Probability Notes

Fall 2017

Sinho Chewi

Contents

Chapter 0

Combinatorics

Before we dive into the fascinating subject of probability theory, we first need to know how to count, since counting will be used to solve a variety of probability questions. Moreover, the study of **combinatorics** is interesting in its own right because the methods developed to count various objects are often creative and fun!

0.1 Basic Rules of Counting

The goal of this chapter is to build up a useful bag of tricks which will eventually allow us to count a wide range of scenarios. First, we begin with the **Multiplication Rule**. Suppose you have to make a two-stage decision: first, you must pick one choice out of m total choices (where m is a positive integer); then, regardless of your first choice, you must pick one choice out of n total choices (where n is a positive integer). For example, you might first choose one of your m hats, and then one of your n scarves. How many total possible choices do you have?

If you choose your first hat, then you have n choices for your scarf. If you choose your second hat, then again you have n choices for your scarf. In fact, based on the way that the problem is stated, *regardless of your choice of hat, you will always have n choices for your scarf.* So, the total number of choices is $\sum_{i=1}^{m}(\text{number of choices when you choose hat } i) = \sum_{i=1}^{m} n = mn$. See Figure 1 for a visualization.

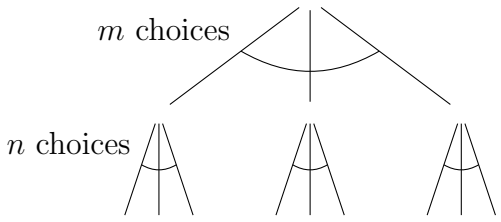


Figure 1: Multiplication Rule. When there are m choices at the first stage and n choices at the second stage, there are a total of mn possibilities.

The multiplication rule extends easily to the case when there are more than two stages. If there are k stages, and on each stage i, there are n_i possible choices (where k and n_i are

positive integers, $i \in \{1, \ldots, k\}$), then the total number of possibilities is $\prod_{i=1}^{k} n_i$.

Example 0.1 (Cardinality of the Power Set)**.** Let X be a non-empty finite set. How many elements are in the power set of X? In other words, what is the number of distinct subsets of X?

To count the number of subsets of X, we visualize the process of "building" a subset $X' \subseteq X$. We start with $X' = \varnothing$. For each element $x \in X$, we must choose whether or not we want to put x in our subset X'. Hence, we can view the construction of X' as a decision process with $|X|$ stages, one stage for each $x \in X$, where each stage has two possible choices: put x in X', or exclude x from X'. So, by the multiplication rule, the total number of subsets $X' \subseteq X$ is $\prod_{i=1}^{|X|} 2 = 2^{|X|}$.

Example 0.2 (Bit Strings)**.** Let $n \in \mathbb{Z}_+$. What is the number of length-n bit strings?

In fact, we have already solved this problem in Example 0.1! If X is a set with $|X| = n$, then we can think of each subset $X' \subseteq X$ as a length-n bit string. To see this, first we arbitrarily order the elements of X as $x_1, \ldots, x_n$. Then, the ith character of the bit string associated with X' is 1 if and only if $x_i \in X'$. Hence, the number of length-n bit strings is the same as the number of subsets of X, which is 2^n.

It is worth thinking a little more about what we did in Example 0.2. Essentially, we gave a bijection $\{\text{all length-}n \text{ bit strings}\} \leftrightarrow \{X' : X' \subseteq X\}$. This illustrates an important principle in combinatorics:

> **Bijection Principle**: If there is a bijection between two sets, the two sets have the same number of elements.

Although the above sentence may seem like it does not contain much content, we will see that it is actually extremely useful. Sometimes, by thinking about a set in a different way, it becomes much easier to count.

0.1.1 Distinguishable Balls, Distinguishable Bins

Here, we introduce an analogy that we will use repeatedly: balls and bins. Let $m, n \in \mathbb{Z}_+$. The scenario is that we are throwing n balls into m bins, and we would like to count the number of ways this can happen. First, we will assume that all of the balls and bins are distinguishable (they are numbered, for example, so we can tell them apart). We can think of the process of throwing the n balls as a n-stage decision process; at each stage, each ball has m choices for its destination bin. By the multiplication rule, there are m^n total possibilities.

> There are m^n ways to throw n distinguishable balls into m distinguishable bins.

Notice that this example actually includes the length-n bit strings (Example 0.2) as a special case: we can think of the length-n bit strings as throwing n balls into two bins, where the bins represent the bits 0 and 1.

0.1.2 Permutations

From the multiplication rule, we can now count the number of permutations of a finite set. Let $n \in \mathbb{Z}_+$. A **permutation** of $\{1, \ldots, n\}$ is a way of rearranging the numbers $(1, \ldots, n)$. Formally, a permutation σ is a bijection $\sigma : \{1, \ldots, n\} \leftrightarrow \{1, \ldots, n\}$. To count the number of permutations, we can again think of building a permutation via a n-stage decision process, where at the ith step (for $i \in \{1, \ldots, n\}$) we are deciding on the value of $\sigma(i)$.

We start with $\sigma(1)$, which is allowed to be anything in $\{1, \ldots, n\}$, so there are n choices here. Next is $\sigma(2)$, which is allowed to be anything in $\{1, \ldots, n\}$ *except* $\sigma(1)$ (since σ must be one-to-one), so the number of choices for $\sigma(2)$ is $n - 1$. Similarly, at the ith step, $\sigma(i)$ is allowed to be anything in $\{1, \ldots, n\}$ except the $i - 1$ values we have already chosen, so there are $n - (i - 1)$ choices for $\sigma(i)$. Putting it together, the total number of possibilities is $\prod_{i=1}^{n}(n-(i-1)) = \prod_{i=1}^{n} i = 1 \cdot 2 \cdots (n-1) \cdot n$. This quantity appears a lot in combinatorics, so we give it a special symbol. We define $n! := \prod_{i=1}^{n} i$ (read "n **factorial**") for $n \in \mathbb{Z}_+$, and we define $0! = 1$ by convention. [1]

The number of permutations of $\{1, \ldots, n\}$ is $n!$.

Another way to think about permutations is that we are counting the number of *ordered* subsets of size n from $\{1, \ldots, n\}$. Extending this, we can try to count the number of ordered subsets of size k from $\{1, \ldots, n\}$, where $k \in \{1, \ldots, n\}$. We can follow the same procedure as before, except that instead of n stages, we only have k stages (we stop once we have selected k elements for our ordered subset). Hence:

The number of ordered subsets of size k from $\{1, \ldots, n\}$ is

$$\prod_{i=1}^{k}\big(n - (i-1)\big) = \frac{n!}{(n-k)!}.$$

0.1.3 Combinations

We now proceed to count the number of *unordered* subsets of size k from $\{1, \ldots, n\}$. We already know that the number of ordered subsets of size k is $n!/(n-k)!$; the key here is to observe that each *unordered* subset gives rise to exactly $k!$ *ordered* subsets. To understand this, observe that the number of orderings of a set of size k is the same as the number of ways to permute the k elements, which we found to be $k!$. So, the number of ordered subsets must be $k!$ times as numerous as the number of unordered subsets, which gives us a formula to count the latter: $n!/(k!(n-k)!)$. This is another quantity which appears so frequently that it has another name: it is denoted $\binom{n}{k}$ and it is called the **binomial coefficient** (we will see the reason for this name later). Observe that $\binom{n}{k} = \binom{n}{n-k}$.

The number of unordered subsets of size k from $\{1, \ldots, n\}$ is $\binom{n}{k} = \binom{n}{n-k}$.

[1] Conventions like these are purely for convenience when stating theorems.

Exercise 1 For a positive integer n, suppose that n schools send their badminton teams to a tournament. In the tournament, each school must play every other school exactly once. How many games are played in total?

Example 0.3 (Bit Strings with a Fixed Number of Ones)**.** Suppose $n \in \mathbb{Z}_+$ and $m \in \{0, \ldots, n\}$. What is the number of length-n bit strings with exactly m ones?

Observe that once we specify the locations of the m ones, the remaining positions must be 0, and so we have specified the entire bit string. Therefore, the number of such bit strings is exactly the number of ways to choose m locations out of n total locations (since a length-n string has n places where we can write a character). This, in turn, is equal to the number of unordered subsets of size m from a set of size n (we do not care about the order because we cannot tell the 1s apart from each other), which is $\binom{n}{m} = \binom{n}{n-m}$.

0.1.4 Indistinguishable Balls, Distinguishable Bins: Stars & Bars

Again, we return to the scenario of throwing balls into bins, only now we assume that the balls are all indistinguishable. This will change the number of total configurations: some configurations which were previously distinct (because we could tell the balls apart) will no longer be distinct, and so we expect the total number of configurations to decrease.

To count the number of configurations, we can use a very clever trick known as **Stars & Bars**. The idea is to construct a bijection between the set of all configurations of n indistinguishable balls in m distinguishable bins with the set of length-$(n + m - 1)$ strings from the alphabet $\{\star, |\}$ (this is the reason behind the name "stars and bars"). We will represent the n balls using n star symbols. It is natural to then think that the m bins will be represented using m bar symbols, but this is not the case; instead, we will represent the $m-1$ *spaces between the bins* using $m-1$ bar symbols. A picture is worth a thousand words, so here is what the bijection looks like for $m = n = 3$.

balls and bins configuration	stars and bars string
3 balls in bin 1	$\star\star\star \,\|\|$
2 balls in bin 1, 1 ball in bin 2	$\star\star \,\|\, \star \,\|$
2 balls in bin 1, 1 ball in bin 3	$\star\star \,\|\|\, \star$
1 ball in bin 1, 2 balls in bin 2	$\star \,\|\, \star\star \,\|$
1 ball in bin 1, 1 ball in bin 2, 1 ball in bin 3	$\star \,\|\, \star \,\|\, \star$
1 ball in bin 1, 2 balls in bin 3	$\star \,\|\|\, \star\star$
3 balls in bin 2	$\|\, \star\star\star \,\|$
2 balls in bin 2, 1 ball in bin 3	$\|\, \star\star \,\|\, \star$
1 ball in bin 2, 2 balls in bin 3	$\|\, \star \,\|\, \star\star$
3 balls in bin 3	$\|\|\, \star\star\star$

It may seem strange at first, but we really do need only $m - 1$ bar symbols, because $m - 1$ bars create m different slots in which the stars can be placed! Given any balls and bins configuration, we can map it to a length-$(n + m - 1)$ string with exactly n stars and $m - 1$

bars. Conversely, any length-$(n+m-1)$ string with n stars and $m-1$ bars corresponds to a configuration with n balls and m bins. You should convince yourself of these facts.

It remains to count the number of length-$(n+m-1)$ bit strings with exactly n stars and $m-1$ bars; but by Example 0.3, this is $\binom{n+m-1}{n} = \binom{n+m-1}{m-1}$.

> There are $\binom{n+m-1}{n} = \binom{n+m-1}{m-1}$ ways to throw n indistinguishable balls into m distinguishable bins.

Warning: The counting techniques in the previous sections will all find their uses in probability theory. However, stars and bars is almost *never* useful in probability questions. Keep this in mind when we study probability theory.

0.2 Combinatorial Proofs

We have already seen that the number of length-n bit strings is 2^n (Example 0.2). However, there is a different way to count the number of length-n bit strings. The number of length-n bit strings with exactly m ones is $\binom{n}{m}$ (Example 0.3). Therefore, if we sum over all possible values of m, then we should also be counting the total number of length-n bit strings. This actually provides a proof of the following formula: $\sum_{m=0}^{n} \binom{n}{m} = 2^n$.

This is known as a **combinatorial proof**, because we have proven two quantities to be equal by showing that both quantities count the same objects. Combinatorial proofs are rather magical: often, proving a combinatorial identity via algebra can be very difficult, but giving the quantities on both sides of the claimed identity a combinatorial interpretation proves the identity rather easily!

The above example is a special case of the next theorem, which is quite important. We give a combinatorial proof.

Theorem 0.4 (Binomial Theorem)**.** *Let $n \in \mathbb{Z}_+$. Then, $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$.*

Proof. We can write $(x+y)^n$ as the product of n identical factors $(x+y)$. Any term in the expansion of $(x+y)^n$ is obtained by choosing exactly one of the two terms, x or y, from each of the n factors $(x+y)$. In order to obtain the term $x^k y^{n-k}$, we must choose x from exactly k of the $(x+y)$ factors, and we must choose y from the remaining $n-k$ factors. So, the coefficient of $x^k y^{n-k}$ in the expansion of $(x+y)^n$ is the number of ways to choose x from exactly k of the n factors; but this is precisely $\binom{n}{k}$. □

The binomial theorem (Theorem 0.4) explains the name "binomial coefficient": the numbers $\binom{n}{k}$ are exactly the coefficients that appear in the binomial theorem.

Exercise 2 **Pascal's Identity** Let $k \leq n$ be positive integers. Prove $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$, both algebraically and via a combinatorial proof. Note that this identity gives a method for computing the binomial coefficients, see Figure 2.

$$\begin{array}{ccccccccccccccccc}
 & & & & & & 1 & & & & & & & & 1 & & \\
 & & & & \binom{1}{0} & & & & \binom{1}{1} & & & & & & 1 & & 1 \\
 & & & \binom{2}{0} & & & \binom{2}{1} & & & & \binom{2}{2} & & & = & & 1 & & 2 & & 1 \\
 & & \binom{3}{0} & & & \binom{3}{1} & & & \binom{3}{2} & & & \binom{3}{3} & & & 1 & & 3 & & 3 & & 1 \\
 & \binom{4}{0} & & & \binom{4}{1} & & & \binom{4}{2} & & & \binom{4}{3} & & & \binom{4}{4} & 1 & & 4 & & 6 & & 4 & & 1
\end{array}$$

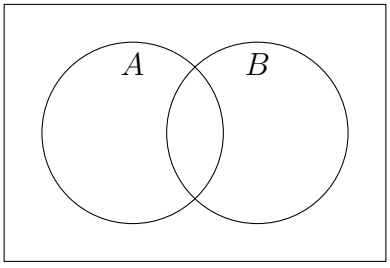
Figure 2: Pictured are the first five rows of Pascal's Triangle. Each entry is computed by summing the two entries above it.

Exercise 3 Use Pascal's identity (Exercise 2) to prove the binomial theorem via induction.

0.3 Inclusion-Exclusion Principle

Let p and q be distinct prime numbers. How many numbers in the set $\{1, \ldots, pq\}$ are relatively prime to both p and q? To tackle this question, we might first observe that it is probably easier to think about the numbers which are *not* relatively prime to p. Since p is prime, the numbers that are not relatively prime to p are precisely the multiples of p, $p, 2p, \ldots, (q-1)p, qp$. Similarly, the numbers that are not relatively prime to q are the multiples of q: $q, 2q, \ldots, (p-1)q, pq$. Since there are q multiples of p and p multiples of q, we are tempted to say that there are $p+q$ numbers which are not relatively prime to p or q. However, we have counted the number pq twice since it is a multiple of both p and q, so we need to subtract one to correct for the overcounting. Thus, there are $p+q-1$ numbers which are not relatively prime to p or q, so there are $pq-(p+q-1) = (p-1)(q-1)$ numbers which are relatively prime to p and q.

The example above illustrates a general principle, known as the **Inclusion-Exclusion Principle**. Suppose we have two finite sets A and B, and we wish to count the number of elements in their union. Now look at the following picture:



If we take $|A| + |B|$, the sum of the numbers of elements in each set individually, then we will double-count the elements which are common to both sets. Therefore, if we subtract the

number of elements in the intersection, we obtain the formula $|A \cup B| = |A| + |B| - |A \cap B|$. This formula is frequently useful because it is often easier to compute the number of items in an intersection rather than a union.

We would now like to state and prove the general formula. However, we will see another version of the inclusion-exclusion principle when we study probability theory, so to avoid having to prove the same idea twice, we will state the principle generally enough to apply to both counting and probability.

Let X be a set and let $\mathcal{X}$ be a collection of subsets of X. Let $f : \mathcal{X} \to [0, \infty)$ be a function which assigns a non-negative number to each set $X' \in \mathcal{X}$. In the context of counting, think of X as a finite set, $\mathcal{X}$ as the power set of X, and f is the map which assigns each set $X' \subseteq X$ its cardinality: $f : X' \mapsto |X'|$. We say that f is **finitely additive** if for every positive integer n and any *disjoint* sets $X_1, \ldots, X_n \in \mathcal{X}$, we have $f(X_1 \cup \cdots \cup X_n) = \sum_{i=1}^{n} f(X_i)$. Note that the cardinality map is indeed finitely additive, because if the different sets are disjoint, then the number of elements in the union of the sets is the sum of the numbers of elements in each set individually. Finally, we will use the notation $(X')^{\mathsf{c}}$ for the **complement** of X' in X, that is, $(X')^{\mathsf{c}} := X \setminus X'$.

Theorem 0.5 (Generalized Inclusion-Exclusion Principle)**.** *Let X be a set and $\mathcal{X}$ be a collection of subsets of X, and let $f : \mathcal{X} \to [0, \infty)$ be a finitely additive function. Let $X_1, \ldots, X_n \in \mathcal{X}$ be sets, not necessarily disjoint. Then:*

$$f(X_1 \cup \cdots \cup X_n) = \sum_{k=1}^{n} \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ |I|=k}} (-1)^{k+1} f\Big(\bigcap_{i \in I} X_i\Big). \tag{1}$$

The inner summation ranges over all k-tuples of indices $(i_1, \ldots, i_k)$ satisfying the condition $1 \leq i_1 < \cdots < i_k \leq n$. In words, the generalized inclusion-exclusion principle prescribes the following rule for calculating $|X_1 \cup \cdots \cup X_n|$: First, sum the number of elements in each X_i. Next, subtract the number of elements in each pairwise intersection of sets, $|X_i \cap X_j|$. Then, add the number of elements in each 3-wise intersections of sets, $|X_i \cap X_j \cap X_k|$, and continue with the same pattern (with alternating signs).

Proof. First, study the diagram for the case of $n = 2$, which reveals why the inclusion-exclusion principle is necessary. To avoid dealing with overlapping sets, we can write $X_1 \cup X_2$ as the *disjoint* union of three sets $X_1 \cap X_2^{\mathsf{c}}$, $X_1 \cap X_2$, and $X_1^{\mathsf{c}} \cap X_2$. We can write these sets as length-2 bit strings: the first digit is 1 if X_1 appears in the intersection, and the second digit is 1 if X_2 appears in the intersection. Then, the three disjoint sets correspond to the strings 10, 11, and 01, respectively. The three strings above are all length-2 bit strings which contain at least one 1 (the string 00 corresponds to $X_1^{\mathsf{c}} \cap X_2^{\mathsf{c}}$, which is not contained in $X_1 \cup X_2$).

Generalizing, $X_1 \cup \cdots \cup X_n$ can be written as the disjoint union of $2^n - 1$ sets, where

each set is encoded as a length-n bit string. For each $i \in \{1, \ldots, n\}$, the ith digit is 1 if and only if A_i appears in the intersection, and in total we have every length-n bit string except the string of all 0s.

Consider a set B corresponding to a bit string with exactly m 1s, where $m \in \{1, \ldots, n\}$. Let us examine the RHS of (1) and see how many times we count $f(B)$. If we consider one set at a time, then B is contained in exactly m of the sets X_i, so $f(B)$ appears m times with a positive sign. If we consider pairwise intersections of sets, then B is contained in exactly $\binom{m}{2}$ of the sets $X_i \cap X_j$, and here $f(B)$ appears with a negative sign. Continuing on, if we consider k-wise intersections of sets for $k \in \{1, \ldots, n\}$, then $f(B)$ appears $\binom{m}{k}$ times, with the sign $(-1)^{k+1}$. Therefore, the total number of times that $f(B)$ is counted is

$$\sum_{k=1}^{m} (-1)^{k+1} \binom{m}{k} = 1 - \sum_{k=0}^{m} (-1)^k \binom{m}{k}.$$

However, by the binomial theorem (Theorem 0.4), $\sum_{k=0}^{m} (-1)^k \binom{m}{k} = (1-1)^m = 0$, so we see that $f(B)$ is counted exactly one time! Since this applies to each of the length-n bit strings (except the string of all 0s), we see that the RHS of (1) sums the values of each of the $2^n - 1$ disjoint sets which make up $X_1 \cup \cdots \cup X_n$ exactly once, which establises the formula. □

Exercise 4 Prove the generalized inclusion-exclusion principle by induction.

0.4 Solutions to Exercises

Exercise 1 The number of games is the number of unordered pairs of schools, which is $\binom{n}{2} = n(n-1)/2$.

Exercise 2 **Pascal's Identity** Algebraically:

$$\begin{aligned}\binom{n-1}{k-1} + \binom{n-1}{k} &= \frac{(n-1)!}{(k-1)!(n-k)!} + \frac{(n-1)!}{k!(n-k-1)!} \\ &= \frac{(n-1)!}{(k-1)!(n-k-1)!}\left(\frac{1}{n-k} + \frac{1}{k}\right) = \frac{n!}{k!(n-k)!} = \binom{n}{k}.\end{aligned}$$

For the combinatorial proof, let X be a set with $|X| = n$ and let $x \in X$. To choose a subset of size k from X, note that there are $\binom{n-1}{k-1}$ ways to choose a subset including x and $\binom{n-1}{k}$ ways to choose a subset not including x.

Exercise 3 There is nothing to prove when $n = 1$. Suppose that the theorem holds for a positive integer n. Then,

$$(x+y)^{n+1} = (x+y)\sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} = \sum_{k=0}^{n} \binom{n}{k} x^{k+1} y^{n-k} + \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k+1}$$

$$= \sum_{k=1}^{n} \left(\binom{n}{k-1} + \binom{n}{k} \right) x^k y^{n-k+1} + x^{n+1} + y^{n+1}$$
$$= \sum_{k=1}^{n} \binom{n+1}{k} x^k y^{n+1-k} + x^{n+1} + y^{n+1} = \sum_{k=0}^{n+1} \binom{n+1}{k} x^k y^{n+1-k}.$$

Exercise 4 There is nothing to prove when $n = 1$. When $n = 2$, write the disjoint union $X_1 \cup X_2 = (X_1 \cap X_2^{\mathsf{c}}) \cup (X_1 \cap X_2) \cup (X_1^{\mathsf{c}} \cap X_2)$. Then,

$$f(X_1 \cup X_2) = f(X_1 \cap X_2^{\mathsf{c}}) + f(X_1 \cap X_2) + f(X_1^{\mathsf{c}} \cap X_2)$$

and

$$f(X_1) + f(X_2) = f(X_1 \cap X_2) + f(X_1 \cap X_2^{\mathsf{c}}) + f(X_1 \cap X_2) + f(X_1^{\mathsf{c}} \cap X_2)$$

so $f(X_1 \cup X_2) = f(X_1) + f(X_2) - f(X_1 \cap X_2)$. Now, suppose that the formula holds for $n-1$ events, for some integer $n \geq 3$. We have

$$f\Big(\bigcup_{i=1}^{n} X_i\Big) = f\Big(X_n \cup \bigcup_{i=1}^{n-1} X_i\Big) = f(X_n) + f\Big(\bigcup_{i=1}^{n-1} X_i\Big) - f\Big(X_n \cap \bigcup_{i=1}^{n-1} X_i\Big)$$
$$= f(X_n) + \sum_{k=1}^{n-1} \sum_{\substack{I \subseteq \{1,\ldots,n-1\} \\ |I|=k}} (-1)^{k+1} f\Big(\bigcap_{i \in I} X_i\Big) - f\Big(\bigcup_{i=1}^{n-1} (X_n \cap X_i)\Big),$$

where we have applied the inclusion-exclusion principle twice (once for two sets and once for $n-1$ sets). The last term in the expression above is

$$-f\Big(\bigcup_{i=1}^{n-1} (X_n \cap X_i)\Big) = -\sum_{k=1}^{n-1} \sum_{\substack{I \subseteq \{1,\ldots,n-1\} \\ |I|=k}} (-1)^{k+1} f\Big(X_n \cap \bigcap_{j=1}^{k} X_j\Big),$$

by applying inclusion-exclusion yet again. Now combine all of the terms together.

Chapter 1

Probability Theory

We introduce the axioms of probability and illustrate the fundamentals.

1.1 Probability Axioms

Not everything in life goes as smoothly as anticipated: packets are dropped during transmission, light bulbs burn out prematurely, and doctors must diagnose patients who display symptoms of illness. (By the way, we will discuss all of these situations using probabilistic models soon enough.) We have to face uncertainty in the outcomes of our actions, and we have to reason logically about incomplete information. Probability theory isn't always depressing though: we also use probability to gamble intelligently and design robust systems.

It turns out that a wide variety of phenomena can be modeled effectively by the following mathematical objects:

- a set of all possible outcomes of interest, which we call Ω (also known as the **probability space** or **sample space**),
- subsets of the probability space, which we call **events** (sometimes denoted $\mathcal{F}$),
- a function that assigns values to sets, which will be our **probability measure** $\mathbb{P}$.

Again, with less formality: Ω contains all possible outcomes, and events are *sets of possibilities*. We discuss events instead of individual outcomes because we are often interested in more than one outcome at a time. (When we discuss continuous probability, we will also see that it is not sufficient to *only* talk about individual outcomes.) Finally, our probability measure is our way of assigning *likelihoods* to the events, with higher numbers corresponding to more likely outcomes.

The question we are primarily interested in right now is: what properties do we want our probability measure $\mathbb{P}$ to satisfy? It is natural to say that the likelihood of nothing happening at all is 0, which we write as:

$$\boxed{\mathbb{P}(\varnothing) = 0} \tag{1.1}$$

(Recall that $\varnothing$ is the empty set.)

The second condition is by convention: it is extremely useful to restrict the probability values to lie in the range $[0, 1]$. This condition can be written as:

$$\boxed{\mathbb{P}(\Omega) = 1} \tag{1.2}$$

An important consequence of this choice is that we may interpret the probability value of an event as the long-term proportion that the event occurs. As a first example, consider the archetypal fair coin toss. The probability that the coin comes up heads is $1/2$, which can be interpreted as follows: if we flipped the coin from now until forever (an infinite number of times), then we would expect a fraction $1/2$ of the flips to come up heads. This is the **frequentist** view of statistics. Briefly, the other major view of statistics is known as the **subjectivist** view, in which the probability value is interpreted as the *degree of your belief* in the outcome. Regardless of which view you adopt, the laws of probability are the same.

Finally, the last probability axiom is **countable additivity**, which is stated as follows: if for each $i \in \mathbb{N}$ we have an event A_i, such that the events A_i are pairwise *disjoint* (i.e. no two events have an outcome in common), then the probabilities of the events must add:

$$\boxed{\mathbb{P}\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) = \sum_{i\in\mathbb{N}} \mathbb{P}(A_i)} \tag{1.3}$$

In other words, *likelihoods add* for disjoint events. Perhaps it is not so clear right now why countable additivity is so fundamental that we make it an axiom, but without this axiom we would not have much of a theory at all. Before we move on, we can verify a quick consequence of countable additivity: if A_1 and A_2 are disjoint events, then $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2)$. This is known as **finite additivity**, and it follows from countable additivity and (1.1) by taking $A_3 = A_4 = \cdots = \varnothing$.

Exercise 5 Prove via induction that for $n \in \mathbb{Z}_+$, if $A_1, \ldots, A_n$ are disjoint events, then finite additivity implies $\mathbb{P}(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^{n} \mathbb{P}(A_i)$. (Note that finite additivity alone *does not* imply countable additivity.)

1.2 Fundamental Probability Facts

From the axioms, we can derive a number of rules which are invaluable to our calculations. The first is that for any event A, we can write $\Omega = A \cup A^{\mathsf{c}}$ (where A^{c} is the **complement** of A). Applying finite additivity, we obtain $1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^{\mathsf{c}}) = \mathbb{P}(A) + \mathbb{P}(A^{\mathsf{c}})$, which is also written as:

$$\boxed{\mathbb{P}(A^{\mathsf{c}}) = 1 - \mathbb{P}(A)} \tag{1.4}$$

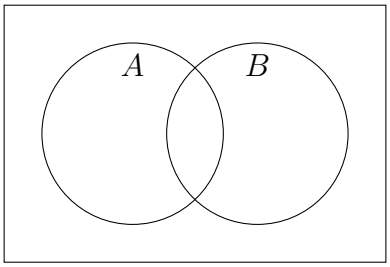
The next is known as **subadditivity**: if $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$. For if $A \subseteq B$, then we can write $B = A + (B \setminus A)$, where A and $B \setminus A$ are disjoint sets. Then, we have

$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A)$, and noting that probabilities are non-negative, we can drop the second term and turn it into an inequality.

The next is the **inclusion-exclusion principle**, which we have already encountered in Theorem 0.5 (observe that $\mathbb{P}$ is a finitely additive function on the set of events). For the case of two events A and B, the inclusion-exclusion principle says:

$$\boxed{\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)} \tag{1.5}$$

As a reminder, the picture is as follows:



In the general case, if n is any positive integer, and $A_1, \ldots, A_n$ are events, then:

$$\boxed{\mathbb{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ |I|=k}} \mathbb{P}\Big(\bigcap_{i \in I} A_i\Big)} \tag{1.6}$$

If we take the inclusion-exclusion principle for two events, (1.5), and drop the last term on the right, then we have a simple inequality: $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$. With induction, we extend this to the **union bound**. If n is a positive integer and $A_1, \ldots, A_n$ are events, then:

$$\boxed{\mathbb{P}(A_1 \cup \cdots \cup A_n) \leq \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n)} \tag{1.7}$$

It's a crude bound, but nevertheless useful.

Exercise 6 Union Bound Prove the union bound for finitely many events using induction.

Exercise 7 Union Bound for Countably Many Events Let $A_1, A_2, A_3, \ldots$ be a countably infinite sequence of events. Prove that the union bound still holds:

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

(Note that induction does not work to prove the statement here since there are now infinitely many events.)

Exercise 8 Bonferroni's Inequality Prove that for $n \in \mathbb{N}$, $n \geq 2$, and events $A_1, \ldots, A_n$, it holds that $\mathbb{P}(A_1 \cap \cdots \cap A_n) \geq \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) - (n-1)$.

1.3 Discrete Probability

We now consider the case when Ω is countable. Recall that the probability measure $\mathbb{P}$ is a function which assigns real numbers to *sets of outcomes*; in general, this statement is far more expressive than if we had simply said that $\mathbb{P}$ assigns real numbers to *individual outcomes*. However, we will not need the full power of this statement yet. In the case of discrete probability, *the probability of any event is completely determined once we specify the probability of each outcome*. Let us make this explicit:

$$\boxed{\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\})} \tag{1.8}$$

What we are saying here is that the probability of the event A is the sum of the probabilities of each outcome in A, which follows from the countable additivity axiom. Perhaps this is intuitively clear and we are being overly fussy about the details, but attention to detail will be quite important later in the course.

1.3.1 Uniform Sample Space

An important case is when every outcome is equally likely to occur, i.e. $\mathbb{P}(\omega) = 1/|\Omega|$. We say that Ω is a **uniform probability space**. The advantage of having a uniform probability space is that we may employ methods of combinatorics to compute probabilities:

$$\boxed{\mathbb{P}(A) = \frac{|A|}{|\Omega|}} \tag{1.9}$$

To compute the probability of the event A, we count the number of ways in which A is achieved, and divide by the total number of elements in Ω.

Example 1.1 (Sampling without Replacement)**.** Suppose that you have n objects in a bag, and you draw k items one by one without looking, where $k \leq n$ are positive integers. After each draw, you do *not* place the item back in the bag. This is known as **sampling without replacement**. We model sampling without replacement as a uniform probability space, where we view each of the $\binom{n}{k}$ possible combinations of k items to be equally likely. [a]

$$|\Omega| = \binom{n}{k},$$
$$\mathbb{P}(\omega) = \frac{1}{\binom{n}{k}} \qquad \forall \omega \in \Omega.$$

As a specific example, often we discuss drawing a five-card hand from a shuffled deck of 52 playing cards. Here, we are sampling the cards without replacement, where $n = 52$

and $k = 5$. Each of the possible five-card hands is equally likely to be drawn.

[a] We will often write $\mathbb{P}(\omega)$ for notational simplicity rather than $\mathbb{P}(\{\omega\})$.

1.4 Conditional Probability

First, we state the important **law of total probability**, which allows us to break down the probability of one event into the probabilities of smaller events. Formally, let n be a positive integer and suppose that the events $A_1, \ldots, A_n$ *partition* the sample space, that is, they are disjoint and $A_1 \cup \cdots \cup A_n = \Omega$. Then, to find the probability of an event B, we can write:

$$\boxed{\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \cap A_i)} \tag{1.10}$$

The utility of this statement is that $\mathbb{P}(B)$ may be more difficult to compute than $\mathbb{P}(B \cap A_i)$, because for the latter, we have more information: we know that both B and A_i have occurred, and the knowledge of A_i may aid us in the computation of $\mathbb{P}(B)$. This line of reasoning is also useful for calculating probabilities associated with sequential processes, which can be represented in a tree structure. Each node of the tree represents a decision, with the different branches from a node representing the different possible outcomes from the decision. To fully analyze such situations, we will need another tool called **conditional probability**.

Conditional probability is the idea that events can affect each other; for example, observing that there are more clouds in the sky today informs us that there is a greater chance for rain. How do we formalize this idea mathematically? Suppose that $\mathbb{P}$ represents our beliefs in the likelihoods of various events (adopting the subjectivist view for a moment), and then we observe an event A. In a sense, $\mathbb{P}$ is *outdated information*; $\mathbb{P}$ is a probability law that was formulated before we learned about A, and knowing A might change our beliefs. What we need is a *new* probability law, which we will call the *conditional probability law given the event* A. Since we have a new probability law, we may call it by a different name, such as $\mathbb{P}_A$, but this is not standard. Instead, we write $\mathbb{P}(B \mid A)$, read as "the probability of B given A", to mean the probability assigned to the event B under the new law.

It is important to recognize that the new law is a true probability law, i.e. it must satisfy the three probability axioms. In addition, we would like the following property: $\mathbb{P}(A \mid A) = 1$. This is equivalent to saying that knowing the occurence of A, we are certain that A has indeed occurred. These statements are intended to motivate the following definition. Let A be an event with $\mathbb{P}(A) > 0$. Then:

$$\boxed{\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}} \tag{1.11}$$

The interpretation of this equation is that the conditional probability law is formed from the previous probability law in two stages: first, we restrict our attention only to the events which are subsets of A; second, we *normalize* the probability, that is, we divide by $\mathbb{P}(A)$ to ensure that the probability of A itself under this new law is now 1.

Exercise 9 Conditional Probability Prove that the conditional probability law $\mathbb{P}(\cdot \mid A)$ satisfies the probability axioms, where the sample space is now restricted to A and the events are restricted to events which are subsets of A.

If $\mathbb{P}(B) > 0$ as well, the expression above can be written in a more symmetric way:

$$\mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(A \cap B) = \mathbb{P}(B \mid A)\mathbb{P}(A). \tag{1.12}$$

Often, we can write down conditional probability laws quite easily in a cause-and-effect fashion, such as the probability that a patient has a fever given that the patient has the flu. Then, the equation $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B \mid A)$ states that we can directly apply our cause-and-effect knowledge, by multiplying the unconditional probability $\mathbb{P}(A)$ with the conditional probability $\mathbb{P}(B \mid A)$, to obtain the probability that both events A and B occur. The rule $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B \mid A)$ is often known as the **chain rule** or **product rule** of probability. Look at Figure 1.1 for a visualization.

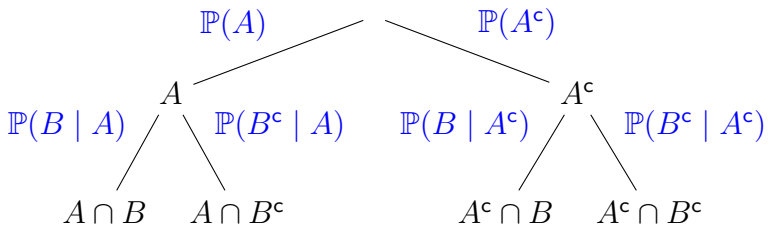


Figure 1.1: Conditional probability can be visualized sequentially as a tree. To obtain the probability of a node, such as $\mathbb{P}(A \cap B)$, multiply the probabilities along the edges leading up to the node: $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B \mid A)$. To obtain $\mathbb{P}(B)$, sum up the probabilities of the leaves which correspond to the event B: $\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B)$.

Example 1.2 (Birthday Problem I)**.** Suppose that n and m are positive integers and we throw n balls into m bins. What is the smallest n such that the probability of a collision (two balls landing in the same bin) is at least $1/2$?

When $m = 365$ and we view the "balls" as people and the "bins" as birthdays, the question becomes: how many people do you need to gather in a room before the probability that two people share a birthday is at least $1/2$? This is commonly known as the **birthday problem**.

Let A_i denote the event that ball i does not collide with balls $1, \ldots, i-1$. We have $\mathbb{P}(A_1) = 1$ since the first ball cannot produce a collision. We would like to compute $\mathbb{P}(A_1 \cap \cdots \cap A_n)$, the probability that we do not have any collisions. To do so, we will use the chain rule of probability,

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \prod_{i=2}^{n} \mathbb{P}(A_i \mid A_1 \cap \cdots \cap A_{i-1}). \tag{1.13}$$

Here, $\mathbb{P}(A_i \mid A_1 \cap \cdots \cap A_{i-1})$ is the conditional probability that ball i does not collide with the first $i-1$ balls, given that none of the first $i-1$ balls produced a collision. If none of the first $i-1$ balls produced a collision, then they must each be in separate bins; therefore, $\mathbb{P}(A_i \mid A_1 \cap \cdots \cap A_{i-1})$ is the probability of ball i landing in one of the other $m-(i-1)$ bins, which is $(m-i+1)/m$.

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \frac{m}{m} \cdot \frac{m-1}{m} \cdots \frac{m-n+1}{m} = \frac{m!}{m^n(m-n)!}$$

Although we have obtained an exact answer, many times it is worth the effort to investigate if we can approximate the solution in order to better understand its properties. We rewrite the desired probability in the following way:

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = 1 \cdot \Big(1 - \frac{1}{m}\Big) \cdot \cdots \cdot \Big(1 - \frac{n-1}{m}\Big) = \prod_{i=0}^{n-1} \Big(1 - \frac{i}{m}\Big)$$

Next, we can use the approximation $\exp x \approx 1 + x$ to replace the factor $1 - i/m$ with $\exp(-i/m)$. ($1 + x$ is the first-order series expansion of $\exp x$.)

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) \approx \prod_{i=0}^{n-1} \exp\Big(-\frac{i}{m}\Big) = \exp\Big(-\frac{1}{m}\sum_{i=0}^{n-1} i\Big)$$

Evaluating the summation, we have $\sum_{i=0}^{n-1} i = n(n-1)/2 \approx n^2/2$, so

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) \approx \exp\Big(-\frac{n^2}{2m}\Big).$$

Setting $\mathbb{P}(A_1 \cap \cdots \cap A_n) = 1/2$, we find that $n^2 = (2 \ln 2)m$, or $n = \sqrt{(2 \ln 2)m}$. If we ignore constants, then in the language of computational complexity, we have $m \in O(n^2)$.

As another application, suppose that we are designing a hash function

$$h : \{0, \ldots, n-1\} \to \{0, \ldots, m-1\}.$$

Here, the "balls" are inputs and the "bins" are the number of buckets in our hash table. What we have found is that if we want the probability of collision between two inputs to be less than $1/2$ (in other words, for distinct inputs i and j, we want $\mathbb{P}(h(i) \neq h(j)) \leq 1/2$), then we want the size of our hash table, m, to be roughly n^2.

Example 1.3 (Birthday Problem II)**.** Using the union bound, one can obtain the results of Example 1.2 more easily. Since there are n balls, there are $\binom{n}{2} = n(n-1)/2$ different pairs of balls. Let B_i denote the event that the ith pair of balls collides, and write

$B = \bigcup_{i=1}^{n(n-1)/2} B_i$ as the event of a collision. By the union bound,

$$\mathbb{P}(B) \leq \sum_{i=1}^{n(n-1)/2} \mathbb{P}(B_i) = \frac{n(n-1)}{2} \cdot \frac{1}{m} \approx \frac{n^2}{2m}.$$

To calculate $\mathbb{P}(B_i)$, we observed that regardless of where the first ball lands, the probability that the second ball in the pair lands in the same bin as the first ball is $1/m$. Again, if we set $\mathbb{P}(B) = 1/2$, we find that $m \in O(n^2)$.

Although the union bound may seem crude, it is usually simple to apply and often yields good results.

1.4.1 Bayes's Law

Consider the above situation, where we have the conditional probability that a patient has the fever knowing that the patient has the flu. Actually, in reality, the situation is backwards: we can observe that the patient has a fever, but we are not yet sure of the diagnosis. In fact, there may be competing explanations for the phenomena, ranging from pneumonia to infections and even cancer. How do we use probabilistic reasoning to make an informed diagnosis?

We use the law of total probability, (1.10), along with the definition of conditional probability, to write the following derivation of **Bayes's rule**:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A \cap B) + \mathbb{P}(A^{\mathsf{c}} \cap B)} = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B \mid A)\mathbb{P}(A) + \mathbb{P}(B \mid A^{\mathsf{c}})\mathbb{P}(A^{\mathsf{c}})}$$

In one sense, we have derived a way to relate $\mathbb{P}(A|B)$ to $\mathbb{P}(B|A)$, so we have a computational tool: whenever we have the "wrong" conditional probabilities for a problem we face, use Bayes's rule to turn the conditional probabilities around. More generally, if n is a positive integer and $A_1, \ldots, A_n$ partition the sample space, then:

$$\boxed{\mathbb{P}(A_k \mid B) = \frac{\mathbb{P}(B \mid A_k)\mathbb{P}(A_k)}{\sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)}, \qquad k = 1, \ldots, n} \tag{1.14}$$

The derivation of Bayes's rule was not so difficult, so one may wonder why such a big deal is made out of the formula. Let us examine another way to view Bayes's rule: given multiple possible explanations for an observation we have just made, we can discern the most likely explanation. This is called *probabilistic inference*, and it concerns problems such as patient diagnosis, or determining whether a blood test result is a false positive.

Another interpretation is gained when we view conditional probability in the framework of taking an old probability law ("outdated information") and producing a new probability law after the observation of an event. In this view, Bayes's rule can be viewed as an *update rule*, which tells us the correct way to respond to new information. Sometimes, it is said that Bayes's rule is the foundation for the field of artificial intelligence. After all, a rational agent is one that collects information and best figures out how to utilize the new information, a task for which Bayes's rule excels.

1.5 Independence

Speaking of information, what if we observe an event A, but knowing that A occurs tells us exactly nothing about whether B will occur? In other words, our belief that B occurs is unchanged after the observation of A. In this case, we say that A and B are **independent**: $\mathbb{P}(B \mid A) = \mathbb{P}(B)$. In order for this definition to make sense, we need $\mathbb{P}(A) > 0$, but actually it is possible to write the condition for independence in a more symmetric way which does not require $\mathbb{P}(A) > 0$.

$$\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)} \tag{1.15}$$

Independence is a truly subtle concept. If we view the independence of A and B as the *information* about B you receive from observing A (or vice versa), then we should have that A^{c} and B are also independent, and A and B^{c} should also be independent, and so forth. Pinning down the concept of independence as "information" requires a more formal setting than we have described so far, but for now, concepts such as "knowledge" and "information" provide a good intuitive guide. For example, we will make statements such as: if I know the outcome of one coin toss, I still have no clue what the next coin toss will be, so we say that coin tosses are independent of each other.

In the above example, you may have noticed that the independence of coin tosses requires talking about the independence of multiple events, and it is not fully clear what we mean. In fact, there are different forms of independence of the events $A_1, \ldots, A_n$, when n is a positive integer. We say that the events $A_1, \ldots, A_n$ are **pairwise independent** if every pair of events is independent. A stronger statement is the statement of **mutual independence**. If we have two sets of events, such that the sets are disjoint, then mutual independence requires that any combination of events from the first set is independent of any combination of events from the second event. For example, $A_1 \cap A_3$ should be independent of $A_2 \cup A_5 \cup A_6$, and so forth. An equivalent way of stating this condition is:

$$\boxed{\mathbb{P}\Big(\bigcap_{i=1}^{n} A_i'\Big) = \prod_{i=1}^{n} \mathbb{P}(A_i')} \tag{1.16}$$

where each A_i' is allowed to be either A_i or Ω. The reason we allow A_i' to be Ω is because we want to allow combinations such as $A_1 \cap A_2$ (combinations which do not use *every* A_i), which we can write as $A_1 \cap A_2 \cap \Omega \cap \cdots \cap \Omega$. This is just a technical detail though. Importantly, you should recognize that mutual independence is *stronger* than pairwise independence. In the case of coin tosses, we would like the stronger condition of mutual independence to hold, because we want to talk about sets of coin tosses as well.

How about an infinite set of events $\{A_\alpha\}_{\alpha \in \mathcal{A}}$ for some indexing set $\mathcal{A}$, maybe even uncountably many events? The answer is that we say $\{A_\alpha\}_{\alpha \in \mathcal{A}}$ is independent if every *finite subcollection* of the events is independent.

Example 1.4. Here is an example to demonstrate that mutual independence is indeed stronger than pairwise independence. Consider two flips of a fair coin, and define the following events: H_1 is the event that the first coin is heads, H_2 is the event that the second coin is heads, and S is the event that the two tosses produced the same outcome (two heads or two tails).

We can check that the three events are pairwise independent. Conditioned on H_i, $i = 1, 2$, $\mathbb{P}(S \mid H_i)$ is the probability that the other coin also lands heads, which is $1/2$. Hence, $\mathbb{P}(S \mid H_i) = \mathbb{P}(S) = 1/2$, so H_i and S are independent. Clearly, H_1 and H_2 are independent, so we have pairwise independence.

However, we do not have mutual independence:

$$\frac{1}{4} = \mathbb{P}(H_1 \cap H_2 \cap S) \neq \mathbb{P}(H_1)\mathbb{P}(H_2)\mathbb{P}(S) = \frac{1}{8}$$

Example 1.5. One might ask if $\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)$ alone is sufficient for mutual independence. The answer is no.

Consider the probability space of a fair die, $\Omega = \{1, 2, 3, 4, 5, 6\}$, and define the following events: $A_1 = A_2 = \{1, 2, 3\}$ and $A_3 = \{3, 4, 5, 6\}$. We have

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(\{3\}) = \frac{1}{6} = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3),$$

but clearly A_1 and A_2 are not independent. Therefore, when checking for mutual independence of events A_i, one must check that *each* of the conditions (1.16) holds.

1.5.1 Correlated Events

We say that the events A and B are **positively correlated** if $\mathbb{P}(A \cap B) > \mathbb{P}(A)\mathbb{P}(B)$. Similarly, we say that A and B are **negatively correlated** if $\mathbb{P}(A \cap B) < \mathbb{P}(A)\mathbb{P}(B)$. Of course, if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, then A and B are independent.

Intuitively, if $\mathbb{P}(A \cap B)$ is large relative to $\mathbb{P}(A)\mathbb{P}(B)$, then events A and B tend to occur together; another way to see this is that $\mathbb{P}(A \cap B) > \mathbb{P}(A)\mathbb{P}(B)$ implies that $\mathbb{P}(A \mid B) > \mathbb{P}(A)$ and $\mathbb{P}(B \mid A) > \mathbb{P}(B)$. Hence, observing one of A or B increases the likelihood of observing the other. If $\mathbb{P}(A \cap B)$ is small relative to $\mathbb{P}(A)\mathbb{P}(B)$, then observing one of A or B will decrease the likelihood of observing the other; an extreme example is when A and B are mutually exclusive (i.e. disjoint).

1.6 Solutions to Exercises

Exercise 5 The base case when $n = 1$ is trivial (here, trivial is used to mean "there is nothing to show"). Otherwise, suppose that the result holds for n disjoint events and let $A_1, \ldots, A_{n+1}$ be disjoint. By applying finite additivity to the events $A_1 \cup \cdots \cup A_n$ and A_{n+1} (which are disjoint), we have $\mathbb{P}(A_1 \cup \cdots \cup A_{n+1}) = \mathbb{P}(A_1 \cup \cdots \cup A_n) + \mathbb{P}(A_{n+1})$. But now, we can apply the inductive claim to the first term and conclude

$$\mathbb{P}(A_1 \cup \cdots \cup A_{n+1}) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_{n+1}).$$

Exercise 6 Union Bound The case of $n = 2$ was already proven. So, suppose that the union bound is true for some positive integer n. Then,

$$\mathbb{P}(A_1 \cup \cdots \cup A_{n+1}) \leq \mathbb{P}(A_1 \cup \cdots \cup A_n) + \mathbb{P}(A_{n+1})$$

by the union bound for two events, and then applying the inductive hypothesis to the first term on the right we obtain $\mathbb{P}(A_1 \cup \cdots \cup A_{n+1}) \leq \sum_{i=1}^{n+1} \mathbb{P}(A_i)$.

Exercise 7 Union Bound for Countably Many Events Define $A_1' := A_1$ and for each $i = 2, 3, \ldots$, define $A_i' := A_i \setminus \bigcup_{j=1}^{i-1} A_j$. Therefore, for each positive integer i, we have $A_i' \subseteq A_i$, so $\mathbb{P}(A_i') \leq \mathbb{P}(A_i)$. Notice that $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} A_i'$, and the sequence of events $A_1', A_2', A_3', \ldots$ is disjoint, so we can apply the countable additivity axiom. Thus, $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \mathbb{P}(\bigcup_{i=1}^{\infty} A_i') = \sum_{i=1}^{\infty} \mathbb{P}(A_i') \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Exercise 8 Bonferroni's Inequality From the inclusion-exclusion principle (1.5), we have $1 \geq \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$, which can be rearranged to yield the case of $n = 2$: $\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$. Now, the rest follows by induction. Suppose that the result holds for $n = k$. Then,

$$\mathbb{P}(A_1 \cap \cdots \cap A_{k+1}) \geq \mathbb{P}(A_1 \cap \cdots \cap A_k) + \mathbb{P}(A_{k+1}) - 1 \geq \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_{k+1}) - k.$$

Exercise 9 Conditional Probability We check that $\mathbb{P}(\varnothing \mid A) = \mathbb{P}(\varnothing \cap A)/\mathbb{P}(A) = 0$ since $\mathbb{P}(\varnothing \cap A) = \mathbb{P}(\varnothing) = 0$; also, $\mathbb{P}(A \mid A) = \mathbb{P}(A \cap A)/\mathbb{P}(A) = 1$. If $A_1, A_2, A_3, \ldots$ is a sequence of disjoint events in the new sample space, then

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i \;\Big|\; A\Big) = \frac{\mathbb{P}(A \cap \bigcup_{i=1}^{\infty} A_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(\bigcup_{i=1}^{\infty} (A \cap A_i))}{\mathbb{P}(A)} = \frac{\sum_{i=1}^{\infty} \mathbb{P}(A \cap A_i)}{\mathbb{P}(A)} = \sum_{i=1}^{\infty} \mathbb{P}(A_i \mid A)$$

since the events $\{A \cap A_i\}_{i=1}^{\infty}$ are disjoint.

Chapter 2

Discrete Random Variables

Random variables are the tool of choice for modeling probability spaces for which the outcomes are associated with a *numerical value* of interest. For instance, stock market prices and the daily temperature are two examples of "random numbers" in our lives. Each random variable has an associated *distribution* which describes its behavior; we will introduce common families of discrete distributions. We will see that the *expectation* of a random variable is a useful summary statistic of its distribution which satisfies a crucial property: linearity.

2.1 Random Variables

Definition 2.1. A **random variable** is a function $X : \Omega \to \mathbb{R}$ that assigns a real number to every outcome ω in the probability space.

We typically denote random variables by capital letters. See Figure 2.1 for an abstract view of a random variable.

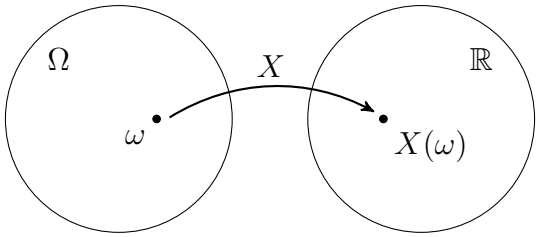


Figure 2.1: Abstract diagram of a random variable.

We typically view random variables as numerical outcomes associated with a random experiment. In the study of probability theory, random variables are the main objects of interest. Eventually, we will place less emphasis on the probability space lurking in the background, and we will focus on the properties of random variables.

2.1.1 Functions of Random Variables

We define addition of random variables in the following way: the random variable $X + Y$ is the random variable that maps ω to $X(\omega) + Y(\omega)$. Similarly, the random variable XY is the

random variable that maps ω to $X(\omega)Y(\omega)$. More generally, let n be any positive integer and $f : \mathbb{R}^n \to \mathbb{R}$ be any function. Then $f(X_1, \ldots, X_n)$ is defined to be the random variable that maps ω to $f(X_1(\omega), \ldots, X_n(\omega))$. See Figure 2.2 for an illustration for a single-variable function.

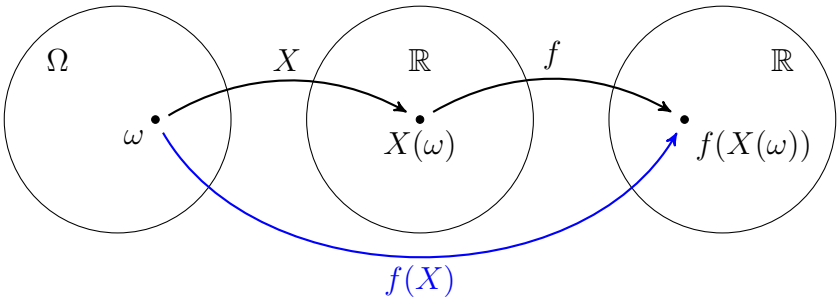


Figure 2.2: Function of a random variable.

2.1.2 The Distribution of a Random Variable

Random variables assign probabilities to real numbers. To see how this is done, consider the sample space of two tosses of a fair coin: $\Omega = \{HH, HT, TH, TT\}$. Let us define the random variable X in the following way:

$$\begin{aligned} X(HH) &= 2 \\ X(HT) &= 1 \\ X(TH) &= 1 \\ X(TT) &= 0 \end{aligned}$$

(Here, X represents the number of heads that we see in the two coin tosses.) What is the probability that X takes on a particular value? We can write:

$$\begin{aligned} \mathbb{P}(X = 0) &= \mathbb{P}(\{TT\}) = \frac{1}{4} \\ \mathbb{P}(X = 1) &= \mathbb{P}(\{HT, TH\}) = \frac{1}{2} \\ \mathbb{P}(X = 2) &= \mathbb{P}(\{HH\}) = \frac{1}{4} \end{aligned}$$

We can reasonably say that X assigns the probability $1/4$ to the real number 0, the probability $1/2$ to the real number 1, and the probability $1/4$ to the real number 2. In this way, X induces a probability measure on the real line, which we call the **distribution** of X. The distribution of X satisfies the probability axioms; for example, (1.2) applied to the distribution of X is the equation:

$$\boxed{\sum_x \mathbb{P}(X = x) = 1} \tag{2.1}$$

where the sum ranges over all values of x in the range of X.

In the example above, we were explicit about the probability space Ω in order to show that X is a function from Ω to $\mathbb{R}$. However, the utility of random variables is that we can often *forget* about the underlying probability space Ω and focus our attention on the distribution of X. For a discrete random variable X, we can specify the distribution of X simply by giving the probabilities $\mathbb{P}(X = x)$ for all x in the range of X, without reference to the original probability space Ω. Indeed, that is what we will proceed to do from here onwards.

There are many common distributions (described in detail below), which have special names. We use the symbol $\sim$ to denote that a random variable has a known distribution, e.g. $X \sim \text{Binomial}(n, p)$ indicates that the distribution of X is the $\text{Binomial}(n, p)$ distribution.

Equivalently, we can describe a probability distribution by its **cumulative distribution function (CDF)**, which is defined as $F_X(x) := \mathbb{P}(X \leq x)$. The CDF contains exactly the same information as the distribution of X. To see this fact, observe that we can recover the probability distribution function (also known as the PDF) from the CDF by the following formula:

$$\boxed{\mathbb{P}(X = x) = \mathbb{P}(X \leq x) - \mathbb{P}(X \leq x - 1)} \tag{2.2}$$

for $x \in \mathbb{Z}$, assuming X takes on integer values.

Exercise 10 Suppose you have m bins, where m is a positive integer. You keep throwing balls into the bins uniformly at random until one of the bins has at least two balls. What is the distribution of X?

2.1.3 Multiple Random Variables

The **joint distribution** of two random variables X and Y is the probability distribution $\mathbb{P}(X = x, Y = y)$ for all possible pairs of values (x, y). The joint distribution must satisfy the normalization condition:

$$\boxed{\sum_x \sum_y \mathbb{P}(X = x, Y = y) = 1} \tag{2.3}$$

We can recover the distribution of X separately (known as the **marginal distribution** of X) by summing over all possible values of Y:

$$\boxed{\mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y)} \tag{2.4}$$

Similarly:

$$\boxed{\mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x, Y = y)} \tag{2.5}$$

See Figure 2.3 for a visualization of the joint distribution.

	y_1	y_2	y_3	$\cdots$	
x_1	$\mathbb{P}(X = x_1, Y = y_1)$	$\mathbb{P}(X = x_1, Y = y_2)$	$\mathbb{P}(X = x_1, Y = y_3)$	$\cdots$	$\mathbb{P}(X = x_1)$
x_2	$\mathbb{P}(X = x_2, Y = y_1)$	$\mathbb{P}(X = x_2, Y = y_2)$	$\mathbb{P}(X = x_2, Y = y_3)$	$\cdots$	$\mathbb{P}(X = x_2)$
x_3	$\mathbb{P}(X = x_3, Y = y_1)$	$\mathbb{P}(X = x_3, Y = y_2)$	$\mathbb{P}(X = x_3, Y = y_3)$	$\cdots$	$\mathbb{P}(X = x_3)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$\mathbb{P}(Y = y_1)$	$\mathbb{P}(Y = y_2)$	$\mathbb{P}(Y = y_3)$	$\cdots$	

Figure 2.3: Joint distributions are often represented as tables, by specifying the value of $\mathbb{P}(X = x, Y = y)$ for all x and all y. The name "marginal" distribution arises because the distribution of X alone is written in the margins of the table: by summing the probabilities across the ith row, we obtain $\mathbb{P}(X = x_i)$. Similarly, by summing probabilities down the jth column, we obtain $\mathbb{P}(Y = y_j)$.

The joint distribution contains all of the information about X and Y. From the joint distribution, we can recover the marginal distributions of X and Y. The converse is not true: the marginal distributions are usually *not* sufficient to recover the joint distribution. The reason for this is that the joint distribution captures information about the *dependence* of X and Y. This leads us to formulate the following definition:

Definition 2.2. We say that two discrete random variables are **independent** if

$$\forall x, y \in \mathbb{R}, \qquad \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y). \tag{2.6}$$

If X and Y are independent, then we can write their joint probability as a product of their marginal probabilities (sometimes, we say that the joint distribution *factors*). As a result, calculations involving independent random variables are much simpler. All of the results in this section generalize easily to multiple random variables.

2.2 Expectation

Knowing the full probability distribution gives us a lot of information, but sometimes it is helpful to have a summary of the distribution.

Definition 2.3. The **expectation** (or **expected value**) of a discrete random variable X is defined to be:

$$\boxed{\mathbb{E}[X] := \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega) = \sum_{x} x\mathbb{P}(X = x)} \tag{2.7}$$

Technical Remark: The expectation is not always well-defined. To avoid these issues, throughout the course we will assume $\mathbb{E}[|X|] < \infty$.

The expected value has an interpretation as the *long-run average* of an experiment in which you measure the values of X. Precisely, if $X_1, X_2, X_3, \dots$ are independent copies of X, then we will later see that

$$\lim_{N\to\infty} \frac{X_1 + \cdots + X_N}{N} \to \mathbb{E}[X]$$

(in a certain sense that we will formalize later).

Often, the expectation is easier to work with than the full probability distributions because it satisfies **linearity**:

Theorem 2.4 (Linearity of Expectation)**.** *Suppose X, Y are random variables, $a \in \mathbb{R}$ is a constant, and c is the constant random variable (i.e. $\forall \omega \in \Omega\ c(\omega) = c$). Then:*

1. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
2. $\mathbb{E}[aX + c] = a\,\mathbb{E}[X] + c$

Proof. The proof essentially follows from the linearity of summations.

1.
$$\begin{aligned}\mathbb{E}[X+Y] &= \sum_{\omega\in\Omega}\big(X(\omega)\mathbb{P}(\omega) + Y(\omega)\mathbb{P}(\omega)\big) = \sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\omega) + \sum_{\omega\in\Omega} Y(\omega)\mathbb{P}(\omega)\\ &= \mathbb{E}[X] + \mathbb{E}[Y].\end{aligned}$$

2.
$$\begin{aligned}\mathbb{E}[aX+c] &= \sum_{\omega\in\Omega}\big(aX(\omega) + c(\omega)\big)\mathbb{P}(\omega) = a\sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\omega) + c\sum_{\omega\in\Omega}\mathbb{P}(\omega)\\ &= a\,\mathbb{E}[X] + c.\end{aligned}$$

□

Important: Notice that we did *not* assume that X and Y are independent. We will use these properties repeatedly to solve complicated problems.

In the previous section, we noted that if X is a random variable and $f : \mathbb{R} \to \mathbb{R}$ is a function, then $f(X)$ is a random variable. The expectation of $f(X)$ is:

$$\boxed{\mathbb{E}[f(X)] = \sum_{\omega\in\Omega} f\big(X(\omega)\big)\mathbb{P}(\omega) = \sum_{x} f(x)\mathbb{P}(X = x)} \tag{2.8}$$

The definition can be extended easily to functions of multiple random variables using the

joint distribution:

$$\boxed{\mathbb{E}[f(X_1,\ldots,X_n)] = \sum_{x_1,\ldots,x_n} f(x_1,\ldots,x_n)\mathbb{P}(X_1 = x_1,\ldots,X_n = x_n)} \tag{2.9}$$

Next, we prove an important fact about the expectation of independent random variables.

Theorem 2.5 (Expectation of Independent Random Variables)**.** *Let X and Y be independent random variables. Then the random variable XY satisfies*

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]. \tag{2.10}$$

Proof.

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{x,y} xy\mathbb{P}(X = x, Y = y) = \sum_{x,y} xy\mathbb{P}(X = x)\mathbb{P}(Y = y)\\ &= \Big(\sum_x x\mathbb{P}(X = x)\Big)\Big(\sum_y y\mathbb{P}(Y = y)\Big) = \mathbb{E}[X]\,\mathbb{E}[Y].\end{aligned}$$

The definition of independent random variables was used in the first line of the proof. It is crucial to remember that *the theorem does not hold true when X and Y are not independent!* □

2.2.1 Tail Sum Formula

Next, we derive an important formula for computing the expectation of a random variable that only takes on values in the natural numbers.

Theorem 2.6 (Tail Sum Formula)**.** *Let X be a random variable that only takes on values in $\mathbb{N}$. Then*

$$\mathbb{E}[X] = \sum_{x=1}^{\infty} \mathbb{P}(X \geq x).$$

Proof. We manipulate the formula for the expectation: [a]

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=1}^{\infty} k\mathbb{P}(X = k) = \sum_{k=1}^{\infty}\sum_{x=1}^{k} \mathbb{P}(X = k)\\ &= \sum_{x=1}^{\infty}\sum_{k=x}^{\infty} \mathbb{P}(X = k) = \sum_{x=1}^{\infty} \mathbb{P}(X \geq x).\end{aligned}$$

□

[a] Although the proof is short, the interchange of summations can be very confusing at first. One way

to visualize the interchange of summations is to write out the (k, x) pairs over which we are summing:

$$\begin{array}{llll|llll} (k=1,x=1) & & & & (x=1,k=1) & (x=1,k=2) & (x=1,k=3) & \cdots \\ (k=2,x=1) & (k=2,x=2) & & & & (x=2,k=2) & (x=2,k=3) & \cdots \\ (k=3,x=1) & (k=3,x=2) & (k=3,x=3) & & & & (x=3,k=3) & \cdots \\ \vdots & \vdots & \vdots & & & & & \ddots \end{array}$$

First, convince yourself that the left side and the right side are both writing out the same (k, x) pairs. On the left side, the pairs are grouped first by the k-value. On the right side, the pairs are grouped first by the x-value. Next, observe that the left side represents the summation in the first line of the proof, while the right side represents the summation in the second line of the proof.

The formula is known as the **tail sum formula** because we compute the expectation by summing over the tail probabilities of the distribution.

2.3 Discrete Probability Distributions

We will now discuss the common discrete probability distributions.

2.3.1 Uniform Distribution

As a first example of a probability distribution, consider the **uniform distribution over** $\{1, \ldots, n\}$, denoted as Uniform$\{1, \ldots, n\}$. The meaning of *uniform* is that each element of the set is equally likely to be chosen; therefore, the probability distribution is

$$\mathbb{P}(X = x) = \frac{1}{n}, \qquad x \in \{1, \ldots, n\}.$$

The expectation of the uniform distribution is calculated fairly easily from the definition:

$$\mathbb{E}[X] = \sum_{x=1}^{n} x \cdot \frac{1}{n} = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}$$

where to evaluate the sum, we have used the triangular number identity:

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}. \tag{2.11}$$

Example 2.7 (Symmetry)**.** Suppose you have a box with n chocolates, for some positive integer n, where one of the chocolates is a special dark chocolate. You pick out chocolates randomly, one at a time, without putting them back into the box, until you find the special dark chocolate. Let X denote the number of chocolates that you remove from the box. What is the distribution and expectation of X?

By symmetry considerations, $X \sim$ Uniform$\{1, \ldots, n\}$. The argument proceeds as follows: since you are picking chocolates without replacement, we can imagine an equivalent situation in which you number the chocolates from 1 to n, pick a random permutation of $1, \ldots, n$, and draw the chocolates in the order specified by the permutation. A random

permutation must assign the special chocolate to each of the n positions with equal probability, and X, the number of chocolates you remove from the box, equals the random position assigned to the special chocolate; but we just argued that the special chocolate is equally likely to be in any of the n positions. So, X is uniform over $\{1, \ldots, n\}$. As discussed above, the expectation is $(n+1)/2$.

As a consequence, if each person picks out just one chocolate from the box at random, it does not matter whether you are the first or the last person to pick out a chocolate: the probability that you will end up with the special dark chocolate is always the same, $1/n$.

2.3.2 Bernoulli Distribution

The **Bernoulli distribution with parameter** p, $p \in [0,1]$, denoted Bernoulli(p), is a simple distribution that describes the result of performing one experiment which succeeds with probability p. The probability space is $\Omega = \{\text{Success}, \text{Failure}\}$ with $\mathbb{P}(\text{Success}) = p$ and $\mathbb{P}(\text{Failure}) = 1 - p$. Define the random variable X as

$$X(\omega) = \begin{cases} 0, & \omega = \text{Failure}, \\ 1, & \omega = \text{Success}. \end{cases}$$

The distribution of X is

$$\mathbb{P}(X = x) = \begin{cases} 1-p, & x = 0, \\ p, & x = 1, \\ 0, & \text{otherwise}. \end{cases}$$

A concise way to describe the distribution is $\mathbb{P}(X = x) = (1-p)^{1-x}p^x$ for $x \in \{0,1\}$. The expectation of the Bernoulli(p) distribution is

$$\mathbb{E}[X] = 0 \cdot \mathbb{P}(X = 0) + 1 \cdot \mathbb{P}(X = 1) = 0 \cdot (1-p) + 1 \cdot p = p.$$

A quick example: the number of heads in one fair coin flip follows the Bernoulli(1/2) distribution.

2.3.3 Indicator Random Variables

Let $A \subseteq \Omega$ be an event. We define the **indicator of** A, $\mathbb{1}\{A\}$, to be the random variable

$$\mathbb{1}\{A\}(\omega) = \begin{cases} 0, & \omega \notin A, \\ 1, & \omega \in A. \end{cases}$$

Observe that $\mathbb{1}\{A\}$ follows the Bernoulli(p) distribution where $p = \mathbb{P}(A)$.

An important property of indicator random variables (and Bernoulli random variables) is that $X = X^k$ for any positive integer k. To see why this is true, note that X can only take on values in the set $\{0,1\}$. Since $0^k = 0$ and $1^k = 1$, then $X(\omega) = X^k(\omega)$ for all outcomes

$\omega \in \Omega$. We will use this property when we discuss the variance of probability distributions.

The expectation of the indicator random variable is

$$\boxed{\mathbb{E}[\mathbb{1}\{A\}] = \mathbb{P}(A)} \tag{2.12}$$

because it is a Bernoulli random variable with $p = \mathbb{P}(A)$.

2.3.4 Binomial Distribution

The **binomial distribution with parameters** n **and** p (where n is a positive integer and $p \in [0, 1]$), abbreviated Binomial(n, p), describes the number of successes when we conduct n independent trials, where each trial has a probability p of success. The binomial distribution is found by the following argument: the probability of having a series of trials with k successes (and therefore $n - k$ failures) is $p^k(1 - p)^{n-k}$. We need to multiply this expression by the number of ways to achieve k successes in n trials, which is $\binom{n}{k}$. Hence,

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \qquad x \in \{0, \ldots, n\}.$$

Exercise 11 Prove that the probabilities of the binomial distribution sum to 1.

Let us proceed to compute the expectation of this distribution. According to the formula,

$$\mathbb{E}[X] = \sum_x x\mathbb{P}(X = x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1 - p)^{n-x}.$$

This is quite a difficult sum to calculate! (Try it yourself and see if you can make any progress.) To make our work simpler, we will instead make a connection between the binomial distribution and the Bernoulli distribution we defined earlier. Let X_i be the indicator random variable for the event that trial i is a success, for $i = 1, \ldots, n$. (In the language of the previous section, $X_i = \mathbb{1}\{A_i\}$, where A_i is the event that trial i is a success.) The key insight lies in observing

$$X = X_1 + \cdots + X_n.$$

Each indicator variable X_i is 1 or 0 depending on whether trial i is a success, so if we sum up all of the indicator variables, then we obtain the total number of successes in all n trials. Therefore, compute

$$\mathbb{E}[X] = \mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n].$$

Notice that in the last line, we used linearity of expectation (Theorem 2.4). Now we can see why linearity of expectation is so powerful: combined with indicator variables, it allows us to break up the expectation of a complicated random variable into the sum of the expectations of simple random variables. Thus,

$$\mathbb{E}[X_i] = \mathbb{P}(\text{trial } i \text{ is a success}) = p.$$

Each term in the sum is simply p, and there are n such terms, so therefore

$$\mathbb{E}[X] = np.$$

The result should make intuitive sense: if you are conducting n trials, and the probability of success is p, then you expect a fraction p of the trials to be successes, which is saying that you expect np total successes. The expectation matches our intuition.

The random variables $X_1, \ldots, X_n$ are an example of i.i.d. random variables, which is a term that comes up very frequently: i.i.d. stands for **independent and identically distributed**. Since each trial is independent of each other by assumption, the variables $X_1, \ldots, X_n$ are independent, although we did not need independence in order to compute the expectation.

A strategy now emerges for tackling complicated expected value questions: when computing $\mathbb{E}[X]$, try to see if you can break down X into the sum of indicator random variables. Then, computing the expectation becomes much easier because you can take advantage of linearity.

Exercise 12 Throw n balls uniformly at random into m bins, where m and n are positive integers. What is the expected number of empty bins?

2.3.5 Geometric Distribution

The **geometric distribution with parameter** p, $p \in (0, 1)$, abbreviated Geometric(p), describes the number of trials required to obtain the first success, assuming that each trial is independent and has a probability of success p. If it takes exactly x trials to obtain the first success, there were first $x - 1$ failures (each with probability $1 - p$) and one success (with probability p). Hence, the distribution is

$$\mathbb{P}(X = x) = (1 - p)^{x-1} p, \qquad x \in \mathbb{Z}^+.$$

Exercise 13 Prove that the probabilities of the geometric distribution sum to 1.

When working with the geometric distribution, it is often easier to work with the tail probabilities $\mathbb{P}(X > x)$. In order for $X > x$ to hold, there must be at least x failures; hence,

$$\mathbb{P}(X > x) = (1 - p)^x, \qquad x \in \mathbb{N}.$$

Note that the tail probability is related to the CDF in the following way:

$$\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x).$$

The clever way to find the expectation of the geometric distribution uses a method known as the renewal method. $\mathbb{E}[X]$ is the expected number of trials until the first success. Suppose we carry out the first trial, and one of two outcomes occurs. With probability p, we obtain a success and we are done (it only took 1 trial until success). With probability $1 - p$, we

obtain a failure, and we are right back where we started. In the latter case, how many trials do we expect until our first success? The answer is $1+\mathbb{E}[X]$: we have already used one trial, and we expect $\mathbb{E}[X]$ more trials since nothing has changed from our original situation (the geometric distribution is *memoryless*). Hence,

$$\mathbb{E}[X] = p \cdot 1 + (1-p) \cdot (1 + \mathbb{E}[X]).$$

Solving this equation yields

$$\mathbb{E}[X] = \frac{1}{p},$$

which is also intuitive: if we have, say, a 1/100 chance of success on each trial, we would naturally expect 100 trials until our first success. (Note: If the method above does not seem rigorous to you, then worry not. We will revisit the method under the framework of conditional expectation in the future.)

Here is a more computational way to obtain the formula. We want to evaluate the sum

$$\mathbb{E}[X] = \sum_{x=1}^{\infty} x(1-p)^{x-1}p = p\sum_{x=0}^{\infty}(x+1)(1-p)^x.$$

By using the identity [1]

$$\sum_{k=0}^{\infty}(k+1)x^k = \frac{1}{(1-x)^2}$$

we can evaluate our original sum:

$$\mathbb{E}[X] = p \cdot \frac{1}{(1-(1-p))^2} = p \cdot \frac{1}{p^2} = \frac{1}{p}.$$

Exercise 14 Compute $\mathbb{E}[X]$ again, where $X \sim \text{Geometric}(p)$, using the tail sum formula (Theorem 2.6).

We can show that the minimum of independent geometric random variables is geometric:

[1] To derive the identity, we start with the following identity for the sum of an infinite geometric series:

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}, \qquad |x| < 1. \tag{2.13}$$

Multiply both sides of the identity by x:

$$\sum_{k=0}^{\infty} x^{k+1} = \frac{x}{1-x}$$

Then, differentiate both sides with respect to x:

$$\sum_{k=0}^{\infty}(k+1)x^k = \frac{1}{(1-x)^2}$$

Theorem 2.8 (Minimum of Independent Geometric RVs)**.** *Let* $X \sim \text{Geometric}(p)$ *and* $Y \sim \text{Geometric}(q)$ *be independent random variables. Then,*

$$\min\{X, Y\} \sim \text{Geometric}(p + q - pq).$$

Proof. We can use the tail probabilities to simplify the derivation. If the minimum of X and Y is greater than z, then both X and Y are greater than z. Hence,

$$\begin{aligned}\mathbb{P}(\min\{X, Y\} > z) &= \mathbb{P}(X > z, Y > z) = \mathbb{P}(X > z)\mathbb{P}(Y > z) = (1-p)^z(1-q)^z \\ &= \big(1 - (p + q - pq)\big)^z.\end{aligned}$$

We recognize the last expression as the tail probability of a geometric random variable with parameter $p + q - pq$. □

Another way of thinking about the above result is to imagine that if X represents the number of trials until experiment 1 succeeds, and Y represents the number of trials until experiment 2 succeeds, then $\min\{X, Y\}$ represents the number of trials until either experiment 1 or experiment 2 succeeds. On each trial, we have a success from X with probability p and a success from Y with probability q, independently of each other. By the inclusion-exclusion rule, the probability that we have a success from either experiment is $p+q-pq$, and different trials are independent from each other by assumption. Hence, the number of trials until the first success from either experiment is geometric with parameter $p + q - pq$.

Example 2.9 (Coupon Collector's Problem)**.** There are n different coupons that you would like to collect, where n is a positive integer. Every time you buy an item from the store, you receive a random coupon (each of the n coupons is equally likely to appear). What is the expected number of items you must buy before you collect every coupon?

Let T_i be the number of items it requires to collect the ith new coupon, for $i = 1, \ldots, n$. In other words, starting from when you have seen $i - 1$ distinct coupons, T_i represents the additional number of items you must purchase before you see a coupon you have not seen before. T, the total time to collect all coupons, is $T := \sum_{i=1}^{n} T_i$.

Once we have collected $i - 1$ coupons, there are $n - i + 1$ coupons we have not seen yet, so the probability that the next item we buy comes with a coupon we have not seen is $(n - i + 1)/n$. If we regard each object bought as an independent trial, then we see that $T_i \sim \text{Geometric}(p)$, where $p = (n - i + 1)/n$. By linearity of expectation (Theorem 2.4),

$$\mathbb{E}[T] = \sum_{i=1}^{n} \mathbb{E}[T_i] = \sum_{i=1}^{n} \frac{n}{n - i + 1} = n \sum_{i=1}^{n} \frac{1}{i} = nH_n,$$

where H_n is the nth harmonic sum, $H_n := \sum_{i=1}^{n} i^{-1}$. A good approximation to H_n is

$\ln n + \gamma$, where γ is the **Euler-Mascheroni constant**, defined to be

$$\gamma := \lim_{n\to\infty} (H_n - \ln n). \tag{2.14}$$

γ has the numerical value $\gamma \approx 0.577$. Using this approximation, $\mathbb{E}[T] \approx n(\ln n + \gamma)$, or more simply, $\mathbb{E}[T] \in \Theta(n \log n)$.

2.3.6 Memoryless Property

An important property of the geometric distribution is that it is **memoryless**, which is to say that a random variable following the geometric distribution only depends on its current state and not on its past. To make this notion formal, we shall show:

Theorem 2.10 (Memoryless Property)**.** *The geometric distribution satisfies, for all $s, t \in \mathbb{N}$ with $s < t$,*

$$\mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t).$$

Proof.

$$\begin{aligned}\mathbb{P}(X > s + t \mid X > s) &= \frac{\mathbb{P}(X > s + t, X > s)}{\mathbb{P}(X > s)} = \frac{\mathbb{P}(X > s + t)}{\mathbb{P}(X > s)} = \frac{(1-p)^{s+t}}{(1-p)^s} \\ &= (1-p)^t = \mathbb{P}(X > t).\end{aligned}$$

□

Intuitively, the theorem says: suppose you have already tried flipping a coin s times, without success. The probability that it takes you at least t more coin flips until your first success is the *same* as the probability that your friend picks up a coin and it takes him/her at least t coin flips. Moral of the story: the geometric distribution does not care how many times you have already flipped the coin, because it is *memoryless*.

2.3.7 Negative Binomial Distribution

We can consider a slight generalization of the geometric distribution: if we have independent trials, with probability of success p, how many trials do we need until we obtain k successes? Consider the probability that we require x trials: we have k successes and $x - k$ failures, and the probability of any sequence of k successes and $x - k$ failures is $p^k(1-p)^{x-k}$. Now, a counting argument gives the number of such sequences: the last success must occur on the xth trial, and there are $\binom{x-1}{k-1}$ ways to distribute the $k - 1$ remaining successes among the other trials. Hence, the **negative binomial distribution with parameter p of order k**, where $p \in [0, 1]$ and k is a positive integer, is

$$\mathbb{P}(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \qquad x = k, k+1, \ldots .$$

When $k = 1$, we recover the geometric distribution. To compute the expectation, we use linearity of expectation (Theorem 2.4) once again: let X_i be the number of trials it takes to

obtain the ith success, starting after we have already observed $i-1$ successes, for $i = 1, \ldots, n$. Then $X = \sum_{i=1}^{k} X_i$, where $X_i \sim$ Geometric(p). Using the expectation of the geometric distribution,

$$\mathbb{E}[X] = \sum_{i=1}^{k} \mathbb{E}[X_i] = \frac{k}{p}.$$

2.3.8 Poisson Distribution

The **Poisson distribution with parameter** λ, $\lambda > 0$, abbreviated Poisson(λ), can be viewed as an approximation to the binomial distribution in a certain regime: let the number of trials, n, approach infinity and the probability of success per trial, p, approach 0, such that the mean $\mathbb{E}[X] = np$ remains a fixed value λ. These assumptions also tell us when the approximation is reasonable: the probability of success should be low, and the number of trials should be high, such that the product np is roughly between 1 and 10. Under these conditions, x is typically small compared to n. The probability distribution is

$$\begin{aligned}\mathbb{P}(X = x) &= \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \frac{n(n-1)\cdots(n-x+2)(n-x+1)p^x}{x!}(1-p)^{n-x} \approx \frac{n^x p^x}{x!}\Big(1 - \frac{\lambda}{n}\Big)^n\end{aligned}$$

(we neglected the term $(1-p)^{-x}$ since as $n \to \infty$, $p \to 0$, and the term $(1-p)^{-x} \to 1$)

$$\to \frac{\lambda^x \exp(-\lambda)}{x!},$$

where in the last line, we have used the identity

$$\lim_{n\to\infty} \Big(1 + \frac{x}{n}\Big)^n = \exp x. \tag{2.15}$$

The previous section was to motivate the form of the Poisson distribution. We now define the Poisson distribution:

$$\mathbb{P}(X = x) = \frac{\lambda^x \exp(-\lambda)}{x!}, \qquad x \in \mathbb{N}.$$

The fact that the probabilities sum to 1 follows from the identity [2]

$$\exp x = \sum_{k=0}^{\infty} \frac{x^k}{k!}. \tag{2.16}$$

The expectation of the Poisson distribution is, as we would expect, λ. We prove it:

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x \frac{\lambda^x \exp(-\lambda)}{x!} = \lambda \exp(-\lambda) \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda.$$

[2] In many sources, the power series is taken as the definition of $\exp x$. The power series converges everywhere.

2.3.9 Poisson Merging

We prove an important fact about the sums of independent Poisson random variables.

Theorem 2.11 (Poisson Merging)**.** *Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ be independent random variables. Then*

$$X + Y \sim \text{Poisson}(\lambda + \mu).$$

Proof. We will compute the distribution of $X + Y$ and show that it is Poisson (using independence).

$$\begin{aligned}
\mathbb{P}(X+Y=z) &= \sum_{j=0}^{z} \mathbb{P}(X=j, Y=z-j) = \sum_{j=0}^{z} \frac{\lambda^j \exp(-\lambda)}{j!} \frac{\mu^{z-j} \exp(-\mu)}{(z-j)!} \\
&= \frac{\exp(-(\lambda+\mu))}{z!} \sum_{j=0}^{z} \frac{z!}{j!(z-j)!} \lambda^j \mu^{z-j} = \frac{\exp(-(\lambda+\mu))}{z!} \sum_{j=0}^{z} \binom{z}{j} \lambda^j \mu^{z-j} \\
&= \frac{(\lambda+\mu)^z \exp(-(\lambda+\mu))}{z!} = \mathbb{P}\big(\text{Poisson}(\lambda+\mu) = z\big).
\end{aligned}$$

In the last line, we have used the binomial theorem (Theorem 0.4). □

Remarks: Linearity of expectation tells us that $\mathbb{E}[X+Y] = \mathbb{E}[X]+\mathbb{E}[Y] = \lambda+\mu$. This theorem tells us something stronger, namely that the *distribution* of $X+Y$ remains Poisson, with parameter $\lambda+\mu$. By induction, it follows that for any positive integer n, if the random variables $X_i \sim \text{Poisson}(\mu_i)$ for $i \in \{1, \ldots, n\}$ are independent, then $\sum_{i=1}^{n} X_i \sim \text{Poisson}(\sum_{i=1}^{n} \mu_i)$.

2.3.10 Poisson Splitting

Here, we introduce another important property of the Poisson distribution known as the **Poisson splitting** property, which will be proven first and motivated later.

Theorem 2.12 (Poisson Splitting)**.** *Suppose that $X \sim \text{Poisson}(\lambda)$ and that conditioned on $X = x$, Y follows the $\text{Binomial}(x, p)$ distribution and Z follows the $\text{Binomial}(x, 1-p)$ distribution, such that $Y + Z = X$. Then $Y \sim \text{Poisson}(\lambda p)$, $Z \sim \text{Poisson}(\lambda(1-p))$, and Y and Z are independent.*

Proof. We use the definition of conditional probability and show that Y has the correct distribution. Notice that the sum starts from $x = y$ because $X \geq Y$. For $y \in \mathbb{N}$,

$$\mathbb{P}(Y=y) = \sum_{x=y}^{\infty} \mathbb{P}(X=x, Y=y) = \sum_{x=y}^{\infty} \mathbb{P}(X=x)\mathbb{P}(Y=y \mid X=x)$$

$$= \sum_{x=y}^{\infty} \frac{\lambda^x \exp(-\lambda)}{x!} \binom{x}{y} p^y (1-p)^{x-y} = \exp(-\lambda) \sum_{x=y}^{\infty} \frac{\lambda^x}{x!} \frac{x!}{y!(x-y)!} p^y (1-p)^{x-y}$$
$$= \frac{(\lambda p)^y \exp(-\lambda)}{y!} \sum_{x=y}^{\infty} \frac{(\lambda(1-p))^{x-y}}{(x-y)!} = \frac{(\lambda p)^y \exp(-\lambda)}{y!} \exp\big(\lambda(1-p)\big)$$
$$= \frac{(\lambda p)^y \exp(-\lambda p)}{y!} = \mathbb{P}\big(\text{Poisson}(\lambda p) = y\big).$$

This proves that $Y \sim \text{Poisson}(\lambda p)$. Since the above proof holds for all $p \in (0,1)$, it also holds with p replaced by $1-p$, and so we conclude that $Z \sim \text{Poisson}(\lambda(1-p))$. Now, to show that Y and Z are independent, we calculate for $y, z \in \mathbb{N}$:

$$\begin{aligned}
\mathbb{P}(Y = y, Z = z) &= \mathbb{P}(X = y+z, Y = y, Z = z) \\
&= \mathbb{P}(X = y+z)\mathbb{P}(Y = y, Z = z \mid X = y+z) \\
&= \frac{\lambda^{y+z} \exp(-\lambda)}{(y+z)!} \frac{(y+z)!}{y!z!} p^y (1-p)^z \\
&= \frac{\lambda^y \exp(-\lambda p)}{y!} \frac{(\lambda(1-p))^z \exp(-\lambda(1-p))}{z!} \\
&= \mathbb{P}(Y = y)\mathbb{P}(Z = z). \qquad \square
\end{aligned}$$

Think about the Poisson distribution as representing the number of arrivals of some process. Then, for each arrival, we flip a coin with bias p and keep the arrival only if the coin came up heads; the above theorem says that the number of arrivals for this new process is also Poisson, with mean λp.

As an example: suppose that the number of calls that a calling center receives per hour is distributed according to a Poisson distribution with mean λ. Furthermore, suppose that each call that the calling center receives is independently a telemarketer with probability p (therefore, the distribution of telemarketing calls is binomial, conditioned on the number of calls received). Then, the number of telemarketing calls that the calling receives per hour (unconditional) follows a Poisson distribution with mean λp.

Exercise 15 Suppose X, Y are independent Poisson random variables with parameters λ and μ respectively. What is the conditional distribution of Y conditioned on $X + Y = n$, where $n \in \mathbb{N}$?

2.4 Solutions to Exercises

Exercise 10 This is an example where it is easier to work with tail probabilities. Note that $\mathbb{P}(X \geq 2) = 1$ since the minimum number of balls thrown is 2. For $x = 3, \ldots, m+1$, the probability of throwing at least x balls is $\prod_{k=1}^{x-2}(1 - k/m)$. (When you have thrown k balls before, then k bins are occupied, and the probability of landing in an unoccupied bin

is $1 - k/m$. Since you must throw $x - 1$ balls without collisions, the last probability to be multiplied is $1 - (x-2)/m$.) Specifying the tail probabilities is actually enough to fully specify the distribution of X, but let us calculate $\mathbb{P}(X = x)$. For $x = 2, \ldots, m+1$,

$$\mathbb{P}(X = x) = \mathbb{P}(X \geq x) - \mathbb{P}(X \geq x+1) = \frac{x-1}{m} \prod_{k=1}^{x-2} \Big(1 - \frac{k}{m}\Big).$$

Exercise 11 This is an immediate consequence of the binomial theorem (Theorem 0.4), which states: $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$ for $n \in \mathbb{N}$.

Exercise 12 Let X_i be the indicator that the ith bin is empty, for $i = 1, \ldots, m$. The number of empty bins is then $\sum_{i=1}^{m} X_i$. Also, $\mathbb{E}[X_i]$ is the probability that the ith bin is empty. Since the probability that any specific ball avoids bin i is $1 - 1/m$, the probability that all of the balls avoid bin i (that is, the probability that bin i is empty) is $(1 - 1/m)^n$. Hence, by linearity of expectation (Theorem 2.4), $\mathbb{E}[\sum_{i=1}^{m} X_i] = \sum_{i=1}^{m} \mathbb{E}[X_i] = m(1 - 1/m)^n$.

Exercise 13 This follows from the infinite geometric series formula (2.13). Applying this formula, $p \sum_{x=1}^{\infty} (1-p)^{x-1} = p/(1 - (1-p)) = 1$.

Exercise 14 $\sum_{x=1}^{\infty} \mathbb{P}(X \geq x) = \sum_{x=1}^{\infty} (1-p)^{x-1} = 1/(1 - (1-p)) = 1/p$.

Exercise 15 Note that $X + Y \sim \text{Poisson}(\lambda + \mu)$ because of Poisson merging (Theorem 2.11). For $y \in \{0, \ldots, n\}$, using the independence of X and Y,

$$\begin{aligned}
\mathbb{P}(Y = y \mid X + Y = n) &= \frac{\mathbb{P}(Y = y, X + Y = n)}{\mathbb{P}(X + Y = n)} = \frac{\mathbb{P}(X = n - y, Y = y)}{\mathbb{P}(X + Y = n)} \\
&= \frac{\mathbb{P}(X = n - y)\mathbb{P}(Y = y)}{\mathbb{P}(X + Y = n)} \\
&= \frac{\lambda^{n-y} \exp(-\lambda)(n-y)!^{-1} \mu^y \exp(-\mu) y!^{-1}}{(\lambda+\mu)^n \exp(-(\lambda+\mu)) n!^{-1}} \\
&= \binom{n}{y} \Big(\frac{\mu}{\lambda+\mu}\Big)^y \Big(\frac{\lambda}{\lambda+\mu}\Big)^{n-y},
\end{aligned}$$

so we recognize that $Y \mid X + Y = n \sim \text{Binomial}(n, \mu/(\lambda+\mu))$.

Chapter 3

Variance & Inequalities

Previously, we have discussed the expectation of a random variable, which is a measure of the center of the probability distribution. Today, we discuss the variance, which is a measure of the *spread* of the distribution Variance, in a sense, is a measure of how unpredictable your results are. We will also cover some important inequalities for bounding tail probabilities.

3.1 Variance

Suppose your friend offers you a choice: you can either accept \$1 immediately, or you can enter in a raffle in which you have a 1/100 chance of winning a \$100 payoff. The expected value of each of these deals is simply \$1, but clearly the offers are very different in nature! We need another measure of the probability distribution that will capture the idea of *variability* or *risk* in a distribution. We are now interested now in how often a random variable will take on values close to the mean.

Perhaps we could study the quantity $X - \mathbb{E}[X]$ (the difference between what we expected and what we actually measured), but we quickly notice a problem:

$$\mathbb{E}\big[X - \mathbb{E}[X]\big] = \mathbb{E}[X] - \mathbb{E}[X] = 0. \tag{3.1}$$

The expectation of this quantity is always 0, no matter the distribution! Every random variable (except the constant random variable) can take on values above or below the mean by definition, so studying the average of the differences is not interesting. To address this problem, we could study $|X - \mathbb{E}[X]|$ (thereby making all differences positive), but in practice, it becomes much harder to analytically solve problems using this quantity. Instead, we will study a quantity known as the variance of the probability distribution:

Definition 3.1. The **variance** of a probability distribution is:

$$\boxed{\operatorname{var} X = \mathbb{E}[(X - \mathbb{E}[X])^2]} \tag{3.2}$$

Remark: We often denote the mean of the probability distribution as μ, and the variance as σ^2. We call $\sigma = \sqrt{\operatorname{var} X}$ the **standard deviation** of X. The standard deviation is useful

because it has the same units as X, allowing for easier comparison. As an example, if X represents the height of an individual, then σ_X would have units of meters, while $\operatorname{var} X$ has units of meters^2.

3.1.1 The Computational Formula

The explicit formula for variance is:

$$\boxed{\operatorname{var} X = \sum_x (x - \mathbb{E}[X])^2 \mathbb{P}(X = x)} \tag{3.3}$$

In practice, however, we tend to use the following formula to calculate variance:

Theorem 3.2 (Computational Formula for Variance)**.** *The variance of X is:*

$$\boxed{\operatorname{var} X = \mathbb{E}[X^2] - \mathbb{E}[X]^2} \tag{3.4}$$

Proof. We use linearity of expectation (note that $\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X]$ since $\mathbb{E}[X]$ is just a constant):

$$\begin{aligned}\mathbb{E}[(X - \mathbb{E}[X])^2] &= \mathbb{E}\big[X^2 - 2X\,\mathbb{E}[X] + \mathbb{E}[X]^2\big] = \mathbb{E}[X^2] - 2\,\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2.\end{aligned}$$

□

This formula will be extremely useful to us throughout the course, so please memorize it!

3.1.2 Properties of Variance

We next examine some useful properties of the variance.

Theorem 3.3 (Properties of Variance)**.** *Let X be a random variable, $a \in \mathbb{R}$ be a constant, and c be the constant random variable. Then:*

$$\boxed{\operatorname{var}(aX + c) = a^2 \operatorname{var} X} \tag{3.5}$$

Proof. We use the computational formula:

$$\begin{aligned}\operatorname{var}(aX + c) &= \mathbb{E}[(aX + c)^2] - \mathbb{E}[aX + c]^2 = \mathbb{E}[a^2X^2 + 2acX + c^2] - (a\,\mathbb{E}[X] + c)^2 \\ &= a^2\,\mathbb{E}[X^2] + 2ac\,\mathbb{E}[X] + c^2 - a^2\,\mathbb{E}[X]^2 - 2ac\,\mathbb{E}[X] - c^2 \\ &= a^2(\mathbb{E}[X^2] - \mathbb{E}[X]^2) = a^2 \operatorname{var} X\end{aligned}$$

□

Observe that adding a constant does not change the variance. Intuitively, adding a constant shifts the distribution to the left or right, but does not affect its shape (and therefore its spread). On the other hand, scaling by a constant *does* scale the variance.

Corollary 3.4 (Scaling of the Standard Deviation)**.** *Let X be a random variable, $a \in \mathbb{R}$ be a constant, and c be the constant random variable. Then:*

$$\boxed{\sigma_{aX+c} = |a|\sigma_X} \tag{3.6}$$

Proof. The corollary follows immediately from taking the square root of the result in Theorem 3.3. □

We saw that linearity of expectation was an extremely powerful tool for computing expectation values. We would like to have a similar property hold for variance, but additivity of variance does not hold *in general*. However, we have the following useful theorem:

Theorem 3.5 (Variance of Sums of Random Variables)**.** *Let X and Y be random variables. Then:*

$$\boxed{\operatorname{var}(X+Y) = \operatorname{var} X + \operatorname{var} Y + 2(\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y])} \tag{3.7}$$

Proof. As always, we start with the computational formula for variance.

$$\begin{aligned}
\operatorname{var}(X+Y) &= \mathbb{E}[(X+Y)^2] - \mathbb{E}[X+Y]^2 = \mathbb{E}[X^2 + 2XY + Y^2] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\
&= \mathbb{E}[X^2] + 2\,\mathbb{E}[XY] + \mathbb{E}[Y]^2 - \mathbb{E}[X]^2 - 2\,\mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[Y]^2 \\
&= (\mathbb{E}[X^2] - \mathbb{E}[X]^2) + (\mathbb{E}[Y]^2 - \mathbb{E}[Y]^2) + 2(\mathbb{E}[XY] - 2\,\mathbb{E}[X]\,\mathbb{E}[Y]) \\
&= \operatorname{var} X + \operatorname{var} Y + 2(\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y])
\end{aligned}$$

□

We will reveal the importance of the term $\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$ later in the course. For now, we are more interested in the corollary:

Corollary 3.6 (Variance of Independent Random Variables)**.** *Let X and Y be independent. Then:*

$$\operatorname{var}(X+Y) = \operatorname{var} X + \operatorname{var} Y \tag{3.8}$$

Proof. When X and Y are independent, $\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0$ according to Theorem 2.5. □

In general, for any positive integer n, if $X_1, \ldots, X_n$ are pairwise independent random variables, then

$$\operatorname{var}(X_1 + \cdots + X_n) = \operatorname{var} X_1 + \cdots + \operatorname{var} X_n. \tag{3.9}$$

Remark: Observe that the assumption of pairwise independence can be replaced by the condition that $X_1, \ldots, X_n$ are *pairwise uncorrelated*, which means $\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\,\mathbb{E}[X_j]$ for every pair (i, j).

Exercise 16 Let X and Y be independent. Show that

$$\operatorname{var}(XY) = \mathbb{E}[X]^2 \operatorname{var} Y + \mathbb{E}[Y]^2 \operatorname{var} X + (\operatorname{var} X)(\operatorname{var} Y).$$

In particular, observe that $\operatorname{var}(XY) \neq (\operatorname{var} X)(\operatorname{var} Y)$ in general.

3.2 Probability Distributions Revisited

We will revisit the probability distributions introduced last time and proceed to calculate their variances.

3.2.1 Uniform Distribution

Recall that for any positive integer n, if $X \sim \text{Uniform}\{1, \ldots, n\}$, then $\mathbb{E}[X] = (n+1)/2$.

$$\mathbb{E}[X^2] = \sum_{x=1}^{n} x^2 \cdot \frac{1}{n} = \frac{1}{n} \sum_{x=1}^{n} x^2 = \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6},$$

where we have used the identity (verified using induction)

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}. \tag{3.10}$$

The variance is calculated to be (after a little algebra):

$$\operatorname{var} X = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}.$$

3.2.2 Bernoulli Distribution & Indicator Random Variables

Recall that if $X \sim \text{Bernoulli}(p)$, $\mathbb{E}[X] = p$. Additionally, recall that Bernoulli random variables satisfy the important property $X^2 = X$. Hence, we have

$$\mathbb{E}[X^2] = \mathbb{E}[X] = p.$$

The variance of a Bernoulli random variable is

$$\operatorname{var} X = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1-p).$$

In the special case of indicator random variables, $p = \mathbb{P}(A)$ and:

$$\boxed{\operatorname{var} \mathbb{1}\{A\} = \mathbb{P}(A)\big(1 - \mathbb{P}(A)\big)} \tag{3.11}$$

Observe that $p(1-p)$ is maximized when $p = 1/2$, and it attains a maximum vaue of $1/4$ (to convince yourself, draw the parabola). We will make use of this observation later.

3.2.3 Binomial Distribution

Recall that if $X \sim \text{Binomial}(n, p)$, X is the sum of i.i.d. Bernoulli(p) random variables:

$$X = X_1 + \cdots + X_n.$$

Hence,

$$\text{var}\, X = \text{var}(X_1 + \cdots + X_n) = \text{var}\, X_1 + \cdots + \text{var}\, X_n,$$

where we have used the independence of the indicator random variables to apply Corollary 3.6. Since each indicator random variable follows the Bernoulli(p) distribution,

$$\text{var}\, X = n\, \text{var}\, X_1 = np(1-p).$$

3.2.4 Computing the Variance of Dependent Indicators

Unlike when we computed the mean of the binomial distribution (which did not require any assumptions except that the binomial distribution could be written as the sum of indicators), the calculation of the variance of the binomial distribution relied on a crucial fact: the indicator variables were *independent*. In this section, we outline a general method for computing the variance of random variables which can be written as the sum of indicators, even when the indicators are not independent.

Let X be written as the sum of n identically distributed indicators, where n is a positive integer, and the indicators are *not* assumed to be independent:

$$X = \mathbb{1}\{A_1\} + \cdots + \mathbb{1}\{A_n\}$$

We first note that the expectation is easy, thanks to linearity of expectation (which holds regardless of whether the indicator random variables are independent or not):

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[\mathbb{1}\{A_i\}] = \sum_{i=1}^{n} \mathbb{P}(A_i) \tag{3.12}$$

Using the fact that the indicator variables are identically distributed:

$$\boxed{\mathbb{E}[X] = n\mathbb{P}(A_1)} \tag{3.13}$$

Next, we compute $\mathbb{E}[X^2]$:

$$\mathbb{E}[X^2] = \mathbb{E}[(\mathbb{1}\{A_1\} + \cdots + \mathbb{1}\{A_n\})^2]$$

Observe that the square $(\mathbb{1}\{A_1\} + \cdots + \mathbb{1}\{A_n\})^2$ has two types of terms:

1. There are *like-terms*, such as $\mathbb{1}\{A_1\}^2$. However, we know from the properties of indicators that $\mathbb{1}\{A_i\}^2 = \mathbb{1}\{A_i\}$. There are n of these terms in total:

$$\sum_{i=1}^{n} \mathbb{1}\{A_i\}^2 = \sum_{i=1}^{n} \mathbb{1}\{A_i\} = \mathbb{1}\{A_1\} + \cdots + \mathbb{1}\{A_n\} = X \tag{3.14}$$

2. Then, there are *cross-terms*, such as $\mathbb{1}\{A_1\}\,\mathbb{1}\{A_2\}$. There are n^2 total terms in the square, and n of those terms are like-terms, which leaves $n^2-n=n(n-1)$ cross-terms. We usually write the sum:

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{1}\{A_i\}\,\mathbb{1}\{A_j\}=\mathbb{1}\{A_1\}\,\mathbb{1}\{A_2\}+\cdots+\mathbb{1}\{A_{n-1}\}\,\mathbb{1}\{A_n\}$$

We can discover more about the cross-terms by examining their meaning. Consider the term $\mathbb{1}\{A_i\}\,\mathbb{1}\{A_j\}$: it is the product of two indicators. Each indicator $\mathbb{1}\{A_i\}$ is either 0 or 1; therefore, their product is also 0 or 1, which suggests that the product is also an indicator! The product is 1 if and only if each indicator is 1, which in the language of probability is expressed as

$$\mathbb{P}(\mathbb{1}\{A_i\}\,\mathbb{1}\{A_j\}=1)=\mathbb{P}(\mathbb{1}\{A_i\}=1,\mathbb{1}\{A_j\}=1)=\mathbb{P}(A_i\cap A_j). \tag{3.15}$$

We have arrived at a crucial fact: *the product of two indicators* $\mathbb{1}\{A_i\}$ *and* $\mathbb{1}\{A_j\}$ *is itself an indicator for the event* $A_i\cap A_j$. Therefore, we can rewrite the sum:

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{1}\{A_i\}\,\mathbb{1}\{A_j\}=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{1}\{A_i\cap A_j\}. \tag{3.16}$$

Putting it together, we have that

$$X^2=\Big(\sum_{i=1}^{n}\mathbb{1}\{A_i\}\Big)^2=X+\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{1}\{A_i\cap A_j\}. \tag{3.17}$$

The expectation of the square is

$$\mathbb{E}[X^2]=\mathbb{E}[X]+\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}[\mathbb{1}\{A_i\cap A_j\}]=n\mathbb{P}(A_i)+n(n-1)\mathbb{P}(A_i\cap A_j). \tag{3.18}$$

(For simplicity, we made the assumption that all of the intersection probabilities $\mathbb{P}(A_i\cap A_j)$ are the same.) Finally, the variance is $\mathbb{E}[X^2]-\mathbb{E}[X]^2$, or:

$$\boxed{\operatorname{var}X=n\mathbb{P}(A_1)+n(n-1)\mathbb{P}(A_1\cap A_2)-n^2\mathbb{P}(A_1)^2} \tag{3.19}$$

Although the resulting formula looks rather complicated, it is a remarkably powerful demonstration of the techniques we have developed so far. The path we have taken is an amusing one: when the indicators are not independent, additivity of variance fails to hold, so the tool we ended up relying on was... linearity of expectation and indicators!

Example 3.7 (Fixed Points)**.** Suppose we take n hats from n people (n is a positive integer), shuffle the hats around randomly, and then hand a hat back to each person. What is the expected number of people X who receive their own hat back?

We can apply linearity of expectation. If we let X_i denote the indicator that the ith person receives his or her hat back, for $i=1,\ldots,n$, then $\mathbb{E}[X_i]=1/n$ since each person

receives his or her own hat back with probability $1/n$. (To convince yourself of this, note that there are $n!$ possible arrangements of the hats, and if we demand that person i receives his or her hat back, then there are $(n-1)!$ ways to rearrange the other hats. Thus, the probability is $(n-1)!/n! = 1/n$.) By linearity of expectation, $\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = 1$.

What about the variance of X? We can apply (3.19). $A_i \cap A_j$ is the event that person i and person j both receive their hats back, which occurs with probability $1/(n(n-1))$. (Again, think about the permutations: there are $n!$ total arrangements of hats, and if we give person i and person j their hats back, there are $(n-2)!$ ways to rearrange the other hats. The probability is $(n-2)!/n! = 1/(n(n-1))$.) Applying the formula, we have

$$\operatorname{var} X = n \cdot \frac{1}{n} + n(n-1) \cdot \frac{1}{n(n-1)} - n^2 \cdot \frac{1}{n^2} = 1 + 1 - 1 = 1.$$

3.2.5 Geometric Distribution

Next, we compute the variance of the geometric distribution. (It will mostly be an exercise in manipulating series, but we include it for completeness.) Recall that if $X \sim \text{Geometric}(p)$, $\mathbb{E}[X] = 1/p$. We compute $\mathbb{E}[X^2] = p \sum_{x=1}^{\infty} x^2(1-p)^{x-1}$. Recall the identity

$$\sum_{k=1}^{\infty} k x^{k-1} = \frac{1}{(1-x)^2}, \qquad |x| < 1. \tag{3.20}$$

From this formula, we can derive the identity [1]

$$\sum_{k=1}^{\infty} k^2 x^{k-1} = \frac{1+x}{(1-x)^3}, \qquad |x| < 1.$$

Setting $x = 1 - p$ yields

$$\sum_{k=1}^{\infty} k^2 (1-p)^{k-1} = \frac{2-p}{p^3}.$$

[1] Starting with (3.20), we shift the index k by 1 to obtain

$$\sum_{k=0}^{\infty} (k+1) x^k = \frac{1}{(1-x)^2}.$$

Differentiating this equation with respect to x yields

$$\sum_{k=1}^{\infty} k(k+1) x^{k-1} = \frac{2}{(1-x)^3}. \tag{3.21}$$

Subtracting (3.20) from (3.21) now yields

$$\sum_{k=1}^{\infty} k^2 x^{k-1} = \frac{2}{(1-x)^3} - \frac{1}{(1-x)^2} = \frac{1+x}{(1-x)^3}.$$

So, we obtain

$$\mathbb{E}[X^2] = p \cdot \frac{2-p}{p^3} = \frac{2-p}{p^2}$$

and the variance is computed to be

$$\operatorname{var} X = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

3.2.6 Negative Binomial Distribution

Recall that the negative binomial distribution is the number of trials until the kth success for a positive integer k, where each trial is independent with probability of success p. To calculate the expectation, we wrote $X = \sum_{i=1}^{k} X_i$, where X_i is the number of trials it takes to observe the ith success, starting from $i-1$ successes. Each X_i has an independent Geometric(p) distribution, so we can write the variance of X as the sum of the variances:

$$\operatorname{var} X = \sum_{i=1}^{k} \operatorname{var} X_i = \frac{k(1-p)}{p^2}.$$

3.2.7 Poisson Distribution

Once again, computing the variance of the Poisson distribution will be an exercise in manipulating complicated sums (with one clever trick). Recall that if $X \sim \text{Poisson}(\lambda)$, $\mathbb{E}[X] = \lambda$. We will proceed to calculate $\mathbb{E}[X(X-1)]$ instead.

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{x=2}^{\infty} x(x-1)\frac{\lambda^x \exp(-\lambda)}{x!} = \lambda^2 \exp(-\lambda) \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} = \lambda^2 \exp(-\lambda) \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\ &= \lambda^2.\end{aligned}$$

Hence, by linearity of expectation,

$$\operatorname{var} X = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

3.3 Inequalities

Often, probability distributions can be difficult to compute exactly, so we will cover a few important bounds.

3.3.1 Markov's Inequality

The following inequality is quite flexible because it can be applied with any non-negative increasing function f. The only information we require is $\mathbb{E}[f(X)]$.

Theorem 3.8 (Markov's Inequality)**.** *Let X be a random variable, f be an increasing, non-negative function, and $a \in \mathbb{R}$ such that $f(a) \neq 0$. Then:*

$$\boxed{\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[f(X)]}{f(a)}} \tag{3.22}$$

Proof. Let $\mathbb{1}\{X \geq a\}$ be the indicator that $X \geq a$ and define $h(X) = f(a)\,\mathbb{1}\{X \geq a\}$. We claim that $h(X) \leq f(X)$ always:

1. If $X < a$, then $\mathbb{1}\{X \geq a\} = 0$, so $h(X) = 0 \leq f(X)$ (since f is non-negative).
2. If $X \geq a$, then $h(X) = f(a) \leq f(X)$ (since f is increasing).

Then, we have

$$\mathbb{E}[f(X)] \geq \mathbb{E}[h(X)] = \mathbb{E}[f(a)\,\mathbb{1}\{X \geq a\}] = f(a)\,\mathbb{E}[\mathbb{1}\{X \geq a\}] = f(a)\mathbb{P}(X \geq a). \quad \square$$

Corollary 3.9 (Weak Markov's Inequality)**.** *Let X be a non-negative random variable and $a > 0$. Then*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}. \tag{3.23}$$

Proof. The proof is immediate because $f(x) = x$ is an increasing function. $\square$

Example 3.10. Here is a quick example showing that Markov's inequality is tight. Let X take on the value $a > 0$ with probability p, and 0 otherwise. Then $\mathbb{E}[X] = ap$ and Markov's inequality gives $\mathbb{P}(X \geq a) \leq p$, which is tight.

3.3.2 Chebyshev's Inequality

We will use Markov's inequality to derive another inequality which uses the variance to bound the probability distribution. Chebyshev's inequality is useful for deriving confidence intervals and estimating sample sizes.

Theorem 3.11 (Chebyshev's Inequality)**.** *Let X be a random variable and $a > 0$. Then:*

$$\boxed{\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\operatorname{var} X}{a^2}} \tag{3.24}$$

Proof. Let $Y = |X - \mathbb{E}[X]|$ and $f(y) = y^2$. Since f is an increasing function, apply Markov's inequality:

$$\mathbb{P}(Y \geq a) \leq \frac{\mathbb{E}[Y^2]}{a^2}$$

Note, however, that $\mathbb{E}[Y^2] = \mathbb{E}[|X - \mathbb{E}[X]|^2] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \operatorname{var} X$. This completes the proof. □

Corollary 3.12 (Another Look at Chebyshev's Inequality)**.** *Let X be a random variable with standard deviation σ and $k > 0$. Then:*

$$\boxed{\mathbb{P}(|X - \mathbb{E}[X]| \geq k\sigma) \leq \frac{1}{k^2}} \tag{3.25}$$

Proof. Set $a = k\sigma$ in Chebyshev's inequality. □

Notably, Chebyshev's inequality justifies why we call the variance a measure of the spread of the distribution. The probability that X lies more than k standard deviations away from the mean is bounded by $1/k^2$, which is to say that a larger standard deviation means X is more likely to be found away from its mean, while a low standard deviation means X will remain fairly close to its mean.

3.3.3 Cauchy-Schwarz Inequality

The next inequality is also quite useful in certain situations, and the proof is really cute.

Theorem 3.13 (Cauchy-Schwarz Inequality)**.** *We have*

$$\boxed{|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}} \tag{3.26}$$

provided that $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$.

Proof. Consider

$$0 \leq \mathbb{E}[(X + xY)^2] = \underbrace{\mathbb{E}[Y^2]}_{a>0} x^2 + \underbrace{2\,\mathbb{E}[XY]}_{b} x + \underbrace{\mathbb{E}[X^2]}_{c}.$$

Observe that the above expression is a quadratic equation in the variable x, which is always non-negative. Visualize the parabola: the parabola is always non-negative, which means it has either 0 or 1 root. This is equivalent to the condition $b^2 \leq 4ac$, or

$$4\,\mathbb{E}[XY]^2 \leq 4\,\mathbb{E}[X^2]\,\mathbb{E}[Y^2].$$

Dividing by 4 and taking square roots yields the desired inequality. □

3.4 Weak Law of Large Numbers

Now, we can justify why the expectation is called the *long-run average* of a sequence of values. Suppose that $X_1, X_2, X_3, \ldots$ are i.i.d. random variables, which we can think of as successive

measurements of a true variable X. The idea is that X is some quantity which we wish to measure, and X follows some probability distribution with unknown parameters: mean μ and variance σ^2. For each positive integer i, the random variable X_i is a measurement of X, which is to say that each X_i *follows the same probability distribution as* X. In particular, this means that each X_i also has mean μ and variance σ^2. We are interested in the *average* of the samples we collect:

$$\bar{X}_n := \frac{X_1 + \cdots + X_n}{n} \tag{3.27}$$

What is the expectation of $\bar{X}_n$? Since we would like to measure the parameter μ, we are hoping that $\mathbb{E}[\bar{X}_n] = \mu$. (In other words, we are hoping that the average of our successive measurements will be close to the true parameter μ.) We can use linearity of expectation to quickly check that this holds:

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n}(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]) = \frac{1}{n} \cdot n\mu = \mu$$

We therefore call $\bar{X}_n$ an **unbiased estimator** of μ.

The next question to ask is: on average, we expect $\bar{X}_n$ to estimate μ. But for a given experiment, *how close to* μ *do we expect* $\bar{X}_n$ *to be? How long will it take for* $\bar{X}_n$ *to converge to its mean,* μ? These are questions that involve the variance of the distribution. First, let us compute

$$\operatorname{var} \bar{X}_n = \frac{1}{n^2}(\operatorname{var} X_1 + \cdots + \operatorname{var} X_n) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}. \tag{3.28}$$

(Notice that the answer is *not* σ^2!) In calculating our answer, we have used the scaling property of the variance and the assumption of independence. The dependence of $\sigma_{\bar{X}_n}$, the standard deviation of $\bar{X}_n$, is

$$\sigma_{\bar{X}_n} \in O\Big(\frac{1}{\sqrt{n}}\Big), \tag{3.29}$$

a dependence that is well-worth remembering. In particular, this result states that *the more samples we collect, the smaller our standard deviation becomes*! This result is what allows the scientific method to work: without it, gathering more samples would not make us any more certain of our results. We next state and prove a famous result, which shows that $\bar{X}_n$ converges to its expected value μ after enough samples are drawn.

Theorem 3.14 (Weak Law of Large Numbers (WLLN))**.** *For all* $\varepsilon > 0$*, in the limit as* $n \to \infty$*,*

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) \to 0. \tag{3.30}$$

Proof. We will use Chebyshev's inequality:

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\operatorname{var} \bar{X}_n}{\varepsilon^2}$$

Filling in what we know about $\operatorname{var} \bar{X}_n$, we have

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2},$$

which tends to 0 as $n \to \infty$. □

Intuitively, the theorem is asking: what is the probability that $\bar{X}_n$ is ε-far away from μ? The answer is: as $n \to \infty$, the probability becomes 0. *Increasing the number of samples decreases the probability that the sample average will be far from the true average.* When the WLLN holds, we say that $\bar{X}_n$ **converges in probability** to μ as $n \to \infty$.

3.5 Confidence Intervals

Often, we have a sample of n data points $X_1, \ldots, X_n$ (where n is a positive integer) and the distribution of $(X_1, \ldots, X_n)$ depends on an unknown parameter θ. Our goal is to estimate θ. The idea behind a confidence interval is to construct an interval $C(X)$ (which is a function of our sample; notice that $C(X)$ is random) such that $\mathbb{P}(\theta \in C(X)) \geq 1 - \delta$, where $\delta > 0$, and $1 - \delta$ is called the **confidence level**.

Example 3.15. We will describe how Chebyshev's inequality can be used to produce a confidence interval. Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. In the notation of the preceding section, we will prove

$$\boxed{\mathbb{P}\big(\mu \in (\bar{X} - a, \bar{X} + a)\big) \geq 1 - \delta} \tag{3.31}$$

where $a > 0$ is a constant. The interpretation of the statement is that we collect n samples and compute the sample average $\bar{X}$. With $1 - \delta$ confidence, we believe that the true mean μ lies in the interval $(\bar{X} - a, \bar{X} + a)$.

We use Chebyshev's Inequality:

$$\begin{aligned}\mathbb{P}\big(\mu \in (\bar{X} - a, \bar{X} + a)\big) &= \mathbb{P}(|\bar{X} - \mu| < a) = 1 - \mathbb{P}(|\bar{X} - \mu| \geq a) \\ &\geq 1 - \frac{\operatorname{var} \bar{X}}{a^2} = 1 - \frac{\sigma^2}{na^2}\end{aligned}$$

If we set our confidence level to be $1 - \delta$, then we can solve for a.

$$1 - \frac{\sigma^2}{na^2} = 1 - \delta \quad \Longrightarrow \quad a = \frac{1}{\sqrt{\delta n}}\sigma$$

Therefore, our confidence interval is $(\bar{X} - \sigma/\sqrt{\delta n}, \bar{X} + \sigma/\sqrt{\delta n})$.

Example 3.16. As a specialization of the previous example, we will consider the problem of estimating the bias of a coin. A coin is biased with $\mathbb{P}(H) = p$, and our goal is

to find a $1-\delta$ confidence interval for p. How can we construct a confidence interval here?

Flip the coin n times and let X_i be the indicator that the ith flip came up heads for $i = 1, \ldots, n$. Observe that in this case, $\mu = \mathbb{E}[X_i] = p$ and $\sigma^2 = \operatorname{var} X_i = p(1-p)$, since we are in the setting of $X_i \sim \text{Bernoulli}(p)$. Now, we can apply our bound from the previous example:

$$a = \frac{1}{\sqrt{\delta n}}\sqrt{p(1-p)} \leq \frac{1}{2\sqrt{\delta n}}$$

where we have used $p(1-p) \leq 1/4$. Hence, our $1-\delta$ confidence interval for the bias p is

$$\Big(\bar{X} - \frac{1}{2\sqrt{\delta n}}, \bar{X} + \frac{1}{2\sqrt{\delta n}}\Big).$$

The confidence intervals using Chebyshev's inequality are fairly large because Chebyshev's inequality is not very sharp. Later, we will see how to construct better approximate confidence intervals.

Exercise 17 Another situation of interest when constructing confidence intervals is to fix the confidence level $1-\delta$ and the size of the interval $(\bar{X} - \varepsilon, \bar{X} + \varepsilon)$ (where $\varepsilon > 0$), and then to ask what the sample size n must be in order to achieve this confidence interval. In the confidence interval for the bias of the coin (Example 3.16), solve for the required sample size in order for $(\bar{X} - \varepsilon, \bar{X} + \varepsilon)$ to be a valid $1-\delta$ confidence interval, and explain how the required sample size changes if ε is halved or if δ is halved.

3.6 Bonus: Chernoff Bounds

Consider what happens when we apply Markov's inequality to the increasing non-negative function $f(x) = \exp(\theta x)$ (for $\theta > 0$). We obtain the statement:

$$\boxed{\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}[\exp(\theta X)]}{\exp(\theta x)}} \tag{3.32}$$

What value of θ should we choose? The obvious answer is: the best possible one! In other words, we can *optimize* over the values of θ in search of the best possible bound. This is known as a **Chernoff bound** and it can be used to prove that the tail probabilities of certain distributions decay *exponentially fast*.

Exercise 18 Use (3.32) to bound $\mathbb{P}(X \geq x)$ when $X \sim \text{Poisson}(\lambda)$. What is your bound when $\lambda = 1$ and $x > 1$?

3.7 Solutions to Exercises

Exercise 16 We have

$$\begin{aligned}\operatorname{var}(XY) &= \mathbb{E}[X^2Y^2] - \mathbb{E}[X]^2\,\mathbb{E}[Y]^2 = \mathbb{E}[X^2]\,\mathbb{E}[Y^2] - \mathbb{E}[X]^2\,\mathbb{E}[Y]^2 \\ &= (\operatorname{var} X + \mathbb{E}[X]^2)(\operatorname{var} Y + \mathbb{E}[Y]^2) - \mathbb{E}[X]^2\,\mathbb{E}[Y]^2 \\ &= \mathbb{E}[X]^2 \operatorname{var} Y + \mathbb{E}[Y]^2 \operatorname{var} X + (\operatorname{var} X)(\operatorname{var} Y).\end{aligned}$$

Exercise 17 By Chebyshev's inequality,

$$\mathbb{P}(|\bar{X} - p| \geq \varepsilon) \leq \frac{\operatorname{var} \bar{X}}{\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}$$

and we need the RHS to be $\leq \delta$, so we must take $n \geq (4\delta\varepsilon^2)^{-1}$. If we halve ε, then the required sample size quadruples; if we halve δ, then the required sample size doubles.

Exercise 18 First, we compute

$$\begin{aligned}\mathbb{E}[\exp(\theta X)] &= \sum_{x=0}^{\infty} \exp(\theta x) \frac{\lambda^x \exp(-\lambda)}{x!} = \exp(-\lambda) \sum_{x=0}^{\infty} \frac{(\lambda \exp\theta)^x}{x!} = \exp(-\lambda)\exp(\lambda \exp\theta) \\ &= \exp\big(\lambda(\exp\theta - 1)\big).\end{aligned}$$

The Chernoff bound gives

$$\mathbb{P}(X \geq x) \leq \frac{\exp(\lambda(\exp\theta - 1))}{\exp(\theta x)} = \exp(-\lambda)\exp(\lambda \exp\theta - \theta x). \tag{3.33}$$

To optimize this bound, we differentiate with respect to θ and we set the result to 0.

$$0 = \frac{\mathrm{d}}{\mathrm{d}\theta} \exp(-\lambda)\exp(\lambda\exp\theta - \theta x) = \exp(-\lambda)\exp(\lambda\exp\theta - \theta x)(\lambda \exp\theta - x).$$

We can see that the optimal value of θ is given by $\exp\theta = x/\lambda$. Plug this into (3.33) to obtain the bound

$$\mathbb{P}(X \geq x) \leq \exp(-\lambda)\exp\Big(\lambda \cdot \frac{x}{\lambda} - x \ln \frac{x}{\lambda}\Big) = \exp\Big(-x \ln \frac{x}{\lambda} + x - \lambda\Big).$$

Consider the special case where $\lambda = 1$ and $x > 1$. Then $\mathbb{P}(X \geq x) \leq x^{-x}\exp(-(1-x))$, which does indeed decrease exponentially fast. See Figure 3.1 for a visualization.

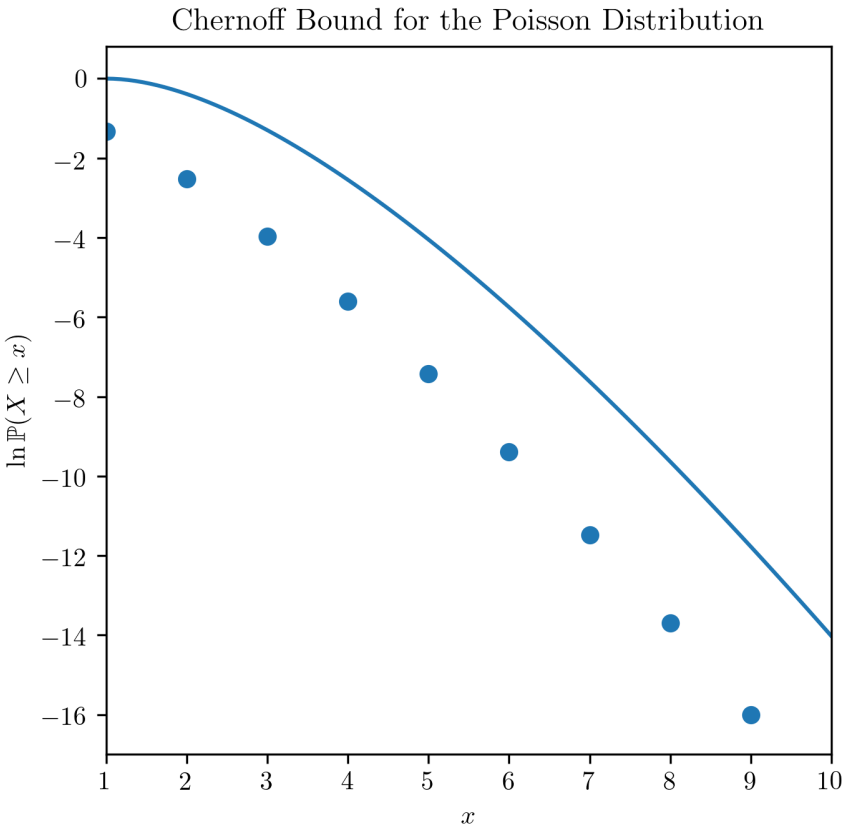


Figure 3.1: The upper bound on $\mathbb{P}(X \geq x)$ using the Chernoff bound, where $X \sim \text{Poisson}(1)$, is plotted as a continuous curve on a logarithmic scale. The true probability $\mathbb{P}(X \geq x)$ of the Poisson distribution appears as the solid dots in the figure. The upper bound is quite good and appears to capture the decay of $\mathbb{P}(X \geq x)$.

Chapter 4

Regression & Conditional Expectation

A fundamental question in statistics is: given a set of data points $\{(X_i, Y_i)\}_{i=1}^n$, how can we *estimate* the value of Y as a function of X? First, we will discuss the covariance and correlation of two random variables, and use these quantities to derive the best *linear* estimator of Y, known as the LLSE. Then, we will define the conditional expectation, and proceed to derive the best *general* estimator of Y, known as the MMSE. The notion of conditional expectation will turn out to be an immensely useful concept with further applications.

4.1 Covariance

We have already discussed the case of independent random variables, but many of the variables in real life are dependent upon each other, such as the height and weight of an individual. Now we will consider how to quantify the *linear* dependence of random variables, starting with the definition of covariance.

Definition 4.1. The **covariance** of two random variables X and Y is defined as:

$$\boxed{\text{cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]} \tag{4.1}$$

The covariance is the product of the deviations of the two variables from their respective means. Suppose that whenever X is larger than its mean, Y is also larger than its mean; then, the covariance will be positive, and we say the variables are *positively correlated*. On the other hand, if whenever X is larger than its mean, Y is smaller than its mean, then the covariance is negative and we say that the variables are *negatively correlated*. In other words: *positive correlation means that X and Y tend to fluctuate in the same direction, while negative correlation means that X and Y tend to fluctuate in opposite directions.*

Recall that we said that two *events* A and B are

1. positively correlated if $\mathbb{P}(A \cap B) > \mathbb{P}(A)\mathbb{P}(B)$,
2. independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$,
3. negatively correlated if $\mathbb{P}(A \cap B) < \mathbb{P}(A)\mathbb{P}(B)$.

Let us connect the correlation of events with the correlation of random variables. Let $\mathbb{1}\{A\}$ and $\mathbb{1}\{B\}$ be the indicators for events A and B respectively. Note that

$$\begin{aligned}\mathbb{E}[\mathbb{1}\{A\}\,\mathbb{1}\{B\}] &= \mathbb{P}(A \cap B),\\ \mathbb{E}[\mathbb{1}\{A\}] &= \mathbb{P}(A),\\ \mathbb{E}[\mathbb{1}\{B\}] &= \mathbb{P}(B)\end{aligned}$$

since if X and Y are indicator random variables, then XY is either 0 or 1, so by the definition of expectation, $\mathbb{E}[XY] = 1 \cdot \mathbb{P}(XY = 1)$.

1. If A and B are positively correlated, then $\mathbb{E}[\mathbb{1}\{A\}\,\mathbb{1}\{B\}] > \mathbb{E}[\mathbb{1}\{A\}]\,\mathbb{E}[\mathbb{1}\{B\}]$, so $\operatorname{cov}(\mathbb{1}\{A\}, \mathbb{1}\{B\}) > 0$.

2. If A and B are independent, then the corresponding indicators are uncorrelated: $\mathbb{E}[\mathbb{1}\{A\}\,\mathbb{1}\{B\}] = \mathbb{E}[\mathbb{1}\{A\}]\,\mathbb{E}[\mathbb{1}\{B\}]$, so $\operatorname{cov}(\mathbb{1}\{A\}, \mathbb{1}\{B\}) = 0$. (**Caution**: In this case, $\mathbb{1}\{A\}$ and $\mathbb{1}\{B\}$ are independent, but in general, $\operatorname{cov}(X, Y) = 0$ only means that X and Y are *uncorrelated*, which is *weaker* than independence. We will see an example illustrating this below.)

3. If A and B are negatively correlated, then $\mathbb{E}[\mathbb{1}\{A\}\,\mathbb{1}\{B\}] < \mathbb{E}[\mathbb{1}\{A\}]\,\mathbb{E}[\mathbb{1}\{B\}]$, so $\operatorname{cov}(\mathbb{1}\{A\}, \mathbb{1}\{B\}) < 0$.

The three cases above demonstrate that *correlation of events corresponds to correlation of the corresponding indicator random variables.*

Just as we had a computational formula for variance, we have a computational formula for covariance.

Theorem 4.2 (Computational Formula for Covariance)**.** *Let X and Y be random variables. Then:*

$$\boxed{\operatorname{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]} \tag{4.2}$$

Proof. We take the definition of covariance and expand it out:

$$\begin{aligned}\operatorname{cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}\big[XY - X\,\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\,\mathbb{E}[Y]\big]\\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] + \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\end{aligned}$$ □

Corollary 4.3 (Independent Implies Uncorrelated)**.** *Let X and Y be independent random variables. Then* $\operatorname{cov}(X, Y) = 0$.

Proof. By the assumption of independence, we have $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$ (Theorem 2.5); the corollary follows immediately. □

Example 4.4. Pick a point uniformly randomly from $\{(1,0),(0,1),(-1,0),(0,-1)\}$, see Figure 4.1. Let X be the x-coordinate and Y be the y-coordinate of the point that we choose. Observe that $XY = 0$ always, so $\mathbb{E}[XY] = 0$. By symmetry about the origin, we have $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. Hence, $\operatorname{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0$. However, X and Y are *not* independent. In particular,

$$0 = \mathbb{P}(X = 0, Y = 0) \neq \mathbb{P}(X = 0)\mathbb{P}(Y = 0) = \frac{1}{4}.$$

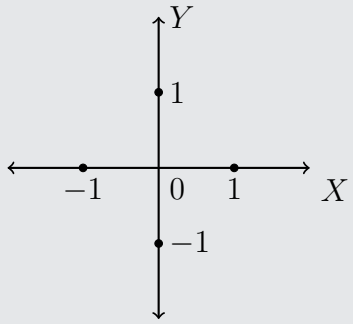


Figure 4.1: Choose one of the four points shown uniformly at random.

Important: Corollary 4.3 shows that independence of X and Y implies $\operatorname{cov}(X,Y) = 0$. Example 4.4 shows that the converse is *not* true, i.e. $\operatorname{cov}(X,Y) = 0$ is *not* sufficient to say that X and Y are independent. Study the above example carefully!

Next, we will show how the covariance and variance are related.

Corollary 4.5 (Covariance & Variance)**.** *Let X be a random variable. Then:*

$$\boxed{\operatorname{var} X = \operatorname{cov}(X,X)} \tag{4.3}$$

Proof. The proof is straightforward.

$$\operatorname{cov}(X,X) = \mathbb{E}[X^2] - \mathbb{E}[X]\,\mathbb{E}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \operatorname{var} X \qquad \square$$

Corollary 4.5 is not very useful for calculating $\operatorname{var} X$, but it shows that variance can be seen as a special case of covariance.

Corollary 4.6 (Variance of Sums of Random Variables)**.** *Let X and Y be random variables. Then:*

$$\boxed{\operatorname{var}(X+Y) = \operatorname{var} X + \operatorname{var} Y + 2\operatorname{cov}(X,Y)} \tag{4.4}$$

Proof. Since $\operatorname{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$, this is a restatement of the familiar fact

$$\operatorname{var}(X + Y) = \operatorname{var} X + \operatorname{var} Y + 2(\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y])$$

(Theorem 3.5). □

The utility of the last result is that it holds true for *any* random variables, even ones that are not independent.

4.1.1 Symmetry & Bilinearity of Covariance

Next, we show that the covariance is symmetric and bilinear, that is, linear in each of its arguments.

Theorem 4.7 (Symmetry & Bilinearity of Covariance)**.** *Suppose X_i, Y_i are random variables, $i = 1, 2$, and $a \in \mathbb{R}$ is a constant. Then the following properties hold:*

$$\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X), \tag{4.5}$$
$$\operatorname{cov}(X_1 + X_2, Y) = \operatorname{cov}(X_1, Y) + \operatorname{cov}(X_2, Y) \tag{4.6}$$
$$\operatorname{cov}(X, Y_1 + Y_2) = \operatorname{cov}(X, Y_1) + \operatorname{cov}(X, Y_2) \tag{4.7}$$
$$a \operatorname{cov}(X, Y) = \operatorname{cov}(aX, Y) = \operatorname{cov}(X, aY) \tag{4.8}$$

Proof. The proofs are straightforward using linearity of expectation (Theorem 2.4). The symmetry of covariance (4.5) follows because the expression $\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$ is symmetric in X and Y. To prove (4.6),

$$\begin{aligned}\operatorname{cov}(X_1 + X_2, Y) &= \mathbb{E}[(X_1 + X_2)Y] - \mathbb{E}[X_1 + X_2]\,\mathbb{E}[Y] \\ &= \mathbb{E}[X_1 Y] - \mathbb{E}[X_1]\,\mathbb{E}[Y] + \mathbb{E}[X_2 Y] - \mathbb{E}[X_2]\,\mathbb{E}[Y] \\ &= \operatorname{cov}(X_1, Y) + \operatorname{cov}(X_2, Y).\end{aligned}$$

Now, (4.7) follows from (4.6) and symmetry. Similarly, to prove (4.8),

$$\operatorname{cov}(aX, Y) = \mathbb{E}[aXY] - \mathbb{E}[aX]\,\mathbb{E}[Y] = a(\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]) = a \operatorname{cov}(X, Y)$$

and $\operatorname{cov}(X, aY) = a \operatorname{cov}(X, Y)$ follows from symmetry. □

4.1.2 Standardized Variables

Sometimes, we would like to write random variables in a standard form for easier computations. "Standard form" means zero mean and unit variance.

Definition 4.8. Let X be a non-constant random variable. Then

$$\boxed{X^* := \frac{X - \mathbb{E}[X]}{\sigma_X}} \tag{4.9}$$

is called the **standard form** of X.

Next, we explain why X^* is called the standard form.

Theorem 4.9 (Mean & Variance of Standardized Random Variables)**.** *Let X^* be a standardized random variable. Then the mean and variance of X^* are:*

$$\begin{aligned} \mathbb{E}[X^*] &= 0, \\ \operatorname{var} X^* &= \mathbb{E}[(X^*)^2] = 1. \end{aligned}$$

Proof. First, we prove that the mean is 0 using linearity of expectation (Theorem 2.4).

$$\mathbb{E}[X^*] = \frac{\mathbb{E}[X] - \mathbb{E}[X]}{\sigma_X} = 0$$

Next, since $\mathbb{E}[X^*] = 0$, then $\operatorname{var} X^* = \mathbb{E}[(X^*)^2] - \mathbb{E}[X^*]^2 = \mathbb{E}[(X^*)^2]$. Using the properties of variance,

$$\operatorname{var} X^* = \operatorname{var} \frac{X - \mathbb{E}[X]}{\sigma_X} = \frac{\operatorname{var} X}{\sigma_X^2} = \frac{\operatorname{var} X}{\operatorname{var} X} = 1. \qquad \square$$

Standardizing the random variable X is equivalent to shifting the distribution so that its mean is 0, and scaling the distribution so that its standard deviation is 1. The random variable X^* is rather convenient because it is dimensionless: for example, if X and Y represent measurements of the temperature of a system in degrees Fahrenheit and degrees Celsius respectively, then $X^* = Y^*$.

4.1.3 Correlation

We next take a slight detour in order to define the correlation of two random variables, which appears frequently in statistics. Although we will not use correlation, the exposition to correlation presented here should allow you to interpret studies in journals.

Definition 4.10. The **correlation** of two random variables X and Y is:

$$\boxed{\operatorname{correlation}(X, Y) := \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}} \tag{4.10}$$

The correlation is often denoted by ρ or r, and is sometimes referred to as Pearson's correlation coefficient.

The next result provides an interpretation of the correlation.

Theorem 4.11 (Covariance & Correlation)**.** *The correlation of two random variables X and Y is:*

$$\boxed{\text{correlation}(X,Y) = \text{cov}(X^*, Y^*) = \mathbb{E}[X^*Y^*]} \tag{4.11}$$

Proof. We calculate the covariance of X^* and Y^* using the properties of covariance.

$$\text{cov}(X^*, Y^*) = \text{cov}\Big(\frac{X - \mathbb{E}[X]}{\sigma_X}, \frac{Y - \mathbb{E}[Y]}{\sigma_Y}\Big) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

(Remember that the covariance of a constant with a random variable is 0.) The second equality follows easily because $\text{cov}(X^*, Y^*) = \mathbb{E}[X^*Y^*] - \mathbb{E}[X^*]\,\mathbb{E}[Y^*]$ and we have $\mathbb{E}[X^*] = \mathbb{E}[Y^*] = 0$. □

The result states that we can view the correlation as a *standardized* version of the covariance. As a result, we can also prove a result about the possible values of the correlation:

Theorem 4.12 (Magnitude of Correlation)**.** *If X, Y are non-constant random variables, then*

$$-1 \leq \text{correlation}(X,Y) \leq 1 \tag{4.12}$$

with equality if and only if Y is a linear function of X.

Proof. The expectation of the random variable $(X^* \mp Y^*)^2$ is non-negative. [a] So,

$$0 \leq \mathbb{E}[(X^* \mp Y^*)^2] = \mathbb{E}[(X^*)^2] \mp 2\,\mathbb{E}[X^*Y^*] + \mathbb{E}[(Y^*)^2] = 2 \mp 2\,\mathbb{E}[X^*Y^*].$$

We have the inequality

$$\pm\,\text{correlation}(X,Y) = \pm\,\mathbb{E}[X^*Y^*] \leq 1$$

and the result follows by considering the two cases. Note that equality holds if and only if $0 = \mathbb{E}[(X^* \mp Y^*)^2]$, which can only happen if $X^* \mp Y^* = 0$. This implies that $Y^* = \pm X^*$, which is true if and only if $Y = aX + b$ for constants $a, b \in \mathbb{R}$. □

[a] The notation $\pm$ means "plus or minus", and the notation $\mp$ means "minus or plus". Think of it as shorthand for considering both cases at the same time, so that I do not have to write out the same steps twice.

Exercise 19 Suppose $\text{correlation}(X,Y) = \pm 1$. What are the $a, b \in \mathbb{R}$ so that $Y = aX + b$?

Now, we can see that correlation is a useful measure of the degree of *linear* dependence between two variables X and Y. If $\text{correlation}(X,Y)$ is -1 or 1, then X and Y are perfectly linearly correlated, i.e. a plot of Y versus X would be a straight line. The closer the correlation is to ± 1, the closer the data resembles a straight-line relationship. If X and Y are independent, then the correlation is 0 (but the converse is not true). As a final remark,

the square of the correlation coefficient is called the **coefficient of determination** (usually denoted R^2). The coefficient of determination appears frequently next to best-fit lines on scatter plots as a measure of how well the best-fit line fits the data.

4.2 LLSE

We will immediately apply our development of the covariance to the problem of finding the best linear predictor of Y given X. We begin by presenting the main result, and then proceed to prove that the result satisfies the properties we desire. [1]

Definition 4.13. Let X and Y be random variables. The **least linear squares estimate (LLSE)** of Y given X is defined as:

$$\boxed{L(Y \mid X) := \mathbb{E}[Y] + \frac{\operatorname{cov}(X,Y)}{\operatorname{var} X}(X - \mathbb{E}[X])} \tag{4.13}$$

Observe that the LLSE is a random variable: in fact, it is a function of X.

4.2.1 Orthogonality Property

Theorem 4.14 (Orthogonality Property of LLSE)**.** *The LLSE satisfies:*

$$\mathbb{E}[Y - L(Y \mid X)] = 0 \tag{4.14}$$
$$\mathbb{E}\big[\big(Y - L(Y \mid X)\big)X\big] = 0 \tag{4.15}$$

Proof. Proof of (4.14):

$$\begin{aligned}\mathbb{E}[Y - L(Y \mid X)] &= \mathbb{E}\Big[Y - \mathbb{E}[Y] - \frac{\operatorname{cov}(X,Y)}{\operatorname{var} X}(X - \mathbb{E}[X])\Big] \\ &= \mathbb{E}[Y] - \mathbb{E}[Y] - \frac{\operatorname{cov}(X,Y)}{\operatorname{var} X}(\mathbb{E}[X] - \mathbb{E}[X]) = 0\end{aligned}$$

Proof of (4.15):

$$\begin{aligned}\mathbb{E}\big[\big(Y - L(Y \mid X)\big)X\big] &= \mathbb{E}\Big[X\Big(Y - \mathbb{E}[Y] - \frac{\operatorname{cov}(X,Y)}{\operatorname{var} X}(X - \mathbb{E}[X])\Big)\Big] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] - \frac{\operatorname{cov}(X,Y)}{\operatorname{var} X}(\mathbb{E}[X^2] - \mathbb{E}[X]^2) \\ &= \operatorname{cov}(X,Y) - \frac{\operatorname{cov}(X,Y)}{\operatorname{var} X} \cdot \operatorname{var} X = 0\end{aligned}$$

□

[1] The material for the sections on regression rely heavily on Professor Walrand's notes, although I have inserted my own interpretation of the material wherever appropriate.

If you have not studied linear algebra, then the rest of the section can be safely skipped. Linear algebra is not necessary to understand the properties of the LLSE, although linear algebra certainly enriches the theory of linear regression. We discuss linear algebra concepts solely to motivate the orthogonality property.

Given a probability space Ω, then the space of random variables over Ω with finite variance is a vector space V (that is, random variables satisfy the vector space axioms). Indeed, we have already introduced how to add and scalar-multiply random variables.

The map $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ given by $\langle X, Y \rangle := \mathbb{E}[XY]$ is an inner product, which makes V into an inner product space. We say that two random variables X and Y are *orthogonal* if $\mathbb{E}[XY] = 0$. According to the orthogonality property, we see that the random variable $Y - L(Y \mid X)$ is orthogonal to both the constant random variable 1 and the random variable X. Hence, $Y - L(Y \mid X)$ is orthogonal to the plane $\mathcal{L}(X) := \text{span}\{1, X\}$. Note that $\mathcal{L}(X)$ is the subspace of V containing all linear functions $aX + b$ of X, and $L(Y \mid X) \in \mathcal{L}(X)$. Geometrically, we have that *the projection of Y onto $\mathcal{L}(X)$ is $L(Y \mid X)$*, which is to say that $L(Y \mid X)$ is, in a sense, the *closest* linear function of X to Y. We will make this notion more precise in the next section.

4.2.2 Optimality of the LLSE

Next, we show that the orthogonality property implies that $L(Y \mid X)$ is the best linear predictor of Y, in a least-squares sense.

Theorem 4.15 (Least-Squares Property)**.** *Define $\mathcal{L}(X) := \{aX + b : a, b \in \mathbb{R}\}$ to be the set of linear functions of X Then for any $aX + b \in \mathcal{L}(X)$,*

$$\mathbb{E}\big[\big(Y - L(Y \mid X)\big)^2\big] \leq \mathbb{E}[(Y - aX - b)^2]. \tag{4.16}$$

In other words, when we estimate Y, $L(Y \mid X)$ has the lowest mean squared error out of any other linear function of X.

Proof. According to the orthogonality property, we can combine (4.14) and (4.15) to obtain

$$\mathbb{E}\big[\big(Y - L(Y \mid X)\big)(aX + b)\big] = 0$$

for any $aX + b \in \mathcal{L}(X)$. Let $\hat{Y} := L(Y \mid X)$ to simplify the notation. We calculate:

$$\begin{aligned}
\mathbb{E}[(Y - aX - b)^2] &= \mathbb{E}\big[\big((Y - \hat{Y}) + (\hat{Y} - aX - b)\big)^2\big] \\
&= \mathbb{E}[(Y - \hat{Y})^2] + 2\underbrace{\mathbb{E}[(Y - \hat{Y})(\hat{Y} - aX - b)]}_{=0} + \mathbb{E}[(\hat{Y} - aX - b)^2] \\
&= \mathbb{E}[(Y - \hat{Y})^2] + \mathbb{E}[(\hat{Y} - aX - b)^2]
\end{aligned}$$

In the second line of the proof, the term $\mathbb{E}[(Y - L(Y \mid X))(L(Y \mid X) - aX - b)]$ vanishes by the orthogonality property because $L(Y \mid X) - aX - b \in \mathcal{L}(X)$. The quantity above represents the mean squared error of $aX + b$ as a predictor of Y, so we seek to minimize this quantity. Since the expectation of a non-negative random variable is always non-negative, the mean squared error is minimized when $aX + b = L(Y \mid X)$. □

The least-squares line is used everywhere as a visual summary of a trend. Now you have seen the theory behind this powerful tool!

Exercise 20 What is the squared error of the LLSE, that is, what is $\mathbb{E}[(Y - L(Y \mid X))^2]$?

4.3 Quadratic Regression

We can extend the ideas of the previous section to find the best *quadratic* estimator of Y given X. A quadratic function of X has the form $f(X) = aX^2 + bX + c$. Our goal is to make $Y - f(X)$ *orthogonal* to any other quadratic function of X, that is, we want $f(X)$ to be the *projection* of Y onto $\mathcal{Q}(X)$ (where $\mathcal{Q}(X)$ is the space of quadratic functions of X).

To achieve this, we want $Y - f(X)$ to be orthogonal to 1, X, and X^2. We have the equations

$$0 = \mathbb{E}[Y - aX^2 - bX - c], \tag{4.17}$$
$$0 = \mathbb{E}[(Y - aX^2 - bX - c)X], \tag{4.18}$$
$$0 = \mathbb{E}[(Y - aX^2 - bX - c)X^2]. \tag{4.19}$$

We can solve these equations for a, b, and c in order to find the best quadratic function of X as an estimator of Y. We will not work through the details, but hopefully you can see how the general procedure goes.

4.4 Conditional Expectation

Next, we will search for an even more powerful predictor of Y given X. [2] If $\mathbb{P}(Y = y) > 0$, we define the conditional expectation $\mathbb{E}[X \mid Y = y]$ using the following formula:

$$\boxed{\mathbb{E}[X \mid Y = y] = \sum_x x\mathbb{P}(X = x \mid Y = y)} \tag{4.20}$$

In other words, $\mathbb{E}[X \mid Y = y]$ is the expectation of X *with respect to the probability distribution of X conditioned on $Y = y$*. It is important to stress that $\mathbb{E}[X \mid Y = y]$ is just a real number, just like any other expectation.

[2] The material presented in this section appears slightly differently from the presentation in Professor Walrand's lecture notes, but the lecture notes are still the source for these discussion notes.

Notice that for every possible value of Y, we can assign a real number $\mathbb{E}[X \mid Y = y]$. In other words, this is a *function* of y:

$$f : \mathbb{R} \to \mathbb{R} \quad \text{given by} \quad f(y) = \mathbb{E}[X \mid Y = y]$$

Hence, let us define $\mathbb{E}(X \mid Y)$ to be a *function* of Y in the following manner:

Definition 4.16. Let X and Y be random variables. Then $\mathbb{E}(X \mid Y)$ is also a random variable, called the **conditional expectation** of X given Y, which has the value $\mathbb{E}[X \mid Y = y]$ with probability $\mathbb{P}(Y = y)$.

Observe that $\mathbb{E}(X \mid Y)$ is a function of Y, i.e. $\mathbb{E}(X \mid Y) = f(Y)$. This point cannot be stressed enough: $\mathbb{E}(X \mid Y)$ is a *random variable*! Although the conditional expectation may seem mysterious at first, there is an easy rule for writing down $\mathbb{E}(X \mid Y)$. Let us consider an example for concreteness.

Example 4.17. Suppose that we roll a die N times, where N is some non-negative integer-valued random variable. Let X be the sum of the dice rolls. Conditioned on $N = 1$ (that is, we roll one die), then the expected value is $7/2$, i.e. $\mathbb{E}[X \mid N = 1] = 7/2$. Similarly, conditioned on $N = 2$, we roll two dice, so $\mathbb{E}[X \mid N = 2] = 7$ (the expected sum of two dice is 7). In general, conditioned on $N = n$, we roll n dice and we have $\mathbb{E}[X \mid N = n] = 7n/2$.

If $N = n$, then the random variable $\mathbb{E}(X \mid N)$ has the value $\mathbb{E}[X \mid N = n] = 7n/2$; hence, we can write $\mathbb{E}(X \mid N) = 7N/2$ (which is a function of N, following the discussion above).

At first, it may appear that in going from the expression for $\mathbb{E}[X \mid N = n]$ to $\mathbb{E}(X \mid N)$, we merely replaced the n with N. Of course, there is more going on than just a simple substitution ($\mathbb{E}(X \mid Y)$ is a random variable and $\mathbb{E}[X \mid Y = y]$ is just a real number), but *substituting $y \mapsto Y$ is exactly the procedure for writing down* $\mathbb{E}(X \mid Y)$. Don't worry if conditional expectation is difficult to grasp at first. Mastery of this concept requires practice, and we will soon see how to apply conditional expectation to the problem of prediction.

4.4.1 The Law of Iterated Expectation

First, we prove an amazing fact about conditional expectation. We noted that $\mathbb{E}(X \mid Y)$ is a random variable, and of course, we are always interested in the expectation values of random variables. Hence, we can ask the question: what is the expectation of $\mathbb{E}(X \mid Y)$? To answer this question, recall that $\mathbb{E}(X \mid Y)$ is a function of Y, that is, $\mathbb{E}(X \mid Y) = f(Y)$. Then, to calculate the expectation of $\mathbb{E}(X \mid Y)$, we see that *we should compute the expectation with respect to the probability distribution of* Y. Once you understand this point, the following proof is straightforward in its details.

Theorem 4.18 (Law of Iterated Expectation)**.** *Let X and Y be random variables.*

$$\boxed{\mathbb{E}[\mathbb{E}(X \mid Y)] = \mathbb{E}[X]} \tag{4.21}$$

Proof. As noted above, we compute $\mathbb{E}[\mathbb{E}(X \mid Y)]$ with respect to the probability distribution of Y.

$$\begin{aligned}\mathbb{E}[\mathbb{E}(X \mid Y)] &= \sum_y \mathbb{E}[X \mid Y = y]\mathbb{P}(Y = y) = \sum_y \Big(\sum_x x\mathbb{P}(X = x \mid Y = y)\Big)\mathbb{P}(Y = y) \\ &= \sum_y \sum_x x\mathbb{P}(X = x, Y = y) = \sum_x x\Big(\sum_y \mathbb{P}(X = x, Y = y)\Big) \\ &= \sum_x x\mathbb{P}(X = x) = \mathbb{E}[X].\end{aligned}$$

In the first line, we use $\mathbb{E}[f(Y)] = \sum_y f(y)\mathbb{P}(Y = y)$; then, we use the definition of $\mathbb{E}[X \mid Y = y]$, which was given in the previous section. In the second line, we use our knowledge of conditional probability; then, we recall that summing over all possible values of Y in the joint distribution $\mathbb{P}(X = x, Y = y)$ yields the marginal distribution $\mathbb{P}(X = x)$. Finally, in the last line, we come back to the definition of $\mathbb{E}[X]$. □

Example 4.19. Let us return to Example 4.17, where we found that $\mathbb{E}(X \mid N) = 7N/2$. For concreteness, suppose that $N \sim \text{Geometric}(p)$ (there is no particular reason to pick the geometric distribution here; think of the random variable N as simply providing you with a random number of rolls, so we could have chosen any random variable which takes on values in the positive integers). We have

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}(X \mid N)] = \frac{7}{2}\,\mathbb{E}[N] = \frac{7}{2p}.$$

Observe that we have found the expectation of a sum of a *random* number of random variables, a task that would be far more difficult without the tool of conditioning.

Example 4.20 (Random Walk)**.** For a second example, let us consider a drunk individual is walking along the number line. At the start, the drunkard starts at the origin ($X_0 = 0$); at successive time steps, the drunkard is equally likely to take a step forward ($+1$) or backward (-1). In other words, given that the drunkard is at position k at time step n, at time step $n + 1$ the drunkard will be equally likely to be found at either position $k + 1$ or $k - 1$.

Denoting the drunkard's position at time step n as X_n, we first compute $\mathbb{E}[X_n]$. We can write $X_{n+1} = X_n + \mathbb{1}_+ - \mathbb{1}_-$, where $\mathbb{1}_+$ is the indicator that the drunkard takes a step forward and $\mathbb{1}_-$ is the indicator that the drunkard takes a step backward. Taking the expectation of both sides yields $\mathbb{E}[X_{n+1}] = \mathbb{E}[X_n] + 1/2 - 1/2 = \mathbb{E}[X_n]$ (remembering

that the probability of moving forward, which is the same as the probability of moving backward, is 1/2). We have found that the expected position of the drunkard is always the same for all n, so the expected position of the drunkard at any later time is the same as the starting position: $\mathbb{E}[X_n] = \mathbb{E}[X_0] = 0$.

Next, we calculate $\mathbb{E}[X_n^2]$. Conditioned on the drunkard being located at position k at time step n, then the drunkard will be equally likely to be found at position $k+1$ or $k-1$. Hence, we have

$$\mathbb{P}\big(X_{n+1}^2 = (k+1)^2 \mid X_n = k\big) = \mathbb{P}\big(X_{n+1}^2 = (k-1)^2 \mid X_n = k\big) = \frac{1}{2}.$$

Taking the expectation yields

$$\mathbb{E}[X_{n+1}^2 \mid X_n = k] = \frac{1}{2}(k+1)^2 + \frac{1}{2}(k-1)^2 = k^2 + 1.$$

Substituting X_n for k, we have

$$\mathbb{E}(X_{n+1}^2 \mid X_n) = X_n^2 + 1.$$

Using the law of total expectation,

$$\mathbb{E}[X_{n+1}^2] = \mathbb{E}[\mathbb{E}(X_{n+1}^2 \mid X_n)] = \mathbb{E}[X_n^2] + 1.$$

$\mathbb{E}[X_0^2] = 0$ provides the initial condition, so $\mathbb{E}[X_n^2] = n$. In particular, we have explicitly shown that $\operatorname{var} X_n = \mathbb{E}[X_n^2] - \mathbb{E}[X_n]^2 = n$ for positive integers n, so the standard deviation of the drunkard's walk is $\sqrt{n}$.

Example 4.21 (Galton-Watson Branching Process I)**.** Let us consider a simple model of population growth. Suppose we start with a population of $X_0 = N$ people (where N is a positive integer), and let X_n be the number of people in the population on the nth time step for $n \in \mathbb{N}$. Furthermore, assume:

1. Each person survives for only one time step.
2. At each time step, independently of the rest of the population, each individual is expected to leave behind μ offspring.

What do we expect the population to look like at time n?

The challenge of this problem is that the number of people in the population at any given time step is a random variable, so we have a random number of individuals who reproduce randomly. However, we have already seen in Example 4.17 a systematic method for dealing with this kind of randomness! First, we *condition* on the number of individuals in the population at time $n-1$, and we compute the population size at time n.

Conditioned on $X_{n-1} = m$, we have m people, who are each expected to leave behind μ offspring. Therefore, $\mathbb{E}[X_n \mid X_{n-1} = m] = m\mu$ and we have $\mathbb{E}(X_n \mid X_{n-1}) = \mu X_{n-1}$. Then, we can use the law of iterated expectation:

$$\mathbb{E}[X_n] = \mathbb{E}[\mathbb{E}(X_n \mid X_{n-1})] = \mu\,\mathbb{E}[X_{n-1}]$$

We have obtained a *recursive* solution to the problem: at each time step, we expect the population size to be multiplied by μ. Therefore,

$$\mathbb{E}[X_n] = \mu^n\,\mathbb{E}[X_0] = \mu^n N.$$

We can see that if $\mu < 1$, then the expected population size tends to 0, whereas if $\mu > 1$, the expected population size grows exponentially fast.

4.5 MMSE

4.5.1 Orthogonality Property

Before we apply the conditional expectation to the problem of prediction, we first note a few important properties of conditional expectation. The first is that the conditional expectation is linear, which is a crucial property that carries over from ordinary expectations. For example, $\mathbb{E}(X + Y \mid Z) = \mathbb{E}(X \mid Z) + \mathbb{E}(Y \mid Z)$. The justification for this, briefly, is that $\mathbb{E}[X + Y \mid Z = z] = \mathbb{E}[X \mid Z = z] + \mathbb{E}[Y \mid Z = z]$.

The second important property is: let $f(X)$ be any function of X and $g(Y)$ any function of Y. Then the conditional expectation $\mathbb{E}(f(X)g(Y) \mid Y) = g(Y)\,\mathbb{E}(f(X) \mid Y)$. The intuitive idea behind this property is that when we condition on Y, then any function of Y is treated as a constant and can be moved outside of the expectation by linearity. In other words, $\mathbb{E}[f(X)g(Y) \mid Y = y] = \mathbb{E}[f(X)g(y) \mid Y = y] = g(y)\,\mathbb{E}[f(X) \mid Y = y]$. Then, to obtain $\mathbb{E}(f(X)g(Y) \mid Y)$, we perform our usual procedure of substituting $y \mapsto Y$.

Let us now prove the analoge of the orthogonality property for $\mathbb{E}(Y \mid X)$.

Theorem 4.22 (Orthogonality Property)**.** *Let X and Y be random variables, and let $\phi(X)$ be any function of X. Then we have:*

$$\boxed{\mathbb{E}\big[\big(Y - E(Y \mid X)\big)\phi(X)\big] = 0} \tag{4.22}$$

Proof. We first calculate $\mathbb{E}((Y - \mathbb{E}(Y \mid X))\phi(X) \mid X)$ using the properties of conditional expectation.

$$\begin{aligned}\mathbb{E}\big\{\big(Y - \mathbb{E}(Y \mid X)\big)\phi(X) \;\big|\; X\big\} &= \phi(X)\,\mathbb{E}\big(Y - \mathbb{E}(Y \mid X) \;\big|\; X\big) \\ &= \phi(X)\big\{\mathbb{E}(Y \mid X) - \mathbb{E}\big(\mathbb{E}(Y \mid X) \;\big|\; X\big)\big\}\end{aligned}$$

$$= \phi(X)\big(\mathbb{E}(Y \mid X) - \mathbb{E}(Y \mid X)\big) = 0$$

Note: Why is $\mathbb{E}(\mathbb{E}(Y \mid X) \mid X) = \mathbb{E}(Y \mid X)$? We know that $\mathbb{E}(Y \mid X)$ is a function of X, and conditioned on the value of X, $\mathbb{E}(Y \mid X)$ is essentially a constant. Now observe that Theorem 4.18 gives us

$$\mathbb{E}\big[\big(Y - \mathbb{E}(Y \mid X)\big)\phi(X)\big] = \mathbb{E}\big[\mathbb{E}\big\{\big(Y - \mathbb{E}(Y \mid X)\big)\phi(X) \;\big|\; X\big\}\big] = 0. \qquad \square$$

In our proof, we used the useful trick of conditioning on a variable first in order to use the law of iterated expectation. Compare with the orthogonality property for the LLSE: the orthogonality property for conditional expectation is stronger in that $\phi(X)$ is allowed to be any function of X, whereas the orthogonality property for the LLSE was proven for linear functions of X.

4.5.2 Minimizing Mean Squared Error

Definition 4.23. The **minimum mean square error (MMSE)** estimator of Y given X is the random variable $f(X)$ which minimizes the mean squared error, i.e. for any function g,

$$\mathbb{E}\big[\big(Y - f(X)\big)^2\big] \leq \mathbb{E}\big[\big(Y - g(X)\big)^2\big].$$

Compared to the task of finding the best *linear* estimator of Y given X, finding the best *general* estimator of Y given X seems to be an even more difficult task. However, the MMSE will simply turn out to be our new friend, the conditional expectation. In fact, the proof is virtually the same as the proof for the LLSE.

Theorem 4.24 (MMSE)**.** *Let X and Y be random variables. Then the MMSE of Y given X is $\mathbb{E}(Y \mid X)$, i.e. for any function g,*

$$\mathbb{E}\big[\big(Y - E(Y \mid X)\big)^2\big] \leq \mathbb{E}\big[\big(Y - g(X)\big)^2\big]. \tag{4.23}$$

Proof. Let $\hat{Y} = \mathbb{E}(Y \mid X)$ for simplicity of notation. We have, using the orthogonality property,

$$\begin{aligned}
\mathbb{E}\big[\big(Y - g(X)\big)^2\big] &= \mathbb{E}[(Y - \hat{Y} + \hat{Y} - g(X))^2] \\
&= \mathbb{E}[(Y - \hat{Y})^2] + 2\underbrace{\mathbb{E}\big[(Y - \hat{Y})\big(\hat{Y} - g(X)\big)\big]}_{=0} + \mathbb{E}\big[\big(\hat{Y} - g(X)\big)^2\big] \\
&= \mathbb{E}[(Y - \hat{Y})^2] + \mathbb{E}\big[\big(\hat{Y} - g(X)\big)^2\big].
\end{aligned}$$

The term $\mathbb{E}[(Y - \mathbb{E}(Y \mid X))(\mathbb{E}(Y \mid X) - g(X))]$ vanishes by the orthogonality property Theorem 4.22 applied to $\phi(X) = \mathbb{E}(Y \mid X) - g(X)$. $\square$

We have come a long way, and the answer is surprisingly intuitive. We have just found the *best estimator* of Y given X (in the mean squared error sense) is simply the expected value of Y given X!

4.6 Conditional Variance

Definition 4.25. Let X and Y be random variables. We define $\operatorname{var}(X \mid Y = y)$ to be the variance of the conditional probability distribution $\mathbb{P}(X = x \mid Y = y)$. Furthermore, the **conditional variance** $\operatorname{var}(X \mid Y)$ is defined to be the random variable that takes on the value $\operatorname{var}(X \mid Y = y)$ with probability $\mathbb{P}(Y = y)$.

Note that $\operatorname{var}(X \mid Y)$ is a function of Y, analogously to $\mathbb{E}(X \mid Y)$.

Theorem 4.26 (Law of Total Variance)**.** *Let X and Y be random variables. Then:*

$$\boxed{\operatorname{var} X = \mathbb{E}[\operatorname{var}(X \mid Y)] + \operatorname{var} \mathbb{E}(X \mid Y)} \tag{4.24}$$

Proof. First, we use the computational formula for the variance.

$$\operatorname{var} X = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

We calculate each term by the law of iterated expectation.

$$\begin{aligned}
\operatorname{var} X &= \mathbb{E}[\mathbb{E}(X^2 \mid Y)] - \mathbb{E}[\mathbb{E}(X \mid Y)]^2 \\
&= \mathbb{E}[\mathbb{E}(X^2 \mid Y)] - \mathbb{E}[\mathbb{E}(X \mid Y)^2] + \mathbb{E}[\mathbb{E}(X \mid Y)^2] - \mathbb{E}[\mathbb{E}(X \mid Y)]^2 \\
&= \mathbb{E}\big[\mathbb{E}(X^2 \mid Y) - \mathbb{E}(X \mid Y)^2\big] + \operatorname{var} \mathbb{E}(X \mid Y) \\
&= \mathbb{E}[\operatorname{var}(X \mid Y)] + \operatorname{var} \mathbb{E}(X \mid Y)
\end{aligned}$$

□

The formula above is the analogue of the law of iterated expectation for variance.

Example 4.27 (Variance of a Random Sum)**.** Suppose $X = X_1 + \cdots + X_N$, where N is a non-negative integer-valued random variable. What is $\operatorname{var} X$?

We compute $\mathbb{E}(X \mid N) = N\,\mathbb{E}[X_i]$ and $\operatorname{var}(X \mid N) = N \operatorname{var} X_i$. Then,

$$\operatorname{var} X = \mathbb{E}[\operatorname{var}(X \mid N)] + \operatorname{var} \mathbb{E}(X \mid N) = \mathbb{E}[N] \operatorname{var} X_i + (\operatorname{var} N)\,\mathbb{E}[X_i]^2.$$

Example 4.28 (Galton-Watson Branching Process II)**.** Let us revisit Example 4.21 with the additional assumption that the number of offspring has variance σ^2. Using our new tools, we can now compute $\operatorname{var} X_n$.

As we computed above, $\mathbb{E}(X_n \mid X_{n-1}) = \mu X_{n-1}$. Given that there are m individuals in the population, the variance of the population at the next time step is $m\sigma^2$ by linearity of variance (recall that we assume that the individuals' offsprings are independent). Hence, $\operatorname{var}(X_n \mid X_{n-1}) = \sigma^2 X_{n-1}$. Now, we may apply (4.24):

$$\operatorname{var} X_n = \mathbb{E}[\sigma^2 X_{n-1}] + \operatorname{var}(\mu X_{n-1}) = \sigma^2 \mu^{n-1} N + \mu^2 \operatorname{var} X_{n-1}$$

First, suppose that $\mu = 1$. Then, the recurrence simplifies to $\operatorname{var} X_n = \sigma^2 N + \operatorname{var} X_{n-1}$, which means that the variance increases linearly: $\operatorname{var} X_n = \sigma^2 N n$. (Here, we are assuming that the initial population is fixed, so $\operatorname{var} X_0 = 0$.)

For $\mu \neq 1$, the solution to the recurrence is obtained by finding a pattern after a few iterations:

$$\begin{aligned}\operatorname{var} X_n &= \sigma^2 \mu^{n-1} N + \mu^2 \operatorname{var} X_{n-1} = \sigma^2 \mu^{n-1} N + \sigma^2 \mu^n N + \mu^4 \operatorname{var} X_{n-2} \\ &= \cdots = \sigma^2 \mu^{n-1} N \sum_{k=0}^{n-1} \mu^k = \sigma^2 \mu^{n-1} N \frac{1-\mu^n}{1-\mu}\end{aligned}$$

We have used the formula for a finite geometric series.

4.7 Solutions to Exercises

Exercise 19 From the proof of Theorem 4.12, we see that $Y^* = \pm X^*$. Substituting in the definitions of X^* and Y^*, we have $(Y - \mu_Y)/\sigma_Y = \pm(X - \mu_X)/\sigma_X$, so

$$Y = \underbrace{\pm \frac{\sigma_Y}{\sigma_X}}_{a} X + \underbrace{\mu_Y \mp \frac{\sigma_Y}{\sigma_X} \mu_X}_{b}.$$

Exercise 20 Since $Y - L(Y \mid X)$ is zero mean,

$$\mathbb{E}\big[\big(Y - L(Y \mid X)\big)^2\big] = \operatorname{cov}\big(Y - L(Y \mid X), Y - L(Y \mid X)\big).$$

Note that $\operatorname{cov}(Y - L(Y \mid X), L(Y \mid X)) = 0$ by the orthogonality property, and thus the squared error is $\operatorname{cov}(Y - L(Y \mid X), Y) = \operatorname{var} Y - \operatorname{cov}(L(Y \mid X), Y)$. Plugging in the expression for $L(Y \mid X)$,

$$\operatorname{cov}\big(L(Y \mid X), Y\big) = \operatorname{cov}\Big(\mathbb{E}[Y] + \frac{\operatorname{cov}(X,Y)}{\operatorname{var} X} X, Y\Big) = \frac{\operatorname{cov}(X,Y)^2}{\operatorname{var} X},$$

so the squared error is $\operatorname{var} Y - \operatorname{cov}(X,Y)^2/(\operatorname{var} X)$.

Chapter 5

Markov Chains

Markov chains are an important class of probabilistic models that lend themselves favorably to analysis. First, we describe how to calculate useful properties of Markov chains: expected hitting times and absorption probabilities. Then, we discuss the limiting behavior of Markov chains and classify their long-term behavior.

5.1 Introduction

Markov chains are an important probabilistic method of modeling processes over time. Concretely, consider a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$. (Note that we are only considering *discrete-time* Markov chains.) We say that X_n represents the state of a system at time n, and we are interested in the behavior of the system. In general, a system could depend on the entire history of the system until that point, that is, the distribution of X_n could depend on $X_0, \ldots, X_{n-1}$. It is very difficult to analyze such systems, however, so we make a powerful simplifying assumption: for each positive integer n and each sequence of states $x_0, \ldots, x_n$,

$$\mathbb{P}(X_n = x_n \mid X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) = \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}). \tag{5.1}$$

We call this assumption the **Markov property**. A common way to describe the Markov property is: *the future depends on the past only through the present.* In many situations, the Markov property is a very reasonable assumption to make. Even when the validity of the assumption is questionable, the Markov property may still be necessary to make calculations tractable and analysis possible.

Let us be explicit in what a Markov chain entails. A Markov chain consists of:

- A **state space** $\mathcal{X}$. We consider only finite discrete-time Markov chains in this course, that is, Markov chains with a finite number of states.
- An **initial probability distribution** over the states, π_0.
- The probabilities of transitioning between states: for every pair of states i and j, we must specify $P(i, j)$, the probability of moving from state i to state j. Once we are in

state i, we know that we must transition to *some* state at the next time step (even if we transition back to state i), so we require $\sum_{j\in\mathcal{X}} P(i,j) = 1$.

From the three components of a Markov chain, we can define a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ such that for all states i, j,

$$\mathbb{P}(X_0 = i) = \pi_0(i) \tag{5.2}$$

and

$$\mathbb{P}(X_{n+1} = j \mid X_n = i) = P(i,j). \tag{5.3}$$

Notice that (5.3) uses the Markov property discussed above. The interpretation of the random variables is that the system transitions from state to state at every time step, and X_n represents the state of the system at time n.

There are $|\mathcal{X}|^2$ transition probabilities $P(i,j)$, and we often organize them into a matrix P such that the (i,j) entry of the matrix is $P(i,j)$. We call P the **transition probability matrix**. The condition $\sum_{j\in\mathcal{X}} P(i,j) = 1$ means that each row of P must sum to 1. (**Warning**: There is another convention in which the *columns* of the transition matrix must sum to 1, but this is only done in linear algebra courses.)

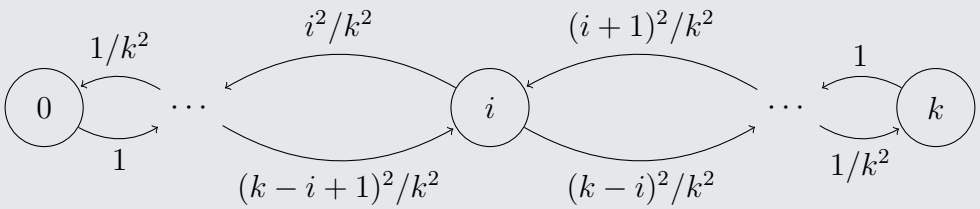
Exercise 21 Let $\mathcal{X}$ and $\mathcal{Y}$ be finite state spaces, and let $f : \mathcal{X} \to \mathcal{Y}$ be a function. Show that if $\{X_n\}_{n\in\mathbb{N}}$ is a Markov chain on $\mathcal{X}$, then $\{f(X_n)\}_{n\in\mathbb{N}}$ may not be a Markov chain (that is, $\{f(X_n)\}_{n\in\mathbb{N}}$ may not satisfy the Markov property (5.1)).

Instead of writing the cumbersome notation $\mathbb{P}(X_n = \cdot)$ for the distribution of X_n, we will use the notation π_n to mean the distribution of X_n: $\pi_n(\cdot) = \mathbb{P}(X_n = \cdot)$.

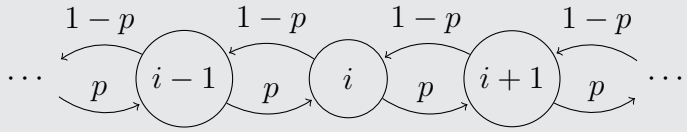
Example 5.1 (Bernoulli-Laplace Diffusion)**.** As a simple model for diffusion, suppose there are k white marbles and k black marbles in two boxes (k marbles in each box, where k is a positive integer). At each time step, we pick one marble from each box and exchange them. We can model the situation as a Markov chain, where the states are $\mathcal{X} = \{0, \ldots, k\}$, representing the number of white marbles in the first box.

Suppose that there are currently $i \in \{1, \ldots, k-1\}$ white marbles in the first box. Then, the probability that the number of white marbles will decrease to $i-1$ is i/k (the probability of choosing a white marble from the first box), multiplied by i/k (the probability of choosing a black marble from the second box). By a similar argument, the probability that the number of white marbles will increase to $i+1$ is $(k-i)^2/k^2$, so our transition probabilities are

$$P(i,j) = \begin{cases} i^2/k^2, & j = i-1, \\ 2i(k-i)/k^2, & j = i, \\ (k-i)^2/k^2, & j = i+1. \end{cases}$$



Example 5.2 (Random Walk I)**.** A **random walk** is an example of a Markov chain in which the state space $\mathcal{X}$ is infinite. Take $\mathcal{X} = \mathbb{Z}$ and let $P(i, i-1) = 1-p$ and $P(i, i+1) = p$ for all $i \in \mathcal{X}$. If $p = 1/2$, we say the Markov chain is a **simple random walk**.



Example 5.3 (Queueing)**.** As another example of an infinite Markov chain, consider a model for **queueing** in which X_n represents the number of customers in the queue at time n. At time n, one customer is removed from the queue, and Y_n new customers arrive, where the Y_n are i.i.d. and take values in $\mathbb{N}$. The state space is $\mathcal{X} = \mathbb{N}$, representing the number of customers in the queue. For $i \in \mathcal{X} \setminus \{0\}$, the number of customers is reduced to $i-1$, and then the probability of transitioning to $j \in \{i-1, i, i+1, \dots\}$ is the probability that $j - (i-1)$ customers arrive.

$$P(i, j) = \mathbb{P}(Y_n = j - i + 1)$$

Along with $P(0, j) = P(1, j)$ for $j \in \mathcal{X}$, we have specified the transition probabilities.

5.2 Transition of Distribution

The first question we can ask is: given a transition matrix P, how can we compute the distribution at time $n \in \mathbb{Z}_+$, π_n? Here is where the Markov property shines: we can easily compute π_n if we know the distribution at one time step prior. Therefore, suppose we are given π_{n-1}. We can compute, for $j \in \mathcal{X}$,

$$\begin{aligned}\pi_n(j) &= \sum_{i \in \mathcal{X}} \mathbb{P}(X_{n-1} = i, X_n = j) = \sum_{i \in \mathcal{X}} \mathbb{P}(X_{n-1} = i)\mathbb{P}(X_n = j \mid X_{n-1} = i) \\ &= \sum_{i \in \mathcal{X}} \pi_{n-1}(i) P(i, j).\end{aligned}$$

Actually, the equation above has exactly the same form as matrix multiplication. If we write π_n as a row vector and P as a matrix, then we have found:

$$\boxed{\pi_n = \pi_{n-1} P} \tag{5.4}$$

Iterating this, we can obtain the distribution at time n starting from the initial distribution.

$$\boxed{\pi_n = \pi_0 P^n} \tag{5.5}$$

Passing from one time step to the next is simply multiplication by the transition matrix. In practice, it is extremely easy for computers to carry out this matrix multiplication, so Markov chain models are often very tractable. In theory, we can use the power of linear algebra in order to analyze a probabilistic model. Hopefully, you can begin to see why Markov chains are such a useful family of models.

5.3 Markov Chain Computations

5.3.1 Hitting Time

Let $x, z \in \mathcal{X}$. We will define the following random variables:

$$T_z = \min\{n \in \mathbb{N} : X_n = z\},$$
$$T_z^+ = \min\{n \in \mathbb{Z}_+ : X_n = z\},$$

that is, T_z is the *first time we reach state* z. [The only difference between T_z and T_z^+ is that if the Markov chain starts from state z, then $T_z = 0$, but $T_z^+ > 0$.] We are interested in computing $\mathbb{E}_x[T_z]$, the *expected hitting time of* z *starting from state* x. [The subscript on $\mathbb{E}_x$ indicates that we are starting from state x.] We observe that $\mathbb{E}_z[T_z] = 0$.

Although we are interested in computing $\mathbb{E}_x[T_z]$, we will find that computing $\mathbb{E}_x[T_z]$ requires computing $\mathbb{E}_y[T_z]$ for *all* states $y \in \mathcal{X}$. The **first-step equations (FSE)** are:

$$\boxed{\begin{aligned} &\mathbb{E}_z[T_z] = 0, \\ &\mathbb{E}_y[T_z] = 1 + \sum_{y' \in \mathcal{X}} P(y, y')\, \mathbb{E}_{y'}[T_z], \qquad \forall y \in \mathcal{X}. \end{aligned}} \tag{5.6}$$

The first-step equations say: in order to reach z from y, we must first take a step (that is the $+1$ in the equations above). After we take a step, we transition to some other step y', and once we are in state y', we expect to take $\mathbb{E}_{y'}[T_z]$ more steps until we reach z. Of course, we must weight each possiblity by the probability of transitioning to the state y', and we sum over all states to account for all possibilities.

When we carry out this procedure, we end up with $|\mathcal{X}|$ equations, one for each state. Each equation is *linear* in the $\{\mathbb{E}_y[T_z]\}_{y \in \mathcal{X}}$, so we can solve the system just as we solve any other linear system (either through repeated substitution or Gaussian elimination). Once we solve the system, we have $\mathbb{E}_x[T_z]$, which is the expected time to reach z from x.

Example 5.4. Suppose $X \sim \text{Geometric}(p)$. Previously, we have computed $\mathbb{E}[X] = 1/p$. Here, we will show how to model X as a Markov chain and compute the expectation.

Recall that a geometric random variable represents the number of flips of a biased coin

until we see heads. Therefore, we will define a two-state Markov chain with states $\mathcal{X} = \{1, 2\}$, where state 1 represents the state of "we are still flipping the coin" and state 2 represents the terminal state ("we are done flipping the coin"). Then, we can see that the expected time to reach state 2 from state 1 is precisely the expectation of X. We have $\mathbb{E}_2[T_2] = 0$, and

$$\mathbb{E}_1[T_2] = 1 + (1-p)\,\mathbb{E}_1[T_2] + p\,\mathbb{E}_2[T_2] = 1 + (1-p)\,\mathbb{E}_1[T_2],$$

since with probability $1-p$, we see tails and we have to keep flipping, and with probability p, we see heads and we move to the terminal state. Now, we can solve the equation for $\mathbb{E}_1[T_2]$ and we obtain $p\,\mathbb{E}_1[T_2] = 1$, or $\mathbb{E}_1[T_2] = 1/p$.

We can generalize the above method. Suppose that each state x gives you a reward r_x for visiting that state, and we want to find the expected reward that we accumulate before we reach any of the nodes in a set S. Observe that the first-step equations above correspond to the special case where each state gives a reward of 1 and the goal is the set $S = \{z\}$. In general, define R_S to be the expected sum of rewards that we accumulate until we reach a state in S. Then, we have the **expected reward equations**:

$$\boxed{\begin{aligned} &\mathbb{E}_x[R_S] = 0, && \forall x \in S, \\ &\mathbb{E}_x[R_S] = r_x + \sum_{y \in \mathcal{X}} P(x,y)\,\mathbb{E}_y[R_S], && \forall x \notin S. \end{aligned}} \tag{5.7}$$

Here, we are saying that the accumulated rewards consist of the reward we obtain now, plus the reward that we expect at state y (weighted by the probability of transitioning to y).

Caution: Do *not* memorize the equations in this section! Instead, carefully examine their meaning and try to understand why they solve the problem we are trying to solve. Make these equations part of your intuition.

5.3.2 Probability of S before S'

We have found how to compute the expected time to reach state z from state x. Now, consider the problem of reaching S *before* S', where S and S' are disjoint sets of states. We introduce the random variable $T_S = \min\{n \in \mathbb{N} : X_n \in S\}$, the first time that the Markov chain reaches a state in S, and we use the notation $\mathbb{P}_x$ to indicate that the Markov chain is starting from state x. Then, $\mathbb{P}_x(T_S < T_{S'})$ the probability of reaching S before S', starting from state x. The equations for $\mathbb{P}_x(T_S < T_{S'})$ are:

$$\boxed{\begin{aligned} &\mathbb{P}_x(T_S < T_{S'}) = 0, && \forall x \in S', \\ &\mathbb{P}_x(T_S < T_{S'}) = 1, && \forall x \in S, \\ &\mathbb{P}_x(T_S < T_{S'}) = \sum_{y \in \mathcal{X}} P(x,y)\mathbb{P}_y(T_S < T_{S'}), && \forall x \notin S \cup S'. \end{aligned}} \tag{5.8}$$

These equations are similar to the first-step equations in that they "look ahead" one time step. For each possible state y, weighted by the probability of transitioning to y, we sum

up the probability of reaching S before S' starting at y. Solving these equations amounts to solving another system of linear equations.

Example 5.5 (Gambler's Ruin)**.** Consider a model of a gambler who starts with $a \in \mathbb{N}$ dollars and wishes to leave the casino with $b > a$ dollars ($b \in \mathbb{N}$). The set of states is $\mathcal{X} = \{0, \ldots, b\}$ representing the amount of money that the gambler possesses. The gambler wins a bet with probability p, and all bets are for one dollar. What is the probability of $\{b\}$ before $\{0\}$, that is, what is the probability that the gambler will succeed before going bankrupt?

For this example, we will use the simpler notation $\alpha(i)$ for the probability of reaching b before 0, starting with i dollars (for $i \in \mathcal{X}$). We have the equations:

$$\begin{aligned} &\alpha(0) = 0, \\ &\alpha(b) = 1, \\ &\alpha(i) = (1-p)\alpha(i-1) + p\alpha(i+1), \qquad i \notin \{0, b\}. \end{aligned} \tag{5.9}$$

The equations (5.9) above is known as a **difference equation** or a **linear recurrence relation**. The equations (5.9) are difficult to solve directly, but we will solve them by judicious guessing. In other words, we will conjecture the form of $\alpha(i)$, and then plug in our guess for $\alpha(i)$ into (5.9) to verify that our guess was indeed correct. If you have never worked with differential equations or difference equations before, the guesses may seem strange and unmotivated, but unfortunately we do not have the time to give a full treatment of the solutions of such equations.

First, we assume that $p = 1/2$ (the odds are even). We search for a solution of the form $\alpha(i) = c_1 i + c_0$, where c_0 and c_1 are constants that are currently unknown. We will solve for what c_0 and c_1 must be by plugging in our guess for $\alpha(i)$ into (5.9). Indeed, from $\alpha(0) = 0$, we see that $c_0 = 0$, and from $\alpha(b) = 1$, we see that $c_1 = 1/b$. Hence, our refined guess is $\alpha(i) = i/b$. To verify that our guess for $\alpha(i)$ really represents a valid solution, we plug the result into (5.9).

$$\frac{i}{b} = \frac{1}{2}\Big(\frac{i-1}{b} + \frac{i+1}{b}\Big). \qquad \checkmark$$

Hence, we have found the correct solution: the probability of success increases linearly with the amount of money with which we start.

Now, suppose that $p \neq 1/2$. We search for a solution of the form $\alpha(i) = c_1\rho^i + c_0$. Plugging in our guess for $\alpha(i)$ into (5.9), we obtain:

$$c_1\rho^i + c_0 = (1-p)(c_1\rho^{i-1} + c_0) + p(c_1\rho^{i+1} + c_0)$$

After cancelling terms and simplifying, we obtain the equation

$$p\rho^2 - \rho + 1 - p = 0.$$

The equation is quadratic in ρ, and it is solved by choosing the solution corresponding to $\rho < 1$. (One can verify that the solution $\rho = 1$ does not satisfy $\alpha(0) = 0$ and $\alpha(b) = 1$.)

$$\rho = \frac{1 - \sqrt{1 - 4p(1-p)}}{2p} = \frac{1 - (2p-1)}{2p} = \frac{1-p}{p}$$

Setting $\alpha(0) = 0$, we have $c_0 + c_1 = 0$. Setting $\alpha(b) = 1$, we have $c_0 + c_1\rho^b = 1$. We solve to find $c_0 = -c_1 = (1-\rho^b)^{-1}$, from which we extract our desired solution:

$$\alpha(i) = \frac{1-\rho^i}{1-\rho^b}$$

If $p < 1/2$, then $\rho < 1$, and this spells bad news for the gambler. (You should try plugging in various values of ρ, i, and b above to obtain numerical results.)

Now, we can repeat the entire argument above, switching p and q (which amounts to replacing ρ with ρ^{-1}) and i with $b-i$. This corresponds to computing the probability of reaching 0 before b starting from state i, which we denote as $\bar{\alpha}(i)$. We have the results

$$\bar{\alpha}(i) = \begin{cases} \dfrac{b-i}{b}, & p = 1/2, \\ \dfrac{1-\rho^{-(b-i)}}{1-\rho^{-b}}, & p \neq 1/2. \end{cases}$$

It can be seen that $\alpha(i) + \bar{\alpha}(i) = 1$, which means that we are guaranteed to hit b before 0 or 0 before b. It is impossible for the gambler to be stuck at the casino forever.

Example 5.6 (Gambler's Ruin II)**.** In Example 5.5, we found that $\alpha(i)$, the probability of reaching b before 0, starting from state i, is

$$\alpha(i) = \begin{cases} \dfrac{i}{b}, & p = 1/2, \\ \dfrac{1-\rho^i}{1-\rho^b}, & p \neq 1/2, \end{cases}$$

where $\rho = (1-p)/p$. Equivalently, we can think of $\alpha(i)$ as the probability of winning $b-i$ additional dollars before losing i dollars. Here, money is measured relative to i, the

amount of money that was brought into the casino. By replacing b with $b+i$, we obtain

$$\tilde{\alpha}(i) = \begin{cases} \dfrac{i}{b+i}, & p = 1/2, \\ \dfrac{1-\rho^i}{1-\rho^{b+i}}, & p \neq 1/2, \end{cases}$$

the probability of winning b dollars before losing i dollars. Think of $\tilde{\alpha}(i)$ as the probability that, starting with no money, we reach b before $-i$. Now, let $i \to \infty$: we are allowing the gambler to have infinite capital (so the gambler is allowed to incur any amount of debt), and we are interested in the probability that the gambler will ever profit b dollars.

Case 1. $p > 1/2$: Here, $\rho < 1$, so $\rho^i \to 0$ and $\tilde{\alpha}(i) \to 1$.

Case 2. $p = 1/2$: The ratio $i/(b+i)$ also tends to 1 as $i \to \infty$.

Case 3. $p < 1/2$: Since $\rho > 1$, we can rearrange the formula to

$$\tilde{\alpha}(i) = \frac{1-\rho^i}{1-\rho^{b+i}} = \frac{\rho^{-i}-1}{\rho^{-i}-\rho^b} \to \frac{1}{\rho^b} = \left(\frac{p}{1-p}\right)^b$$

as $i \to \infty$, since $\rho^{-i} \to 0$.

In summary, a gambler with infinite capital is guaranteed to profit b dollars if $p \geq 1/2$; otherwise, the probability is $(p/(1-p))^b$.

Example 5.7. There is a simpler way of computing the probability of one event before another which is conceptually interesting. As a concrete example, suppose that we have two machines which fail with probabilities p_1 and p_2 respectively. When either one of the machine fails, the entire factory shuts down. What is the probability that the first machine has failed when the factory shuts down?

The question is really asking: *given* that the factory shuts down at some time step, what is the probability that the first machine has failed? We can apply the law of conditional probability here. Let M_i denote the event that machine i fails on the specified time step.

$$\mathbb{P}(M_1 \mid M_1 \cup M_2) = \frac{\mathbb{P}(M_1)}{\mathbb{P}(M_1 \cup M_2)} = \frac{p_1}{p_1 + p_2 - p_1 p_2}$$

(Apply the inclusion-exclusion rule.) See if you can obtain the same answer using a Markov chain model.

Exercise 22 Consider a symmetric random walk on the vertices of the three-dimensional hypercube Q_3, that is, at each step of the walk we choose one of the neighboring vertices on Q_3 uniformly at random and transition to that vertex. Suppose that vertices 000 and 111

are made absorbing (i.e., $P(000, 000) = P(111, 111) = 1$). For each vertex v, what is the probability that the random walk starting at v is absorbed in 000?

5.4 Long-Term Behavior of Markov Chains

Now that we have seen how to solve some problems with Markov chains, let us analyze their long-term behavior.

5.4.1 Classification of States

First, we observe that there are only two types of states: *transient* states and *recurrent* states. Transient states are states from which it it is possible to leave and never return. For example, suppose that the state i has a non-zero probability of transitioning to state j, but once we are in state j, then there is no way to return to state i. Since a Markov chain runs forever, eventually we must transition from i to j, and for all time steps afterwards, we will never return to i. Therefore, we must have $\pi_n(i) \xrightarrow[n\to\infty]{} 0$.

The other possibility is a recurrent state, which is the opposite of the situation described above. Suppose that we start from state i, and regardless of the next state, there is always a positive probability that we will return to state i. In this case, since a Markov chain runs forever, we must necessarily return to state i infinitely many times, which means we visit state i infinitely often during the course of our Markov chain.

We now formalize our observations. Let $f_n(i, j)$ denote the probability that a system starting at state i first visits the state j at time n, $n \in \mathbb{N}$. Let $\rho(i, j) := \sum_{n=1}^{\infty} f_n(i, j)$ denote the probability that a system starting at state i eventually visits state j.

Definition 5.8. In a Markov chain, a state i is **recurrent** if $\rho(i, i) = 1$ and **transient** otherwise.

Our next observation is beautifully summarized by the phrase "recurrence is contagious".

Lemma 5.9. *Consider a Markov chain on the countable state space $\mathcal{X}$. If $x, y \in \mathcal{X}$ are such that x is recurrent and $\rho(x, y) > 0$, then y is also recurrent and $\rho(y, x) = 1$.*

Proof. Suppose, for the sake of contradiction, that $\rho(y, x) < 1$. Since $\rho(x, y) > 0$, the Markov chain can reach y from x with positive probability; and once the chain has reached y, since we are assuming that $\rho(y, x) < 1$, there is a positive probability that the chain will never return to x. In other words, we have a positive probability that the chain will leave x and never return to x, but this contradicts the fact that x is a recurrent state (which says $\rho(x, x) = 1$). So, we must have $\rho(y, x) = 1$.

Now, we can prove the first claim, namely, that y is a recurrent state. From the previous part, we know that the chain starting from y is certain to eventually reach x. Once the chain has reached x, there is some positive probability that the chain will eventually reach y in, say, K steps (where $K \in \mathbb{N}$), since $\rho(x, y) > 0$. If the chain has not reached y in K steps, then we know that the chain will eventually return to x (since x is recurrent), and again the chain has a positive probability of reaching y in K steps. Here is the logic: Each time the chain returns to x, it has a positive probability of eventually reaching y, and the chain returns to x infinitely often, so the chain must be guaranteed to eventually reach y. We have proved that the chain starting from y is guaranteed to return to y, that is, $\rho(y, y) = 1$, and so y is recurrent. □

Something strange happens with Lemma 5.9. We start with the assumption that x is recurrent and $\rho(x, y) > 0$. The conclusion of Lemma 5.9 says that y is recurrent. But now we can apply Lemma 5.9 to the recurrent state y and we conclude that $\rho(x, y) = 1$. We started by assuming $\rho(x, y) > 0$, but we were actually able to prove $\rho(x, y) = 1$, amazingly enough! As a consequence, both x and y will be visited infinitely many times.

Moreover, Lemma 5.9 does indeed say that recurrence is contagious. If x is a recurrent state, then all other states that x can eventually reach with positive probability are in fact reached with certainty; these states are also recurrent; and starting from those other states, they are also guaranteed to reach x eventually. Therefore, we have the following decomposition:

Theorem 5.10 (Markov Chain Decomposition)**.** *Consider a Markov chain on the countable state space $\mathcal{X}$. The state space $\mathcal{X}$ can be partitioned into classes of states such that if x, y belong to class C, then x can reach y with positive probability and vice versa. Also, for any given class, exactly one of the following occurs:*

1. *All states in the class are recurrent. Then, the class is called a* ***recurrent class****.*
2. *All states in the class are transient. Then, the class is called a* ***transient class****.*

We now see that recurrence and transience are *class properties*, that is, properties that are shared among all states within a class in the decomposition of a Markov chain.

Remark: The classification of states does not require the Markov chain to be finite.

5.4.2 Invariant Distribution

Essentially, a Markov chain must eventually leave its transient classes and become "trapped" in the recurrent classes. On the other hand, if the Markov chain has only one recurrent class, then it cannot be decomposed any further, so we say that the Markov chain is **irreducible**.

Definition 5.11. A Markov chain is **irreducible** if from any state we can reach any

other state with positive probability. Equivalently, the decomposition of the Markov chain has only one recurrent class and no transient classes.

Irreducibility has consequences for the distribution of a Markov chain. We say that a probability distribution π is an **invariant distribution** if:

$$\boxed{\pi = \pi P} \tag{5.10}$$

Observe that if at any time $n \in \mathbb{N}$, $\pi_n = \pi$, then the distribution does not change at the next time step, which implies that the distribution will forever be the same at every time step from that point onwards (by induction, of course). If we start off with the invariant distribution ($\pi_0 = \pi$), then $\pi_n = \pi$ for all $n \in \mathbb{N}$.

Theorem 5.12 (Existence of the Invariant Distribution)**.** *A finite, irreducible Markov chain has a unique invariant distribution.*

Proof. Fix any recurrent state x and define

$$\mu_x(y) := \mathbb{E}_x\Big[\sum_{n=0}^{T_x^+-1} \mathbb{1}\{X_n = y\}\Big] = \sum_{n=0}^{\infty} \mathbb{P}_x(X_n = y, T_x^+ > n)$$

to be the expected number of visits to state y before returning to state x. We claim that this is invariant in the sense that $\mu_x P = \mu_x$. To verify this, we calculate

$$\sum_{y\in\mathcal{X}} \mu_x(y)P(y,z) = \sum_{n=0}^{\infty}\sum_{y\in\mathcal{X}} \mathbb{P}_x(X_n = y, T_x^+ > n)P(y,z)$$

and we must check that this equals $\mu_x(z)$. For $z = x$,

$$\begin{aligned}\sum_{n=0}^{\infty}\sum_{y\in\mathcal{X}} \mathbb{P}_x(X_n = y, T_x^+ > n)P(y,x) &= \sum_{n=0}^{\infty}\sum_{y\in\mathcal{X}} \mathbb{P}_x(X_n = y, X_{n+1} = x, T_x^+ > n)\\ &= \sum_{n=0}^{\infty} \mathbb{P}_x(X_{n+1} = x, T_x^+ > n) = \sum_{n=0}^{\infty} \mathbb{P}_x(T_x^+ = n+1)\\ &= 1,\end{aligned}$$

(since $\mathbb{P}_x(T_x^+ = 0) = 0$ by the definition of T_x^+), but note that

$$\mu_x(x) = \mathbb{E}_x\Big[\sum_{n=0}^{T_x^+-1} \mathbb{1}\{X_n = x\}\Big] = 1$$

since $X_0 = x$ (the chain starts from state x), and x is not visited between time 1 and $T_x^+ - 1$ by the definition of T_x^+. Now, suppose $z \neq x$.

$$\begin{aligned}\sum_{n=0}^{\infty}\sum_{y\in\mathcal{X}}\mathbb{P}_x(X_n = y, T_x^+ > n)P(y,z) &= \sum_{n=0}^{\infty}\sum_{y\in\mathcal{X}}\mathbb{P}_x(X_n = y, X_{n+1} = z, T_x^+ > n)\\ &= \sum_{n=0}^{\infty}\mathbb{P}(X_{n+1} = z, T_x^+ > n)\\ &= \sum_{n=0}^{\infty}\mathbb{P}(X_{n+1} = z, T_x^+ > n+1)\end{aligned}$$

(since if $T_x^+ > n$ and $X_{n+1} = z \neq x$, then also $T_x^+ > n+1$)

$$= \sum_{n=0}^{\infty}\mathbb{P}(X_n = z, T_x^+ > n) = \mu_x(z).$$

So, we have found a solution to (5.10); the only problem is that μ_x is not a probability distribution (the sum of its entries is not 1). But if we normalize μ_x and define

$$\pi(y) := \frac{\mu_x(y)}{\sum_{z\in\mathcal{X}}\mu_x(z)},$$

then π is an invariant probability distribution. We will not prove uniqueness. □

The idea of the above proof is that $\mu_x(y)$ represents the expected number of visits to y in $0, \ldots, T_x^+ - 1$, whereas $(\mu_x P)(y)$ represents the expected number of visits to y in $1, \ldots, T_x^+$; but the former starts at state x, while the latter ends at state x, and so the expected number of visits to y is the same in each case.

Of course, this has the hidden implication that a Markov chain that is not irreducible need not have a unique invariant distribution. Consider, for example, $P = I_2$ (the 2×2 identity matrix); this is the transition probability matrix for a Markov chain which is *not* irreducible. Notice that *any* distribution is an invariant distribution.

It turns out that the invariant distribution tells us more. For $n \in \mathbb{Z}_+$, let $v_n(i)$ be the number of times that the state i is visited in the first n time steps, so $v_n(i) = \sum_{j=0}^{n-1}\mathbb{1}\{X_j = i\}$. Then the *fraction* of time spent in state i is $v_n(i)/n$, and we have the following theorem:

Theorem 5.13. *Suppose we have a finite, irreducible Markov chain with invariant distribution π. Then, for all states i,*

$$\frac{v_n(i)}{n} \xrightarrow[n\to\infty]{} \pi(i).$$

Thus, the invariant distribution also tells us the long-term fraction of time that we will spend in each state.

5.4.3 Convergence of Distribution

Next, we are interested in studying the problem of whether the distribution of a Markov chain *converges*. We make the following definition:

Definition 5.14. The **limiting distribution** $\tilde{\pi}$ of a Markov chain, if it exists, is the probability distribution to which the Markov chain converges:

$$\tilde{\pi} := \lim_{n\to\infty} \pi_n.$$

In other words, for every state x,

$$\tilde{\pi}(x) := \lim_{n\to\infty} \pi_n(x).$$

In giving the above definition, we used the notation $\tilde{\pi}$ to avoid confusion with the invariant distribution π. The *invariant distribution* and the *limiting distribution* are not the same, but we will see that they are related (Theorem 5.19). An invariant distribution is a probability distribution which satisfies $\pi = \pi P$. The limiting distribution is the distribution to which the Markov chain converges. A Markov chain may not have a limiting distribution if its distribution does not converge; we will see examples of this presently.

If $P = I$, then $\pi_n = \pi_0$ for all $n \in \mathbb{N}$. This means that in the long-run, the behavior of the Markov chain depends heavily on the initial distribution. Is this always the case? Are there Markov chains for which we can make statements about the limiting distribution, regardless of the initial distribution? [1] *Yes*, and the relevant property here is **aperiodicity**.

Example 5.15. Consider the Markov chain on state space $\mathcal{X} = \{1, 2\}$ with transition probability matrix

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

Clearly, the distribution will not tend toward a limit, unless the initial distribution is the stationary distribution $\pi = \begin{bmatrix} 1/2 & 1/2 \end{bmatrix}$. The reason for this is that $\pi_n(1)$ and $\pi_n(2)$ will swap at every time step, so if $\pi_0(1) \neq \pi_0(2)$, then we will not have convergence.

The amazing fact is that, essentially, as long as we can rule out this case, we are guaranteed convergence to a limiting distribution!

Define the **period** of a state i to be the largest integer $d(i) \geq 1$ (if such a number exists) such that the number of time steps it takes to return to state i is necessarily a multiple of

[1] There is a reason why we can hope that such a statement might be true. Recall the law of large numbers for i.i.d. random variables, which says that regardless of the initial behavior of the sequence of random variables, the sample average will "settle down" and converge to the expectation. The hope is that we can make similar statements about the limiting distribution of a Markov chain regardless of the initial conditions of the system.

d. We say that a Markov chain is **aperiodic** if the period if every state is 1. Intuitively, aperiodicity is related to the tendency of the Markov chain to *mix* its states. To give an idea of how aperiodicity is used, we can prove an important lemma using a little number theory.

Lemma 5.16. *Let S be a set of positive integers which is closed under addition, such that $\gcd S = 1$. There exists an integer n_0 such that for all integers $n \geq n_0$, $n \in S$.*

Proof. If $\gcd S = 1$, there is a sequence of positive integers $\{a_n\}_{n\in\mathbb{N}}$ in S with greatest common divisor 1. For $n \in \mathbb{N}$, let $d_n = \gcd\{a_1, \ldots, a_n\}$. Note that d_n is non-increasing with n and $d_n \xrightarrow[n\to\infty]{} 1$, so there exists some $m \in \mathbb{N}$ such that $d_m = 1$. Therefore, we have found a finite subset $\{a_1, \ldots, a_m\} \subseteq S$ with $\gcd\{a_1, \ldots, a_m\} = 1$. We can find integers $c_1, \ldots, c_m$ such that

$$c_1 a_1 + \cdots + c_m a_m = 1.$$

By relabeling the a_n, we can assume that the the first k coefficients are positive and the remaining coefficients are negative, for some $k \in \{0, \ldots, n\}$.

$$\underbrace{c_1 a_1 + \cdots + c_k a_k}_{b_1} - \underbrace{(-c_{k+1} a_{k+1} - \cdots - c_m a_m)}_{b_2} = 1$$

Since S is closed under addition, we have found $b_1, b_2 \in S$ with $b_1 - b_2 = 1$.

Take $n_0 = b_2(b_2 - 1)$. For any integer $n \geq n_0$, we can write $n = qb_2 + r$, where $r \in \{0, \ldots, b_2 - 1\}$ and $q \in \mathbb{N}$, $q \geq b_2 - 1 \geq r$.

$$n = qb_2 + r = qb_2 + r(b_1 - b_2) = (q - r)b_2 + rb_1,$$

which is in S because S is closed under addition. □

Lemma 5.17. *For $i, j \in \mathcal{X}$ for an irreducible, aperiodic Markov chain, there exists a positive integer n' such that $P^n(i, j) > 0$ for all integers $n \geq n'$.*

Proof. Let S be the set $\{n \in \mathbb{Z}_+ : P^n(j, j) > 0\}$. If $n_1, n_2 \in S$, then $n_1 + n_2 \in S$ since $P^{n_1+n_2}(j, j) \geq P^{n_1}(j, j)P^{n_2}(j, j) > 0$, so S is closed under addition. Since the Markov chain is aperiodic, $\gcd S = 1$. Now, Lemma 5.16 gives an integer n_0 such that for all $n \geq n_0$, $P^n(j, j) > 0$.

Since the Markov chain is irreducible, there is some n_1 such that $P^{n_1}(i, j) > 0$. Take $n' = n_0 + n_1$. For any integer $n \geq n'$, observe that $n - n_1 \geq n_0$, so from the result above, $P^n(i, j) \geq P^{n_1}(i, j)P^{n-n_1}(j, j) > 0$, which completes the proof. □

The lemma says that if we continue to raise the transition matrix to successively higher powers, then eventually we will reach a transition matrix with all positive entries. A transition

matrix with this property is known as a **regular transition matrix**; we have just proven that the transition matrix for an irreducible, aperiodic Markov chain is regular.

The fundamental result about finite irreducible aperiodic Markov chains is summarized in the next two theorems, which we will not prove:

Theorem 5.18. *For a finite, irreducible Markov chain, the period of every state is the same.*

In other words, the period is also a class property (just like recurrence and transience).

Theorem 5.19 (Markov Chain Convergence Theorem)**.** *For a finite, irreducible, and aperiodic Markov chain, the distribution of the chain converges. Moreover, the limiting distribution is the invariant distribution.*

5.4.4 A Complete Analysis of the Asymptotic Distribution of Finite Markov Chains

Combining the Markov chain convergence theorem (Theorem 5.19) with a few observations allows us to fully analyze the asymptotic distribution of finite Markov chains. First, if the Markov chain is irreducible and aperiodic; then, we can apply the convergence theorem Theorem 5.19 directly and we have convergence to the unique invariant distribution (which exists, by Theorem 5.12). If the Markov chain is irreducible and periodic, then we saw in Example 5.15 that the distribution may not converge at all.

Now, we tackle the trickier case when the Markov chain is not irreducible. Any probability that enters a recurrent class will remain in that recurrent class forever. If the Markov chain has any transient classes, then eventually all of the probability which was concentrated on the transient classes will flow into the recurrent classes. (Observe also that this is essentially the main obstruction which prevents convergence independent of the initial distribution. The fraction of the probability which eventually comes to rest in a particular recurrent class may change depending on how we initially allocate the probability among the different classes at the start of the chain.)

Since the transient classes "die out", we focus our attention on the recurrent classes now. If any of the recurrent classes is periodic, then again we cannot have convergence of the distribution at all. If all of the recurrent classes are aperiodic, then the distribution will indeed converge, but here the convergence again depends on the initial distribution (because of the comments in the previous paragraph).

We should consider the theory of finite Markov chains to be an astounding success! For a model which has wide-reaching applications and such ease of applicability, it is remarkable that we can also obtain such a complete understanding of its theoretical properties.

Example 5.20. Consider the state space $\mathcal{X} = \{1, 2\}$ and transition probability matrix

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

Then, the Markov chain with transition probability matrix P converges to its unique invariant distribution $\pi = \begin{bmatrix} 0 & 1 \end{bmatrix}$ in one transition, regardless of its initial distribution.

Notice that this Markov chain is not irreducible. The chain has one transient class, $\{1\}$, and one recurrent class, $\{2\}$. All of the probability which was initially in the transient class flows into the recurrent class, and since there is only one recurrent class, after just one transition the Markov chain is essentially irreducible (state 1 is no longer relevant).

5.4.5 Balance Equations

Now that we see the importance of the invariant distribution, we can turn towards calculation. Recall that the invariant distribution satisfies $\pi = \pi P$, which is equivalent to $\pi(P - I) = 0$. This is a matrix-vector equation, so it actually forms a linear system of N equations in the unknowns $\pi(i)$. Using some linear algebra knowledge, it can be shown that one of the equations is redundant, that is, linearly dependent on the rest of the equations (since 1 is an eigenvalue of P, the matrix $P - I$ does not have full rank). Therefore, we augment the system of equations with the normalization condition $\sum_{i \in \mathcal{X}} \pi(i) = 1$, which uniquely specifies the invariant distribution. The system of equations $\pi(P - I) = 0$ is called the **balance equations**.

Example 5.21 (Random Walk on a Circle)**.** Consider i.i.d. random variables $(\xi_n)_{n \in \mathbb{N}}$ with distribution $\mathbb{P}(\xi_n = i) = p_i$ for $i \in \{0, \ldots, N-1\}$. We have $\sum_{i=0}^{N-1} p_i = 1$. Let $X_n = \sum_{i=0}^{n} \xi_i \bmod N$. Then X_n describes a random walk on a circle. It is possible to model the X_n as a Markov chain.

Starting from 0, the probability of transitioning to state i is p_i. Starting from state 1, the probability of transitioning to state i is p_{i-1} (where the index $i - 1$ is also taken modulo N). Generalizing, we see that $P(i, j) = p_{j-i}$ for all $i, j \in \mathcal{X}$, where $j - 1$ is interpreted modulo N. The transition matrix has a special property: each row is a distinct permutation of $\begin{bmatrix} p_0 & \cdots & p_{N-1} \end{bmatrix}$. In particular, the columns sum to 1: $\sum_{i=0}^{N-1} P(i, j) = 1$, for all j. If we assume that each $p_i > 0$, then the Markov chain is irreducible and aperiodic, and we can conjecture a uniform stationary distribution. Indeed, if $\pi(i) = N^{-1}$ for each $i \in \mathcal{X}$,

$$\sum_{i=0}^{N-1} \pi(i) P(i, j) = \frac{1}{N} \underbrace{\sum_{i=0}^{N-1} P(i, j)}_{=1} = \frac{1}{N} = \pi(j), \qquad j \in \mathcal{X}$$

so $\pi = \begin{bmatrix} N^{-1} & \cdots & N^{-1} \end{bmatrix}$ is the unique stationary distribution. Since finite, irreducible, and aperiodic Markov chains converge, we have shown that the distribution of X_n converges to π as $n \to \infty$, which is to say that if the transition probabilities are positive, a random walk on a circle will always converge to the uniform distribution over $\{0, \ldots, N-1\}$.

5.5 Solutions to Exercises

Exercise 21 Consider $\mathcal{X} = \{0, 1, 2\}$, $\mathcal{Y} = \{0, 1\}$, and $f : \mathcal{X} \to \mathcal{Y}$ defined by

$$\begin{aligned} f(0) &= 0, \\ f(1) &= 1, \\ f(2) &= 0. \end{aligned}$$

Let $\{X_n\}_{n \in \mathbb{N}}$ be a Markov chain on $\mathcal{X}$ which we start with a uniform distribution over the three states, with transition probabilities $P(0,1) = P(1,2) = P(2,0) = 1$, that is, the Markov chain moves deterministically $0 \to 1 \to 2 \to 0$. Then, observe that

$$\mathbb{P}\big(f(X_2) = 0 \,\big|\, f(X_1) = 0, f(X_0) = 1\big) = \mathbb{P}(X_2 \in \{0,2\} \mid X_1 \in \{0,2\}, X_0 = 1) = 1$$

since if $X_0 = 1$, then we must have $X_1 = 2$ and $X_2 = 0$, so $f(X_2) = 0$. On the other hand, $\mathbb{P}(f(X_2) = 0 \mid f(X_1) = 0) = \mathbb{P}(X_2 \in \{0,2\} \mid X_1 \in \{0,2\})$, but X_1 is equally likely to be 0 or 2, and $X_1 = 0$ implies $f(X_2) = 1$, while $X_1 = 2$ implies $f(X_2) = 0$; hence, $\mathbb{P}(f(X_2) = 0 \mid f(X_1) = 0) = 1/2$, and we see that $\{f(X_n)\}_{n \in \mathbb{N}}$ does not satisfy the Markov property (5.1).

Exercise 22 The probability of absorption in 000 starting from 111 is 0. By symmetry, the probability of absorption in 000 is the same from 001, 010, and 100, so let this probability be p; similarly, let the probability of absorption in 000 from 011, 101, and 110 be q. Thus,

$$\begin{aligned} p &= \frac{1}{3} \cdot 1 + \frac{2}{3} \cdot q, \\ q &= \frac{1}{3} \cdot 0 + \frac{2}{3} \cdot p, \end{aligned}$$

so $p = 3/5$, $q = 2/5$.

Chapter 6

Continuous Probability I

Our study of discrete random variables has allowed us to model coin flips and dice rolls, but often we would like to study random variables that take on a *continuous* range of values (i.e. an uncountable number of values). At first, understanding continuous random variables will require a conceptual leap, but most of the results from discrete probability carry over into their continuous analogues, with sums replaced by integrals.

6.1 Continuous Probability: A New Intuition

Pick a random number in the interval $[0, 1] \subseteq \mathbb{R}$. What is the probability that we pick *exactly* the number $2/3$? There are uncountably many real numbers in the interval $[0, 1]$, so it seems overwhelmingly unlikely that we pick any particular number. We must find that the probability of choosing exactly $2/3$ is 0 (it is impossible). In fact, for any real number $a \in [0, 1]$, the probability of choosing a must also be 0. But we are guaranteed to choose *some* number in $[0, 1]$, so how is that possible that whatever number we chose was chosen with probability 0? Furthermore, if I consider the probability of choosing a number less than $1/2$, then intuitively, we would like to say the probability is $1/2$. Does this not imply that when we add up a bunch of zero probabilities, we manage to get a non-zero probability? Clearly, our theory of discrete probability breaks down.

Therefore, let us begin with a few definitions. It is natural if you find that you cannot interpret these definitions immediately, since our intuition from discrete probability will require some updates. Over the course of working with continuous probability, you will start to build a new intuition.

The key to resolving the issues raised above is to recall our original definition of a probability measure: a probability measure assigns real numbers to *sets* (not individual outcomes). Even in this new setting, as long as we can consistently assign probability values to sets, then probability theory will continue to work as before. One way of assigning consistent probabilities to sets is through a density function, which will be the focus of our studies.

The **density function** of a continuous random variable X (also known as the probability

density function, or PDF), is a real-valued function f_X such that

1. f_X is non-negative: $\forall x \in \mathbb{R}\ f_X(x) \geq 0$.
2. f_X is **normalized**, which is to say that f_X satisfies:

$$\boxed{\int_{\mathbb{R}} f_X(x)\,\mathrm{d}x = 1} \tag{6.1}$$

Compare the normalization condition here to the normalization condition in discrete probability, which says

$$\sum_x \mathbb{P}(X = x) = 1.$$

We can interpret the continuous normalization condition to mean that "the probability that $X \in \mathbb{R}$ is 1". Similarly, we can define the probability that X lies in some interval $[a, b]$ as:

$$\boxed{\mathbb{P}(X \in [a,b]) := \int_a^b f_X(x)\,\mathrm{d}x} \tag{6.2}$$

Remark: It does not matter whether I write the interval as $[a, b]$ (including the endpoints) or (a, b) (excluding the endpoints). The endpoints themselves do not contribute to the probability, since the probability of a single point is 0 (as discussed above).

The probability of an interval in $\mathbb{R}$ is interpreted as the *area under the density function above the interval*. Similarly, we can calculate the probability of the union of disjoint intervals by adding together the probabilities of each interval.

When we discuss continuous probability, it is also extremely useful to use the **cumulative distribution function** of X, or the CDF, defined as:

$$\boxed{F_X(x) := \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(x')\,\mathrm{d}x'} \tag{6.3}$$

Remark: Once again, it makes no difference whether I write $\mathbb{P}(X \leq x)$ or $\mathbb{P}(X < x)$, since we have that $\mathbb{P}(X = x) = 0$. From now on, I will be sloppy and use the two interchangeably.

To obtain the PDF from the CDF, we use the Fundamental Theorem of Calculus:

$$\boxed{f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x)} \tag{6.4}$$

Exercise 23 Suppose X has density $f(x) = cx(1-x)$ for $x \in [0, 1]$, and $f(x) = 0$ elsewhere, where $c > 0$ is a constant. Find the constant c and the CDF F.

We have given an intepretation of the area under the density function f_X as a probability. The natural question is: what is the interpretation of f_X itself? In the next two sections, we present an interpretation of f_X and introduce two ways of computing distributions in continuous probability.

6.1.1 Differentiate the CDF

To motivate the discussion, we will consider the following example: suppose you pick a random point within a circle of radius 1. Let R be the random variable which denotes the radius of the chosen point (i.e. the distance away from the center of the circle). What is f_R?

The first method is to simply work with the CDF and to obtain f_R by differentiating F_R. Since $F_R(r) := \mathbb{P}(R < r)$, and we have chosen a point uniformly randomly inside of the circle, then the probability we are looking for is the ratio of area of the inner circle (which has radius r) to the area of the total circle (which has radius 1).

$$\mathbb{P}(R < r) = \frac{\text{area inside a circle of radius } r}{\text{total area inside circle}} = \frac{\pi r^2}{\pi} = r^2, \qquad 0 < r < 1$$

Differentiating quickly yields $f_R(r) = 2r$ for $0 < r < 1$. Often, differentiating the CDF is a fast way of finding the density function.

6.1.2 The Differential Method

The second method works with the density function directly, and therefore involves manipulation of differential elements (such as $\mathrm{d}r$). If you have taken physics courses before, then you may already be familiar with the method. If you do not feel comfortable with the manipulations in this section, you can always differentiate the CDF instead.

To briefly motivate the procedure, let us consider $F_R(r + \mathrm{d}r)$. Using a Taylor expansion,

$$F_R(r + \mathrm{d}r) = F_R(r) + \Big(\frac{\mathrm{d}}{\mathrm{d}r} F_R(r)\Big)\,\mathrm{d}r + O\big((\mathrm{d}r)^2\big),$$

where the notation $O((\mathrm{d}r)^2)$ includes terms of order $(\mathrm{d}r)^2$ or higher. Recalling that we have $F_R(r) = \mathbb{P}(R < r)$ and the derivative of the CDF is the density function,

$$\mathbb{P}(R < r + \mathrm{d}r) - \mathbb{P}(R < r) = f_R(r)\,\mathrm{d}r + O\big((\mathrm{d}r)^2\big).$$

Let us immediately apply the formula we have derived to the motivating example. From the CDF, we have that the probability of picking a point with $R < r + \mathrm{d}r$ is $(r + \mathrm{d}r)^2$, and the probability of picking a point with $R < r$ is r^2. Therefore,

$$\mathbb{P}(R < r + \mathrm{d}r) - \mathbb{P}(R < r) = (r + \mathrm{d}r)^2 - r^2 = r^2 + 2r\,\mathrm{d}r + (\mathrm{d}r)^2 - r^2 = 2r\,\mathrm{d}r + (\mathrm{d}r)^2.$$

The expression above must equal $f_R(r)\,\mathrm{d}r + O((\mathrm{d}r)^2)$. Therefore, by looking at the term which is proportional to $\mathrm{d}r$, we can identify $f_R(r) = 2r$ and we obtain the same answer!

Initially, the differential method seems to require more calculations and is not as formal as working with the CDF instead. However, through this discussion we have obtained an *interpretation* for the density function:

$$\boxed{\mathbb{P}(X < x + \mathrm{d}x) - \mathbb{P}(X < x) = \mathbb{P}\big(X \in (x, x + \mathrm{d}x)\big) = f_X(x)\,\mathrm{d}x} \tag{6.5}$$

In words: the probability that the random variable X is found in the interval $(x, x + \mathrm{d}x)$ is proportional to both the length of the interval $\mathrm{d}x$ and the density function evaluated in the interval. The interpretation can be rephrased in the following way: the density function $f_X(x)$ is the *probability per unit length* near x.

The basic procedure for obtaining the density function directly is:

1. Use the information given in the problem to find $\mathbb{P}(X \in (x, x + \mathrm{d}x))$.
2. Drop terms of order $(\mathrm{d}x)^2$ or higher.
3. Identify the term multiplied by $\mathrm{d}x$ as the density function $f_X(x)$.

6.2 Continuous Analogues of Discrete Results

Now, we will return to familiar ground by re-introducing the results from discrete probability in the continuous case. In most cases, replacing summations with integration over $\mathbb{R}$ will work as intended.

The **expectation** of a continuous random variable X is:

$$\boxed{\mathbb{E}[X] := \int_{\mathbb{R}} x f_X(x)\,\mathrm{d}x} \tag{6.6}$$

Similarly, the expectation of a function of X is:

$$\boxed{\mathbb{E}[g(X)] := \int_{\mathbb{R}} g(x) f_X(x)\,\mathrm{d}x} \tag{6.7}$$

We can continue to use the formula $\operatorname{var} X = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ to obtain the variance of a continuous random variable. Since integrals are linear, linearity of expectation still holds.

The **joint density** of two continuous random variables X and Y is $f_{X,Y}$. The joint distribution represents everything there is to know about the two random variables. The joint distribution must satisfy the normalization condition:

$$\boxed{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1} \tag{6.8}$$

We say that X and Y are **independent** if and only if the joint density factorizes:

$$\boxed{f_{X,Y}(x,y) = f_X(x) f_Y(y)} \tag{6.9}$$

To obtain the **marginal distribution** of X from the joint distribution, integrate out the unnecessary variables (in this case, we integrate out Y):

$$\boxed{f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y} \tag{6.10}$$

The joint density can be extended easily to multiple random variables $X_1, \ldots, X_n$, where n is any positive integer. The joint density satisfies the normalization condition:

$$\boxed{\int_{\mathbb{R}^n} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n = 1} \tag{6.11}$$

How do we use the joint density to compute probabilities? Consider a region $J \subseteq \mathbb{R}^2$. The probability that $(X, Y) \in J$ is:

$$\boxed{\mathbb{P}\big((X,Y) \in J\big) := \int_J f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y} \tag{6.12}$$

In other words, *to find the probability that (X, Y) is in a region J, we integrate the joint density over the region J.* As in multivariable calculus, it is often immensely helpful to draw the region of integration (J) before actually computing the integral.

To calculate the expectation of a function of many random variables, we compute:

$$\boxed{\mathbb{E}[g(X_1,\ldots,X_n)] = \int_{\mathbb{R}^n} g(x_1,\ldots,x_n) f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n} \tag{6.13}$$

6.2.1 Tail Sum Formula

There is a continuous analogue of the tail sum formula.

Theorem 6.1 (Continuous Tail Sum Formula)**.** *Let X be a non-negative random variable. Then:*

$$\boxed{\mathbb{E}[X] = \int_0^\infty \big(1 - F_X(x)\big)\,\mathrm{d}x} \tag{6.14}$$

Proof.

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty x f_X(x)\,\mathrm{d}x = \int_0^\infty \int_0^x f_X(x)\,\mathrm{d}t\,\mathrm{d}x = \int_0^\infty \int_t^\infty f_X(x)\,\mathrm{d}x\,\mathrm{d}t \\ &= \int_0^\infty \mathbb{P}(X > t)\,\mathrm{d}t = \int_0^\infty \big(1 - F_X(t)\big)\,\mathrm{d}t\end{aligned}$$

The proof is quite similar to the discrete case. Interchanging the order of integration is justified by Fubini's theorem because the integrand is non-negative. □

Exercise 24 Suppose that $c > 0$ and $X \geq 0$. What is $\mathbb{E}[\min(c, X)]$? (In this example, X is partly discrete and partly continuous. We say that X is an example of a **mixed random variable**.)

6.3 Important Continuous Distributions

6.3.1 Uniform Distribution

The first distribution we will look at in detail is the Uniform$[0,1]$ **distribution**, which means that X is chosen uniformly randomly in the interval $[0,1]$. The property of being *uniform* means that the probability of choosing a number within an interval should only depend on the length of the interval. We can produce this by requiring the density function to equal a constant c in the interval $[0,1]$. Of course, since the density must be normalized to 1, then $c = 1$ and the density function is

$$f_X(x) = 1, \qquad 0 < x < 1. \tag{6.15}$$

The CDF is found by integrating (see Figure 6.1):

$$F_X(x) = \begin{cases} 0, & x \leq 0 \\ x, & 0 < x \leq 1 \\ 1, & x > 1 \end{cases} \tag{6.16}$$

Similarly, suppose that X and Y are i.i.d. Uniform$[0,1]$ random variables. Since they are independent, the joint distribution is simply the product of their respective density functions:

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) = 1, \qquad 0 < x < 1,\ 0 < y < 1$$

The uniform distribution is especially simple because the density is constant. Suppose we want to find the probability that (X,Y) will lie in a region $J \subseteq \mathbb{R}^2$. We can integrate the joint density to obtain:

$$\mathbb{P}\big((X,Y) \in J\big) = \int_J f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_J \mathrm{d}x\,\mathrm{d}y = \text{area } J.$$

The procedure for computing probabilities is therefore very simple. *To find the probability of an event involving two i.i.d.* Uniform$[0,1]$ *random variables X and Y, draw the unit square, and shade in the region in the square which corresponds to the given event. The area of the shaded region is the desired probability.* As a result, many questions involving uniform distributions have very geometrical solutions.

To compute the expectation, we can simply observe that the distribution is symmetric about $x = 1/2$, so we can immediately write down $1/2$ as the expectation. To make the argument slightly more formal, consider the random variable $Y = 1 - X$. Observe that Y and X have identical distributions (but they are *not* independent). Therefore, $\mathbb{E}[X] = \mathbb{E}[Y] = \mathbb{E}[1 - X]$, or $\mathbb{E}[X] = 1 - \mathbb{E}[X]$ using linearity. Hence,

$$\mathbb{E}[X] = \frac{1}{2}. \tag{6.17}$$

One question you may have is: why are X and Y identically distributed? To answer that question in a rigorous way, we will need the change of variables formula in order to show

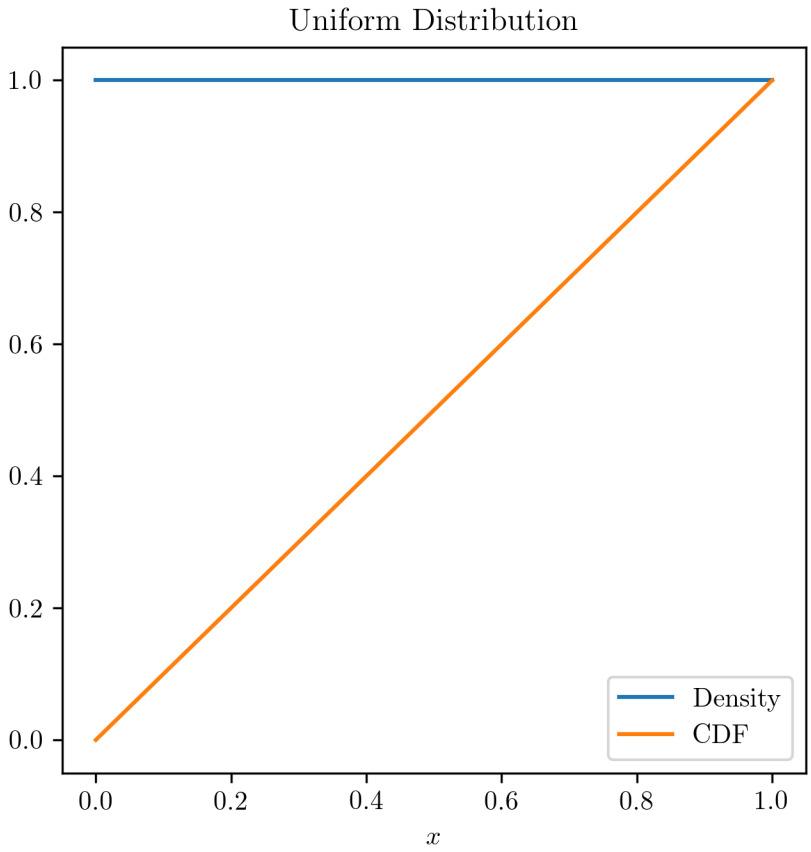


Figure 6.1: The density and CDF of the Uniform$[0, 1]$ distribution are plotted.

that X and Y have the same density function (and hence the same distribution). For now, accept the intuition! (Alternatively, carry out the integral $\mathbb{E}[X] = \int_0^1 x \, \mathrm{d}x$, but that's boring.)

We compute variance in the standard way:

$$\mathbb{E}[X^2] = \int_0^1 x^2 \, \mathrm{d}x = \frac{1}{3} x^3 \Big|_0^1 = \frac{1}{3}.$$

Then, use $\operatorname{var} X = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 1/3 - 1/4$ to show that

$$\operatorname{var} X = \frac{1}{12}. \tag{6.18}$$

Exercise 25 The Uniform$[a, b]$ distribution, where $a < b$, has density $f(x) = 1/(b - a)$ for $x \in [a, b]$, and $f(x) = 0$ for $x \notin [a, b]$.

1. What is the CDF of the Uniform$[a, b]$ distribution? What is the tail probability $\mathbb{P}(\text{Uniform}[a, b] > x)$?

2. If $X \sim \text{Uniform}[a, b]$, prove that X has the same distribution as $a + (b - a)U$, where $U \sim \text{Uniform}[0, 1]$ random variable. In other words, the Uniform$[a, b]$ distribution can be obtained from the Uniform$[0, 1]$ distribution by scaling and shifting.

3. Using the previous part, compute $\mathbb{E}[X]$ and $\operatorname{var} X$.

6.3.2 Exponential Distribution

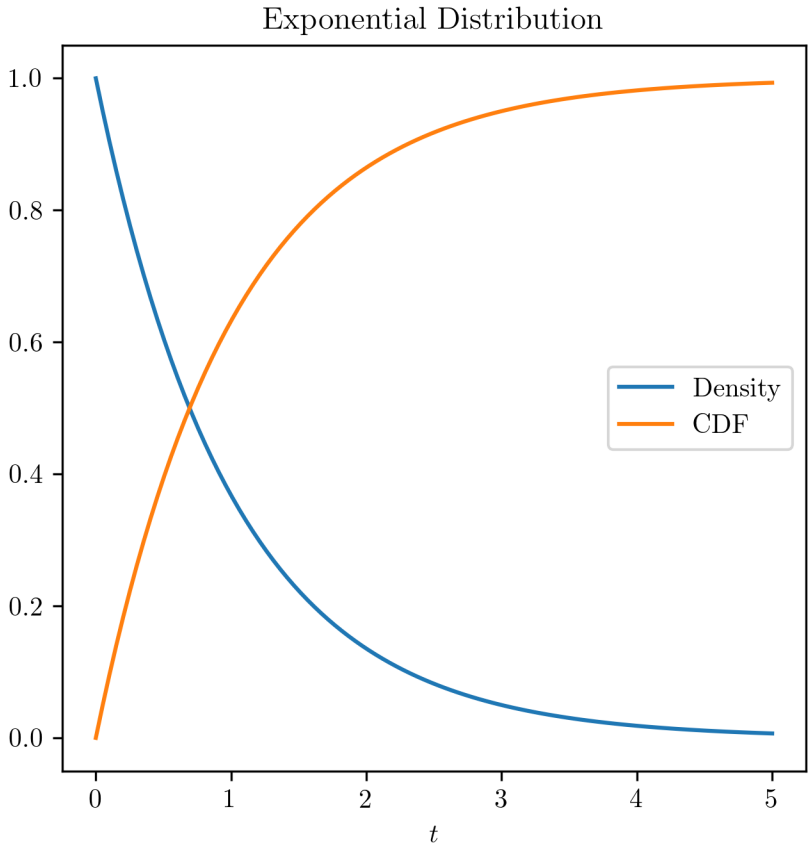


Figure 6.2: The density and CDF of the Exponential(1) distribution are plotted.

The **exponential distribution with parameter** λ, where $\lambda > 0$, is given by the density function

$$f_T(t) := \lambda \exp(-\lambda t), \qquad t > 0.$$

The CDF of the exponential distribution is found by integrating the density (see Figure 6.2):

$$F_T(t) = \begin{cases} 0, & t < 0 \\ 1 - \exp(-\lambda t), & t \geq 0 \end{cases}$$

We will proceed to compute the basic properties of the exponential distribution. First, check for yourself that the exponential distribution is properly normalized:

$$\int_0^\infty \lambda \exp(-\lambda t)\,\mathrm{d}t = 1 \tag{6.21}$$

We claim that the normalization condition itself is already most of the work necessary to find the expectation of T. Normally, we would need to calculate

$$\mathbb{E}[T] = \int_0^\infty t\lambda \exp(-\lambda t)\,\mathrm{d}t$$

which can be solved using integration by parts. Instead, here is a trick. We observe that

$$\frac{\partial}{\partial \lambda}\exp(-\lambda t) = -t\exp(-\lambda t),$$

so[1]

$$\begin{aligned}\mathbb{E}[T] &= \lambda\int_0^\infty t\exp(-\lambda t)\,\mathrm{d}t = \lambda\int_0^\infty -\frac{\partial}{\partial\lambda}\exp(-\lambda t)\,\mathrm{d}t = -\lambda\frac{\mathrm{d}}{\mathrm{d}\lambda}\int_0^\infty \exp(-\lambda t)\,\mathrm{d}t \\ &= -\lambda\frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{1}{\lambda} = -\lambda\cdot -\frac{1}{\lambda^2}.\end{aligned}$$

Therefore,

$$\mathbb{E}[T] = \frac{1}{\lambda}. \tag{6.22}$$

Similarly, the variance is computed by finding

$$\begin{aligned}\mathbb{E}[T^2] &= \lambda\int_0^\infty t^2\exp(-\lambda t)\,\mathrm{d}t = \lambda\int_0^\infty \frac{\partial^2}{\partial\lambda^2}\exp(-\lambda t)\,\mathrm{d}t = \lambda\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\int_0^\infty \exp(-\lambda t)\,\mathrm{d}t \\ &= \lambda\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\frac{1}{\lambda} = \lambda\cdot\frac{2}{\lambda^3} = \frac{2}{\lambda^2}.\end{aligned}$$

Hence, the variance is

$$\operatorname{var} T = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}. \tag{6.23}$$

Exercise 26 Calculate the **median** of the exponential distribution, that is, find t such that $\mathbb{P}(T \le t) = 1/2 = \mathbb{P}(T \ge t)$. In the subject of chemistry, this is also known as the **half-life** of a radioactivate substance.

Exercise 27 Let $T \sim \text{Exponential}(\lambda)$. Compute $\mathbb{E}[T]$ again using the tail sum formula (6.14).

[1] A short word on notation: Inside the integral, we are thinking of $\exp(-\lambda t)$ as a function of both λ and t, and hence we use the partial derivative $\frac{\partial}{\partial\lambda}$ notation. Outside the integral, $\int_0^\infty \exp(-\lambda t)\,\mathrm{d}t$ is a function of λ only, and hence we use the regular derivative $\frac{\mathrm{d}}{\mathrm{d}\lambda}$ notation.

6.3.3 Memoryless Property

Recall that the geometric distribution satisfied the memoryless property: if $X \sim \text{Geometric}(p)$ and $m, n \in \mathbb{N}$, then $\mathbb{P}(X > m + n \mid X > m) = \mathbb{P}(X > n)$. The memoryless property characterizes the geometric distribution, and the exponential distribution is the analogue of the geometric distribution in continuous time. We will in fact prove that the memoryless property *uniquely* characterizes the exponential distribution.

Theorem 6.2 (Memoryless Property of the Exponential Distribution)**.** *Let $T > 0$ be a random variable with a continuous CDF. Then, T satisfies the* ***memoryless property****, that is, for all $s, t > 0$,*

$$\mathbb{P}(T > s + t \mid T > s) = \mathbb{P}(T > t),$$

if and only if $T \sim \text{Exponential}(\lambda)$ for some $\lambda > 0$.

Proof. ($\Longleftarrow$) Suppose $T \sim \text{Exponential}(\lambda)$ for $\lambda > 0$. Then, for $s, t > 0$,

$$\mathbb{P}(T > s + t \mid T > s) = \frac{\mathbb{P}(T > s + t)}{\mathbb{P}(T > s)} = \frac{\exp(-\lambda(s + t))}{\exp(-\lambda s)} = \exp(-\lambda t) = \mathbb{P}(T > t).$$

($\Longrightarrow$) Suppose T satisfies the memoryless property. The memoryless property implies that for $s, t > 0$, $\mathbb{P}(T > s + t) = \mathbb{P}(T > s)\mathbb{P}(T > t)$. By repeatedly applying this equation, we obtain $\mathbb{P}(T > n) = \mathbb{P}(T > 1)^n$ for $n \in \mathbb{N}$. We can extend this equation to hold for all positive rational numbers: suppose $a/b \in \mathbb{Q}$, $a, b > 0$. Then,

$$\mathbb{P}\Big(T > \frac{a}{b}\Big)^b = \mathbb{P}(T > a) = \mathbb{P}(T > 1)^a,$$

which implies $\mathbb{P}(T > a/b) = \mathbb{P}(T > 1)^{a/b}$ for all positive rationals a/b. Therefore, once we have specified a value for $\mathbb{P}(T > 1)$, we have specified $\mathbb{P}(T > q)$ (and thus the CDF) for all $q \in \mathbb{Q}$. By continuity of the CDF, this uniquely determines the CDF everywhere. [a] Now, we cannot have $\mathbb{P}(T > 1) = 0$, or else we would have $\mathbb{P}(T > t) = 0$ for all $t > 0$, which is impossible (by the assumption that $T > 0$). So, let $\lambda > 0$ be defined by $\exp(-\lambda) = \mathbb{P}(T > 1) > 0$. From the memoryless property, for $t > 0$ we have $\mathbb{P}(T > t) = \mathbb{P}(T > 1)^t = \exp(-\lambda t)$, which is an Exponential(λ) distribution. □

[a] The principle that is being used here is that a continuous function is uniquely determined by its values at the rational numbers. We do not have the tools to discuss why this is true, but intuitively, there are too many rational numbers near each real number, so continuity forces the function value at $x \in \mathbb{R}$ to be close to the function values at the rational numbers near x.

6.3.4 The Minimum & Maximum of Exponentials

Theorem 6.3. *Let n be a positive integer and $T_1, \ldots, T_n$ be independent exponential random variables with parameters $\lambda_1, \ldots, \lambda_n > 0$ respectively. Then the minimum of the*

random variables is also exponentially distributed:

$$\min\{T_1, \ldots, T_n\} \sim \text{Exponential}(\lambda_1 + \cdots + \lambda_n) \tag{6.24}$$

Proof. The easiest way to prove this is to once again consider the tail probabilities.

$$\begin{aligned}\mathbb{P}(\min\{T_1, \ldots, T_n\} > t) &= \mathbb{P}(T_1 > t, \ldots, T_n > t) = \mathbb{P}(T_1 > t) \cdots \mathbb{P}(T_n > t) \\ &= \exp(-\lambda_1 t) \cdots \exp(-\lambda_n t) = \exp\big(-(\lambda_1 + \cdots + \lambda_n)t\big).\end{aligned}$$

We have the survival function of an exponential distribution with parameter $\lambda_1 + \cdots + \lambda_n$. □

Example 6.4. Let n be a positive integer and we have n i.i.d. Exponential(1) random variables. What is the expectation of the maximum?

View the exponential random variables as representing the lifetimes of n light bulbs. The expectation of the maximum is the expected time for all n light bulbs to die. This is simply the expected time until the first light bulb dies, plus the expected time it takes for the remaining $n-1$ light bulbs to die. We can compute each of these quantities separately.

The first light bulb to die is the minimum of n Exponential(1) random variables. Since the minimum of independent exponential random variables is exponentially distributed (Theorem 6.3), the minimum life of the light bulbs is Exponential(n), with mean $1/n$. Let S_n be the time for all n light bulbs to die. Once the first light bulb dies, we wait for $n-1$ light bulbs to die. By the memoryless property (Theorem 6.2), however, this is the same as if we had started with $n-1$ i.i.d. Exponential(1) random variables and asked for the maximum. In other words, $\mathbb{E}[S_n] = 1/n + \mathbb{E}[S_{n-1}]$. By solving this recurrence, we obtain

$$\mathbb{E}[S_n] = \sum_{k=1}^{n} \frac{1}{k} = H_n \approx \ln n + \gamma.$$

Exercise 28 Imagine a store with two service lines. Two customers are already waiting in each of the two lines; call these customers A and B respectively. A third customer, C, arrives, and waits for the first available teller. Assume that the two lines have service times which are exponentially distributed, with parameter λ. What is the probability that customer C is the last to leave the store?

6.4 Solutions to Exercises

Exercise 23 Since $\int_0^1 x(1-x)\,\mathrm{d}x = 1/2 - 1/3 = 1/6$ and the density must integrate to 1, then $c = 6$. For $x \in [0,1]$, $F(x) = \int_0^x 6s(1-s)\,\mathrm{d}s = 6(x^2/2 - x^3/3)$, so the full expression for the CDF is:

$$F(x) = \begin{cases} 0, & x < 0 \\ 3x^2 - 2x^3, & x \in [0,1] \\ 1, & x > 1 \end{cases}$$

Exercise 24 We use the continuous tail sum formula (6.14). Notice that if $x < c$, then $\mathbb{P}(\min(c, X) \geq x) = 1 - F_X(x)$, and if $x \geq c$, then $\mathbb{P}(\min(c, X) \geq x) = 0$.

$$\mathbb{E}[\min(c, X)] = \int_0^\infty \mathbb{P}\big(\min(c, X) \geq x\big)\,\mathrm{d}x = \int_0^c \big(1 - F_X(x)\big)\,\mathrm{d}x.$$

Exercise 25

1. The CDF is found by integrating the density:

$$F(x) = \begin{cases} 0, & x < a \\ (x-a)/(b-a), & x \in [a,b] \\ 1, & x > b \end{cases} \tag{6.19}$$

   The tail probability is:

$$\mathbb{P}(\text{Uniform}[a,b] > x) = \begin{cases} 1, & x < a \\ (b-x)/(b-a), & x \in [a,b] \\ 0, & x > b \end{cases} \tag{6.20}$$

2. Since U takes on values in $[0,1]$, it follows that $a + (b-a)U$ takes on values in $[a,b]$. Let $x \in [a,b]$. Then,

$$\mathbb{P}\big(a + (b-a)U \leq x\big) = \mathbb{P}\big((b-a)U \leq x - a\big) = \mathbb{P}\Big(U \leq \frac{x-a}{b-a}\Big) = \frac{x-a}{b-a}$$

   (for the last equality, we used the fact that $(x-a)/(b-a) \in [0,1]$ and U has the CDF given by (6.16)). Therefore, we see that $a + (b-a)U$ has the same CDF as X.

3. Using the previous part, $\mathbb{E}[X] = \mathbb{E}[a + (b-a)U] = a + (b-a)/2 = (a+b)/2$ and $\operatorname{var} X = \operatorname{var}(a + (b-a)U) = (b-a)^2/12$.

Exercise 26 $1/2 = \exp(-\lambda t)$, so $-\ln 2 = -\lambda t$, so the median is $(\ln 2)/\lambda$. Compared to the mean, there is an additional $\ln 2$ factor.

Exercise 27 Since $\mathbb{P}(T \geq t) = \exp(-\lambda t)$ for $t > 0$, $\mathbb{E}[X] = \int_0^\infty \exp(-\lambda t)\,\mathrm{d}t = \lambda^{-1}$.

Exercise 28 Consider the time at which the first customer leaves the store. The other customer who is still waiting in line, by the memoryless property, has a service time distributed as a brand-new Exponential(λ) distribution. Customer C enters the service line which is now free, so the service time for customer C is also Exponential(λ). By symmetry, the probability that customer C will leave first is $1/2$.

Chapter 7

Continuous Probability II

We continue our study of continuous random variables with more advanced tools: conditional probability, change of variables, convolution. We then introduce more distributions, including the all-important normal distribution (with the associated central limit theorem).

7.1 Conditional Probability

7.1.1 Law of Total Probability

Suppose we have an event A and we would like to calculate $\mathbb{P}(A)$ by using the law of total probability. Specifically, we would like to condition on $X = x$. In the discrete case, we would compute $\mathbb{P}(A) = \sum_x \mathbb{P}(A \mid X = x)\mathbb{P}(X = x)$, but we cannot take a summation over uncountably many values. The solution is to instead integrate over the density:

$$\boxed{\mathbb{P}(A) = \int_{-\infty}^{\infty} \mathbb{P}(A \mid X = x) f_X(x)\,\mathrm{d}x} \tag{7.1}$$

Exercise 29 Suppose $X \sim \text{Uniform}[0, 1]$ and $Y \sim \text{Uniform}[0, X]$. That is, conditioned on $X = x$, Y has a Uniform$[0, x]$ distribution. What is $\mathbb{P}(Y > 1/2)$?

7.1.2 Conditional Density

The **conditional density** of Y given $X = x$ is:

$$\boxed{f_{Y|X}(y \mid x) := \frac{f_{X,Y}(x, y)}{f_X(x)}} \tag{7.2}$$

Similarly, the **conditional CDF** of Y given $X = x$ is:

$$\boxed{F_{Y|X}(y \mid x) := \mathbb{P}(Y \le y \mid X = x) = \int_{-\infty}^{y} f_{Y|X}(y' \mid x)\,\mathrm{d}y'} \tag{7.3}$$

We can compute the probability of an event $[a, b]$, where $a < b$, by

$$\mathbb{P}(Y \in [a, b] \mid X = x) = F_{Y|X}(b \mid x) - F_{Y|X}(a \mid x) = \int_a^b f_{Y|X}(y \mid x)\,\mathrm{d}y. \tag{7.4}$$

7.2 Functions of Random Variables

7.2.1 Change of Variables

Often, we wish to find the density of a function of a random variable, such as the density of X^2 (assuming that we already know f_X). The problem can be solved in a satisfying and general way.

Theorem 7.1 (Change of Variables)**.** *Let X be a random variable and let $g : \mathbb{R} \to \mathbb{R}$ be differentiable and one-to-one. Let h be the inverse of g (h exists since g is one-to-one). Then:*

$$\boxed{f_Y(y) = f_X\big(h(y)\big)|h'(y)|} \tag{7.5}$$

Proof. The method I prefer to use is to first manipulate the CDF and then differentiate. First, we consider the case when g is strictly increasing.

$$F_Y(y) = \mathbb{P}(Y < y) = \mathbb{P}\big(g(X) < y\big) = \mathbb{P}\big(X < h(y)\big) = F_X\big(h(y)\big)$$

Differentiate both sides:

$$f_Y(y) = f_X\big(h(y)\big)h'(y)$$

Since h is strictly increasing, then $h'(y) > 0$ and the change of variables equation holds. Now consider the case in which g is strictly decreasing. Now that h is strictly decreasing, applying h to both sides of the inequality also flips the direction of the inequality. Hence

$$F_Y(y) = \mathbb{P}\big(X > h(y)\big) = 1 - \mathbb{P}\big(X < h(y)\big) = 1 - F_X\big(h(y)\big).$$

Differentiate both sides:

$$f_Y(y) = -f_X\big(h(y)\big)h'(y)$$

Here, since h is strictly decreasing, then $h'(y) < 0$ and therefore $|h'(y)| = -h'(y)$. Hence, the change of variables formula still holds. □

My advice is to not bother remembering the change of variables formula. Instead, remember the basic outline of the proof: write down the CDF of Y, then write the expression in terms of the CDF of X, and then differentiate.

Why did we assume that the function g was one-to-one? Only out of convenience: the condition that g is one-to-one is not necessary for change of variables to work, although the

change of variables formula is somewhat more complicated:

$$\boxed{f_Y(y) = \sum_{x:g(x)=y} f_X(x)|h'(y)|} \tag{7.6}$$

The idea is that since g is no longer one-to-one, there may be many values of x such that $g(x) = y$, so we must sum up over all x such that $g(x) = y$. Can we define change of variables in the discrete case? Actually, the discrete case is easier.

$$\boxed{\mathbb{P}(Y = y) = \sum_{x:g(x)=y} \mathbb{P}(X = x)} \tag{7.7}$$

If you think about the above equation, you will realize that we have been using the formula all along without knowing it.

We will use change of variables in the section about the normal distribution.

Exercise 30 If X has density f, what is the density of X^2?

7.2.2 Convolution

Here, we explain how to compute the density of $Z := X + Y$.

$$\mathbb{P}\big(Z \in (z, z + \mathrm{d}z)\big) = f_Z(z)\,\mathrm{d}z$$

Note that $X + Y = Z \in (z, z + \mathrm{d}z)$ is equivalent to the event that $Y \in (z - x, z - x + \mathrm{d}z)$, where x ranges over all values. Hence, we can find $\mathbb{P}(Z \in (z, z + \mathrm{d}z))$ by integrating over the joint density of X and Y.

$$f_Z(z)\,\mathrm{d}z = \int_{-\infty}^{\infty} \int_{z-x}^{z-x+\mathrm{d}z} f_{X,Y}(x, y)\,\mathrm{d}y\,\mathrm{d}x$$

Since $\mathrm{d}z$ is an infinitesimal length, we can assume that $f_{X,Y}$ does not change over the interval of integration in the inner integral. The inner integral therefore equals the value of the function, $f_{X,Y}(x, z - x)$, multiplied by the length of the interval, $\mathrm{d}z$.

$$f_Z(z)\,\mathrm{d}z = \int_{-\infty}^{\infty} f_{X,Y}(x, z - x)\,\mathrm{d}z\,\mathrm{d}x$$

Therefore, we can identify:

$$\boxed{f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z - x)\,\mathrm{d}x} \tag{7.8}$$

When X and Y are independent, the formula simplifies to:

$$\boxed{f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\,\mathrm{d}x} \tag{7.9}$$

This is known as the **convolution formula**.

7.2.3 Ratios of Random Variables

Next, we will compute the density of $Z := X/Y$. $X/Y = Z \in (z, z + \mathrm{d}z)$ is equivalent to $X \in (yz, yz + y\,\mathrm{d}z)$ when $y > 0$, and $X \in (yz + y\,\mathrm{d}z, yz)$ when $y < 0$ (since $y < 0$ implies that $yz + y\,\mathrm{d}z < yz$). Hence, we split up the integral over the joint density into two pieces:

$$f_Z(z)\,\mathrm{d}z = \int_{-\infty}^{0}\int_{yz+y\,\mathrm{d}z}^{yz} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y + \int_{0}^{\infty}\int_{yz}^{yz+y\,\mathrm{d}z} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

We assume that $\mathrm{d}z$ is small so that $f_{X,Y}$ is effectively constant over the inner integral.

$$f_Z(z)\,\mathrm{d}z = \int_{-\infty}^{0} f_{X,Y}(yz,y)(-y\,\mathrm{d}z)\,\mathrm{d}y + \int_{0}^{\infty} f_{X,Y}(yz,y)(y\,\mathrm{d}z)\,\mathrm{d}y = \int_{-\infty}^{\infty} |y| f_{X,Y}(yz,y)\,\mathrm{d}y\,\mathrm{d}z$$

Therefore, we obtain:

$$\boxed{f_Z(z) = \int_{-\infty}^{\infty} |y| f_{X,Y}(yz,y)\,\mathrm{d}y} \tag{7.10}$$

In the special case when X and Y are independent, we have:

$$\boxed{f_Z(z) = \int_{-\infty}^{\infty} |y| f_X(yz) f_Y(y)\,\mathrm{d}y} \tag{7.11}$$

Exercise 31 Let X and Y be i.i.d. Exponential(λ). What is the density of $Z := X/Y$?

7.3 Normal Distribution

We turn our attention to one of the most important probability distributions: the **standard normal distribution** (also called a **Gaussian**), which we denote $\mathcal{N}(0,1)$. (The first parameter is the mean, and the second parameter is the variance). The density $f_X(x)$ will be proportional to $\exp(-x^2/2)$ (defined over all of $\mathbb{R}$), which has no known elementary antiderivative. Therefore, we devote the next section to integrating this function.

7.3.1 Integrating the Normal Distribution

Let us find the normalization constant, that is, we seek $c \in \mathbb{R}$ such that

$$c\int_{\mathbb{R}} \exp\Big(-\frac{x^2}{2}\Big)\,\mathrm{d}x = 1.$$

In fact, we will solve a slightly more general integral by replacing the integrand with $\exp(-\alpha x^2/2)$ instead. (We can always set $\alpha = 1$ at the end of our computations, after all.) The reason for doing this is so we can differentiate under the integral.

We will need to use another trick (hopefully you find all of these tricks somewhat interesting). The trick here is to consider the square of the integral:

$$I(\alpha)^2 := \int_{\mathbb{R}} \exp\Big(-\frac{\alpha x^2}{2}\Big)\,\mathrm{d}x \int_{\mathbb{R}} \exp\Big(-\frac{\alpha y^2}{2}\Big)\,\mathrm{d}y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\Big(-\frac{\alpha(x^2+y^2)}{2}\Big)\,\mathrm{d}x\,\mathrm{d}y$$

Notice that the integrand depends only on the quantity x^2+y^2, which you may recognize as the square of the distance from the origin. (The variables x and y were chosen suggestively to bring to mind the picture of integration on the plane $\mathbb{R}^2$.) Therefore, it is natural to change to polar coordinates with the substitutions

$$x^2 + y^2 = r^2, \tag{7.12}$$
$$\mathrm{d}x\,\mathrm{d}y = r\,\mathrm{d}r\,\mathrm{d}\theta. \tag{7.13}$$

(Don't forget the extra factor of r that arises due to the Jacobian. For more information, consult a multivariable calculus textbook which develops the theory of integration under change of coordinates.)

In polar coordinates, the integral can now be evaluated.

$$\begin{aligned} I(\alpha)^2 &= \int_0^{2\pi}\int_0^{\infty} \exp\Big(-\frac{\alpha r^2}{2}\Big) r\,\mathrm{d}r\,\mathrm{d}\theta = \int_0^{2\pi} \mathrm{d}\theta \int_0^{\infty} r\exp\Big(-\frac{\alpha r^2}{2}\Big)\,\mathrm{d}r \\ &= 2\pi\cdot -\frac{1}{\alpha}\exp\Big(-\frac{\alpha r^2}{2}\Big)\Big|_0^{\infty} = \frac{2\pi}{\alpha} \end{aligned}$$

We obtain the surprising result that $I(\alpha) = \sqrt{2\pi/\alpha}$. Set $\alpha = 1$. The normalized density is

$$f_X(x) = \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{x^2}{2}\Big). \tag{7.14}$$

As noted before, there is no elementary antiderivative of the density function, so we cannot write down the CDF in terms of familiar functions. The CDF of the standard normal distribution is often denoted:

$$\boxed{\Phi(z) := \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{x^2}{2}\Big)\,\mathrm{d}x} \tag{7.15}$$

See Figure 7.1.

Exercise 32 Gaussian Tail Bounds Prove the following bounds on the tail probability of the standard Gaussian:

$$(x^{-1} - x^{-3})\exp\Big(-\frac{x^2}{2}\Big) \le \int_x^{\infty} \exp\Big(-\frac{z^2}{2}\Big)\,\mathrm{d}z \le x^{-1}\exp\Big(-\frac{x^2}{2}\Big) \tag{7.16}$$

Hint: For the left inequality, integrate $\int_x^{\infty}(1-3z^{-4})\exp(-z^2/2)\,\mathrm{d}z$. For the right inequality, change variables to $z := x + y$.

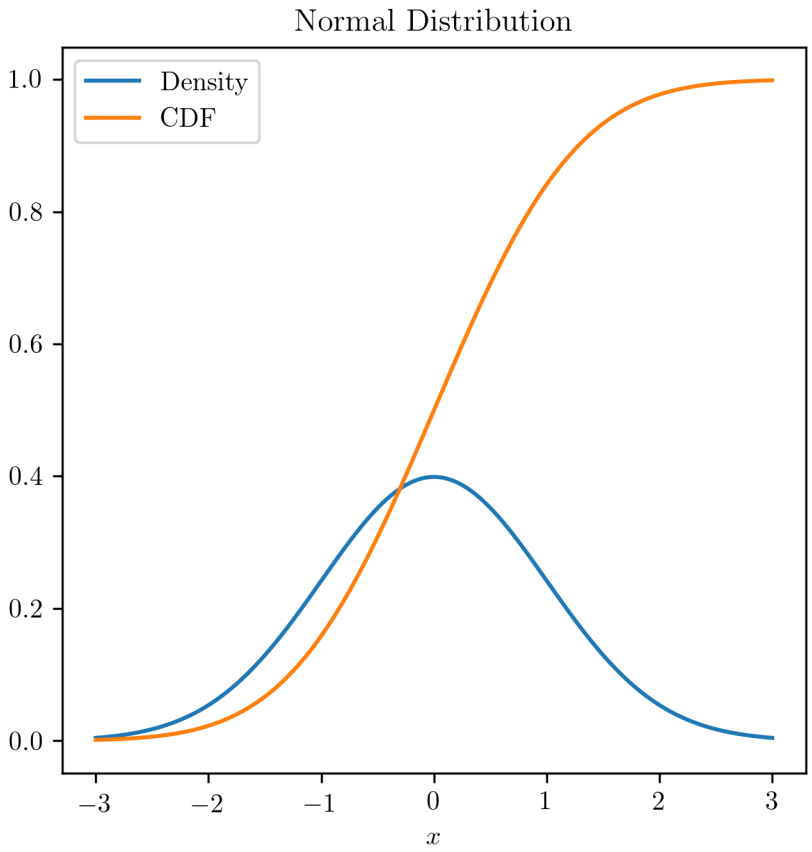


Figure 7.1: The density and CDF of the standard normal distribution are plotted.

7.3.2 Mean & Variance of the Normal Distribution

We hope to find that the mean and variance of the $\mathcal{N}(0,1)$ distribution are 0 and 1, as we claimed. Here, we verify that this is the case. First, notice that the density $f_X(x)$ depends only on x^2, so interchanging $x \mapsto -x$ leaves the density unchanged. The density is therefore symmetric about $x = 0$, and we write down

$$\mathbb{E}[X] = 0.$$

The variance is slightly trickier.

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x$$

Although the integral smells like integration by parts, recall how we managed to avoid integration by parts in a similar integral when we computed the mean and variance of the exponential distribution. We would like to apply a similar trick here, so let us instead consider the integral $\int_{-\infty}^{\infty} x^2 \exp(-\alpha x^2/2)\,\mathrm{d}x$ (with the intention of setting $\alpha = 1$ at the end

of our computations).

$$\begin{aligned}\mathbb{E}[X^2] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 \exp\Big(-\frac{\alpha x^2}{2}\Big)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(-2)\frac{\partial}{\partial\alpha}\exp\Big(-\frac{\alpha x^2}{2}\Big)\,\mathrm{d}x \\ &= \frac{-2}{\sqrt{2\pi}}\frac{\mathrm{d}}{\mathrm{d}\alpha}\int_{-\infty}^{\infty}\exp\Big(-\frac{\alpha x^2}{2}\Big)\,\mathrm{d}x = \frac{-2}{\sqrt{2\pi}}\frac{\mathrm{d}}{\mathrm{d}\alpha}\sqrt{\frac{2\pi}{\alpha}} = -2\cdot -\frac{1}{2}\alpha^{-3/2} = \alpha^{-3/2}\end{aligned}$$

Again, set $\alpha = 1$, and since $\mathbb{E}[X]^2 = 0$,

$$\operatorname{var} X = 1.$$

Hopefully, it should be clear by now why we need the α: it is simply a tool that we use to integrate more easily, and then we discard it after we finish the actual integration. In any case, we have verified that the mean and variance are indeed 0 and 1 respectively.

We now apply the change of variables technique to the standard normal distribution, both as an illustration of the technique, and also to obtain the general form of the normal distribution. Consider the function $g(x) = \mu + \sigma x$ (with $\mu \in \mathbb{R}$ and $\sigma > 0$). Let $Y = g(X)$. We proceed to find the density of Y. First, note that the inverse function $h = g^{-1}$ is

$$h(y) = \frac{y-\mu}{\sigma}$$

and

$$|h'(y)| = \frac{1}{\sigma}.$$

Then we have that

$$f_Y(y) = f_X\big(h(y)\big)|h'(y)| = \frac{1}{\sigma}f_X\Big(\frac{y-\mu}{\sigma}\Big).$$

Plugging in $(y-\mu)/\sigma$ into the standard normal density yields

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{(y-\mu)^2}{2\sigma^2}\Big). \tag{7.17}$$

What are the mean and variance of Y? Recall that $Y = \mu + \sigma X$. Using the basic properties of linearity and scaling,

$$\begin{aligned}\mathbb{E}[Y] &= \mu, \\ \operatorname{var} Y &= \sigma^2.\end{aligned}$$

Y a normal random variable with mean μ and variance σ^2, which is denoted $Y \sim \mathcal{N}(\mu, \sigma^2)$. We have seen that any normal distribution is found from the standard normal distribution by the following two-step procedure: scale by σ and shift by μ.

Exercise 33 Prove the identity $(2\pi)^{-1/2}\int_{-\infty}^{\infty}\exp(-tx^2/2)\,\mathrm{d}x = t^{-1/2}$ for $t > 0$ and use it to compute the moments $\mathbb{E}[X^k]$ for $k \in \mathbb{Z}_+$, where $X \sim \mathcal{N}(0,1)$. *Hint*: Differentiate the identity k times with respect to t.

7.3.3 Sums of Independent Normal Random Variables

In this section, we prove the crucial fact that the sum of independent normal random variables is also normally distributed. The theorem is breathtaking, but let us precede the theorem with some discussion about the joint density of two normal random variables.

Let X, Y be i.i.d. standard normal random variables. Since they are independent, their joint density is the product of the individual densities.

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{x^2}{2}\Big)\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{y^2}{2}\Big) = \frac{1}{2\pi}\exp\Big(-\frac{x^2+y^2}{2}\Big).$$

Observe that the joint density only depends on $x^2 + y^2 = r^2$, which is the square of the distance from the origin. (In fact, we already noticed this property when we were integrating the Gaussian.) In polar coordinates, the density only has a dependence on r, not on θ, which exhibits an important geometric property: the joint density is *rotationally symmetric*, that is, the Gaussian looks exactly the same if you rotate your coordinate axes. How can we utilize this geometric property to prove our result?

Let $Z = X + Y$. Consider $F_Z(z) = \mathbb{P}(Z < z)$. We can write

$$F_Z(z) = \mathbb{P}(Z < z) = \mathbb{P}(X + Y < z).$$

To compute this probability, we integrate the joint density

$$F_Z(z) = \int_J f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

where J is the region

$$J := \{(x,y) : x + y < z\} \subseteq \mathbb{R}^2.$$

We know that $x + y = z$ is a line in $\mathbb{R}^2$. Perhaps we can align our coordinate axes with the line $x + y = z$, and the integral will be simplified. Consider the coordinate system

$$\begin{aligned} x &= \frac{1}{\sqrt{2}}x' - \frac{1}{\sqrt{2}}y', \\ y &= \frac{1}{\sqrt{2}}x' + \frac{1}{\sqrt{2}}y'. \end{aligned}$$

Under these coordinates, the region J is

$$J = \{(x',y') : \sqrt{2}x' < z\} \subseteq \mathbb{R}^2.$$

Since the joint density is invariant under rotations, changing variables from $x \mapsto x'$ and $y \mapsto y'$ should *not* affect the value of the integral. Hence,

$$F_Z(z) = \mathbb{P}(Z < z) = \int_J f_{X,Y}(x',y')\,\mathrm{d}x'\,\mathrm{d}y' = \int_{-\infty}^{z/\sqrt{2}} f_X(x')\,\mathrm{d}x' \int_{-\infty}^{\infty} f_Y(y')\,\mathrm{d}y'$$

$$= \int_{-\infty}^{z/\sqrt{2}} f_X(x')\,\mathrm{d}x' = \mathbb{P}\Big(X < \frac{z}{\sqrt{2}}\Big) = \mathbb{P}(\sqrt{2}X < z).$$

We see that Z has the same distribution as $\sqrt{2}X$, where $X \sim \mathscr{N}(0,1)$. From our previous section, we saw that scaling a standard normal random variable by $\sqrt{2}$ yields the $\mathscr{N}(0,2)$ distribution. However, $X \sim \mathscr{N}(0,1)$ and $Y \sim \mathscr{N}(0,1)$, and we have just found that $Z = X+Y \sim \mathscr{N}(0,2)$. Could summing up independent normal random variables really be as easy as adding their parameters?

Lemma 7.2 (Rotational Invariance of the Gaussian)**.** *Let X, Y be i.i.d. standard normal random variables and let $\alpha, \beta \in [0,1]$ so that $\alpha^2 + \beta^2 = 1$. Then $\alpha X + \beta Y \sim \mathscr{N}(0,1)$.*

Proof. We simply extend the previous argument to any arbitrary rotation. By assumption, we can write down

$$\sin\theta = \alpha$$
$$\cos\theta = \beta$$

for some $\theta \in [0, \pi/2]$. We can write the CDF of $Z = \alpha X + \beta Y$ as

$$F_Z(z) = \mathbb{P}(Z < z) = \mathbb{P}(\alpha X + \beta Y < z) = \int_J f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

where

$$J = \{(x,y) : x\cos\theta + y\sin\theta < z\} \subseteq \mathbb{R}^2.$$

Under the change of coordinates

$$x = x'\cos\theta - y'\sin\theta$$
$$y = x'\sin\theta + y'\cos\theta$$

the area of integration becomes

$$J = \{(x',y') : x' < z\} \subseteq \mathbb{R}^2.$$

Hence we conclude that

$$F_Z(z) = \mathbb{P}(Z < z) = \int_J f_{X,Y}(x',y')\,\mathrm{d}x'\,\mathrm{d}y' = \int_{-\infty}^{z} f_X(x')\,\mathrm{d}x' \int_{-\infty}^{\infty} f_Y(y')\,\mathrm{d}y'$$
$$= \int_{-\infty}^{z} f_X(x')\,\mathrm{d}x' = \mathbb{P}(X < z).$$

Z has the same distribution as X, and $X \sim \mathscr{N}(0,1)$, so we're done. □

Theorem 7.3 (Sums of Independent Gaussians)**.** *Let n be any positive integer and let $X_1, \ldots, X_n$ be independent random variables with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$. Then:*

$$\boxed{X := X_1 + \cdots + X_n \sim \mathcal{N}\Big(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\Big)} \tag{7.18}$$

Proof. We will prove the result for the sum of two independent normal random variables. The general result follows as a quick exercise in induction. Let $Z_i = (X_i - \mu_i)/\sigma_i$ be the standardized form of X_i. Then $Z_i \sim \mathcal{N}(0, 1)$. Additionally, observe that

$$Z = \sqrt{\frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}} \frac{X_1 - \mu_1}{\sigma_1} + \sqrt{\frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}} \frac{X_2 - \mu_2}{\sigma_2} = \frac{X_1 + X_2 - (\mu_1 + \mu_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}.$$

Apply Lemma 7.2 to Z to obtain that $Z \sim \mathcal{N}(0, 1)$. Finally, since

$$X = X_1 + X_2 = \mu_1 + \mu_2 + \sqrt{\sigma_1^2 + \sigma_2^2} Z,$$

it follows from change of variables that

$$Z \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2). \qquad \square$$

The theorem is beautiful, so do not misuse it! The most common mistake that students make is that they forget the important rule: *variances add, standard deviations do not.*

7.4 Central Limit Theorem

Why do we care so much about the normal distribution? It certainly has nice properties, but the exponential distribution is also easy to work with and models many real-life situations very well (e.g. particle decay). The normal distribution, however, goes even farther: I claim that it allows us to model *any* situation whatsoever. In what sense can such a bold statement possibly be true?

First, let us make the statement precise. Suppose $\{X_i\}_{i \in \mathbb{N}}$ is a sequence of i.i.d. random variables with mean μ and finite variance. Define

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then as $n \to \infty$, $\sqrt{n}(\bar{X}_n - \mu)$ *converges in distribution* to the normal distribution.

Remember that the WLLN tells us that as we increase the number of samples, the sample mean *converges in probability* to the expected value. The CLT states that the distribution of the sample mean also converges to a particular distribution, and it is our good friend the

normal distribution! The power of the theorem is that we can start from any distribution at all, discrete or continuous, yet the sample mean will still converge to a single distribution. In fact, there are even versions of the CLT with weaker assumptions. Of all of the theorems in mathematics, the CLT is one of my favorites.

The CLT is the basis for much of modern statistics. When statisticians carry out hypothesis testing or construct confidence intervals, they do not care that they do not know the exact distribution of their data. They simply collect enough samples (30 is usually sufficient) until their sampling distributions are roughly normally distributed.

We have not answered questions such as "what does it mean for a distribution to converge?" and "why is the CLT true?" The former question requires a more rigorous study of distributions. The latter question will only be answered partially below. We sketch the proof here.

Theorem 7.4 (Central Limit Theorem)**.** *Let $\{X_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$. Then, for all $z \in \mathbb{R}$,*

$$\mathbb{P}\Big(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le z\Big) \xrightarrow[n\to\infty]{} \Phi(z).$$

Proof Sketch. If we can prove that the CLT holds when X_1 has mean zero and unit variance, then for general X_1, we could apply the CLT to the random variables $Z_i := (X_i - \mu)/\sigma$ for $i \in \mathbb{N}$ (which have mean zero and unit variance) to conclude that

$$\mathbb{P}\Big(\frac{Z_1 + \cdots + Z_n}{\sqrt{n}} \le z\Big) = \mathbb{P}\Big(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le z\Big) \xrightarrow[n\to\infty]{} \Phi(z)$$

as desired. So, we will now assume that X_1 has mean zero and unit variance.

The idea is to consider the **characteristic function** $\phi_X(t) := \mathbb{E}[\exp(\mathsf{i}tX)]$. [a] The reason why the characteristic function is so useful for studying sums of random variables is because it converts the sum of random variables (where the density is given by convolution) into multiplication of characteristic functions:

$$\begin{aligned}\phi_{X_1+\cdots+X_n}(t) &= \mathbb{E}\big[\exp\big(\mathsf{i}t(X_1 + \cdots + X_n)\big)\big] = \mathbb{E}[\exp(\mathsf{i}tX_1)\cdots\exp(\mathsf{i}tX_n)] \\ &= \mathbb{E}[\exp(\mathsf{i}tX_1)]\cdots\mathbb{E}[\exp(\mathsf{i}tX_n)] = \phi_{X_1}(t)\cdots\phi_{X_n}(t) = \phi_X(t)^n.\end{aligned}$$

where we used the i.i.d. assumption. For a standard Gaussian random variable Z, one can calculate $\phi_Z(t) = \exp(-t^2/2)$. It turns out that there is an *inversion formula* which allows us to recover the density function from the characteristic function, which is to say that proving that $(X_1 + \cdots + X_n)/\sqrt{n}$ converges to the normal distribution is equivalent

to proving that $\phi_{(X_1+\cdots+X_n)/\sqrt{n}}(t) \xrightarrow[n\to\infty]{} \exp(-t^2/2)$. So, note that

$$\phi_{(X_1+\cdots+X_n)/\sqrt{n}}(t) = \mathbb{E}\Big[\exp\Big\{\mathsf{i}\Big(\frac{t}{\sqrt{n}}\Big)(X_1+\cdots+X_n)\Big\}\Big] = \phi_X\Big(\frac{t}{\sqrt{n}}\Big)^n.$$

However, from the definition of ϕ_X,

$$\phi_X(t) = \mathbb{E}[\exp(\mathsf{i}tX)] = \mathbb{E}\Big[1 + \mathsf{i}tX - \frac{t^2X^2}{2} + \cdots\Big] = 1 - \frac{t^2}{2} + \cdots$$

by taking the Taylor expansion and recalling that we are assuming $\mathbb{E}[X] = 0$, $\mathbb{E}[X^2] = 1$. We now truncate the Taylor expansion at the second-order term, and we find that

$$\phi_{(X_1+\cdots+X_n)/\sqrt{n}}(t) = \phi_X\Big(\frac{t}{\sqrt{n}}\Big)^n = \Big(1 - \frac{t^2}{2n} + \cdots\Big)^n \xrightarrow[n\to\infty]{} \exp\Big(-\frac{t^2}{2}\Big). \qquad \square$$

[a] In contexts outside of probability theory, this is usually known as the **Fourier transform**.

The formal proof is not so different; it involves rigorously justifying some of the statements we have asserted (such as the fact that the characteristic function can be inverted to find the density) and using more careful estimates to justify the Taylor expansion.

7.4.1 Confidence Intervals Revisited

Using the CLT, we can obtain much tighter confidence intervals than Chebyshev's inequality.

Example 7.5. Suppose that n is a positive integer and $X_1, \ldots, X_n$ are i.i.d. with mean μ and standard deviation σ. Let $\bar{X}$ be the sample average as before. The CLT tells us that $Z_n := (\bar{X} - \mu)/(\sigma/\sqrt{n}) \to \mathcal{N}(0,1)$ as $n \to \infty$. Now, view $X_1, \ldots, X_n$ as observations, and we will construct a 95% confidence interval for μ. First, we relate the desired probability to Z_n, and then make use of the fact that $Z_n \sim \mathcal{N}(0,1)$ (approximately).

$$\begin{aligned}\mathbb{P}\big(\mu \in (\bar{X} - a, \bar{X} + a)\big) &= \mathbb{P}(|\bar{X} - \mu| \le a) = \mathbb{P}\Big(\Big|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\Big| \le \frac{a}{\sigma/\sqrt{n}}\Big) \\ &= \mathbb{P}\Big(Z_n \le \frac{a}{\sigma/\sqrt{n}}\Big) \approx \int_{-a/(\sigma/\sqrt{n})}^{a/(\sigma/\sqrt{n})} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{z^2}{2}\Big)\,\mathrm{d}z\end{aligned}$$

We set the above probability to 0.95. Although there is no closed-form expression for the probability, we can plug it into a calculator, which tells us that $a/(\sigma/\sqrt{n}) \approx 1.96$. Therefore, $a \approx 1.96\sigma/\sqrt{n}$ and our confidence interval is $(\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n})$.

7.4.2 de Moivre-Laplace Approximation

Recall that the binomial distribution is the sum of i.i.d. Bernoulli trials, so we may apply the CLT. We can approximate the binomial distribution fairly well by the normal distribution, especially when the number of trials is large. Suppose that $X \sim \text{Binomial}(n, p)$. A reasonable

approach to approximating the probability $\mathbb{P}(a \le X \le b)$ would be

$$\begin{aligned}\mathbb{P}(a \le X \le b) &= \mathbb{P}(a - np \le X - np \le b - np) \\ &= \mathbb{P}\Big(\frac{a-np}{\sqrt{np(1-p)}} \le \frac{X-np}{\sqrt{np(1-p)}} \le \frac{b-np}{\sqrt{np(1-p)}}\Big) \\ &\approx \Phi\Big(\frac{b-np}{\sqrt{np(1-p)}}\Big) - \Phi\Big(\frac{a-np}{\sqrt{np(1-p)}}\Big).\end{aligned}$$

Under this approximation, $\mathbb{P}(X = x) = \mathbb{P}(x \le X \le x) = 0$, which is clearly not correct. The problem lies in our attempt to approximate a discrete distribution with a continuous distribution. It turns out that a better approximation is formed by treating the point $\mathbb{P}(X = x)$ as an interval of width 1 in the continuous distribution. In order words, we should treat the point $\{x\}$ as an interval $[x - 1/2, x + 1/2]$ when we apply our approximation. The correction amounts to

$$\mathbb{P}(a \le X \le b) \approx \Phi\Big(\frac{b-np+1/2}{\sqrt{np(1-p)}}\Big) - \Phi\Big(\frac{a-np-1/2}{\sqrt{np(1-p)}}\Big). \tag{7.19}$$

This is commonly known as the **de Moivre-Laplace approximation** to the binomial distribution. The correction factor of $1/2$ becomes negligible when n is large.

7.5 Order Statistics

Let n be a positive integer and $X_1, \ldots, X_n$ be i.i.d. random variables with common density f_X and CDF F_X. Let $X^{(k)}$ be the kth smallest of $X_1, \ldots, X_n$, for $k = 1, \ldots, n$, so that $X^{(1)}$ is the minimum of the points and $X^{(n)}$ is the maximum of the points. These are known as the **order statistics** of the distribution. We can derive the density of $X^{(k)}$: recall that $f_{X^{(k)}}(x)$ has the interpretation $\mathbb{P}(X^{(k)} \in (x, x + \mathrm{d}x)) \approx f_{X^{(k)}}(x) \cdot \mathrm{d}x$. In order for the kth smallest point to lie in the interval $(x, x + \mathrm{d}x)$, we must have the following:

1. $k - 1$ points must lie in the interval $(-\infty, x)$, which has probability $F_X(x)^{k-1}$.
2. One point must lie in the interval $(x, x + \mathrm{d}x)$, which has probability $f_X(x)\,\mathrm{d}x$.
3. $n - k$ points must lie in the interval $(x + \mathrm{d}x, \infty)$, which has probability $(1 - F_X(x))^{n-k}$.

In addition, there are n ways to choose which of the n points lies in the interval $(x, x + \mathrm{d}x)$, and out of the remaining $n - 1$ points, there are $\binom{n-1}{k-1}$ ways to choose which points lie in the interval $(-\infty, x)$. Hence, we have

$$f_{X^{(k)}}(x)\,\mathrm{d}x = \mathbb{P}\big(X^{(k)} \in (x, x + \mathrm{d}x)\big) = n\binom{n-1}{k-1} f_X(x) F_X(x)^{k-1}\big(1 - F_X(x)\big)^{n-k} \cdot \mathrm{d}x,$$

so our result is:

$$\boxed{f_{X^{(k)}}(x) = n\binom{n-1}{k-1} f_X(x) F_X(x)^{k-1}\big(1 - F_X(x)\big)^{n-k}} \tag{7.20}$$

In the special case when $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Uniform}[0,1]$, we have $f_X(x) = 1$ and $F_X(x) = x$ for $0 < x < 1$, so the density of $X^{(k)}$ is

$$f_{X^{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k}, \qquad 0 < x < 1. \tag{7.21}$$

7.6 Beta Distribution

We define the **beta distribution** by specifying the density function. Let $\text{Beta}(r, s)$ denote the beta distribution with parameters $r, s > 0$ (we will see the significance of the parameters soon). If $X \sim \text{Beta}(r, s)$, then the density of X is

$$f_X(x) = \frac{1}{B(r,s)} x^{r-1}(1-x)^{s-1}, \qquad 0 < x < 1, \tag{7.22}$$

where $B(r, s)$ is a normalizing constant. First, we observe that if $X^{(k)}$ is the kth order statistic of n i.i.d. Uniform$[0, 1]$ random variables, then $X^{(k)} \sim \text{Beta}(k, n - k + 1)$. This provides an interpretation for the parameters of the beta distribution: there are k points less than or equal to $X^{(k)}$, and $n - k + 1$ points greater than or equal to $X^{(k)}$. By setting $n = r + s - 1$ and $k = r$ in (7.21), we can obtain an expression for $B(r, s)$ for integer values of r and s:

$$\frac{1}{B(r,s)} = \frac{(r+s-1)!}{(r-1)!(s-1)!} \tag{7.23}$$

Remember that $B(r, s)$ is a normalizing constant for the density (7.22), which implies

$$B(r,s) = \int_0^1 x^{r-1}(1-x)^{s-1}\,\mathrm{d}x = \frac{(r-1)!(s-1)!}{(r+s-1)!}. \tag{7.24}$$

Remembering (7.24) can actually be useful in solving integrals quickly.

We can solve for the moments of the beta distribution:

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^1 \frac{1}{B(r,s)} x^r(1-x)^{s-1}\,\mathrm{d}x = \frac{B(r+1,s)}{B(r,s)} = \frac{(r+s-1)!}{(r-1)!\cancel{(s-1)!}} \frac{r!\cancel{(s-1)!}}{(r+s)!} = \frac{r}{r+s}, \\
\mathbb{E}[X^2] &= \int_0^1 \frac{1}{B(r,s)} x^{r+1}(1-x)^{s-1}\,\mathrm{d}x = \frac{B(r+2,s)}{B(r,s)} = \frac{(r+s-1)!}{(r-1)!\cancel{(s-1)!}} \frac{(r+1)!\cancel{(s-1)!}}{(r+s+1)!} \\
&= \frac{r(r+1)}{(r+s)(r+s+1)},
\end{aligned}$$

so we have

$$\text{var}\, X = \frac{r(r+1)}{(r+s)(r+s+1)} - \frac{r^2}{(r+s)^2} = \frac{rs}{(r+s)^2(r+s+1)}.$$

7.6.1 Flipping Coins

Note that the beta distribution is supported on $(0, 1)$ so there is a natural interpretation of the beta distribution as a distribution over *probability values*. Suppose that we have a biased coin, but we do not know what the bias of the coin is. Assume that we have a *prior* distribution over the bias of the coin: $X \sim \text{Beta}(r, s)$. Since the expectation of the beta distribution is $r/(r+s)$, that means that we believe the bias of the coin is close to $r/(r+s)$, but we do not have enough information to say the true bias with any certainty. We decide to flip the coin more times in order to get a better idea of the bias, and we obtain h heads and t tails. The question is: what is our posterior distribution of X, that is, how should we update our belief about the bias of the coin?

Of course, the answer is Bayes Rule, but with continuous random variables. To be more clear, let $X \sim \text{Beta}(r, s)$ be a random variable which represents our prior belief about the bias of the coin, and let A be the event that we flip h heads and t tails. What is the conditional distribution of X after observing A, $f_{X|A}$?

Bayes Rule tells us that the answer is

$$\mathbb{P}\big(X \in (x, x+\mathrm{d}x) \;\big|\; A\big) = \frac{\mathbb{P}(X \in (x, x+\mathrm{d}x) \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(X \in (x, x+\mathrm{d}x))\mathbb{P}(A \mid X = x)}{\mathbb{P}(A)}.$$

When we use the density interpretation, we have

$$f_{X|A}(x)\,\mathrm{d}x = \frac{f_X(x)\mathbb{P}(A \mid X = x)\,\mathrm{d}x}{\mathbb{P}(A)},$$

which gives the equation

$$f_{X|A}(x) = \frac{f_X(x)\mathbb{P}(A \mid X = x)}{\mathbb{P}(A)}.$$

We know that $f_X(x) = B(r, s)^{-1}x^{r-1}(1-x)^{s-1}$ for $0 < x < 1$. $\mathbb{P}(A \mid X = x)$ is the probability of flipping h heads and t tails if the bias of our coin is x, which is the binomial distribution with $h + t$ trials and probability of success x. Therefore,

$$\mathbb{P}(A \mid X = x) = \frac{(h+t)!}{h!t!}x^h(1-x)^t.$$

To obtain $\mathbb{P}(A)$, we should integrate $\mathbb{P}(A \mid X = x)$ against the density of X:

$$\begin{aligned}\mathbb{P}(A) &= \int_{-\infty}^{\infty} \mathbb{P}(A \mid X = x) f_X(x)\,\mathrm{d}x = \int_0^1 \frac{(h+t)!}{h!t!}x^h(1-x)^t \cdot \frac{1}{B(r,s)}x^{r-1}(1-x)^{s-1}\,\mathrm{d}x \\ &= \frac{(h+t)!}{h!t!}\frac{1}{B(r,s)}\int_0^1 x^{r+h-1}(1-x)^{s+t-1}\,\mathrm{d}x = \frac{(h+t)!}{h!t!}\frac{B(r+h, s+t)}{B(r,s)}\end{aligned}$$

Putting these pieces together, we obtain

$$f_{X|A}(x) = \frac{1}{\cancel{B(r,s)}}x^{r-1}(1-x)^{s-1} \cdot \frac{\cancel{(h+t)!}}{\cancel{h!t!}}x^h(1-x)^t \cdot \frac{\cancel{h!t!}}{\cancel{(h+t)!}}\frac{\cancel{B(r,s)}}{B(r+h, s+t)}$$

$$= \frac{1}{B(r+h, s+t)} x^{r+h-1}(1-x)^{s+t-1}, \qquad 0 < x < 1.$$

We have found that the posterior distribution has the Beta$(r+h, s+t)$ distribution! We now turn to the interpretation of this key result.

We stated that the expectation of the Beta(r, s) distribution is $r/(r+s)$, which would be our best guess for the bias of the coin if we had observed r heads and s tails in $r+s$ coin flips. The result above states that if we observe h more heads and t more tails, then we now have the Beta$(r+h, s+t)$ distribution, which is consistent with the following interpretation: the first parameter represents how many heads we have seen, and the second parameter represents how many tails we have seen. The amazing part is that if we start with a belief according to the beta distribution, then we never have to use another distribution: repeated applications of Bayes Rule will always yield another beta distribution.

Furthermore, our result almost provides an algorithm for estimating the bias of a coin: start with a beta distribution, and keep flipping coins. If we observe heads, we increment the first parameter of the beta distribution. If we observe tails, we increment the second parameter of the beta distribution. Of course, this is an extremely cheap computation, which makes the beta distribution a convenient family of distributions for the estimation problem.

This leads to a minor problem, which is: how should we initialize the parameters of the beta distribution? Regardless of the choice of the initial parameters, after enough flips, the beta distribution will converge to the correct probability. It is true that our choice of initial parameters certainly influences how long it takes for our distribution to converge. Instead of worrying about this problem, however, we may view it as another advantage of the beta distribution: by initializing our initial parameters, we can incorporate our prior belief into the algorithm. Here are some examples to illustrate this point:

If we take an ordinary coin, we may be reasonably confident that the coin is fair, so we can start with a Beta$(500, 500)$ distribution. The fact that $r = s$ reflects the fact that we think the coin is fair, and our choice of the number 500 represents the strength of our belief: we belief strongly in the fairness of the coin, so we initialize the beta distribution with 500 observations of heads and 500 observations of tails. On the other hand, if the coin was given to you by a friend, perhaps you believe that your friend is not malicious (so you still believe that the coin is fair), but you do not trust the coin as much as you would an ordinary coin. In this case, you may decide to initialize the algorithm with a Beta$(20, 20)$ distribution. Finally, suppose that the coin was given to you by a gambler, and you suspect that the coin may be loaded (so that it is more likely to come up heads). In this case, you may encode your belief with a Beta$(40, 20)$ distribution (your suspicions amount to the same information as having observed 40 heads and 20 tails prior to flipping the coin).

Example 7.6. Now, let us approach the coin-flipping problem from a different perspective. We flip a coin n times and we are looking for the distribution of the number of

heads. However, we do not know the bias of the coin, so we assume a uniform prior over the bias of the coin. Let $X \sim \text{Binomial}(n, Y)$ be the number of heads, where $Y \sim \text{Uniform}[0,1]$. What is the distribution of X?

$$\mathbb{P}(X = x) = \int_{-\infty}^{\infty} \mathbb{P}(X = x \mid Y = y) f_Y(y)\,\mathrm{d}y = \int_0^1 \binom{n}{x} y^x (1-y)^{n-x}\,\mathrm{d}y$$

We can solve the integral by using (7.24). We obtain

$$\mathbb{P}(X = k) = \frac{n!}{\cancel{x!(n-x)!}} \frac{\cancel{x!(n-x)!}}{(n+1)!} = \frac{1}{n+1},$$

so we have found that $X \sim \text{Uniform}\{0, \ldots, n\}$, which agrees with intuition. If your prior belief about the bias is uniform, then the number of heads is also uniform.

7.7 Solutions to Exercises

Exercise 29 First, we compute $\mathbb{P}(Y > 1/2 \mid X = x)$: Conditioned on $X = x$, Y has the Uniform$[0, x]$ distribution, and the tail probability of the uniform distribution is in (6.20):

$$\mathbb{P}\Big(Y > \frac{1}{2} \;\Big|\; X = x\Big) = \begin{cases} 0, & x < 1/2 \\ (x - 1/2)/x, & x \geq 1/2 \end{cases}$$

So, integrate over values of $x \geq 1/2$ to find $\mathbb{P}(Y > 1/2)$ (note that the upper limit of integration is $x = 1$ since $X \sim \text{Uniform}[0,1]$).

$$\begin{aligned} \mathbb{P}\Big(Y > \frac{1}{2}\Big) &= \int_{-\infty}^{\infty} \mathbb{P}\Big(Y > \frac{1}{2} \;\Big|\; X = x\Big) f_X(x)\,\mathrm{d}x = \int_{1/2}^1 \Big(1 - \frac{1}{2x}\Big)\,\mathrm{d}x \\ &= \Big[x - \frac{1}{2}\ln x\Big]_{x=1/2}^{x=1} = \frac{1}{2}(1 - \ln 2). \end{aligned}$$

Exercise 30 One has to be a little careful here since the function $x \mapsto x^2$ is not one-to-one. Let F be the CDF of X. Then, $\mathbb{P}(X^2 \leq x) = \mathbb{P}(-\sqrt{x} \leq X \leq \sqrt{x}) = F(\sqrt{x}) - F(-\sqrt{x})$. Differentiate to obtain

$$f_{X^2}(x) = \frac{1}{2\sqrt{x}}\big(f(-\sqrt{x}) + f(\sqrt{x})\big).$$

Exercise 31 We carry out the integral, but note that y ranges from 0 to ∞.

$$f_Z(z) = \int_0^\infty y\lambda \exp(-\lambda y z)\lambda \exp(-\lambda y)\,\mathrm{d}y = \frac{\lambda}{z+1}\int_0^\infty y\lambda(z+1)\exp\big(-\lambda(z+1)y\big)\,\mathrm{d}y$$

We recognize the integral as the expectation of an Exponential$(\lambda(z+1))$ distribution, which is $(\lambda(z+1))^{-1}$. Hence,

$$f_Z(z) = \frac{1}{(1+z)^2}, \qquad z > 0.$$

Exercise 32 **Gaussian Tail Bounds** For the left inequality, we integrate by parts:

$$\begin{aligned}
&\int_x^\infty (1-3z^{-4})\exp\Big(-\frac{z^2}{2}\Big)\,\mathrm{d}z \\
&\quad= \int_x^\infty \exp\Big(-\frac{z^2}{2}\Big)\,\mathrm{d}z + \int_x^\infty (-3z^{-4})\exp\Big(-\frac{z^2}{2}\Big)\,\mathrm{d}z \\
&\quad= \int_x^\infty \exp\Big(-\frac{z^2}{2}\Big)\,\mathrm{d}z + z^{-3}\exp\Big(-\frac{z^2}{2}\Big)\Big|_{z=x}^\infty + \int_x^\infty z^{-2}\exp\Big(-\frac{z^2}{2}\Big)\,\mathrm{d}z \\
&\quad= \int_x^\infty \exp\Big(-\frac{z^2}{2}\Big)\,\mathrm{d}z - x^{-3}\exp\Big(-\frac{x^2}{2}\Big) - z^{-1}\exp\Big(-\frac{z^2}{2}\Big)\Big|_{z=x}^\infty - \int_x^\infty \exp\Big(-\frac{z^2}{2}\Big)\,\mathrm{d}z \\
&\quad= (x^{-1}-x^{-3})\exp\Big(-\frac{x^2}{2}\Big)
\end{aligned}$$

Therefore, one has

$$(x^{-1}-x^{-3})\exp\Big(-\frac{x^2}{2}\Big) = \int_x^\infty (1-3z^{-4})\exp\Big(-\frac{z^2}{2}\Big)\,\mathrm{d}z \le \int_x^\infty \exp\Big(-\frac{z^2}{2}\Big)\,\mathrm{d}z.$$

For the right inequality, we change variables to $z = x + y$.

$$\begin{aligned}
\int_x^\infty \exp\Big(-\frac{z^2}{2}\Big)\,\mathrm{d}z &= \int_0^\infty \exp\Big(-\frac{(x+y)^2}{2}\Big)\,\mathrm{d}y = \exp\Big(-\frac{x^2}{2}\Big)\int_0^\infty \exp(-xy)\underbrace{\exp\Big(-\frac{y^2}{2}\Big)}_{\le 1}\,\mathrm{d}y \\
&\le \exp\Big(-\frac{x^2}{2}\Big)\int_0^\infty \exp(-xy)\,\mathrm{d}y = \exp\Big(-\frac{x^2}{2}\Big)\Big(-\frac{1}{x}\exp(-xy)\Big)\Big|_{y=0}^\infty \\
&= x^{-1}\exp\Big(-\frac{x^2}{2}\Big)
\end{aligned}$$

We have shown that, asymptotically, the tail probability of the Gaussian distribution goes as $\mathbb{P}(Z > z) \sim (2\pi)^{-1/2} z^{-1}\exp(-z^2/2)$ as $z \to \infty$.

Exercise 33 The identity follows from the fact that $(t/2\pi)^{1/2}\exp(-tx^2/2)$ is the density of a $\mathcal{N}(0,t)$ random variable and the density must integrate to 1. Now, differentiate the identity k times with respect to t to obtain

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty \frac{(-1)^k x^{2k}}{2^k}\exp\Big(-\frac{tx^2}{2}\Big)\,\mathrm{d}x = \frac{(-1)^k 1\cdot 3\cdots(2k-3)\cdot(2k-1)}{2^k} t^{-(2k+1)/2}.$$

Set $t = 1$ to obtain $\mathbb{E}[X^{2k}] = 1\cdot 3\cdots(2k-3)\cdot(2k-1)$. The latter expression is commonly denoted as $(2k-1)!!$ and it can also be written as

$$(2k-1)!! = \frac{(2k)!}{2\cdot 4\cdots(2k-2)\cdot(2k)} = \frac{(2k)!}{2^k k!}.$$

This handles all of the even moments; the odd moments $\mathbb{E}[X^{2k+1}]$ are all 0 due to the symmetry of the density function around 0.

Chapter 8

Information Theory

The subject of this chapter is a foray into **information theory**, which concerns itself with questions such as: "What is the shortest average description for a random variable?" (the **data compression problem**); and "What is the maximum rate at which we can send information through a noisy channel?" (the **data transmission problem**). Although we will not explore these questions in any depth, we will introduce the fundamental notions used in information theory to capture the illusive meaning of "information".

8.1 Entropy

Let X be a discrete random variable taking values in a finite set $\mathcal{X}$. Throughout this chapter, we will use the abbreviated notation $p_X(x) := \mathbb{P}(X = x)$.

Definition 8.1. The **entropy** of X, $H(X)$, is defined to be:

$$\boxed{H(X) := -\sum_{x\in\mathcal{X}} p_X(x)\log_2 p_X(x) = \mathbb{E}\Big[\log_2 \frac{1}{p_X(X)}\Big].}$$

Observe that the entropy of X really only depends on the *distribution* of X, p_X, so we will also write $H(p_X) := H(X)$.

We use the base-2 logarithm because we think of entropy as being measured in *bits*; indeed, we will give an interpretation of entropy as a measure of the *information* contained in the random variable. For now, though, suppose that we chose to measure entropy using a logarithm with a different base $b > 1$, that is, we define $H_b(p_X) := -\sum_{x\in\mathcal{X}} p_X(x)\log_b p_X(x)$. Using the logarithm change-of-base rule,

$$\log_b x = \frac{\ln x}{\ln b} \qquad \text{and} \qquad \log_2 x = \frac{\ln x}{\ln 2},$$

so we find that

$$\log_b x = \frac{\ln 2}{\ln b}\log_2 x = (\log_b 2)(\log_2 x).$$

Therefore, we conclude that $H_b(p_X) = (\log_b 2)H(p_X)$. The moral of the story is that *using a different base for the logarithm just introduces a different constant factor in front of the entropy*, so regardless of which base we choose for the logarithm, the theory will remain essentially the same. From now on, we will stick to using base 2. [1]

Further, observe that since $p_X \leq 1$, $-\log_2 p_X \geq 0$, and so the entropy is always non-negative:

$$\boxed{H(p_X) \geq 0}$$

(We consider this to be a desirable property of entropy, because we do not really know how to make sense of "negative information".)

Now, we will provide an interpretation of entropy. [2] Suppose that you are interested in measuring the value of a random variable X. You perform an experiment and observe the event $\{X = x\}$, where $x \in \mathcal{X}$. We define the **surprise** of this event to be $-\log_2 p_X(x)$ and we regard it as the amount of information you gained from the observation. For example, if $p_X(x) = 1$, then your surprise is 0; but this makes sense because $p_X(x) = 1$ means the event $\{X = x\}$ was certain from the beginning! On the other hand, if $p_X(x) = 0$, then your surprise is ∞, which again makes sense because $p_X(x) = 0$ means that $\{X = x\}$ should have been impossible... uh-oh!

In this language, we see that the entropy is your *expected surprise*. Entropy is a strangely self-referential concept, where you make an observation, and then think about the probability that you would make the observation which you just made.

Example 8.2. Fix $n \in \mathbb{Z}_+$. Let $X \sim \text{Uniform}\{1, \ldots, n\}$, so $p_X(x) = 1/n$ for each $x \in \{1, \ldots, n\}$.

$$H(X) = -\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = -\log_2 \frac{1}{n} = \log_2 n.$$

In fact, the uniform distribution is the distribution with the *largest* possible entropy over n symbols. (By "n symbols", we refer to the fact that $\{1, \ldots, n\}$ has n elements. Notice that the entropy of X does not depend on *what* possible values it can take, as the entropy only depends on the *probabilities* with which it takes on these values. So, saying that X has the largest entropy out of any random variable which takes values in $\{1, \ldots, n\}$ is equivalent to saying that X has the largest entropy out of any random variable which takes on at most n values, regardless of what those n values are.) Intuitively, this is true because the uniform distribution is "the most random"; in other words, before you measure X, you have the least

[1] Perhaps this could be called a "computer science mindset".

[2] In reality, it is possible that you will not be satisfied by any of the interpretations we give. After all, why should we believe that a concept as illusive as "information" can be captured by mathematics? Personally, I think the *true* justification for the concept of entropy is provided by looking at the various situations in which it arises and the various results we can prove about entropy which align with our intuition.

information about X when X is uniformly distributed on the set $\{1, \ldots, n\}$.

Proving the above assertion (that if X takes on at most n values, then $H(X) \leq \log_2 n$) is tricky, so we will wait until we have more tools.

Example 8.3. Let $X \sim \text{Bernoulli}(p)$. Then,

$$H(X) = -p \log_2 p - (1-p) \log_2 (1-p).$$

This is known as the **binary entropy function** and it is commonly denoted simply as $H(p)$. See Figure 8.1.

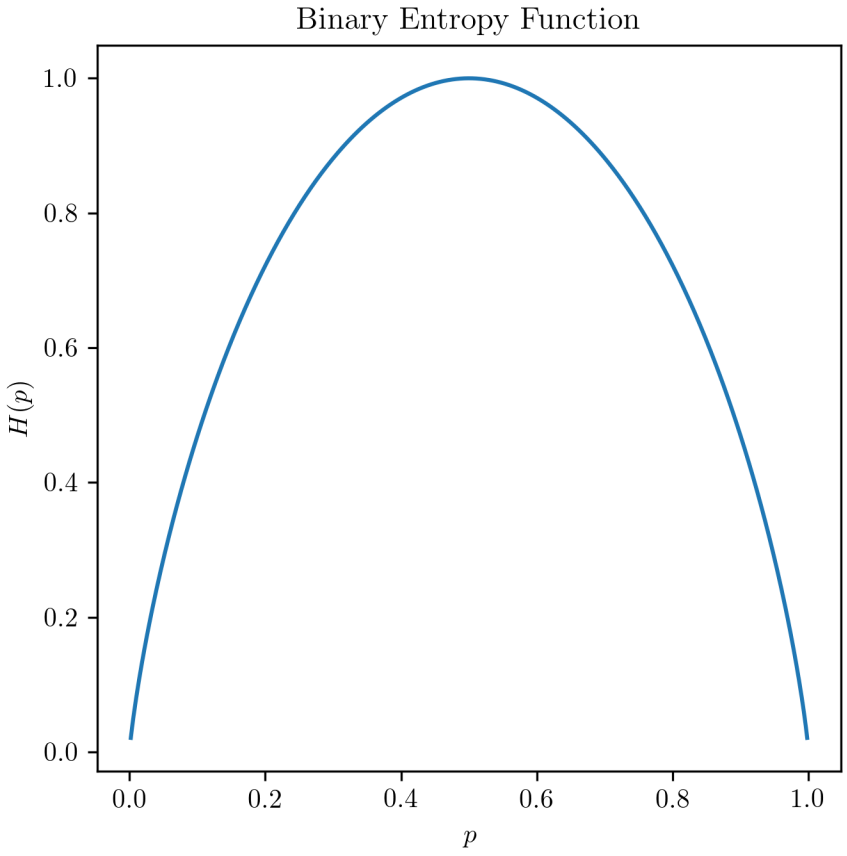


Figure 8.1: Plot of the binary entropy function. Observe that it is non-negative and concave.

Exercise 34 What is the entropy of the Geometric(1/2) distribution?

8.2 Relative Entropy

Look back at the interpretation of surprise mentioned as an interpretation for entropy. Now, consider the situation where you have an incorrect belief about the probability distribution of the random variable X, that is, X has the probability distribution p but you mistakenly believe that X has the probability distribution q. Upon seeing the event $\{X = x\}$, your surprise is now $-\log_2 q(x)$, and your expected surprise overall is $-\sum_{x\in\mathcal{X}} p(x)\log_2 q(x)$.

Since your expected surprise would have been $H(p) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x)$ (if you had correctly known the true distribution p), we see that your *additional* expected surprise from your incorrect belief is

$$-\sum_{x\in\mathcal{X}} p(x)\log_2 q(x) - \Big(-\sum_{x\in\mathcal{X}} p(x)\log_2 p(x)\Big) = \sum_{x\in\mathcal{X}} p(x)\log_2 \frac{p(x)}{q(x)}.$$

We now formulate a definition based on these ideas.

Definition 8.4. The **relative entropy of q from p** is:

$$\boxed{D_{\mathrm{KL}}(p \,\|\, q) := \sum_{x\in\mathcal{X}} p(x)\log_2 \frac{p(x)}{q(x)} = \mathbb{E}_p\Big[\log_2 \frac{p(X)}{q(X)}\Big]}$$

This is also commonly called the **Kullback-Leibler divergence of q from p**, which explains the subscript in the notation.

Observe that if there is any $x \in \mathcal{X}$ such that $p(x) > 0$ but $q(x) = 0$, then the term in the summation for x yields $p(x)\log_2(p(x)/0)$ which we interpret as ∞.

Often, the relative entropy is used as a measure of "distance" between two probability distributions. It is not a true distance function because it is not symmetric: in general, $D_{\mathrm{KL}}(p \,\|\, q) \neq D_{\mathrm{KL}}(q \,\|\, p)$. [3]

Exercise 35 Consider the following two distributions on $\{0, 1\}$:

$$\begin{aligned} p(0) &= 1-p \\ p(1) &= p \end{aligned} \qquad \text{and} \qquad \begin{aligned} q(0) &= 1-q \\ q(1) &= q \end{aligned}$$

Calculate $D_{\mathrm{KL}}(p \,\|\, q)$ and $D_{\mathrm{KL}}(q \,\|\, p)$. [Note: Here, p corresponds to the Bernoulli(p) distribution and q corresponds to the Bernoulli(q) distribution.] Show that for $p = 1/2$ and $q = 1/4$, they are not equal.

However, we can at least show that it is always non-negative:

[3] Moreover, it does not necessarily satisfy the triangle inequality.

Theorem 8.5 (Relative Entropy Inequality)**.** *For all probability distributions p and q on the countable set $\mathcal{X}$, $D_{\mathrm{KL}}(p \,\|\, q) \geq 0$, with equality if and only if $p = q$.*

Proof. We will use the inequality $\ln x \leq x - 1$ for $x > 0$, with equality if and only if $x = 1$. [a] So,

$$\begin{aligned} D_{\mathrm{KL}}(p \,\|\, q) &= \frac{1}{\ln 2} \sum_{x \in \mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} = -\frac{1}{\ln 2} \sum_{x \in \mathcal{X}} p(x) \ln \frac{q(x)}{p(x)} \\ &\geq -\frac{1}{\ln 2} \sum_{x \in \mathcal{X}} p(x) \Big(\frac{q(x)}{p(x)} - 1 \Big) = -\frac{1}{\ln 2} \Big(\sum_{x \in \mathcal{X}} q(x) - \sum_{x \in \mathcal{X}} p(x) \Big) = 0, \end{aligned}$$

with equality if and only if $q(x)/p(x) = 1$ for all $x \in \mathcal{X}$, that is, $p = q$. □

[a] Indeed, $f(x) = x - 1 - \ln x$ has $f''(x) = x^{-2} > 0$ for all $x > 0$, so f is strictly convex and it has a unique minimum. Since $f'(x) = 1 - x^{-1} = 0$ when $x = 1$, it follows that f attains its minimum value of 0 at $x = 1$, which proves that $x - 1 \geq \ln x$ for all $x > 0$, with equality if and only if $x = 1$.

Corollary 8.6 (Maximum Entropy Distribution)**.** *Let X be a random variable taking on values in $\mathcal{X}$, with $|\mathcal{X}| = n$. Then, $H(X) \leq \log_2 n$.*

Proof. Let $p(x) := \mathbb{P}(X = x)$ and q be the uniform distribution on $\mathcal{X}$. Then,

$$D_{\mathrm{KL}}(p \,\|\, q) = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{1/n} = \sum_{x \in \mathcal{X}} p(x) \log_2 n + \sum_{x \in \mathcal{X}} p(x) \log_2 p(x) = \log_2 n - H(X),$$

but from Theorem 8.5, $D_{\mathrm{KL}}(p \,\|\, q) \geq 0$, and so we have $H(X) \leq \log_2 n$. □

Corollary 8.6 verifies our earlier claim that the uniform distribution on n symbols (which has $H(X) = \log_2 n$) has the maximum entropy out of all distributions on n symbols.

8.3 Chernoff Bounds

Take a moment to recall (3.32), which is reproduced here for convenience: For $\theta > 0$,

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}[\exp(\theta X)]}{\exp(\theta x)}. \tag{8.1}$$

We will focus on the simple case of coin flips and demonstrate a connection between the relative entropy and the Chernoff bound.

Theorem 8.7 (Chernoff Bound for Coin Flips)**.** *Let $n \in \mathbb{Z}_+$ and $S_n := X_1 + \cdots + X_n$,*

where the X_i are i.i.d. Bernoulli(p) *random variables. Then, for $\varepsilon > 0$,*

$$\mathbb{P}\Big(\frac{S_n}{n} \geq p + \varepsilon\Big) \leq \exp\big(-nD_{\mathrm{KL}}(p+\varepsilon \,\|\, p)\big),$$

where the relative entropy is calculated using the natural logarithm.

Proof. We will apply (3.32) to the random variable S_n/n:

$$\begin{aligned}\mathbb{P}\Big(\frac{S_n}{n} \geq p+\varepsilon\Big) &= \mathbb{P}\big(S_n \geq (p+\varepsilon)n\big) \leq \frac{\mathbb{E}[\exp(\theta(X_1+\cdots+X_n))]}{\exp(\theta(p+\varepsilon)n)} \\ &= \frac{\mathbb{E}[\exp(\theta X_1)\cdots\exp(\theta X_n)]}{\exp(\theta(p+\varepsilon)n)} = \frac{\mathbb{E}[\exp(\theta X_1)]\cdots\mathbb{E}[\exp(\theta X_n)]}{\exp(\theta(p+\varepsilon)n)} \\ &= \frac{M_X(\theta)^n}{\exp(\theta(p+\varepsilon)n)} = \exp\big\{n\big(\ln M_X(\theta) - \theta(p+\varepsilon)\big)\big\}.\end{aligned}$$

where we have used the i.i.d. assumption, and we used the definition

$$M_X(\theta) := \mathbb{E}[\exp(\theta X_1)] = 1 - p + p\exp\theta$$

(because $\exp(\theta X_1)$ is a random variable which takes on the value 1 with probability $1-p$ and the value $\exp\theta$ with probability p). Now, we seek the best possible bound over all $\theta > 0$, so we differentiate the following quantity with respect to θ:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\big(\ln(1-p+p\exp\theta) - \theta(p+\varepsilon)\big) = \frac{p\exp\theta}{1-p+p\exp\theta} - (p+\varepsilon),$$

and by setting the above quantity to 0, we have:

$$\begin{aligned}\frac{p\exp\theta}{1-p+p\exp\theta} = p+\varepsilon \implies \frac{1-p+p\exp\theta}{p\exp\theta} = \frac{1}{p+\varepsilon} \implies \frac{1-p}{p\exp\theta} + 1 = \frac{1}{p+\varepsilon} \\ \implies \frac{1-p}{p\exp\theta} = \frac{1-p-\varepsilon}{p+\varepsilon} \implies \exp\theta = \frac{p+\varepsilon}{p}\cdot\frac{1-p}{1-p-\varepsilon} \\ \implies \theta = \ln\frac{p+\varepsilon}{p} + \ln\frac{1-p}{1-p-\varepsilon}.\end{aligned}$$

Now, we plug in the above result into $\ln M_X(\theta) - \theta(p+\varepsilon)$:

$$\begin{aligned}&\ln M_X(\theta) + \theta(p+\varepsilon) \\ &\quad= \ln\Big(1-p+(1-p)\frac{p+\varepsilon}{1-p-\varepsilon}\Big) - (p+\varepsilon)\ln\frac{p+\varepsilon}{p} - (p+\varepsilon)\ln\frac{1-p}{1-p-\varepsilon} \\ &\quad= \ln\frac{1-p}{1-p-\varepsilon} - (p+\varepsilon)\ln\frac{p+\varepsilon}{p} - (p+\varepsilon)\ln\frac{1-p}{1-p-\varepsilon}\end{aligned}$$

$$= (1-p-\varepsilon)\ln\frac{1-p}{1-p-\varepsilon} + (p+\varepsilon)\ln\frac{p}{p+\varepsilon} = -D_{\mathrm{KL}}(p+\varepsilon \,\|\, p).$$

We have the remarkable result that $\mathbb{P}(S_n/n \geq p+\varepsilon) \leq \exp(-nD_{\mathrm{KL}}(p+\varepsilon \,\|\, p))$. □

We could go on to upper bound $D_{\mathrm{KL}}(p+\varepsilon \,\|\, p)$, but at this stage we are mainly interested in knowing that there is a bound with decays exponentially with the number of random variables. Therefore, it suffices to observe that for $\varepsilon > 0$, $D_{\mathrm{KL}}(p+\varepsilon \,\|\, p) > 0$ due to Theorem 8.5, and so (8.1) is indeed exponentially decaying.

8.4 Solutions to Exercises

Exercise 34 For the Geometric(1/2) distribution, the distribution p is $p(x) = 2^{-x}$ for $x \in \mathbb{Z}_+$. So, if $X \sim$ Geometric(1/2), then $H(p) = -\mathbb{E}[\log_2 p(X)] = \mathbb{E}[X] = 2$.

Exercise 35 From the definition, we have

$$D_{\mathrm{KL}}(p \,\|\, q) = (1-p)\log_2\frac{1-p}{1-q} + p\log_2\frac{p}{q},$$
$$D_{\mathrm{KL}}(q \,\|\, p) = (1-q)\log_2\frac{1-q}{1-p} + q\log_2\frac{q}{p}.$$

Plugging in $p = 1/2$ and $q = 1/4$, we find (numerically) that $D_{\mathrm{KL}}(p \,\|\, q) \approx 0.208$ and $D_{\mathrm{KL}}(q \,\|\, p) \approx 0.189$, which gives an example where the relative entropy is not symmetric.

Bibliography

[1] Patrick Billingsley. *Probability & Measure*. New York, New York: John Wiley & Sons, Inc., 1995.

[2] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[3] Rick Durrett. *Probability: Theory & Examples*. New York, New York: Cambridge University Press, 2010.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, & Prediction*. Springer, 2008.

[6] Jim Pitman. *Probability*. New York, New York: Springer-Verlag New York, Inc., 1997.

[7] John A. Rice. *Mathematical Statistics & Data Analysis*. Belmont, California: Thomson Higher Education, 2007.

[8] Howard M. Taylor and Samuel Karlin. *An Introduction to Stochastic Modeling*. San Diego, California: Academic Press, 1998.

[9] Jean Walrand. *Probability in Electrical Engineering and Computer Science: An Application-Driven Course*. Quoi?, 2014.

Index