the properties

$$\begin{cases} \operatorname{Var}(aX) = a^2 \operatorname{Var} X, & \text{for all random variables } X \text{ and real numbers } a \\ \operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y), & \text{if } X \text{ and } Y \text{ are independent random variables} \end{cases}$$

Exercise 6.5.8

Consider the distribution which assigns a probability mass of $\frac{c}{n^3}$ to each integer point $n \geq 1$, where c is equal to the reciprocal of $\sum_{n=1}^{\infty} \frac{1}{n^3}$.
Show that this distribution has a finite mean but not a finite variance.

6.5.4 Covariance

Just as mean and variance are summary statistics for the distribution of a single random variable, *covariance* is useful for summarizing how (X, Y) are jointly distributed.

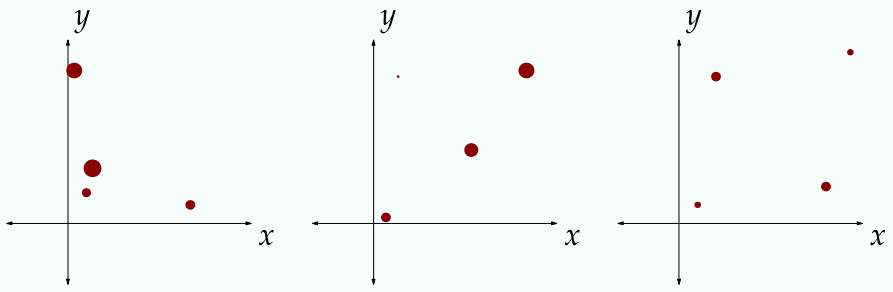
The **covariance** of two random variables X and Y is defined to be the expected product of their deviations from their respective means:

$$\operatorname{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

The covariance of two independent random variables is zero, because the expectation distributes across the product on the right-hand side in that case. Roughly speaking, X and Y tend to deviate from their means positively or negatively together, then their covariance is positive. If they tend to deviate oppositely (that is, X is above its mean and Y is below, or vice versa), then their covariance is negative.

Exercise 6.5.9

Identify each of the following joint distributions as representing positive covariance, zero covariance, or negative covariance. The size of a dot at (x, y) represents the probability that $X = x$ and $Y = y$.



Exercise 6.5.10

Does $\operatorname{Cov}(X, Y) = 0$ imply that X and Y are independent?

Hint: consider Exercise 6.5.9. Alternatively, consider a random variable X which is uniformly distributed on $\{1, 2, 3\}$ and an independent* random variable Z which is uniformly distributed on $\{-1, 1\}$. Set $Y = ZX$. Consider the pair (X, Y).

* This phrasing should be taken to mean that the pair (X, Z) is independent

Exercise 6.5.11

The **correlation** of two random variables X and Y is defined to be their covariance normalized by the product of their standard deviations:

$$\operatorname{Corr}(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sigma(X)\sigma(Y)}$$

In this problem, we will show that the correlation of two random variables is always between -1 and 1. Let $\mu_X = \mathbb{E}[X]$, and let $\mu_Y = \mathbb{E}[Y]$.

(i) Consider the following quadratic polynomial in t:

$$\mathbb{E}[((X - \mu_X) + (Y - \mu_Y)t)^2] = \mathbb{E}[(X - \mu_X)^2] + 2t\mathbb{E}[(X - \mu_X)(Y - \mu_Y)] + t^2\mathbb{E}[(Y - \mu_Y)^2]$$

where t is a variable.

(ii) Explain why this polynomial is nonnegative for all $t \in \mathbb{R}$.

(iii) Recall that a polynomial $at^2 + bt + c$ is nonnegative for all t if and only if the discriminant $b^2 - 4ac$ is nonpositive (this follows from the quadratic formula). Use this fact to show that

$$\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]^2 \leq \operatorname{Var} X \operatorname{Var} Y.$$

(iv) Conclude that $-1 \leq \operatorname{Corr}(X, Y) \leq 1$.

Exercise 6.5.12

Show that

$$\operatorname{Var}(X_1 + X_2 + \cdots + X_n) = \sum_{k=1}^{n} \operatorname{Var} X_k + 2 \sum_{(i,j)\,:\,1 \leq i < j \leq n} \operatorname{Cov}(X_i, X_j).$$

Exercise 6.5.13: Mean and variance of the sample mean

Suppose that $X_1, \ldots, X_n$ are independent random variables with the same distribution. Find the mean and variance of

$$\frac{X_1 + \cdots + X_n}{n}$$

Exercise 6.5.14

The **covariance matrix** of a vector $\mathbf{X} = [X_1, \ldots, X_n]$ of random variables defined on the same probability space is defined to be the matrix Σ whose (i, j)th entry is equal to $\mathrm{Cov}(X_i, X_j)$.

Show that $\Sigma = \mathbb{E}[\mathbf{X}\mathbf{X}']$ if all of the random variables $X_1, \ldots, X_n$ have mean zero. (Note: expectation operations on a matrix or vector of random variables entry-by-entry.)

6.6 Continuous distributions

Not every random phenomenon is ideally modeled using a discrete probability space. For example, we will see that the study of discrete distributions leads us to the *Gaussian distribution*, which smooths its probability mass out across the whole real number line, with most of the mass near the origin and less as you move out toward $-\infty$ or $+\infty$.

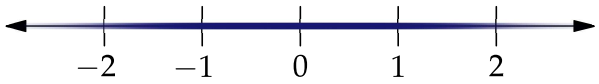


Figure 6.4 The *Gaussian* distribution spreads its probability mass out across the real number line. There is no single point where a positive amount of probability mass is concentrated.

We won't be able to work with such distributions using probability mass functions, since the function which maps each point to the amount of probability mass at that point is the zero function. However, calculus provides us with a smooth way of specifying where stuff is on the number line and how to total it up: **integration**. We can define a function f which is larger where there's more probability mass and smaller where there's less, and we can calculate probabilities by integrating f.

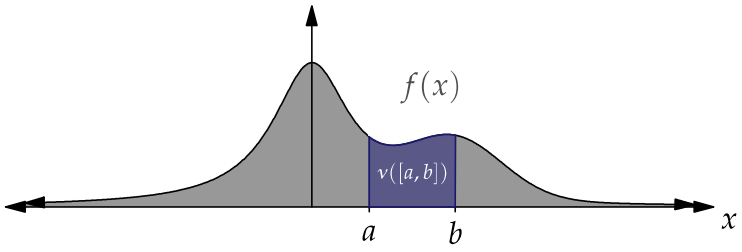


Figure 6.5 The probability measure ν associated with a density f assigns the measure $\int_a^b f(x)\,dx$ to each interval $[a, b]$

The simplest possible choice for f is the function which is 1 on $[0, 1]$ and 0 elsewhere. In this case, the probability mass associated with a set $\mathbb{E} \subset [0, 1]$ is the total length* of E. In higher dimensions, $\Omega = [0, 1]^2$ with the probability measure $\mathbb{P}(E) = \mathrm{area}(E)$ gives us a probability space, as does $\Omega = [0, 1]^3$ with the probability measure $\mathbb{P}(E) = \mathrm{volume}(E)$.

* All the subsets of $[0, 1]$ that we will ever care about have a well-defined notion of length. Some exotic sets are excluded by this caveat, but we can safely ignore them.

Exercise 6.6.1

Consider the probability space $\Omega = [0,1]^2$ with the area probability measure. Show that if $X((\omega_1,\omega_2)) = \omega_1$ and $Y((\omega_1,\omega_2)) = \omega_2$, then the events $\{X \in I\}$ and $\{Y \in J\}$ are independent for any intervals $I \subset [0,1]$ and $J \subset [0,1]$.

Just as a function we integrate to find total mass is called a mass density function, the function we integrate to find total probability is called a **probability density function**. We refer to f as a density because its value at a point may be interpreted as limit as $\epsilon \to 0$ of the probability mass in the ball of radius ϵ around ω divided by the volume (or area/length) of that ball.

Definition 6.6.1

Suppose that $\Omega \subset \mathbb{R}^n$ for some $n \geq 1$, and suppose that $f : \Omega \to [0,\infty)$ has the property that* $\int_\Omega f \mathrm{d}V = 1$. We call f a *probability density function*, abbreviated PDF, and we define

$$\mathbb{P}(E) = \int_E f \,\mathrm{d}V$$

for events $E \subset \Omega$. We call $(\Omega, \mathbb{P})$ a **continuous probability space**.

* We'll denote the volume differential in $\mathbb{R}^n$ by $\mathrm{d}V$, but note that if $n = 2$ it would be more standard to call it the area differential and write it as $\mathrm{d}A$ or $\mathrm{d}x\,\mathrm{d}y$, and if $n = 1$ it would just be $\mathrm{d}x$, the length differential.

Example 6.6.1

Consider the probability space with $\Omega = [0,1]$ and probability measure given by the density $f(x) = 2x$ for $x \in [0,1]$. Find $\mathbb{P}([\frac{1}{2},1])$.

If f is constant on Ω, then we call f the *uniform measure* on Ω. Note that this requires that Ω have finite volume.

All of the tools we developed for discrete probability spaces have analogues for continuous probability spaces. The main idea is to replace sums with integrals, and many of the definitions transfer over with no change. Let's briefly summarize and follow up with some exercises.

1. The distribution of a continuous random variable X is the measure $A \mapsto \mathbb{P}(X \in A)$ on $\mathbb{R}$.
2. The cumulative distribution function F_X of a continuous random variable X is defined by $F_X(x) = \mathbb{P}(X \leq x)$ for all $x \in \mathbb{R}$.
3. The joint distribution of two continuous random variables X and Y is the measure $A \mapsto \mathbb{P}((X,Y) \in A)$ on $\mathbb{R}^2$.
4. If (X,Y) is a continuous pair of random variables with joint density $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$, then the conditional distribution of X given the event $\{Y = y\}$ has density $f_{X \mid Y=y}$ defined by

$$f_{X \mid \{Y=y\}}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

   where $f_Y(y) = \displaystyle\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x$ is the pdf of Y
5. Two continuous random variables X and Y are independent if $\mathbb{P}((X,Y) \in A \times B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$ for all $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$. This is true if and only if (X,Y) has density $(x,y) \mapsto f_X(x) f_Y(y)$, where f_X and f_Y are the densities of X and Y, respectively.

6. The expectation of a continuous random variable X defined on a probability space $(\Omega, \mathbb{P})$

$$\mathbb{E}[X] = \int_\Omega X(\omega) f(\omega)\, \mathrm{d}\omega,$$

where f is $\mathbb{P}$'s density. The expectation is also given by

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x)\, \mathrm{d}x,$$

where f_X is the density of the distribution of X.

Example 6.6.2

Suppose that f is the function which returns 2 for any point in the triangle Ω with vertices $(0,0)$, $(1,0)$, and $(0,1)$ and otherwise returns 0. Suppose that (X,Y) has density f. Find the conditional density of X given $\{Y = y\}$, where y is a number between and 0 and 1.

Exercise 6.6.2

Find the expectation of a random variable whose density is $f(x) = \mathrm{e}^{-x}\mathbf{1}_{x\in[0,\infty)}$.

Exercise 6.6.3

Show that the cumulative distribution function of a continuous random variable is increasing and continuous.

(Note: if f is a nonnegative-valued function on $\mathbb{R}$ satisfying $\int_{\mathbb{R}} f = 1$, then $\lim_{\epsilon\to 0} \int_x^{x+\epsilon} f(t)\mathrm{d}t = 0$ for all $x \in R$.)

Exercise 6.6.4

Suppose that f is a density function on $\mathbb{R}$ and that F is the cumulative distribution function of the associated probability measure on $\mathbb{R}$. Show that F is differentiable and that $F' = f$ wherever f is continuous.

Use this result to show that if U is uniformly distributed on $[0,1]$, then U^2 has density function $f(x) = \frac{1}{2\sqrt{x}}$ on $(0,1]$.

Exercise 6.6.5

Given a cumulative distribution function F, let us define the **generalized inverse** $F^{-1} : [0,1] \to [-\infty, \infty]$ so that $F^{-1}(u)$ is the left endpoint of the interval of points which are mapped by F to a value which is greater than or equal to u.

The generalized inverse is like the inverse function of F, except that if the graph of F has a vertical jump somewhere, then all of the y values spanned by the jump get mapped by F^{-1} to the x-value of the jump, and if the graph of F is flat over a stretch of x-values, then the corresponding y-value gets mapped by F^{-1} back to the left endpoint of the interval of x values.

The remarkably useful **inverse CDF trick** gives us a way of sampling from any distribution whose CDF we can compute a generalized inverse for: it says that if U is uniformly distributed on $[0,1]$, then the cumulative distribution of $X = F^{-1}(U)$ is F.

(i) Confirm that if the graph of F has a jump from (x, y_1) to (x, y_2), then the probability of the event $\{X = x\}$ is indeed $y_2 - y_1$.

(ii) Show that the event $\{X \leq t\}$ has the same probability as the event $\{U \leq F(t)\}$. Conclude that F is in fact the CDF of X. Hint: draw a figure showing the graph of F together with U somewhere on the y-axis and X in the corresponding location on the x-axis.

(iii) Write a Julia function which samples from the distribution whose density function is $2x\mathbf{1}_{0\leq x\leq 1}$.

So far we have discussed probability spaces which are specified with the help of either a probability mass function or a probability density function. These are not the only possibilities. For example, if we produce an infinite sequence of independent bits $B_1, B_2, \ldots$, then the distribution of $B_1/3 + B_2/3^2 + B_3/3^3 + \cdots$ has CDF as shown in Figure 6.6. This function doesn't have jumps, so it does not arise from cumulatively summing a mass function. But it does all of its increasing on a set of total length zero (in other words, there is a set of total length 1 on which the derivative of this function is zero), so it also does not arise from cumulatively integrating a density function.

In general, a person may propose a probability space by specifying any set Ω, a collection of subsets of Ω which supports taking countable unions, intersections, and complements, and a function $\mathbb{P}$ defined on that collection of subsets. We require that certain properties are satisfied:

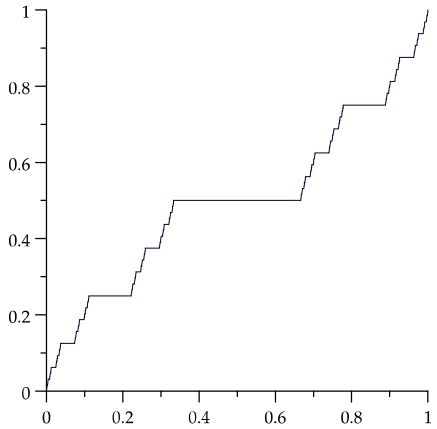


Figure 6.6 The CDF of the uniform measure on the Cantor set.

Definition 6.6.2: Probability space: the general definition

Suppose that Ω is a set and $\mathbb{P}$ is a function defined on a collection of subsets of Ω (called *events*). If

1. $\mathbb{P}(\Omega) = 1$,
2. $\mathbb{P}(E) \geq 0$ for all events E, and
3. $\mathbb{P}(E_1 \cup E_2 \cup \cdots) = \mathbb{P}(E_1) + \mathbb{P}(E_2) + \cdots$ for all sequences of pairwise disjoint events $E_1, E_2, \ldots$,

then we say that $\mathbb{P}$ is a probability measure on Ω, and that Ω together with the given collection of events and the measure $\mathbb{P}$ is a probability space.

6.7 Conditional expectation

The **conditional expectation** of X given $\{Y = y\}$ is defined to be the expectation of X calculated with respect to its conditional distribution given $\{Y = y\}$. For example, if X and Y are continuous random variables, then

$$E[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X \mid \{Y=y\}}(x)\, \mathrm{d}x.$$

Example 6.7.1

Suppose that f is the function which returns 2 for any point in the triangle with vertices $(0,0)$, $(1,0)$, and $(0,1)$ and otherwise returns 0. If (X,Y) has joint pdf f, then the conditional density of X given $\{Y = y\}$ is the mean of the uniform distribution on the segment $[y,1]$, which is $\frac{1+y}{2}$.

The **conditional variance** of X given $\{Y = y\}$ is defined to be the variance of X with respect to its conditional distribution of X given $\{Y = y\}$.

Example 6.7.2

Continuing with Example 6.7.1, the conditional density of X given $\{Y = y\}$ is the variance of the uniform distribution on the segment $[y,1]$, which is $\frac{(1-y)^2}{12}$.

We can regard the conditional expectation of X given Y as a random variable, denoted $\mathbb{E}[X \mid Y]$ by coming up with a formula for $\mathbb{E}[X \mid \{Y = y\}]$ and then substituting Y for y. And likewise for conditional variance.

Example 6.7.3

Continuing further with Example 6.7.1, we have $\mathbb{E}[X \mid Y] = \frac{1+Y}{2}$ and $\mathrm{Var}[X \mid Y] = \frac{(1-Y)^2}{12}$.

Exercise 6.7.1

Find the conditional expectation of X given Y where the pair (X,Y) has density $x + y$ on $[0,1]^2$.

Conditional expectation can be helpful for calculating expectations, because of the **tower law**.

Theorem 6.7.1: Tower law of conditional expectation

If X and Y are random variables defined on a probability space, then

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X].$$

Exercise 6.7.2

Consider a particle which splits into two particles with probability $p \in (0,1)$ at time $t = 1$. At time $t = 2$, each extant particle splits into two particles independently with probability p.

Find the expected number of particles extant just after time $t = 2$. Hint: define Y to be 1 or 0 depending on whether the particle splits at time $t = 1$, and use the tower law with Y.

6.8 Common distributions

6.8.1 Bernoulli distribution

Suppose we conduct an experiment with exactly two outcomes, which we will encode as 0 and 1. For example, consider the following scenarios

- You flip a coin and it comes up heads (1) or tails (0)
- Someone's position on a political issue is either positive (1) or negative (0)
- Someone can either be healthy (1) or sick (0)
- In an online survey, a user answers either true (1) or false (0)

The distribution of the result of such an experiment is governed by a single parameter $p \in [0,1]$, which is the probability of the outcome encoded as 1. The probability of the other outcome is $1 - p$, since one of the two outcomes must occur. It is customary to think of the outcomes 1 and 0 as success and failure, respectively, in which case p may be referred to as the success probability. A sequence of independent Bernoulli random variables with the same success probability p is referred to as a sequence of **Bernoulli trials**.

We write $X \sim \text{Bernoulli}(p)$ to mean that X is Bernoulli distributed with success probability p. The expected value of a $\text{Bernoulli}(p)$ random variable X is

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x \in \{1,0\}} x p_X(x) \\ &= (0)(1-p) + (1)(p) \\ &= p,\end{aligned}$$

and its variance is

$$\mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1-p).$$

Exercise 6.8.1

Consider a sum X of 10 independent Bernoulli random variables with success probability $p = 0.36$.

(i) Find the mean and variance of X.

(ii) Find the value of $k \in \mathbb{Z}$ which maximizes $\mathbb{P}(X = k)$.

Hint: write down an expression for $\mathbb{P}(X = k)$ and then use Julia to find its maximum value.

6.8.2 The binomial distribution

Example 6.8.1

What is the probability of rolling exactly 18 sixes in 100 independent rolls of a fair die?

Generally, n independent trials with success probability p will lead to k total successes with probability

$$\binom{n}{k} p^k (1-p)^{n-k}.$$

This distribution is called the **binomial distribution** and is denoted Binomial(n, p).

Exercise 6.8.2

Stirling's approximation allows us to more easily manipulate factorial expressions algebraically. It says that

$$\lim_{n\to\infty} \frac{n!}{(n/e)^n \sqrt{2\pi n}} = 1.$$

Suppose that n is even and that $p = \frac{1}{2}$. Use Stirling's approximation to show that $\sqrt{n}$ times the probability mass assigned to 0 by the distribution Binomial(n, p) converges to a finite, positive constant as $n \to \infty$. Find the value of this constant.

6.8.3 Geometric distribution

The **geometric distribution** with parameter $p \in (0, 1]$ is the distribution of the index of the first success in a sequence of independent Bernoulli trials.

The probability that the first success occurs on trial k is equal to the probability that the first $k-1$ trials fail and the kth trial succeeds. The probability of this event is $p(1-p)^{k-1}$. Therefore, the probability mass function of the geometric distribution is

$$m(k) = p(1-p)^{k-1}.$$

Exercise 6.8.3

Use Monte Carlo to find the mean and variance of the geometric distribution with parameter $p = 1/3$.

Hint: you can sample from the geometric distribution using the definition: count the number of times you have to run `rand()` until you get a result less than $\frac{1}{3}$.

We can use Taylor series to work out exact expressions for the mean and variance. The mean is equal to

$$p + 2p(1-p) + 3p(1-p)^2 + 4p(1-p)^3 + \cdots,$$

and we recognize all the terms except the first as $-p$ times the derivative of

$$(1-p)^2 + (1-p)^3 + (1-p)^4 + \cdots$$

By the formula for the sum of a geometric series, this expression is equal to

$$\frac{(1-p)^2}{1-(1-p)},$$

and so the mean of the geometric distribution is

$$p - p\frac{\mathrm{d}}{\mathrm{d}p}\left(\frac{(1-p)^2}{p}\right) = \frac{1}{p}.$$

The variance can be worked in a similar but more tedious way, and the result is

$$\operatorname{Var} X = \frac{1-p}{p^2}.$$

These expressions do indeed evaluate to 3 and 6, respectively, when $p = \frac{1}{3}$ is substituted.

Exercise 6.8.4

Suppose that X is geometric with success probability $\frac{1}{2}$, and consider the random variable $Y = 2^X$. What is the expected value of Y?

Exercise 6.8.5

Explain why `ceil(log(rand())/log(1-p))` returns a random variable whose distribution is geometric with success probability p.

Exercise 6.8.6

Every time you visit your favorite restaurant, you choose a meal uniformly at random from the 10 available meals. How many visits will it take on average before you've tried all 10 meals?

Hint: try letting X_k be the number of visits from the time you try the kth unique meal to the time when you try the $(k+1)$st unique meal.

6.8.4 Poisson distribution

The *Poisson distribution* arises as the number of 1's observed in a large number of low-probability Bernoulli random variables. This situation models a surprising variety of real-world scenarios:

1. The number of calls received at a call center in a given hour. Each potential caller has a low probability of calling during that particular hour, and there are many potential callers who are acting independently.
2. The number of meteorites which strike earth in a given year. There are many meteorites which might hit earth, and each one does so with low probability.
3. The number of mutations on a strand of DNA. Each mutation occurs with low probability, but there are many potential sites for a mutation.
4. The number of claims filed with an insurance company in a given month. There are many customers, and they file claims independently and with low probability each month.

Exercise 6.8.7

(i) Find the expected value of S, where S is a sum of 1000 independent Bernoulli random variables with success probability $p = \frac{3}{1000}$.

(ii) Find the probability mass function of S. Hint: find an expression representing the probability mass at each k from 0 to 1000, and then use Julia to evaluate it. You will need to define `n = big(1000)` and `p = big(3)/1000` because arbitrary precision arithmetic is required to avoid overflow issues.

(iii) Compare your results to the probability mass function $m(k) = \frac{3^k}{k!}e^{-3}$ defined on $\{0, 1, 2, \ldots\}$.

Inspired by Exercise 6.8.7, we make the following definition:

Definition 6.8.1: Poisson distribution

The Poisson distribution with mean λ is the distribution whose probability mass function is

$$m(k) = \frac{\lambda^k}{k!}e^{-\lambda}.$$

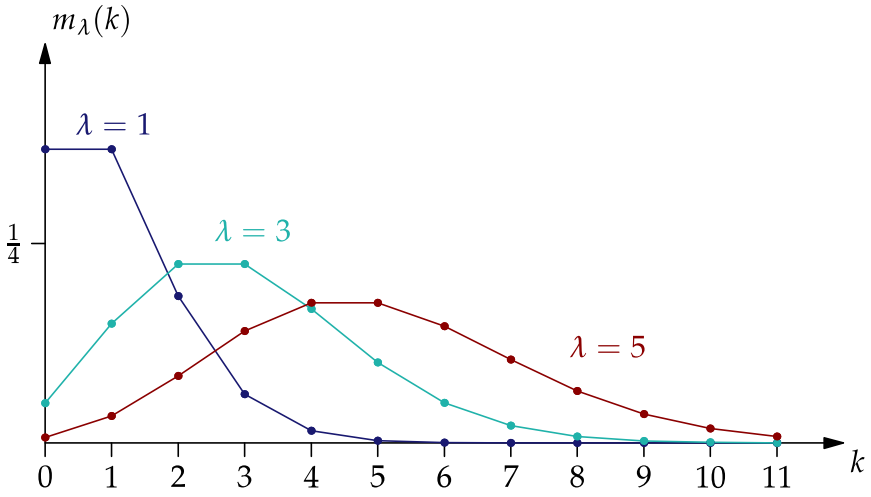


Figure 6.7 The probability mass function m_λ for $\lambda \in \{1, 3, 5\}$

The expression $\frac{\lambda^k}{k!}e^{-\lambda}$ in the definition of the Poisson distribution arises as a limit of the expression

$$\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}.$$

In other words, we use a success probability of $\frac{\lambda}{n}$ so that the expected number of successes remains constant as $n \to \infty$.

The connection between the Poisson and Bernoulli random variables may be used to obtain the mean and variance of the Poisson distribution. The average number of successes in n Bernoulli(λ/n) trials is $(n)(\lambda/n) = \lambda$, by linearity of expectation. Therefore, we expect that the mean of a Poisson random variable with param-

eter λ is equal to λ. Similarly, the variance of the number of successes in n Bernoulli(λ/n) trials is equal to $n\frac{\lambda}{n}\left(1-\frac{\lambda}{n}\right) = \lambda(1-\lambda/n)$. Taking $n \to \infty$, we predict that the variance of a Poisson random variable with parameter λ is also equal to λ. Both of these predictions are accurate:*

* These can be calculated directly from the probability mass function of the Poisson random variable, but we will omit this calculation. To make the derivation via Bernoulli random variables rigorous, we would need to show that mean and variance respect the $n \to \infty$ limit we are taking

Theorem 6.8.1

The mean and variance of a Poisson random variable with parameter λ are λ and λ, respectively.

Exercise 6.8.8

Suppose that the number of typos on a page is a Poisson random variable with mean $\lambda = \frac{1}{3}$.

(i) Provide an explanation for why the Poisson distribution might be a good approximation for the distribution of typos on a page.

(ii) Find the probability that a particular page is typo-free.

6.8.5 Exponential distribution

The exponential distribution also emerges as a limit involving Bernoulli random variables: imagine placing a light bulbs activated by independent Bernoulli(λ/n) random variables at every multiple of $1/n$ on the positive real number line. Consider the position X of the **leftmost lit bulb**. The probability that it occurs to the right of a point $t > 0$ is equal to the probability that all of the $\lfloor nt \rfloor$ bulbs to the left remain unlit:

$$\mathbb{P}(X > t) = \left(1 - \frac{\lambda}{n}\right)^{nt}$$

This probability converges to $e^{-\lambda t}$ as $n \to \infty$.

Definition 6.8.2: Exponential distribution

Let $\lambda > 0$. The exponential distribution with parameter λ is the probability measure on $\mathbb{R}$ which assigns mass $e^{-\lambda t}$ to the interval (t, ∞), for all $t \geq 0$.
Equivalently, the exponential distribution with parameter λ is the probability measure whose density is

$$f(x) = \mathbf{1}_{x \geq 0} \lambda e^{-\lambda x}$$

Exercise 6.8.9

Find the mean of the exponential distribution with parameter λ.

Exercise 6.8.10

Suppose that X is an exponentially distributed random variable with mean λ. Show that

$$\mathbb{P}(X > s + t \mid X > t) = \mathbb{P}(X > s).$$

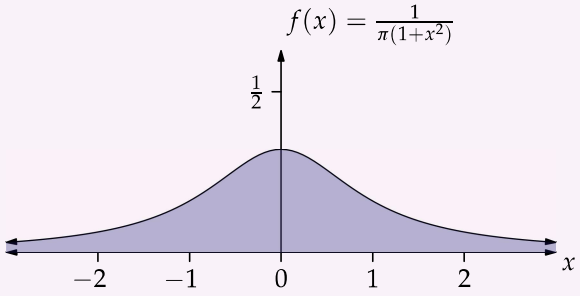
6.8.6 Cauchy distribution

The Cauchy distribution spreads probability mass *way* out on the real number line.

Definition 6.8.3: Cauchy distribution

The **Cauchy distribution** is the probability measure on $\mathbb{R}$ whose density function is

$$f(x) = \frac{1}{\pi} \frac{1}{1 + x^2}.$$



The amount of probability mass assigned by the Cauchy distribution to the interval (x, ∞) is

$$\int_x^\infty \frac{1}{\pi} \frac{1}{1 + t^2} \, dt = \frac{\pi}{2} - \arctan(x) \approx \frac{1}{x}.$$

This mass goes to 0 so slowly that the Cauchy distribution doesn't even have a well-defined mean, let alone a variance. We say that the Cauchy distribution is **heavy-tailed**, and we will use it as an example when we want to study the effects of heavy tails on results like the law of large numbers or the central limit theorem.

Exercise 6.8.11

Show that the mean of the Cauchy distribution is not well-defined.

Exercise 6.8.12

Choose θ uniformly at random from the interval $[0, \pi]$ and fire a ray from the origin at angle θ with respect to the positive x-axis. Let $(X, 1)$ be the point where this ray intersects the line $y = 1$. Show that X is Cauchy-distributed.

6.8.7 Normal distribution

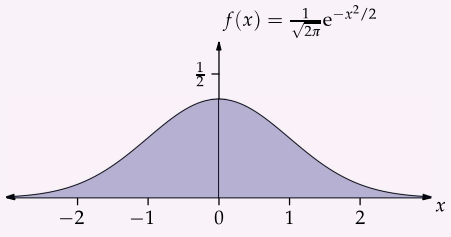
Because of the *central limit theorem*, which we will discuss in Section 6.10, the normal distribution plays a central role in probability and statistics.

Definition 6.8.4: Normal distribution

For $\mu \in \mathbb{R}$ and $\sigma \geq 0$, we define the **normal distribution**, denoted $\mathcal{N}(\mu, \sigma)$, to be the probability measure on $\mathbb{R}$ whose density function is

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

The **standard normal distribution** is $\mathcal{N}(0, 1)$.



Exercise 6.8.13

Show that if Z is a standard normal random variable and $\sigma > 0$, then the distribution of $\sigma Z + \mu$ is $\mathcal{N}(\mu, \sigma)$.

Example 6.8.2

In terms of the cumulative distribution function Φ of the standard normal, express the probability that a normally distributed random variable with mean 1 and variance 3 is between 2 and 4.

If we sum two independent random variables with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the mean and variance of the resulting sum are $\mu_1 + \mu_2$ and $\sigma_1^2 + \sigma_2^2$. Remarkably, if the random variables being summed are normal, then the sum is *also normal*:

Theorem 6.8.2

If X_1 and X_2 are independent normal random variables with distributions $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$, respectively, then the sum $X_1 + X_2$ has distribution $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Exercise 6.8.14

Suppose that $n \geq 1$ and that $X_1, X_2, \ldots, X_n$ are independent standard normal random variables. Find the distribution of $\frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}}$.

6.8.8 The multivariate normal distribution

If $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)$ is an independent sequence of standard normal random variables, A is an $m \times n$ matrix of constants, and μ is an $m \times 1$ vector of constants, then the vector

$$\boldsymbol{X} = A\boldsymbol{Z} + \mu$$

is said to have **multivariate normal distribution**.

If Σ is invertible, then the pdf of $\boldsymbol{X}$ is given by

$$f_X(\boldsymbol{x}) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu)^T\Sigma^{-1}(\boldsymbol{x}-\mu)\right).$$

Figure 6.8 shows a graph of this density as well as 1000 samples from the distribution of $A\boldsymbol{Z} + \mu$, where $A = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$ and $\mu = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$.

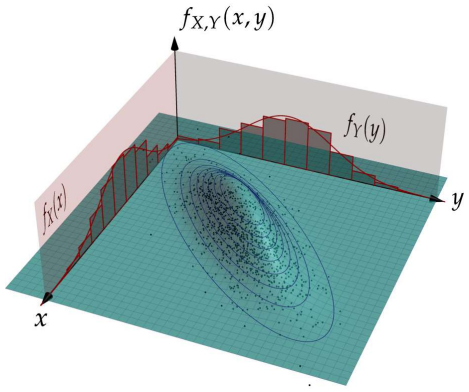


Figure 6.8 A graph of a multivariable normal density

Exercise 6.8.15

Show that the covariance matrix of a multivariate normal random vector $\boldsymbol{X} = A\boldsymbol{Z} + \mu$ is $\Sigma = AA^T$ and that its mean is μ.

Note: you may use the following properties: $\mathrm{Cov}(Y_1, Y_2 + c) = \mathrm{Cov}(Y_1, Y_2)$ for any constant c and any random variables Y_1 and Y_2, and if $\mathbf{Y}$ is an $n \times p$ random matrix and M is an $m \times n$ matrix of constants, then $\mathbb{E}[M\mathbf{Y}] = M\mathbb{E}[\mathbf{Y}]$.

6.9 Law of large numbers

6.9.1 Inequalities

In this section we begin working towards an understanding of the principle that underlies Monte Carlo methods: why is the mean of a random variable close to an average of many independent samples of the random variable? Is this always true, and can we say more precisely how close we should expect the sample average to be to the actual mean of the random variable?

We begin this investigation with a discussion of a two inequalities which relate probabilities and expectations. This can be helpful because sometimes expectations are much easier to calculate than probabilities, and these inequalities will help us use those expectation values to draw conclusions about probabilities.

A nonnegative random variable which has a high probability of being very large necessarily has a high expectation. Therefore, if a random variable's expectation isn't very large, we should be able to say that the random variable is unlikely to be very large. Markov's inequality is a concrete statement of this idea:

Theorem 6.9.1: Markov's inequality

If X is a nonnegative random variable and $a > 0$ is a real number, then

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Exercise 6.9.1

Let X denote the income of a person selected uniformly at random. With *no information whatsoever* about the income distribution, give an upper bound on the probability that X will be more than 3 times larger than the average income.

A nonnegative random variable of particular interest in the context of Monte Carlo techniques is the deviation of a random variable X from its mean: $|X - \mathbb{E}[X]|$. If we apply Markov's inequality to the square of this random variable, we obtain **Chebyshev's inequality**.

Theorem 6.9.2: Chebyshev's inequality

Suppose X is a random variable with finite variance. Then for any $b > 0$,

$$\mathbb{P}(|X - \mathbb{E}[X]| > b) \leq \frac{\mathrm{Var}[X]}{b^2}.$$

Substituting $b = k\sigma$ into Chebyshev's inequality, we obtain the following more memorable version: a finite-variance random variable is k standard deviations from its mean with probability at most $1/k^2$.

Example 6.9.1

The U.S. mint produces dimes with an average diameter of 0.5 inches and standard deviation 0.01. Using Chebyshev's inequality, give a lower bound for the number of coins in a lot of 400 coins that are expected to have a diameter between 0.48 and 0.52.

Exercise 6.9.2

Consider Poisson distributed random variable with mean $\lambda = 3$, and find the exact value of the probability that this random deviates by its mean by more than two standard deviations.

Compare this result to the bound obtained by using Chebyshev's inequality.

6.9.2 Convergence of random variables

A sequence of real numbers $x_1, x_2, x_3, \ldots$ converges to a real number x if the difference $|x_n - x|$ gets as small as desired for sufficiently large n. However, convergence of a sequence of *random variables* $X_1, X_2, \ldots$, to a random variable X is more nuanced, since we have to explain how the probability space is involved. The result of this subtlety is that there are several different notions of convergence.

6.9.2.1 The Borel-Cantelli lemma

We will begin with an useful theorem for handling convergence issues.

Theorem 6.9.3: The Borel-Cantelli lemma

Suppose that $E_1, E_2, \ldots$ is a sequence of events with the property that $\mathbb{P}(E_1) + \mathbb{P}(E_2) + \mathbb{P}(E_3) + \cdots < \infty$. Then the probability that at most finitely many of the events occur is 1.
Furthermore, if $\mathbb{P}(E_1) + \mathbb{P}(E_2) + \mathbb{P}(E_3) + \cdots = \infty$ **and** if the events $\{E_1, E_2, \ldots\}$ are independent, then the probability that infinitely many of the events occur is 1.

Example 6.9.2

Consider an infinite sequence $X_1, X_2, \ldots$ of independent Bernoulli random variables of success probability $p > 0$. Show that the set of ω's for which the sequence $X_1(\omega), X_2(\omega), \ldots$ converges has probability zero.

Exercise 6.9.3

Consider a sequence of independent events $E_1, E_2, \ldots$ for which $\mathbb{P}(E_n) = 1/n$.

Consider also a sequence of independent events $F_1, F_2, \ldots$ for which $\mathbb{P}(F_n) = 1/n^2$.

The Borel-Cantelli lemma implies that infinitely many of the E_i's occur, while only finitely many of the F_i's occur (both with probability 1).

Write some Julia code to simulate a run of the E_i's and a run of the F_i's (in each case, generate at least the first 10,000 samples from each sequence). Comment on whether your results seem consistent with the Borel-Cantelli lemma. Discuss some of the reasons why your simulation results are not definitive.

Exercise 6.9.4

Show that a monkey typing random keys on a typewriter will eventually type the sentence "the quick brown fox jumped over the lazy dog", with probability 1.

For simplicity, assume that the keystrokes are independent and that each keystroke is uniformly distributed among the keys on the keyboard.

Note: the given sentence is 44 characters long.

6.9.2.2 Almost sure convergence

Since the random variables $X_1, X_2, \ldots$, are functions defined on the probability space Ω, we may consider convergence of the numbers $X_1(\omega), X_2(\omega), \ldots$, for each $\omega \in \Omega$ individually.

Definition 6.9.1: Almost sure convergence

We say that $X_1, X_2, \ldots,$ converges almost surely to X if $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$ except on a set of ω's whose probability is zero.

Example 6.9.3

Consider $X_n(\omega) = \omega^n$ on the probability space $[0, 1]$ with measure given by $\mathbb{P}(E) = \text{length}(E)$. Show that X_n converges almost surely to the zero random variable.

6.9.2.3 Convergence in probability

A slightly weaker notion of convergence which is often easier to demonstrate than almost sure convergence is *convergence in probability*: we fix $\epsilon > 0$ and look at the probability that X_n deviates from X by more than ϵ. If this probability goes to zero as $n \to \infty$—no matter how small ϵ is—then we say that X_n converges to X in probability.

Definition 6.9.2: Convergence in probability

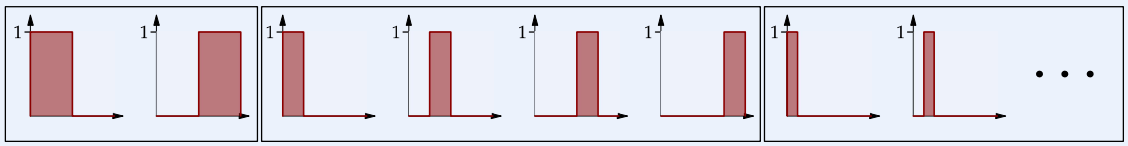
A sequence of random variables $X_1, X_2, \cdots, X_n, \cdots$ is said to **converge in probability** to a random variable X if for all $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X| > \epsilon) = 0$$

It is difficult to appreciate the distinction between convergence in probability and almost sure convergence without seeing an example of a sequence which converges in probability but not almost surely.

Example 6.9.4

Consider the sequence of random variables defined on $[0, 1]$ as follows:



X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8

In other words, the first two random variables are the indicator functions of the first half and the last half of the unit interval. The next four are the indicators of the first quarter, the second quarter, the third quarter, and the fourth quarter of the unit interval. The next eight are indicators of width-one-eighth intervals sweeping across the interval, and so on. For concreteness, suppose that each random variable is equal to 1 at any point ω where the random variable is discontinuous.

Show that this sequence of random variables converges in probability to the zero random variable, but does not converge almost surely.

Example 6.9.4 is admittedly a bit contrived, but it captures the basic idea: we have convergence in probability but not necessarily almost sure convergence if we can control the probability of misbehavior but cannot rule out the possibility that the small misbehaving portion of the probability space migrates forever all around the space.

This phenomenon can arise in more natural examples; for example, this one involving independent random variables:

Example 6.9.5

Consider an independent sequence of Bernoulli random variables X_n, where $\mathbb{P}(X_n = 1) = \frac{1}{n}$. Show that X_n converges to the zero random variable in probability.

There is a reason we have given examples of sequences of random variables which converge in probability but not almost surely, rather than the other way around: *almost sure convergence implies convergence in probability*.

Exercise 6.9.5

Consider a sequence of random variables $X_1, X_2, \ldots$ with the property that $|X_n| \leq 2^{-n}$ for all $n \geq 1$, and define

$$S_n = X_1 + X_2 + \ldots + X_n.$$

Show that S_n converges in probability.

Exercise 6.9.6

Suppose that $X_1, X_2, \ldots$ is a sequence of random variables with the property that $\mathbb{E}[|X_n|] \leq n^{-1/2}$ for all n. Show that X_n converges to 0 in probability.

6.9.2.4 Convergence in distribution

A sequence of random variables $X_1, X_2, \ldots$ converges to a random variable X in distribution if the distributions of $X_1, X_2, \ldots$ converge to the distribution of X. So, we should decide what it means for a sequence of probability measures to converge.

Roughly speaking, we will consider two probability measures close if they put approximately the same amount of probability mass in approximately the same places on the number line. For example, a sequence of continuous probability measures with densities $f_1, f_2, \ldots$ converges to a continuous probability measure with density f if $\lim_{n\to\infty} f_n(x) = f(x)$ for all $x \in \mathbb{R}$ (see Figure 6.9).

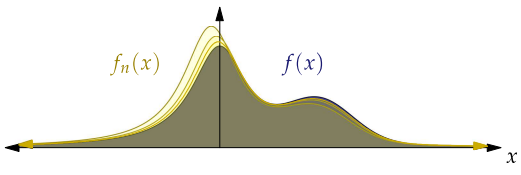


Figure 6.9 The sequence of densities f_n converges to the density f as $n \to \infty$

If the limiting probability measure is not continuous, then the situation is slightly more complicated. For example, we would like to say that the probability measure which puts a mass of $\frac{1}{2} + \frac{1}{n}$ at $\frac{1}{n}$ and a mass of $\frac{1}{2} - \frac{1}{n}$ at $1 + \frac{1}{n}$ converges to Bernoulli $\left(\frac{1}{2}\right)$ as $n \to \infty$. This does not correspond to pointwise convergence of the probability mass functions, since we don't have convergence of probability mass function values at 0 or at 1 in this example.

We can get around this problem by giving ourselves a little space to the left and right of any point where the limiting measure has a positive probability mass. In other words, suppose that ν is a probability measure on $\mathbb{R}$ with probability mass function m, and consider an intervals $I = (a, b)$. Let's call such an interval a *continuity interval* of ν if $m(a)$ and $m(b)$ are both zero.

We will say that a sequence of probability measures $\nu_1, \nu_2, \ldots$ converges to ν if $\nu_n(I)$ converges to $\nu(I)$ for every continuity interval I of ν.

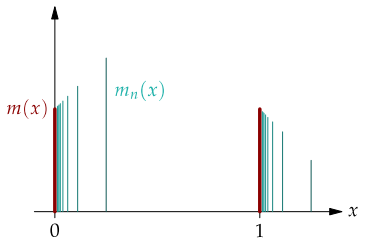


Figure 6.10 The probability measures which assign mass $\frac{1}{2} + \frac{1}{n}$ and $\frac{1}{2} - \frac{1}{n}$ to $\frac{1}{n}$ and $1 + \frac{1}{n}$, respectively, (shown in sea green) converge to the Bernoulli distribution with success probability $\frac{1}{2}$ (shown in red).

We can combine the discrete and continuous definitions into a single definition:

Definition 6.9.3: Convergence of probability measures on $\mathbb{R}$

A sequence $\nu_1, \nu_2, \ldots$ of probability measures on $\mathbb{R}$ converges* to a probability measure ν on $\mathbb{R}$ if $\nu_n(I) \to \nu(I)$ whenever I is an interval satisfying $\nu(\{a, b\}) = 0$, where a and b are the endpoints of I.

* In contexts where more than one notion of probability measure convergence is being used, this would be called *weak* convergence

Exercise 6.9.7

Define $f_n(x) = n\mathbf{1}_{0 \le x \le 1/n}$, and let ν_n be the probability measure with density f_n. Show that ν_n converges to the probability measure ν which puts of all its mass at the origin.

Exercise 6.9.8

Suppose that $X_1, X_2, \ldots$ are independent Exponential(1) random variables and that X is an Exponential(1) random variable defined on the same probability space.

Show that X_n converges to X in distribution, but that X_n does not converge to X almost surely.

6.9.3 Weak law of large numbers

We are now ready to state and prove the weak* law of large numbers. Roughly speaking, it says that the average of n independent, identically distributed finite-variance random variables is very likely to be very close to the mean of the common distribution of the random variables, if n is suitably large.

* *Weak* here is in contrast to the *strong* law of large numbers, which upgrades the mode of convergence to almost-sure

Theorem 6.9.4: Weak law of large numbers

Let ν be a finite-variance distribution on $\mathbb{R}$, and let $X_1, X_2, \ldots$, be an independent sequence of ν-distributed random variables. For $n \geq 1$, define $M_n = \frac{X_1+X_2+\cdots+X_n}{n}$. The weak law of large number states that M_n converges weakly to the mean of ν (regarded as a constant random variable).
In other words, for any $\epsilon > 0$, we have

$$\mathbb{P}\left(\frac{X_1 + \cdots + X_n}{n} \notin [\mu - \epsilon, \mu + \epsilon]\right) \to 0,$$

as $n \to \infty$.

Proof

From Exercise 6.5.13, we know that the mean and variance of M_n are μ and σ^2/n, where σ^2 is the variance of ν.

Therefore, Chebyshev's inequality implies that

$$\mathbb{P}(|M_n - \mu| \geq \varepsilon) \leq \varepsilon^{-2}\sigma^2/n.$$

Since $\varepsilon^{-2}\sigma^2/n$. as $n \to \infty$, we conclude that $\mathbb{P}(|M_n - \mu| \geq \varepsilon)$ converges to 0 as $n \to \infty$.

Example 6.9.6

How many times do we need* to flip a coin which has probability 60% of turning up heads in order to ensure that the proportion of heads is between 59% and 61% with probability at least 99%?

* We'll interpret this question to mean "find a number of flips that works" rather than "find the minimum number of flips that would work".

Note that using Chebyshev's inequality here does not give us a complete answer to our question. We know that 5.76×10^{10} flips are enough, but in fact many fewer flips would work.

Example 6.9.7

Write some code to determine the minimum number of flips required in Example 6.9.6.

Hint: Use the special function `lgamma` to calculate the expression $\binom{n}{k}p^k(1-p)^{n-k}$ without having to resort to arbitrary-precision arithmetic. This function returns the natural logarithm of the function $\Gamma(n) = (n+1)!$.

Thus we see that Chebyshev's inequality can be *quite* pessimistic. In the next section, we will develop a much sharper tool.

Exercise 6.9.9

The finite-variance assumption is necessary for the weak law of large numbers to hold. Repeat the following experiment 100 times: sample from the Cauchy distribution 100,000 times and calculate the sample mean of these samples. Make a histogram of the 100 resulting means. Are these means tightly concentrated?

Note: you can sample from a Cauchy distribution using `tan(π*(rand()-1/2))`.

6.10 Central limit theorem

The law of large numbers tells us that the distribution ν of a mean of many independent, identically distributed finite-variance, mean-μ random variables is concentrated around μ. The **central limit theorem** gives us precise information about *how* the probability mass of ν is concentrated.

Consider a sequence of independent fair coin flips $X_1, X_2, \ldots$, and define the sums

$$S_n = X_1 + \cdots + X_n,$$

for $n \geq 1$. The probability mass functions of the S_n's can be calculated exactly, and they are graphed in Figure 6.11 for several values of n. We see that the graph is becoming increasingly Gaussian-shaped as n increases.

If we repeat this exercise with other distributions in place of Bernoulli$(1/2)$, we obtain similar results. For example, the probability mass functions for sums of the independent Poisson(3) random variables is shown in Figure 6.12. Not only is the shape of the graph stabilizing as n increases, but we're getting the *same* shape as in the Bernoulli example.

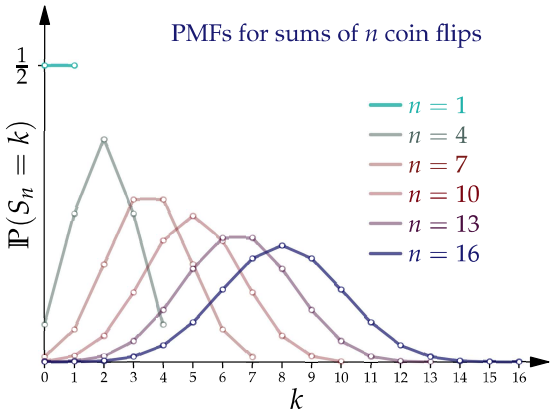


Figure 6.11 Probability mass functions of sums of Bernoulli(1/2) random variables.

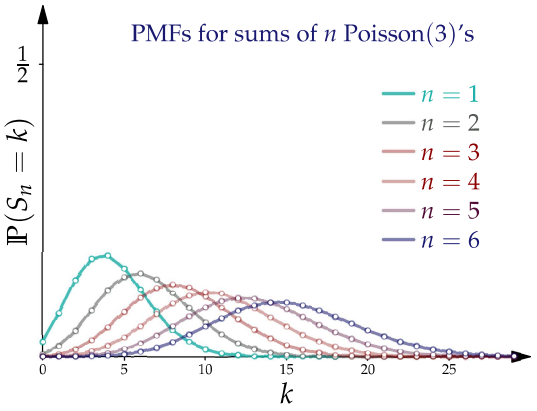


Figure 6.12 Probability mass functions of sums of Poisson(3) random variables.

To account for the shifting and spreading of the distribution of S_n, we *normalize* it: we subtract its mean and then divide by its standard deviation to obtain a random variable with mean zero and variance 1 (see Exercise 6.5.13:

$$S_n \xrightarrow{\text{shift}} S_n - n\mu \xrightarrow{\text{scale}} \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Let's define $S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}}$, which has mean 0 and variance 1. Based on Figures 6.11 and 6.12, we conjecture that the distribution of S_n^* converges as $n \to \infty$ to the distribution $\mathcal{N}(0,1)$. Indeed, that is the case:

Theorem 6.10.1: Central Limit theorem

Suppose that $X_1, X_2, \ldots,$ are independent, identically distributed random variables with mean μ and finite standard deviation σ, and defined the normalized sums $S_n^* = (X_1 + \cdots + X_n - n\mu)/(\sigma\sqrt{n})$ for $n \geq 1$.
For all $-\infty \leq a < b \leq \infty$, we have

$$\lim_{n\to\infty} P(a < S_n^* < b) = P(a < Z < b),$$

where $Z \sim \mathcal{N}(0,1)$. In other words, the sequence $S_1^*, S_2^*, \ldots$ converges in distribution to Z.

The **normal approximation** is the technique of approximating the distribution of S_n^* as $\mathcal{N}(0,1)$.

Example 6.10.1

Suppose we flip a coin which has probability 60% of turning up heads n times. Use the normal approximation to estimate the value of n such that the proportion of heads is between 59% and 61% with probability approximately 99%.

Example 6.10.2

Consider a sum of n independent Bernoulli random variables with $p = 1/2$, and let $m_n : \mathbb{R} \to [0,1]$ be the pmf of S_n^*. Show that $\lim_{n\to\infty} m_n(x) = 0$ for all $x \in \mathbb{R}$, and explain why this does not contradict the central limit theorem.

For simplicity, you may assume that n is even.

Exercise 6.10.1

Suppose that the percentage of residents in favor of a particular policy is 64%. We sample n individuals uniformly at random from the population.

(i) In terms of n, find a interval I centered at 0.64 such that the proportion of residents polled who are in favor of the policy is in I with probability about 95%.

(ii) How many residents must be polled for the proportion of poll participants who are in favor of the policy to be between 62% and 66% with probability at least 95%?

Exercise 6.10.2

Suppose that $X_1, X_2, \ldots$ is a sequence of independent, identically distributed random variables with variance 2 and mean 7. Find the limits of each of the following probabilities $n \to \infty$.

(i) $\mathbb{P}(X_1 + \cdots + X_n = 7n)$

(ii) $\mathbb{P}(6.9n < X_1 + \cdots + X_n < 7.1n)$

(iii) $\mathbb{P}(7n < X_1 + \cdots + X_n < 7n + 3\sqrt{n})$

(iv) $\mathbb{P}(6n < X_1 + \cdots + X_n < 7n + 3\sqrt{n})$