Probability and Statistics for Data Science

Carlos Fernandez-Granda

Preface

These notes were developed for the course *Probability and Statistics for Data Science* at the Center for Data Science in NYU. The goal is to provide an overview of fundamental concepts in probability and statistics from first principles. I would like to thank Levent Sagun and Vlad Kobzar, who were teaching assistants for the course, as well as Brett Bernstein and David Rosenberg for their useful suggestions. I am also very grateful to all my students for their feedback.

While writing these notes, I was supported by the National Science Foundation under NSF award DMS-1616340.

New York, August 2017

Contents

Chapter 1

Basic Probability Theory

In this chapter we introduce the mathematical framework of probability theory, which makes it possible to reason about uncertainty in a principled way using set theory. Appendix A contains a review of basic set-theory concepts.

1.1 Probability spaces

Our goal is to build a mathematical framework to represent and analyze uncertain phenomena, such as the result of rolling a die, tomorrow's weather, the result of an NBA game, etc. To this end we model the phenomenon of interest as an **experiment** with several (possibly infinite) mutually exclusive **outcomes**.

Except in simple cases, when the number of outcomes is small, it is customary to reason about sets of outcomes, called *events*. To quantify how likely it is for the outcome of the experiment to belong to a specific event, we assign a **probability** to the event. More formally, we define a **measure** (recall that a measure is a function that maps sets to real numbers) that assigns probabilities to each event of interest.

More formally, the experiment is characterized by constructing a **probability space**.

Definition 1.1.1 (Probability space)**.** *A probability space is a triple* $(\Omega, \mathcal{F}, \mathrm{P})$ *consisting of:*

- *A* ***sample space*** Ω*, which contains all possible outcomes of the experiment.*
- *A set of events* $\mathcal{F}$*, which must be a* ***σ-algebra*** *(see Definition 1.1.2 below).*
- *A* ***probability measure*** P *that assigns probabilities to the events in* $\mathcal{F}$ *(see Definition 1.1.4 below).*

Sample spaces may be **discrete** or **continuous**. Examples of discrete sample spaces include the possible outcomes of a coin toss, the score of a basketball game, the number of people that show up at a party, etc. Continuous sample spaces are usually intervals of $\mathbb{R}$ or $\mathbb{R}^n$ used to model time, position, temperature, etc.

The term σ-algebra is used in measure theory to denote a collection of sets that satisfy certain conditions listed below. Don't be too intimidated by it. It is just a sophisticated way of stating that if we assign a probability to certain events (for example *it will rain tomorrow* or *it will*

snow tomorrow) we also need to assign a probability to their complements (i.e. *it will not rain tomorrow* or *it will not snow tomorrow*) and to their union (*it will rain or snow tomorrow*).

Definition 1.1.2 (σ-algebra)**.** *A σ-algebra $\mathcal{F}$ is a collection of sets in Ω such that:*

1. *If a set $S \in \mathcal{F}$ then $S^c \in \mathcal{F}$.*
2. *If the sets $S_1, S_2 \in \mathcal{F}$, then $S_1 \cup S_2 \in \mathcal{F}$. This also holds for infinite sequences; if $S_1, S_2, \ldots \in \mathcal{F}$ then $\cup_{i=1}^{\infty} S_i \in \mathcal{F}$.*
3. *$\Omega \in \mathcal{F}$.*

If our sample space is discrete, a possible choice for the σ-algebra is the **power set** of the sample space, which consists of all possible sets of elements in the sample space. If we are tossing a coin and the sample space is

$$\Omega := \{\text{heads}, \text{tails}\}, \tag{1.1}$$

then the power set is a valid σ-algebra

$$\mathcal{F} := \{\text{heads or tails}, \text{heads}, \text{tails}, \emptyset\}, \tag{1.2}$$

where $\emptyset$ denotes the empty set. However, in many cases σ-algebras do not contain every possible set of outcomes.

Example 1.1.3 (Cholesterol)**.** A doctor is interested in modeling the cholesterol levels of her patients probabilistically. Every time a patient visits her, she tests their cholesterol level. Here the *experiment* is the cholesterol test, the outcome is the measured cholesterol level, and the sample space Ω is the positive real line. The doctor is mainly interested in whether the patients to have low, borderline-high, or high cholesterol. The event L (low cholesterol) contains all outcomes below 200 mg/dL, the event B (borderline-high cholesterol) contains all outcomes between 200 and 240 mg/dL, and the event H (high cholesterol) contains all outcomes above 240 mg/dL. The σ-algebra $\mathcal{F}$ of possible events therefore equals

$$\mathcal{F} := \{L \cup B \cup H, L \cup B, L \cup H, B \cup H, L, B, H, \emptyset\}. \tag{1.3}$$

The events are a partition of the sample space, which simplifies deriving the corresponding σ-algebra. △

The role of the probability measure P is to quantify how likely we are to encounter each of the events in the σ-algebra. Intuitively, the probability of an event A can be interpreted as the fraction of times that the outcome of the experiment is in A, as the number of repetitions tends to infinity. It follows that probabilities should always be nonnegative. Also, if two events A and B are disjoint (their intersection is empty), then

$$P(A \cup B) = \frac{\text{outcomes in } A \text{ or } B}{\text{total}} \tag{1.4}$$

$$= \frac{\text{outcomes in } A + \text{outcomes in } B}{\text{total}} \tag{1.5}$$

$$= \frac{\text{outcomes in } A}{\text{total}} + \frac{\text{outcomes in } B}{\text{total}} \tag{1.6}$$

$$= P(A) + P(B). \tag{1.7}$$

Probabilities of unions of disjoint events should equal the sum of the individual probabilities. Additionally, the probability of the whole sample space Ω should equal one, as it contains all outcomes

$$P(\Omega) = \frac{\text{outcomes in } \Omega}{\text{total}} \tag{1.8}$$

$$= \frac{\text{total}}{\text{total}} \tag{1.9}$$

$$= 1. \tag{1.10}$$

These conditions are necessary for a measure to be a valid probability measure.

Definition 1.1.4 (Probability measure)**.** *A probability measure is a function defined over the sets in a σ-algebra $\mathcal{F}$ such that:*

1. $\mathrm{P}(S) \geq 0$ *for any event* $S \in \mathcal{F}$.

2. *If the sets* $S_1, S_2, \ldots, S_n \in \mathcal{F}$ *are disjoint (i.e.* $S_i \cap S_j = \emptyset$ *for* $i \neq j$*) then*

$$\mathrm{P}\left(\cup_{i=1}^{n} S_i\right) = \sum_{i=1}^{n} \mathrm{P}(S_i). \tag{1.11}$$

Similarly, for a countably infinite sequence of disjoint sets $S_1, S_2, \ldots \in \mathcal{F}$

$$\mathrm{P}\left(\lim_{n\to\infty} \cup_{i=1}^{n} S_i\right) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{P}(S_i). \tag{1.12}$$

3. $\mathrm{P}(\Omega) = 1$.

The two first axioms capture the intuitive idea that the probability of an event is a measure such as mass (or length or volume): just like the mass of any object is nonnegative and the total mass of several distinct objects is the sum of their masses, the probability of any event is nonnegative and the probability of the union of several disjoint objects is the sum of their probabilities. However, in contrast to mass, the amount of probability in an experiment cannot be unbounded. If it is highly likely that it will rain tomorrow, then it cannot be also very likely that it will *not* rain. If the probability of an event S is large, then the probability of its complement S^c must be small. This is captured by the third axiom, which normalizes the probability measure (and implies that $\mathrm{P}(S^c) = 1 - \mathrm{P}(S)$).

It is important to stress that the probability measure does *not* assign probabilities to individual outcomes, but rather to events in the σ-algebra. The reason for this is that when the number of possible outcomes is uncountably infinite, then one cannot assign nonzero probability to all the outcomes and still satisfy the condition $\mathrm{P}(\Omega) = 1$. This is not an exotic situation, it occurs for instance in the cholesterol example where any positive real number is a possible outcome. In the case of discrete or countable sample spaces, the σ-algebra may equal the power set of the sample space, which means that we do assign probabilities to events that only contain a single outcome (e.g. the coin-toss example).

Example 1.1.5 (Cholesterol (continued))**.** A valid probability measure for Example 1.1.3 is

$$\mathrm{P}(L) = 0.6, \quad \mathrm{P}(B) = 0.28, \quad \mathrm{P}(H) = 0.12. \tag{1.13}$$

Using the properties, we can determine for instance that $\mathrm{P}(B \cup H) = 0.6 + 0.28 = 0.88$. △

Definition 1.1.4 has the following consequences:

$$\mathrm{P}(\emptyset) = 0, \tag{1.14}$$
$$A \subseteq B \quad \text{implies} \quad \mathrm{P}(A) \leq \mathrm{P}(B), \tag{1.15}$$
$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B). \tag{1.16}$$

We omit the proofs (try proving them on your own).

1.2 Conditional probability

Conditional probability is a crucial concept in probabilistic modeling. It allows us to update probabilistic models when additional information is revealed. Consider a probabilistic space $(\Omega, \mathcal{F}, \mathrm{P})$ where we find out that the outcome of the experiment belongs to a certain event $S \in \mathcal{F}$. This obviously affects how likely it is for any other event $S' \in \mathcal{F}$ to have occurred: we can rule out any outcome not belonging to S. The updated probability of each event is known as the **conditional probability** of S' **given** S. Intuitively, the conditional probability can be interpreted as the fraction of outcomes in S that are also in S',

$$\mathrm{P}\left(S'|S\right) = \frac{\text{outcomes in } S' \text{ and } S}{\text{outcomes in } S} \tag{1.17}$$
$$= \frac{\text{outcomes in } S' \text{ and } S}{\text{total}} \frac{\text{total}}{\text{outcomes in } S} \tag{1.18}$$
$$= \frac{\mathrm{P}(S' \cap S)}{\mathrm{P}(S)}, \tag{1.19}$$

where we assume that $\mathrm{P}(S) \neq 0$ (later on we will have to deal with the case when S has zero probability, which often occurs in continuous probability spaces). The definition is rather intuitive: S is now the new sample space, so if the outcome is in S' then it must belong to $S' \cap S$. However, just using the probability of the intersection would underestimate how likely it is for S' to occur because the sample space has been reduced to S. Therefore we normalize by the probability of S. As a sanity check, we have $\mathrm{P}(S|S) = 1$ and if S and S' are disjoint then $\mathrm{P}(S'|S) = 0$.

The conditional probability $\mathrm{P}(\cdot|S)$ is a valid probability measure in the probability space $(S, \mathcal{F}_S, \mathrm{P}(\cdot|S))$, where $\mathcal{F}_S$ is a σ-algebra that contains the intersection of S and the sets in $\mathcal{F}$. To simplify notation, when we condition on an intersection of sets we write the conditional probability as

$$\mathrm{P}(S|A, B, C) := \mathrm{P}(S|A \cap B \cap C), \tag{1.20}$$

for any events S, A, B, C.

Example 1.2.1 (Flights and rain)**.** JFK airport hires you to estimate how the punctuality of flight arrivals is affected by the weather. You begin by defining a probability space for which the sample space is

$$\Omega = \{\text{late and rain}, \text{late and no rain}, \text{on time and rain}, \text{on time and no rain}\} \tag{1.21}$$

and the σ-algebra is the power set of Ω. From data of past flights you determine that a reasonable estimate for the probability measure of the probability space is

$$\mathrm{P}(\text{late, no rain}) = \frac{2}{20}, \qquad \mathrm{P}(\text{on time, no rain}) = \frac{14}{20}, \tag{1.22}$$

$$\mathrm{P}(\text{late, rain}) = \frac{3}{20}, \qquad \mathrm{P}(\text{on time, rain}) = \frac{1}{20}. \tag{1.23}$$

The airport is interested in the probability of a flight being late if it rains, so you define a new probability space conditioning on the event *rain*. The sample space is the set of all outcomes such that *rain* occurred, the σ-algebra is the power set of $\{\text{on time}, \text{late}\}$ and the probability measure is $\mathrm{P}(\cdot|\text{rain})$. In particular,

$$\mathrm{P}(\text{late}|\text{rain}) = \frac{\mathrm{P}(\text{late, rain})}{\mathrm{P}(\text{rain})} = \frac{3/20}{3/20 + 1/20} = \frac{3}{4} \tag{1.24}$$

and similarly $\mathrm{P}(\text{late}|\text{no rain}) = 1/8$.

△

Conditional probabilities can be used to compute the intersection of several events in a structured way. By definition, we can express the probability of the intersection of two events $A, B \in \mathcal{F}$ as follows,

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B|A) \tag{1.25}$$

$$= \mathrm{P}(B)\,\mathrm{P}(A|B)\,. \tag{1.26}$$

In this formula $\mathrm{P}(A)$ is known as the **prior** probability of A, as it captures the information we have about A before anything else is revealed. Analogously, $\mathrm{P}(A|B)$ is known as the **posterior** probability. These are fundamental quantities in Bayesian models, discussed in Chapter 10. Generalizing (1.25) to a sequence of events gives the *chain rule*, which allows to express the probability of the intersection of multiple events in terms of conditional probabilities. We omit the proof, which is a straightforward application of induction.

Theorem 1.2.2 (Chain rule)**.** *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space and* $S_1, S_2, \ldots$ *a collection of events in* $\mathcal{F}$*,*

$$\mathrm{P}(\cap_i S_i) = \mathrm{P}(S_1)\,\mathrm{P}(S_2|S_1)\,\mathrm{P}(S_3|S_1 \cap S_2)\cdots \tag{1.27}$$

$$= \prod_i \mathrm{P}\left(S_i|\cap_{j=1}^{i-1} S_j\right). \tag{1.28}$$

Sometimes, estimating the probability of a certain event directly may be more challenging than estimating its probability conditioned on simpler events. A collection of disjoint sets $A_1, A_2, \ldots$ such that $\Omega = \cup_i A_i$ is called a **partition** of Ω. The law of total probability allows us to pool conditional probabilities together, weighting them by the probability of the individual events in the partition, to compute the probability of the event of interest.

Theorem 1.2.3 (Law of total probability)**.** *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space and let the collection of disjoint sets* A_1, A_2, $\ldots \in \mathcal{F}$ *be any partition of* Ω*. For any set* $S \in \mathcal{F}$

$$\mathrm{P}(S) = \sum_i \mathrm{P}(S \cap A_i) \tag{1.29}$$

$$= \sum_i \mathrm{P}(A_i)\,\mathrm{P}(S|A_i)\,. \tag{1.30}$$

Proof. This is an immediate consequence of the chain rule and Axiom 2 in Definition 1.1.4, since $S = \cup_i S \cap A_i$ and the sets $S \cap A_i$ are disjoint. □

Example 1.2.4 (Aunt visit)**.** Your aunt is arriving at JFK tomorrow and you would like to know how likely it is for her flight to be on time. From Example 1.2.1, you recall that

$$\mathrm{P}(\text{late}|\text{rain}) = 0.75, \quad \mathrm{P}(\text{late}|\text{no rain}) = 0.125. \tag{1.31}$$

After checking out a weather website, you determine that $\mathrm{P}(\text{rain}) = 0.2$.

Now, how can we integrate all of this information? The events *rain* and *no rain* are disjoint and cover the whole sample space, so they form a partition. We can consequently apply the law of total probability to determine

$$\mathrm{P}(\text{late}) = \mathrm{P}(\text{late}|\text{rain})\,\mathrm{P}(\text{rain}) + \mathrm{P}(\text{late}|\text{no rain})\,\mathrm{P}(\text{no rain}) \tag{1.32}$$

$$= 0.75 \cdot 0.2 + 0.125 \cdot 0.8 = 0.25. \tag{1.33}$$

So the probability that your aunt's plane is late is 1/4.

△

It is crucial to realize that in general $\mathrm{P}(A|B) \neq \mathrm{P}(B|A)$: most players in the NBA probably own a basketball ($\mathrm{P}(\text{owns ball}|\text{NBA})$ is large) but most people that own basketballs are not in the NBA ($\mathrm{P}(\text{NBA}|\text{owns ball})$ is small). The reason is that the prior probabilities are very different: $\mathrm{P}(\text{NBA})$ is much smaller than $\mathrm{P}(\text{owns ball})$. However, it is possible to *invert* conditional probabilities, i.e. find $\mathrm{P}(A|B)$ from $\mathrm{P}(B|A)$, as long as we take into account the priors. This straightforward consequence of the definition of conditional probability is known as Bayes' rule.

Theorem 1.2.5 (Bayes' rule)**.** *For any events* A *and* B *in a probability space* $(\Omega, \mathcal{F}, \mathrm{P})$

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A)\,\mathrm{P}(B|A)}{\mathrm{P}(B)}, \tag{1.34}$$

as long as $\mathrm{P}(B) > 0$.

Example 1.2.6 (Aunt visit (continued))**.** You explain the probabilistic model described in Example 1.2.4 to your cousin Marvin who lives in California. A day later, you tell him that your aunt arrived late but you don't mention whether it rained or not. After he hangs up, Marvin wants to figure out the probability that it rained. Recall that the probability of rain was 0.2, but since your aunt arrived late he should update the estimate. Applying Bayes' rule and the

law of total probability:

$$\begin{aligned} \mathrm{P}\left(\text{rain}|\text{late}\right) &= \frac{\mathrm{P}\left(\text{late}|\text{rain}\right)\mathrm{P}\left(\text{rain}\right)}{\mathrm{P}\left(\text{late}\right)} && (1.35)\\ &= \frac{\mathrm{P}\left(\text{late}|\text{rain}\right)\mathrm{P}\left(\text{rain}\right)}{\mathrm{P}\left(\text{late}|\text{rain}\right)\mathrm{P}\left(\text{rain}\right)+\mathrm{P}\left(\text{late}|\text{no rain}\right)\mathrm{P}\left(\text{no rain}\right)} && (1.36)\\ &= \frac{0.75\cdot 0.2}{0.75\cdot 0.2+0.125\cdot 0.8} = 0.6. && (1.37) \end{aligned}$$

As expected, the probability that it rained increases under the assumption that your aunt is late.

△

1.3 Independence

As discussed in the previous section, conditional probabilities quantify the extent to which the knowledge of the occurrence of a certain event affects the probability of another event. In some cases, it makes no difference: the events are **independent**. More formally, events A and B are independent if and only if

$$\mathrm{P}\left(A|B\right) = \mathrm{P}\left(A\right). \tag{1.38}$$

This definition is not valid if $\mathrm{P}\left(B\right) = 0$. The following definition covers this case and is otherwise equivalent.

Definition 1.3.1 (Independence)**.** *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. Two events $A, B \in \mathcal{F}$ are independent if and only if*

$$\mathrm{P}\left(A \cap B\right) = \mathrm{P}\left(A\right)\mathrm{P}\left(B\right). \tag{1.39}$$

Example 1.3.2 (Congress)**.** We consider a data set compiling the votes of members of the U.S. House of Representatives on two issues in 1984 [1]. The issues are cost sharing for a water project (issue 1) and adoption of the budget resolution (issue 2). We model the behavior of the congressmen probabilistically, defining a sample space where each outcome is a sequence of votes. For instance, a possible outcome is *issue 1 = yes, issue 2 = no*. We choose the σ-algebra to be the power set of the sample space. To estimate the probability measure associated to different events, we just compute the fraction of their occurrence in the data.

$$\begin{aligned} \mathrm{P}\left(\text{issue 1} = \text{yes}\right) &\approx \frac{\text{members voting yes on issue 1}}{\text{total votes on issue 1}} && (1.40)\\ &= 0.597, && (1.41)\\ \mathrm{P}\left(\text{issue 2} = \text{yes}\right) &\approx \frac{\text{members voting yes on issue 2}}{\text{total votes on issue 2}} && (1.42)\\ &= 0.417, && (1.43)\\ \mathrm{P}\left(\text{issue 1} = \text{yes} \cap \text{issue 2} = \text{yes}\right) &\approx \frac{\text{members voting yes on issues 1 and 2}}{\text{total members voting on issues 1 and 2}} && (1.44)\\ &= 0.069. && (1.45) \end{aligned}$$

[1] The data is available here.

Based on these data, we can evaluate whether voting behavior on the two issues was dependent. In other words, if we know how a member voted on issue 1, does this provide information about how they voted on issue 2? The answer is yes, since

$$P\left(\text{issue 1} = \text{yes}\right) \mathrm{P}\left(\text{issue 2} = \text{yes}\right) = 0.249 \tag{1.46}$$

is very different from $\mathrm{P}\left(\text{issue 1} = \text{yes} \cap \text{issue 2} = \text{yes}\right)$. If a member voted yes on issue 1, they were less likely to vote yes on issue 2. △

Similarly, we can define **conditional independence** between two events given a third event. A and B are conditionally independent given C if and only if

$$\mathrm{P}\left(A|B,C\right) = \mathrm{P}\left(A|C\right), \tag{1.47}$$

where $\mathrm{P}\left(A|B,C\right) := \mathrm{P}\left(A|B \cap C\right)$. Intuitively, this means that the probability of A is not affected by whether B occurs or not, *as long as* C *occurs*.

Definition 1.3.3 (Conditional independence)**.** *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space. Two events* $A, B \in \mathcal{F}$ *are conditionally independent given a third event* $C \in \mathcal{F}$ *if and only if*

$$\mathrm{P}\left(A \cap B|C\right) = \mathrm{P}\left(A|C\right) \mathrm{P}\left(B|C\right). \tag{1.48}$$

Example 1.3.4 (Congress (continued))**.** The main factor that determines how members of congress vote is political affiliation. We therefore incorporate it into the probabilistic model in Example 1.3.2. Each outcome now consists of the votes for issues 1 and 2, and also the affiliation of the member, e.g. *issue 1 = yes*, *issue 2 = no*, *affiliation = republican*, or *issue 1 = no*, *issue 2 = no*, *affiliation = democrat*. The σ-algebra is the power set of the sample space. We again estimate the values of the probability measure associated to different events using the data:

$$\mathrm{P}\left(\text{issue 1} = \text{yes} \,|\, \text{republican}\right) \approx \frac{\text{republicans voting yes on issue 1}}{\text{total republican votes on issue 1}} \tag{1.49}$$

$$= 0.134, \tag{1.50}$$

$$\mathrm{P}\left(\text{issue 2} = \text{yes} \,|\, \text{republican}\right) \approx \frac{\text{republicans voting yes on issue 2}}{\text{total republican votes on issue 2}} \tag{1.51}$$

$$= 0.988, \tag{1.52}$$

$$\mathrm{P}\left(\text{issue 1} = \text{yes} \cap \text{issue 2} = \text{yes} \,|\, \text{republican}\right) \approx \frac{\text{republicans voting yes on issues 1 and 2}}{\text{republicans voting on both issues}}$$

$$= 0.134. \tag{1.53}$$

Based on these data, we can evaluate whether voting behavior on the two issues was dependent *conditioned on the member being a republican*. In other words, if we know how a member voted on issue 1 and that they are a republican, does this provide information about how they voted on issue 2? The answer is no, since

$$\mathrm{P}\left(\text{issue 1} = \text{yes} \,|\, \text{republican}\right) \mathrm{P}\left(\text{issue 2} = \text{yes} \,|\, \text{republican}\right) = 0.133 \tag{1.54}$$

is very close to $\mathrm{P}\left(\text{issue 1} = \text{yes} \cap \text{issue 2} = \text{yes} \,|\, \text{republican}\right)$. The votes are approximately independent given the knowledge that the member is a republican. △

As suggested by Examples 1.3.2 and 1.3.4, independence does not imply conditional independence or vice versa. This is further illustrated by the following examples. From now on, to simplify notation, we write the probability of the intersection of several events in the following form

$$\mathrm{P}\left(A, B, C\right) := \mathrm{P}\left(A \cap B \cap C\right). \tag{1.55}$$

Example 1.3.5 (Conditional independence does not imply independence)**.** Your cousin Marvin from Exercise 1.2.6 always complains about taxis in New York. From his many visits to JFK he has calculated that

$$\mathrm{P}\left(\text{taxi}|\text{rain}\right) = 0.1, \quad \mathrm{P}\left(\text{taxi}|\text{no rain}\right) = 0.6, \tag{1.56}$$

where *taxi* denotes the event of finding a free taxi after picking up your luggage. Given the events *rain* and *no rain*, it is reasonable to model the events *plane arrived late* and *taxi* as conditionally independent,

$$\mathrm{P}\left(\text{taxi}, \text{late}|\text{rain}\right) = \mathrm{P}\left(\text{taxi}|\text{rain}\right)\mathrm{P}\left(\text{late}|\text{rain}\right), \tag{1.57}$$

$$\mathrm{P}\left(\text{taxi}, \text{late}|\text{no rain}\right) = \mathrm{P}\left(\text{taxi}|\text{no rain}\right)\mathrm{P}\left(\text{late}|\text{no rain}\right). \tag{1.58}$$

The logic behind this is that the availability of taxis after picking up your luggage depends on whether it's raining or not, but not on whether the plane is late or not (we assume that availability is constant throughout the day). Does this assumption imply that the events are independent?

If they were independent, then knowing that your aunt was late would give no information to Marvin about taxi availability. However,

$$\mathrm{P}\left(\text{taxi}\right) = \mathrm{P}\left(\text{taxi}, \text{rain}\right) + \mathrm{P}\left(\text{taxi}, \text{no rain}\right) \quad \text{(by the law of total probability)} \tag{1.59}$$

$$= \mathrm{P}\left(\text{taxi}|\text{rain}\right)\mathrm{P}\left(\text{rain}\right) + \mathrm{P}\left(\text{taxi}|\text{no rain}\right)\mathrm{P}\left(\text{no rain}\right) \tag{1.60}$$

$$= 0.1 \cdot 0.2 + 0.6 \cdot 0.8 = 0.5, \tag{1.61}$$

$$\mathrm{P}\left(\text{taxi}|\text{late}\right) = \frac{\mathrm{P}\left(\text{taxi}, \text{late}, \text{rain}\right) + \mathrm{P}\left(\text{taxi}, \text{late}, \text{no rain}\right)}{\mathrm{P}\left(\text{late}\right)} \quad \text{(by the law of total probability)}$$

$$= \frac{\mathrm{P}\left(\text{taxi}|\text{rain}\right)\mathrm{P}\left(\text{late}|\text{rain}\right)\mathrm{P}\left(\text{rain}\right) + \mathrm{P}\left(\text{taxi}|\text{no rain}\right)\mathrm{P}\left(\text{late}|\text{no rain}\right)\mathrm{P}\left(\text{no rain}\right)}{\mathrm{P}\left(\text{late}\right)}$$

$$= \frac{0.1 \cdot 0.75 \cdot 0.2 + 0.6 \cdot 0.125 \cdot 0.8}{0.25} = 0.3. \tag{1.62}$$

$\mathrm{P}\left(\text{taxi}\right) \neq \mathrm{P}\left(\text{taxi}|\text{late}\right)$ so the events are *not* independent. This makes sense, since if the airplane is late, it is more probable that it is raining, which makes taxis more difficult to find.

△

Example 1.3.6 (Independence does not imply conditional independence)**.** After looking at your probabilistic model from Example 1.2.1 your contact at JFK points out that delays are often caused by mechanical problems in the airplanes. You look at the data and determine that

$$\mathrm{P}\left(\text{problem}\right) = \mathrm{P}\left(\text{problem}|\text{rain}\right) = \mathrm{P}\left(\text{problem}|\text{no rain}\right) = 0.1, \tag{1.63}$$

so the events *mechanical problem* and *rain in NYC* are independent, which makes intuitive sense. After some more analysis of the data, you estimate

$$\mathrm{P}\left(\text{late}|\text{problem}\right) = 0.7, \quad \mathrm{P}\left(\text{late}|\text{no problem}\right) = 0.2, \quad \mathrm{P}\left(\text{late}|\text{no rain}, \text{problem}\right) = 0.5.$$

The next time you are waiting for Marvin at JFK, you start wondering about the probability of his plane having had some mechanical problem. Without any further information, this probability is 0.1. It is a sunny day in New York, but this is of no help because according to the data (and common sense) the events *problem* and *rain* are independent.

Suddenly they announce that Marvin's plane is late. Now, what is the probability that his plane had a mechanical problem? At first thought you might apply Bayes' rule to compute $\mathrm{P}\left(\text{problem}|\text{late}\right) = 0.28$ as in Example 1.2.6. However, you are not using the fact that it is sunny. This means that the rain was not responsible for the delay, so intuitively a mechanical problem should be more likely. Indeed,

$$\begin{aligned}
\mathrm{P}\left(\text{problem}|\text{late}, \text{no rain}\right) &= \frac{\mathrm{P}\left(\text{late}, \text{no rain}, \text{problem}\right)}{\mathrm{P}\left(\text{late}, \text{no rain}\right)} && (1.64)\\
&= \frac{\mathrm{P}\left(\text{late}|\text{no rain}, \text{problem}\right)\mathrm{P}\left(\text{no rain}\right)\mathrm{P}\left(\text{problem}\right)}{\mathrm{P}\left(\text{late}|\text{no rain}\right)\mathrm{P}\left(\text{no rain}\right)} && \text{(by the Chain Rule)}\\
&= \frac{0.5 \cdot 0.1}{0.125} = 0.4. && (1.65)
\end{aligned}$$

Since $\mathrm{P}\left(\text{problem}|\text{late}, \text{no rain}\right) \neq \mathrm{P}\left(\text{problem}|\text{late}\right)$ the events *mechanical problem* and *rain in NYC* are *not* conditionally independent given the event *plane is late*.

△

Chapter 2

Random Variables

Random variables are a fundamental tool in probabilistic modeling. They allow us to model numerical quantities that are *uncertain*: the temperature in New York tomorrow, the time of arrival of a flight, the position of a satellite... Reasoning about such quantities probabilistically allows us to structure the information we have about them in a principled way.

2.1 Definition

Formally, we define a random variables as a function mapping each outcome in a probability space to a real number.

Definition 2.1.1 (Random variable)**.** *Given a probability space* $(\Omega, \mathcal{F}, \mathrm{P})$*, a random variable* X *is a function from the sample space* Ω *to the real numbers* $\mathbb{R}$*. Once the outcome* $\omega \in \Omega$ *of the experiment is revealed, the corresponding* $X(\omega)$ *is known as a* ***realization*** *of the random variable.*

Remark 2.1.2 (Rigorous definition)**.** *If we want to be completely rigorous, Definition 2.1.1 is missing some details. Consider two sample spaces* Ω_1 *and* Ω_2*, and a* σ*-algebra* $\mathcal{F}_2$ *of sets in* Ω_2*. Then, for* X *to be a random variable, there must exist a* σ*-algebra* $\mathcal{F}_1$ *in* Ω_1 *such that for any set* S *in* $\mathcal{F}_2$ *the inverse image of* S*, defined by*

$$X^{-1}(S) := \{\omega \mid X(\omega) \in S\}, \tag{2.1}$$

belongs to $\mathcal{F}_1$*. Usually, we take* Ω_2 *to be the reals* $\mathbb{R}$ *and* $\mathcal{F}_2$ *to be the Borel* σ*-algebra, which is defined as the smallest* σ*-algebra defined on the reals that contains all open intervals (amazingly, it is possible to construct sets of real numbers that do not belong to this* σ*-algebra). In any case, for the purpose of these notes, Definition 2.1.1 is sufficient (more information about the formal foundations of probability can be found in any book on measure theory and advanced probability theory).*

Remark 2.1.3 (Notation)**.** *We often denote events of the form*

$$\{X(\omega) \in \mathcal{S} \ : \ \omega \in \Omega\} \tag{2.2}$$

for some random variable X *and some set* $\mathcal{S}$ *as*

$$\{X \in \mathcal{S}\} \tag{2.3}$$

to alleviate notation, since the underlying probability space is often of no significance once we have specified the random variables of interest.

A random variable quantifies our uncertainty about the quantity it represents, *not* the value that it happens to finally take once the outcome is revealed. You should *never* think of a random variable as having a fixed numerical value. If the outcome is known, then that determines a *realization* of the random variable. In order to stress the difference between random variables and their realizations, we denote the former with uppercase letters (X, Y, ...) and the latter with lowercase letters (x, y, ...).

If we have access to the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ in which the random variable is defined, then it is straightforward to compute the probability of a random variable X belonging to a certain set S:[1] it is the probability of the event that comprises all outcomes in Ω which X maps to S,

$$\mathrm{P}\left(X \in S\right) = \mathrm{P}\left(\{\omega \mid X\left(\omega\right) \in S\}\right). \tag{2.4}$$

However, we almost never model the probability space directly, since this requires estimating the probability of every possible event in the corresponding σ-algebra. As we explain in Sections 2.2 and 2.3, there are more practical methods to specify random variables, which automatically imply that a valid underlying probability space exists. The existence of this probability space ensures that the whole framework is mathematically sound, but you don't really have to worry about it.

2.2 Discrete random variables

Discrete random variables take values on a *finite or countably infinite* subset of $\mathbb{R}$ such as the integers. They are used to model discrete numerical quantities: the outcome of the roll of a die, the score in a basketball game, etc.

2.2.1 Probability mass function

To specify a discrete random variable it is enough to determine the probability of each value that it can take. In contrast to the case of continuous random variables, this is tractable because these values are countable by definition.

Definition 2.2.1 (Probability mass function)**.** *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $X : \Omega \to \mathbb{Z}$ a random variable. The probability mass function (pmf) of X is defined as*

$$p_X\left(x\right) := \mathrm{P}\left(\{\omega \mid X\left(\omega\right) = x\}\right). \tag{2.5}$$

In words, $p_X\left(x\right)$ is the probability that X equals x.

We usually say that a random variable is **distributed** according to a certain pmf.

If the discrete range of X is denoted by D, then the triplet $\left(D, 2^D, p_X\right)$ is a valid probability space (recall that 2^D is the power set of D). In particular, p_x is a valid probability measure

[1] Strictly speaking, S needs to belong to the Borel σ-algebra. Again, this comprises essentially any subset of the reals that you will ever encounter in probabilistic modeling

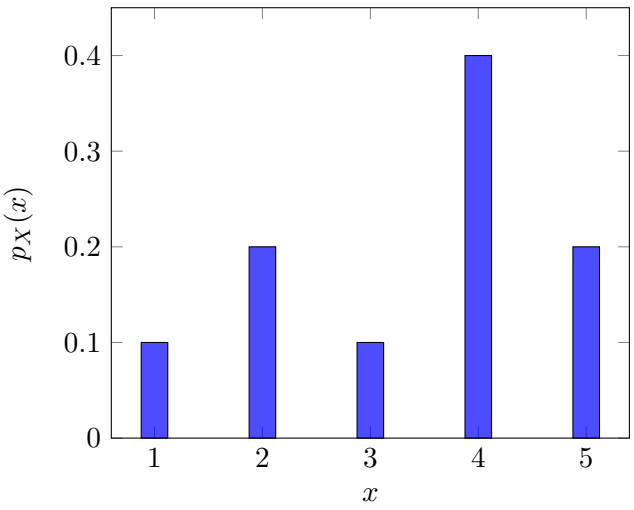


Figure 2.1: Probability mass function of the random variable X in Example 2.2.2.

which satisfies

$$p_X(x) \geq 0 \quad \text{for any } x \in D, \tag{2.6}$$

$$\sum_{x \in D} p_X(x) = 1. \tag{2.7}$$

The converse is also true, if a function defined on a countable subset D of the reals is nonnegative and adds up to one, then it may be interpreted as the pmf of a random variable. In fact, in practice we usually define discrete random variables by just specifying their pmf.

To compute the probability that a random variable X is in a certain set S we take the sum of the pmf over all the values contained in S:

$$\mathrm{P}(X \in S) = \sum_{x \in S} p_X(x). \tag{2.8}$$

Example 2.2.2 (Discrete random variable)**.** Figure 2.1 shows the probability mass function of a discrete random variable X (check that it adds up to one). To compute the probability of X belonging to different sets we apply (2.8):

$$\mathrm{P}(X \in \{1, 4\}) = p_X(1) + p_X(4) = 0.5, \tag{2.9}$$

$$\mathrm{P}(X > 3) = p_X(4) + p_X(5) = 0.6. \tag{2.10}$$

△

2.2.2 Important discrete random variables

In this section we describe several discrete random variables that are useful for probabilistic modeling.

Bernoulli

Bernoulli random variables are used to model experiments that have two possible outcomes. By convention we usually represent an outcome by 0 and the other outcome by 1. A canonical example is flipping a biased coin, such that the probability of obtaining heads is p. If we encode heads as 1 and tails as 0, then the result of the coin flip corresponds to a Bernoulli random variable with parameter p.

Definition 2.2.3 (Bernoulli)**.** *The pmf of a Bernoulli random variable with parameter $p \in [0, 1]$ is given by*

$$p_X(0) = 1 - p, \tag{2.11}$$
$$p_X(1) = p. \tag{2.12}$$

A special kind of Bernoulli random variable is the indicator random variable of an event. This random variable is particularly useful in proofs.

Definition 2.2.4 (Indicator)**.** *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. The indicator random variable of an event $S \in \mathcal{F}$ is defined as*

$$1_S(\omega) = \begin{cases} 1, & \text{if } \omega \in S, \\ 0, & \text{otherwise.} \end{cases} \tag{2.13}$$

By definition the distribution of an indicator random variable is Bernoulli with parameter $\mathrm{P}(S)$.

Geometric

Imagine that we take a biased coin and flip it until we obtain heads. If the probability of obtaining heads is p and the flips are independent then the probability of having to flip k times is

$$\mathrm{P}(k \text{ flips}) = \mathrm{P}(\text{1st flip} = \text{tails}, \ldots, k-1\text{th flip} = \text{tails}, k\text{th flip} = \text{heads}) \tag{2.14}$$
$$= \mathrm{P}(\text{1st flip} = \text{tails}) \cdots \mathrm{P}(k-1\text{th flip} = \text{tails})\, \mathrm{P}(k\text{th flip} = \text{heads}) \tag{2.15}$$
$$= (1-p)^{k-1} p. \tag{2.16}$$

This reasoning can be applied to any situation in which a random experiment with a fixed probability p is repeated until a particular outcome occurs, as long as the independence assumption is met. In such cases the number of repetitions is modeled as a geometric random variable.

Definition 2.2.5 (Geometric)**.** *The pmf of a geometric random variable with parameter p is given by*

$$p_X(k) = (1-p)^{k-1} p, \quad k = 1, 2, \ldots \tag{2.17}$$

Figure 2.2 shows the probability mass function of geometric random variables with different parameters. The larger p is, the more the distribution concentrates around smaller values of k.

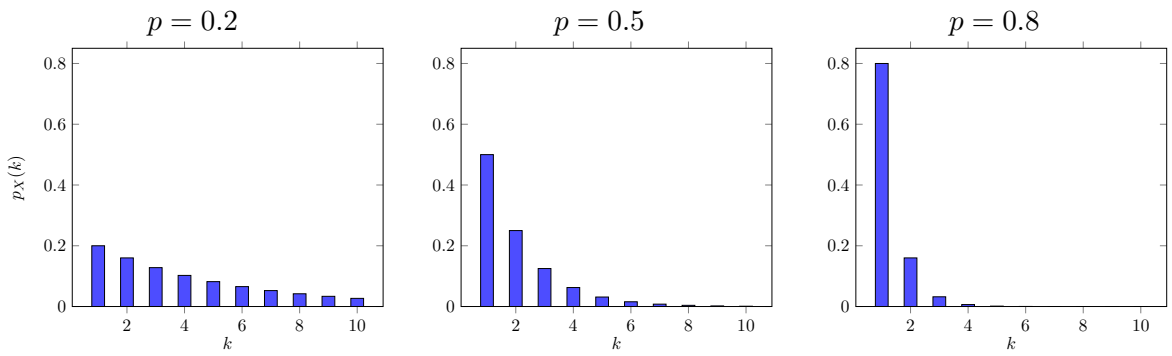


Figure 2.2: Probability mass function of three geometric random variables with different parameters.

Binomial

Binomial random variables are extremely useful in probabilistic modeling. They are used to model the number of positive outcomes of n trials modeled as independent Bernoulli random variables with the same parameter. The following example illustrates this with coin flips.

Example 2.2.6 (Coin flips)**.** If we flip a biased coin n times, what is the probability that we obtain exactly k heads if the flips are independent and the probability of heads is p?

Let us first consider a simpler problem: what is the probability of first obtaining k heads and then $n - k$ tails? By independence, the answer is

$$\begin{aligned} &\mathrm{P}\left(k \text{ heads, then } n-k \text{ tails}\right) && (2.18)\\ &= \mathrm{P}\left(1\text{st flip} = \text{heads}, \ldots, k\text{th flip} = \text{heads}, k+1\text{th flip} = \text{tails}, \ldots, n\text{th flip} = \text{tails}\right)\\ &= \mathrm{P}\left(1\text{st flip} = \text{heads}\right) \cdots \mathrm{P}\left(k\text{th flip} = \text{heads}\right) \mathrm{P}\left(k+1\text{th flip} = \text{tails}\right) \cdots \mathrm{P}\left(n\text{th flip} = \text{tails}\right)\\ &= p^k \left(1-p\right)^{n-k}. && (2.19) \end{aligned}$$

Note that the same reasoning implies that this is also the probability of obtaining exactly k heads *in any fixed order*. The probability of obtaining exactly k heads is the union of all of these events. Because these events are disjoint (we cannot obtain exactly k heads in two different orders simultaneously) we can add their individual to compute the probability of our event of interest. We just need to know the number of possible orderings. By basic combinatorics, this is given by the binomial coefficient $\binom{n}{k}$, defined as

$$\binom{n}{k} := \frac{n!}{k!\left(n-k\right)!}. \tag{2.20}$$

We conclude that

$$\mathrm{P}\left(k \text{ heads out of } n \text{ flips}\right) = \binom{n}{k} p^k \left(1-p\right)^{(n-k)}. \tag{2.21}$$

△

The random variable representing the number of heads in the example is called a binomial random variable.

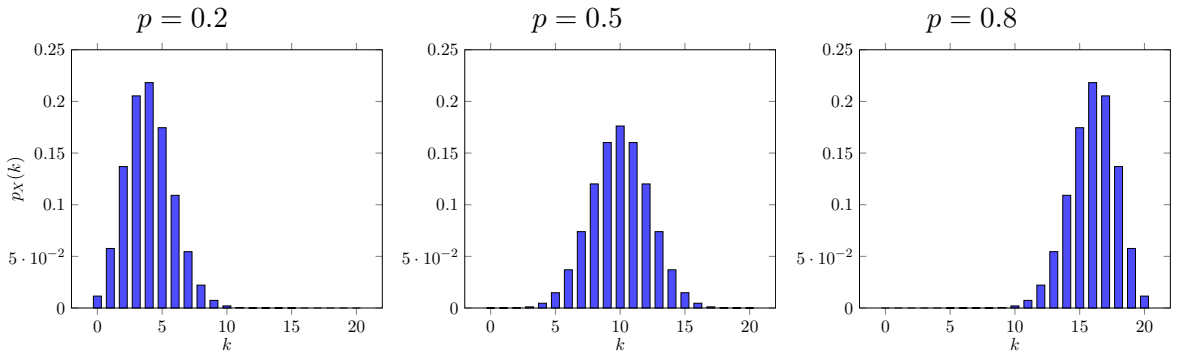


Figure 2.3: Probability mass function of three binomial random variables with different values of p and $n = 20$.

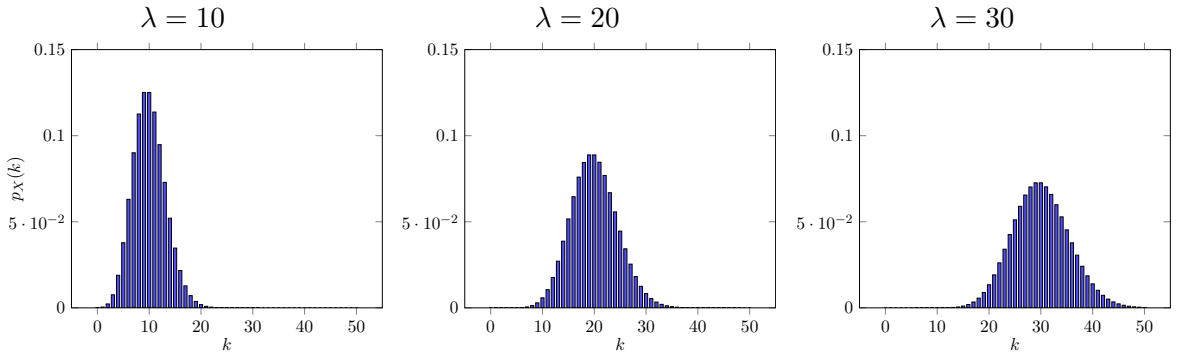


Figure 2.4: Probability mass function of three Poisson random variables with different parameters.

Definition 2.2.7 (Binomial)**.** *The pmf of a binomial random variable with parameters n and p is given by*

$$p_X(k) = \binom{n}{k} p^k (1-p)^{(n-k)}, \quad k = 0, 1, 2, \ldots, n. \tag{2.22}$$

Figure 2.3 shows the probability mass function of binomial random variables with different values of p.

Poisson

We motivate the definition of the Poisson random variable using an example.

Example 2.2.8 (Call center)**.** A call center wants to model the number of calls they receive over a day in order to decide how many people to hire. They make the following assumptions:

1. Each call occurs independently from every other call.
2. A given call has the same probability of occurring at any given time of the day.
3. Calls occur at a rate of λ calls per day.

In Chapter 5, we will see that these assumptions define a Poisson process.

Our aim is to compute the probability of receiving exactly k calls during a given day. To do this we discretize the day into n intervals, compute the desired probability assuming each interval is very small and then let $n \to \infty$.

The probability that a call occurs in an interval of length $1/n$ is λ/n by Assumptions 2 and 3. The probability that $m > 1$ calls occur is $(\lambda/n)^m$. If n is very large this probability is negligible compared to the probability that either one or zero calls are received in the interval, in fact it tends to zero when we take the limit $n \to \infty$. The total number of calls occurring over the whole hour can consequently be approximated by the number of intervals in which a call occurs, as long as n is large enough. Since a call occurs in each interval with the same probability and calls happen independently, the number of calls over a whole day can be modeled as a binomial random variable with parameters n and $p := \lambda/n$.

We now compute the distribution of calls when the intervals are arbitrarily small, i.e. when $n \to \infty$:

$$\begin{aligned}
\mathrm{P}\left(k \text{ calls during the day }\right) &= \lim_{n\to\infty} \mathrm{P}\left(k \text{ calls in } n \text{ small intervals}\right) && (2.23)\\
&= \lim_{n\to\infty} \binom{n}{k} p^k \left(1-p\right)^{(n-k)} && (2.24)\\
&= \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{(n-k)} && (2.25)\\
&= \lim_{n\to\infty} \frac{n!\,\lambda^k}{k!\,(n-k)!\,\left(n-\lambda\right)^k} \left(1-\frac{\lambda}{n}\right)^n && (2.26)\\
&= \frac{\lambda^k\, e^{-\lambda}}{k!}. && (2.27)
\end{aligned}$$

The last step follows from the following lemma proved in Section 2.7.1.

Lemma 2.2.9.

$$\lim_{n\to\infty} \frac{n!}{(n-k)!\,\left(n-\lambda\right)^k} \left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}. \tag{2.28}$$

△

Random variables with the pmf that we have derived in the example are called Poisson random variables. They are used to model situations where something happens from time to time at a constant rate: packets arriving at an Internet router, earthquakes, traffic accidents, etc. The number of such events that occur over a fixed interval follows a Poisson distribution, as long as the assumptions we listed in the example hold.

Definition 2.2.10 (Poisson)**.** *The pmf of a Poisson random variable with parameter λ is given by*

$$p_X\left(k\right) = \frac{\lambda^k\, e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots \tag{2.29}$$

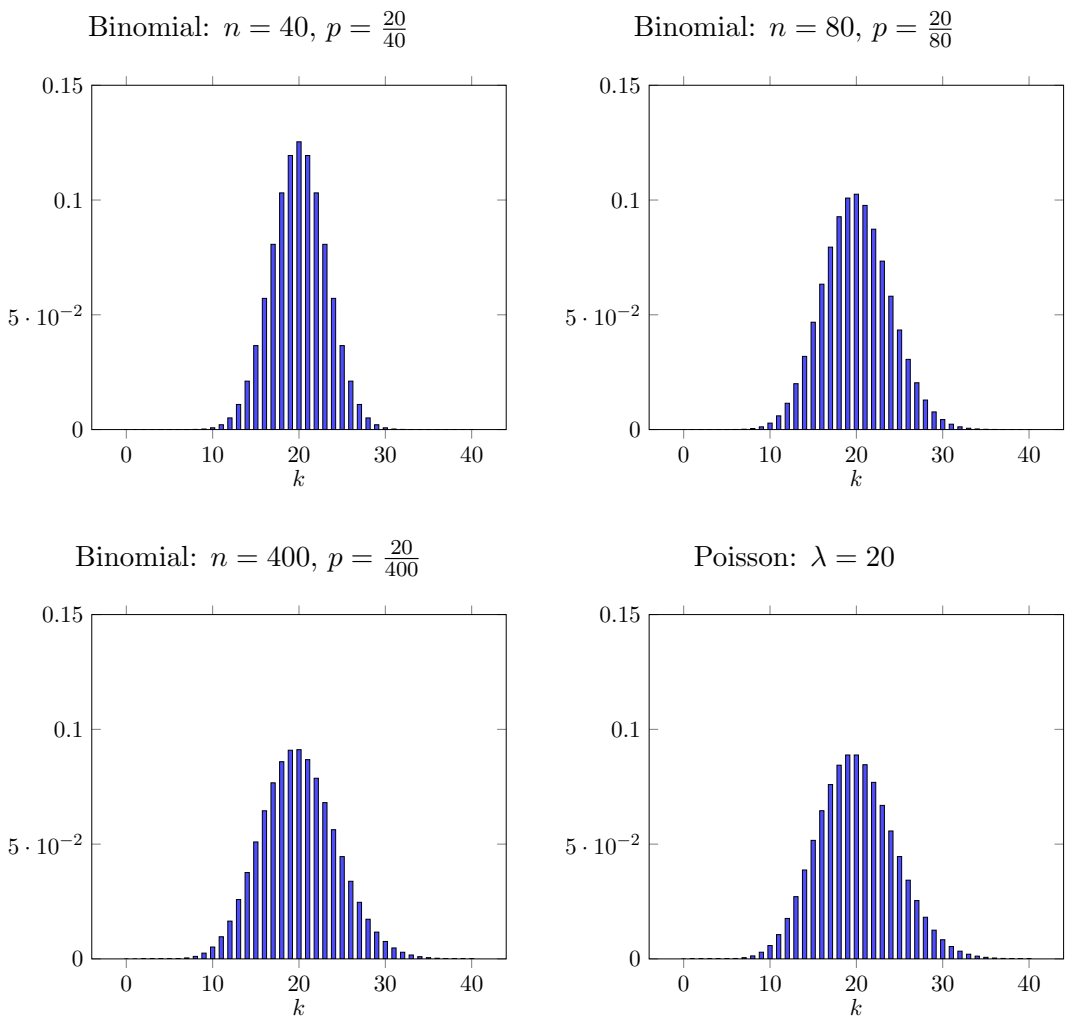


Figure 2.5: Convergence of the binomial pmf with $p = \lambda/n$ to a Poisson pmf of parameter λ as n grows.

Figure 2.4 shows the probability mass function of Poisson random variables with different values of λ. In Example 2.2.8 we prove that as $n \to \infty$ the pmf of a binomial random variable with parameters n and λ/n tends to the pmf of a Poisson with parameter λ (as we will see later in the course, this is an example of *convergence in distribution*). Figure 2.5 shows an example of this phenomenon numerically; the convergence is quite fast.

You might feel a bit skeptical about Example 2.2.8: the probability of receiving a call surely changes over the day and it must be different on weekends! That is true, but the model is actually very useful if we restrict our attention to shorter periods of time. In Figure 2.6 we show the result of modeling the number of calls received by a call center in Israel[2] over an interval of four hours (8 pm to midnight) using a Poisson random variable. We plot the histogram of the number of calls received during that interval for two months (September and October of 1999) together with a Poisson pmf fitted to the data (we will learn how to fit distributions to data later on in the course). Despite the fact that our assumptions do not hold exactly, the model

[2]The data is available here.

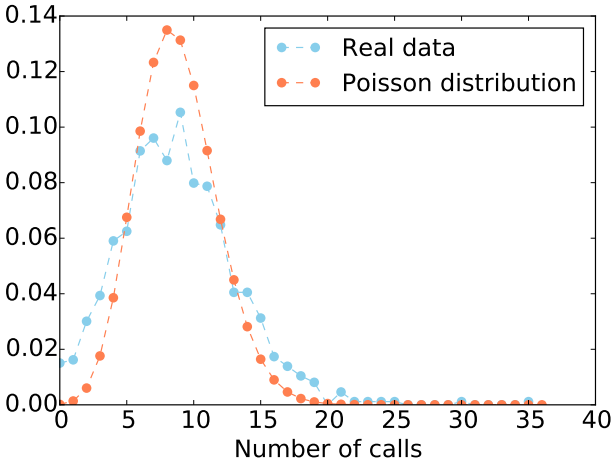


Figure 2.6: In blue, we see the histogram of the number of calls received during an interval of four hours over two months at a call center in Israel. A Poisson pmf approximating the distribution of the data is plotted in orange.

produces a reasonably good fit.

2.3 Continuous random variables

Physical quantities are often best described as continuous: temperature, duration, speed, weight, etc. In order to model such quantities probabilistically we could discretize their domain and represent them as discrete random variables. However, we may not want our conclusions to depend on how we choose the discretization grid. Constructing a continuous model allows us to obtain insights that are valid for *sufficiently fine* grids without worrying about discretization.

Precisely because continuous domains model the limit when discrete outcomes have an arbitrarily fine granularity, we *cannot* characterize the probabilistic behavior of a continuous random variable by just setting values for the probability of X being equal to individual outcomes, as we do for discrete random variables. In fact, we *cannot* assign nonzero probabilities to specific outcomes of an uncertain continuous quantity. This would result in uncountable disjoint outcomes with nonzero probability. The sum of an uncountable number of positive values is infinite, so the probability of their union would be greater than one, which does not make sense.

More rigorously, it turns out that we cannot define a valid probability measure on the power set of $\mathbb{R}$ (justifying this requires measure theory and is beyond the scope of these notes). Instead, we consider events that are composed of *unions of intervals of* $\mathbb{R}$. Such events form a σ-algebra called the Borel σ-algebra. This σ-algebra is granular enough to represent any set that you might be interested in (try thinking of a set that cannot be expressed as a countable union of intervals), while allowing for valid probability measures to be defined on it.

2.3.1 Cumulative distribution function

To specify a random variable on the Borel σ-algebra it suffices to determine the probability of the random variable belonging to all intervals of the form $(-\infty, x)$ for any $x \in \mathbb{R}$.

Definition 2.3.1 (Cumulative distribution function)**.** *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ a random variable. The cumulative distribution function (cdf) of X is defined as*

$$F_X(x) := \mathrm{P}(X \leq x). \tag{2.30}$$

In words, $F_X(x)$ is the probability of X being smaller than x.

Note that the cumulative distribution function can be defined for *both continuous and discrete* random variables.

The following lemma describes some basic properties of the cdf. You can find the proof in Section 2.7.2.

Lemma 2.3.2 (Properties of the cdf)**.** *For any continuous random variable X*

$$\lim_{x \to -\infty} F_X(x) = 0, \tag{2.31}$$

$$\lim_{x \to \infty} F_X(x) = 1, \tag{2.32}$$

$$F_X(b) \geq F_X(a) \quad \text{if } b > a, \quad \text{i.e. } F_X \text{ is nondecreasing.} \tag{2.33}$$

To see why the cdf completely determines a random variable recall that we are only considering sets that can be expressed as unions of intervals. The probability of a random variable X belonging to an interval $(a, b]$ is given by

$$\begin{aligned} \mathrm{P}(a < X \leq b) &= \mathrm{P}(X \leq b) - \mathrm{P}(X \leq a) && (2.34)\\ &= F_X(b) - F_X(a). && (2.35) \end{aligned}$$

Remark 2.3.3. *Since individual points have zero probability, for any continuous random variable X*

$$\mathrm{P}(a < X \leq b) = \mathrm{P}(a \leq X \leq b) = \mathrm{P}(a < X < b) = \mathrm{P}(a \leq X < b). \tag{2.36}$$

Now, to find the probability of X belonging to any particular set, we only need to decompose it into disjoint intervals and apply (2.35), as illustrated by the following example.

Example 2.3.4 (Continuous random variable)**.** Consider a continuous random variable X with a cdf given by

$$F_X(x) := \begin{cases} 0 & \text{for } x < 0, \\ 0.5\,x & \text{for } 0 \leq x \leq 1, \\ 0.5 & \text{for } 1 \leq x \leq 2, \\ 0.5\left(1 + (x-2)^2\right) & \text{for } 2 \leq x \leq 3, \\ 1 & \text{for } x > 3. \end{cases} \tag{2.37}$$

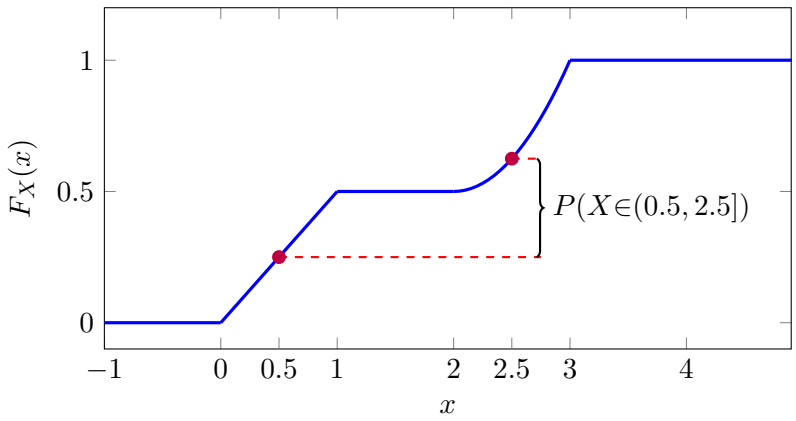


Figure 2.7: Cumulative distribution function of the random variable in Examples 2.3.4 and 2.3.7.

Figure 2.7 shows the cdf on the left image. You can check that it satisfies the properties in Lemma 2.3.2. To determine the probability that X is between 0.5 and 2.5, we apply (2.35),

$$\mathrm{P}\left(0.5 < X \leq 2.5\right) = F_X\left(2.5\right) - F_X\left(0.5\right) = 0.375, \tag{2.38}$$

as illustrated in Figure 2.7. △

2.3.2 Probability density function

If the cdf of a continuous random variable is differentiable, its derivative can be interpreted as a density function. This density can then be integrated to obtain the probability of the random variable belonging to an interval or a union of intervals (and hence to any Borel set).

Definition 2.3.5 (Probability density function)**.** *Let $X : \Omega \to \mathbb{R}$ be a random variable with cdf F_X. If F_X is differentiable then the probability density function or pdf of X is defined as*

$$f_X\left(x\right) := \frac{dF_X\left(x\right)}{d\ x}. \tag{2.39}$$

Intuitively, $f_X\left(x\right)\Delta$ is the probability of X belonging to an interval of width Δ around x as $\Delta \to 0$. By the fundamental theorem of calculus, the probability of a random variable X belonging to an interval is given by

$$\mathrm{P}\left(a < X \leq b\right) = F_X\left(b\right) - F_X\left(a\right) \tag{2.40}$$

$$= \int_a^b f_X\left(x\right) \mathrm{d}x. \tag{2.41}$$

Our sets of interest belong the Borel σ-algebra, and hence can be decomposed into unions of intervals, so we can obtain the probability of X belonging to any such set S by integrating its pdf over S

$$\mathrm{P}\left(X \in S\right) = \int_S f_X\left(x\right) \mathrm{d}x. \tag{2.42}$$

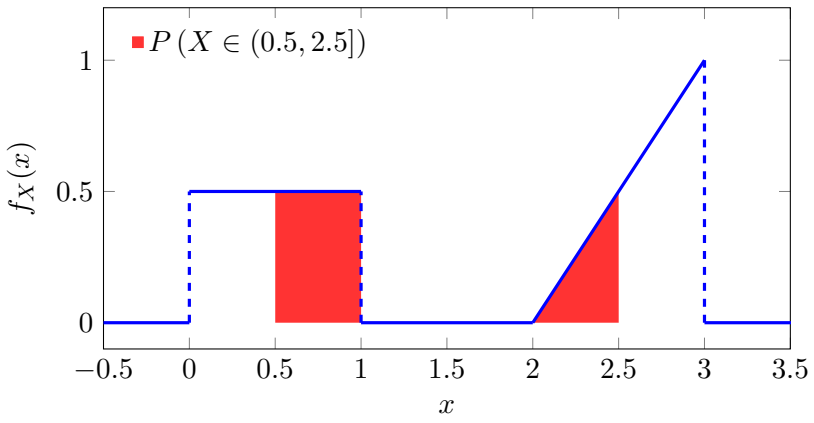


Figure 2.8: Probability density function of the random variable in Examples 2.3.4 and 2.3.7.

In particular, since X belongs to $\mathbb{R}$ by definition

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = \mathrm{P}\left(X \in \mathbb{R}\right) = 1. \tag{2.43}$$

It follows from the monotonicity of the cdf (2.33) that the pdf is nonnegative

$$f_X(x) \geq 0, \tag{2.44}$$

since otherwise we would be able to find two points $x_1 < x_2$ for which $F_X(x_2) < F_X(x_1)$.

Remark 2.3.6 (The pdf is not a probability measure)**.** *The pdf is a* density *which must be integrated to yield a probability. In particular, it is not necessarily smaller than one (for example, take $a = 0$ and $b = 1/2$ in Definition 2.3.8 below).*

Finally, just as in the case of discrete random variables, we often say that a random variable is **distributed** according to a certain pdf or cdf, or that we know its distribution. The reason is that the pmf, pdf or cdf suffice to characterize the underlying probability space.

Example 2.3.7 (Continuous random variable (continued))**.** To compute the pdf of the random variable in Example 2.3.4 we differentiate its cdf, to obtain

$$f_X(x) = \begin{cases} 0 & \text{for } x < 0, \\ 0.5 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{for } 1 \leq x \leq 2 \\ x - 2 & \text{for } 2 \leq x \leq 3 \\ 0 & \text{for } x > 3. \end{cases} \tag{2.45}$$

Figure 2.8 shows the pdf. You can check that it integrates to one. To determine the probability that X is between 0.5 and 2.5, we can just integrate over that interval to obtain the same answer as in Example 2.3.4,

$$\mathrm{P}\left(0.5 < X \leq 2.5\right) = \int_{0.5}^{2.5} f_X(x)\,\mathrm{d}x \tag{2.46}$$

$$= \int_{0.5}^{1} 0.5\,\mathrm{d}x + \int_{2}^{2.5} x - 2\,\mathrm{d}x = 0.375. \tag{2.47}$$

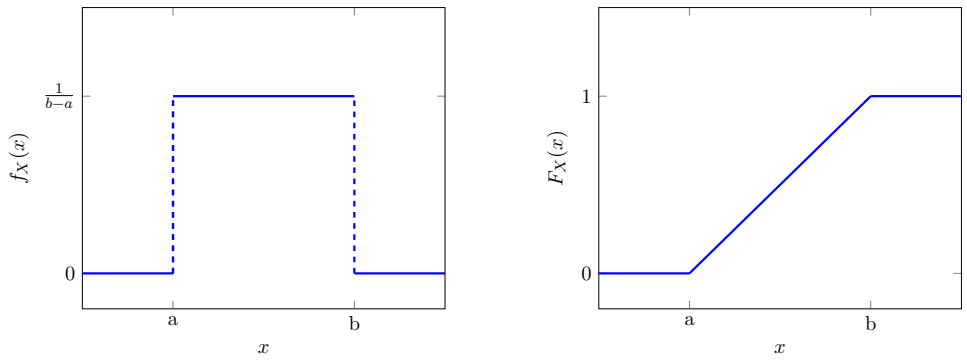


Figure 2.9: Probability density function (left) and cumulative distribution function (right) of a uniform random variable X.

Figure 2.8 illustrates that the probability of an event is equal to the area under the pdf once we restrict it to the corresponding subset of the real line.

△

2.3.3 Important continuous random variables

In this section we describe several continuous random variables that are useful in probabilistic modeling and statistics.

Uniform

A uniform random variable models an experiment in which every outcome within a continuous interval is equally likely. As a result the pdf is constant over the interval. Figure 2.9 shows the pdf and cdf of a uniform random variable.

Definition 2.3.8 (Uniform)**.** *The pdf of a uniform random variable with domain* $[a, b]$*, where* $b > a$ *are real numbers, is given by*

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases} \tag{2.48}$$

Exponential

Exponential random variables are often used to model the time that passes until a certain event occurs. Examples include decaying radioactive particles, telephone calls, earthquakes and many others.

Definition 2.3.9 (Exponential)**.** *The pdf of an exponential random variable with parameter* λ *is given by*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2.49}$$

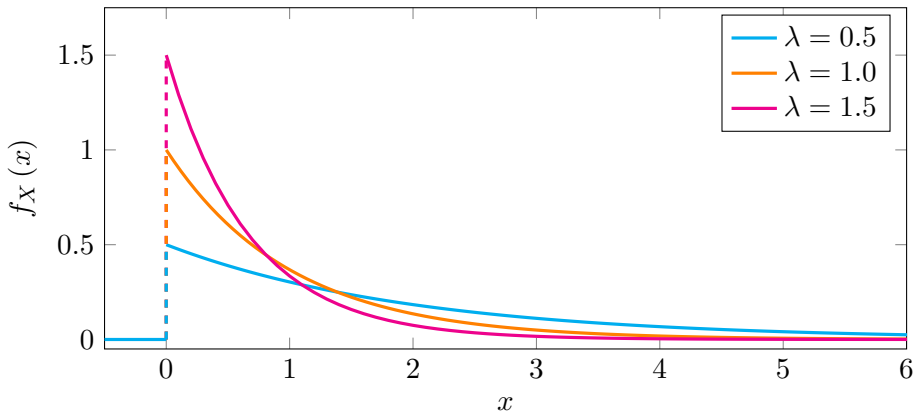


Figure 2.10: Probability density functions of exponential random variables with different parameters.

Figure 2.10 shows the pdf of three exponential random variables with different parameters. In order to illustrate that the potential of exponential distributions for modeling real data, in Figure 2.11 we plot the histogram of inter-arrival times of calls at the same call center in Israel we mentioned earlier. In more detail, these inter-arrival times are the times between consecutive calls occurring between 8 pm and midnight over two days in September 1999. An exponential model fits the data quite well.

An important property of an exponential random variable is that it is *memoryless*. We elaborate on this property, which is shared by the geometric distribution, in Section 2.4.

Gaussian or Normal

The Gaussian or normal random variable is arguably the most popular random variable in all of probability and statistics. It is often used to model variables with unknown distributions in the natural sciences. This is motivated by the fact that sums of independent random variables often converge to Gaussian distributions. This phenomenon is captured by the Central Limit Theorem, which we discuss in Chapter 6.

Definition 2.3.10 (Gaussian)**.** *The pdf of a Gaussian or normal random variable with mean μ and standard deviation σ is given by*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2.50}$$

A Gaussian distribution with mean μ and standard deviation σ is usually denoted by $\mathcal{N}\left(\mu, \sigma^2\right)$.

We provide formal definitions of the mean and the standard deviation of a random variable in Chapter 4. For now, you can just think of them as quantities that parametrize the Gaussian pdf.

It is not immediately obvious that the pdf of the Gaussian integrates to one. We establish this in the following lemma.

Lemma 2.3.11 (Proof in Section 2.7.3)**.** *The pdf of a Gaussian random variable integrates to one.*

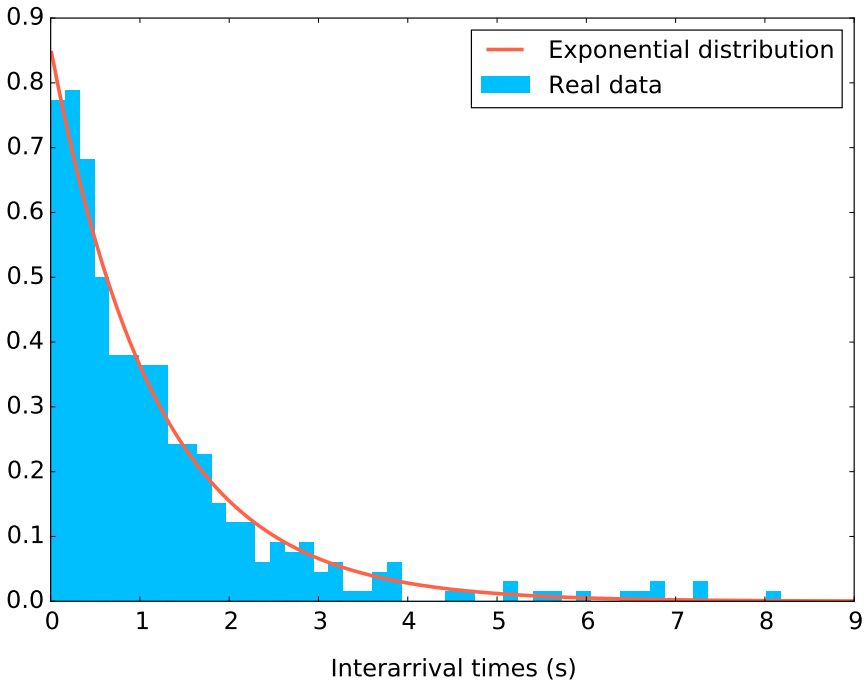


Figure 2.11: Histogram of inter-arrival times of calls at a call center in Israel (red) compared to its approximation by an exponential pdf.

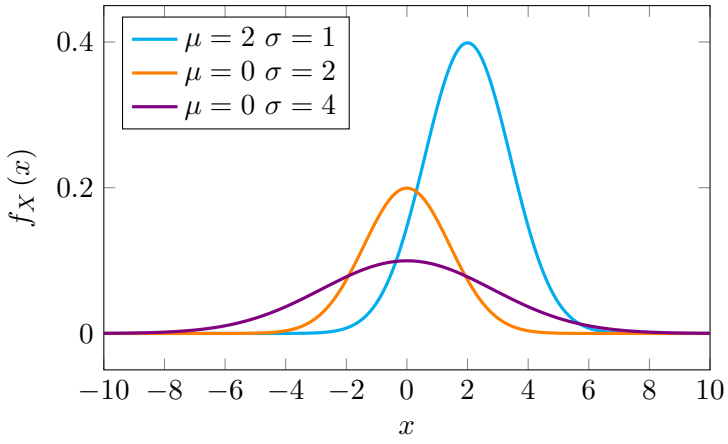


Figure 2.12: Gaussian random variable with different means and standard deviations.

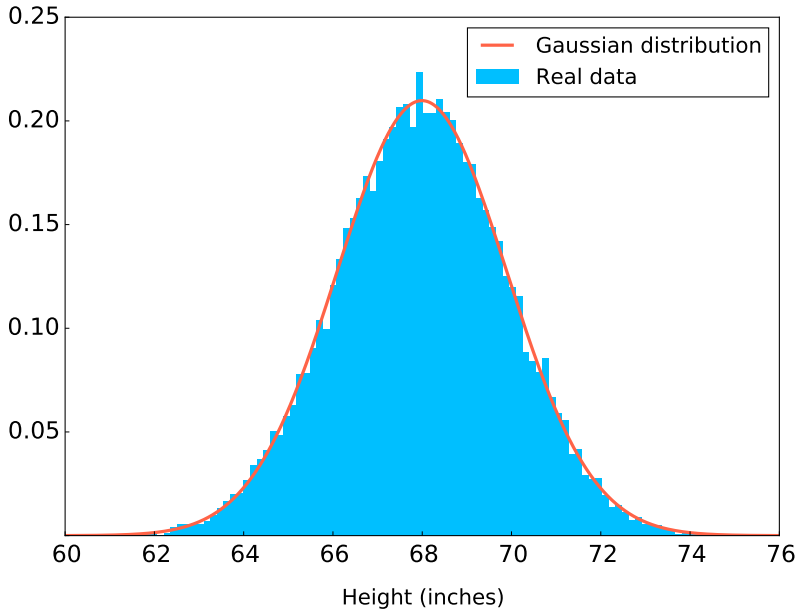


Figure 2.13: Histogram of heights in a population of 25,000 people (blue) and its approximation using a Gaussian distribution (orange).

Figure 2.12 shows the pdfs of two Gaussian random variables with different values of μ and σ. Figure 2.13 shows the histogram of the heights in a population of 25,000 people and how it is very well approximated by a Gaussian random variable[3].

An annoying feature of the Gaussian random variable is that its cdf does not have a closed form solution, in contrast to the uniform and exponential random variables. This complicates the task of determining the probability that a Gaussian random variable is in a certain interval. To tackle this problem we use the fact that if X is a Gaussian random variable with mean μ and standard deviation σ, then

$$U := \frac{X - \mu}{\sigma} \tag{2.51}$$

is a **standard** Gaussian random variable, which means that its mean is zero and its standard deviation equals one. See Lemma 2.5.1 for the proof. This allows us to express the probability of X being in an interval $[a, b]$ in terms of the cdf of a standard Gaussian, which we denote by Φ,

$$\mathrm{P}\left(X \in [a, b]\right) = \mathrm{P}\left(\frac{X - \mu}{\sigma} \in \left[\frac{a - \mu}{\sigma}, \frac{b - \mu}{\sigma}\right]\right) \tag{2.52}$$

$$= \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right). \tag{2.53}$$

As long as we can evaluate Φ, this formula allows us to deal with arbitrary Gaussian random variables. To evaluate Φ people used to resort to lists of tabulated values, compiled by computing the corresponding integrals numerically. Nowadays you can just use Matlab, WolframAlpha, SciPy, etc.

[3]The data is available here.

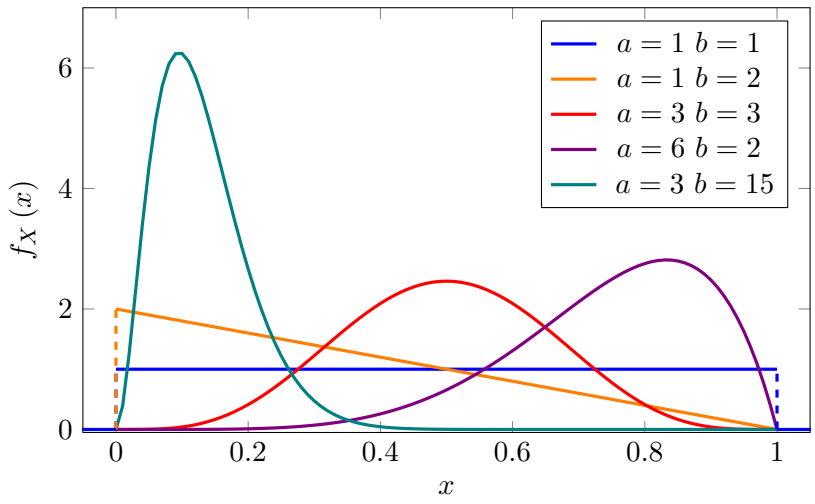


Figure 2.14: Pdfs of beta random variables with different values of the a and b parameters.

Beta

Beta distributions allow us to parametrize unimodal continuous distributions supported on the unit interval. This is useful in Bayesian statistics, as we discuss in Chapter 10.

Definition 2.3.12 (Beta distribution)**.** *The pdf of a beta distribution with parameters a and b is defined as*

$$f_\beta(\theta; a, b) := \begin{cases} \frac{\theta^{a-1}(1-\theta)^{b-1}}{\beta(a,b)}, & \text{if } 0 \leq \theta \leq 1, \\ 0 & \text{otherwise,} \end{cases} \tag{2.54}$$

where

$$\beta(a, b) := \int_u u^{a-1}(1-u)^{b-1}\, du. \tag{2.55}$$

$\beta(a, b)$ is a special function called the beta function or Euler integral of the first kind, which must be computed numerically. The uniform distribution is an example of a beta distribution (where $a = 1$ and $b = 1$). Figure 2.14 shows the pdf of several different beta distributions.

2.4 Conditioning on an event

In Section 1.2 we explain how to modify the probability measure of a probability space to incorporate the assumption that a certain event has occurred. In this section, we review this situation when random variables are involved. In particular, we consider a random variable X with a certain distribution represented by a pmf, cdf or pdf and explain how its distribution changes if we assume that $X \in \mathcal{S}$, for any set $\mathcal{S}$ belonging to the Borel σ-algebra (remember that this includes essentially any useful set you can think of).

If X is discrete with pmf p_X, the conditional pmf of X given $X \in \mathcal{S}$ is

$$p_{X|X\in\mathcal{S}}(x) := \mathrm{P}(X = x|X \in \mathcal{S}) \tag{2.56}$$

$$= \begin{cases} \frac{p_X(x)}{\sum_{s\in\mathcal{S}} p_X(s)} & \text{if } x \in \mathcal{S} \\ 0 \quad \text{otherwise.} \end{cases} \tag{2.57}$$

This is a valid pmf in the new probability space restricted to the event $\{X \in \mathcal{S}\}$.

Similarly if X is continuous with pdf f_X, the conditional cdf of X given the event $X \in \mathcal{S}$ is

$$F_{X|X\in\mathcal{S}}(x) := \mathrm{P}(X \leq x|X \in \mathcal{S}) \tag{2.58}$$

$$= \frac{\mathrm{P}(X \leq x, X \in \mathcal{S})}{\mathrm{P}(X \in \mathcal{S})} \tag{2.59}$$

$$= \frac{\int_{u\leq x, u\in\mathcal{S}} f_X(u)\,\mathrm{d}u}{\int_{u\in\mathcal{S}} f_X(u)\,\mathrm{d}u}, \tag{2.60}$$

again by the definition of conditional probability. One can check that this is a valid cdf in the new probability space. To obtain the conditional pdf we just differentiate this cdf,

$$f_{X|X\in\mathcal{S}}(x) := \frac{\mathrm{d}F_{X|X\in\mathcal{S}}(x)}{\mathrm{d}x}. \tag{2.61}$$

We now apply this ideas to show that the geometric and exponential random variables are memoryless.

Example 2.4.1 (Geometric random variables are memoryless)**.** We flip a coin repeatedly until we obtain heads, but pause after a couple of flips (which were tails). Let us assume that the flips are independent and have the same bias p (i.e. the probability of obtaining heads in every flip is p). What is the probability of obtaining heads in k more flips? Perhaps surprisingly, it is exactly the same as the probability of obtaining a heads after k flips from the beginning.

To establish this rigorously we compute the conditional pmf of a geometric random variable X conditioned on the event $\{X > k_0\}$ (i.e. the first k_0 were tails in our example). Applying (2.56) we have

$$p_{X|X>k_0}(k) = \frac{p_X(k)}{\sum_{m=k_0+1}^{\infty} p_X(m)} \tag{2.62}$$

$$= \frac{(1-p)^{k-1}\,p}{\sum_{m=k_0+1}^{\infty}(1-p)^{m-1}\,p} \tag{2.63}$$

$$= (1-p)^{k-k_0-1}\,p \tag{2.64}$$

if $k > k_0$ and zero otherwise. We have used the fact that the geometric series

$$\sum_{m=k_0+1}^{\infty} \alpha^m = \frac{\alpha^{k_0+1}}{1-\alpha} \tag{2.65}$$

for any $\alpha < 1$.

In the new probability space where the count starts at $k_0 + 1$ the conditional pmf is that of a geometric random variable with the same parameter as the original one. The first k_0 flips don't affect the future, once it is revealed that they were tails.

△

Example 2.4.2 (Exponential random variables are memoryless)**.** Let us assume that the inter-arrival times of your emails follow an exponential distribution (over intervals of several hours this is probably a good approximation, let us know if you check). You receive an email. The time until you receive your next email is exponentially distributed with a certain parameter λ. No email arrives in the next t_0 minutes. Surprisingly, the time from then until you receive your next email is again exponentially distributed with the same parameter, no matter the value of t_0. Just like geometric random variables, exponential random variables are memoryless.

Let us prove this rigorously. We compute the conditional cdf of an exponential random variable T with parameter λ conditioned on the event $\{T > t_0\}$– for an arbitrary $t_0 > 0$– by applying (2.60)

$$F_{T|T>t_0}(t) = \frac{\int_{t_0}^{t} f_T(u)\,\mathrm{d}u}{\int_{t_0}^{\infty} f_T(u)\,\mathrm{d}u} \tag{2.66}$$

$$= \frac{e^{-\lambda t} - e^{-\lambda t_0}}{-e^{-\lambda t_0}} \tag{2.67}$$

$$= 1 - e^{-\lambda(t-t_0)}. \tag{2.68}$$

Differentiating with respect to t yields an exponential pdf $f_{T|T>t_0}(t) = \lambda e^{-\lambda(t-t_0)}$ starting at t_0. △

2.5 Functions of random variables

Computing the distribution of a function of a random variable is often very useful in probabilistic modeling. For example, if we model the current in a circuit using a random variable X, we might be interested in the power $Y := rX^2$ dissipated across a resistor with deterministic resistance r. If we apply a deterministic function $g : \mathbb{R} \to \mathbb{R}$ to a random variable X, then the result $Y := g(X)$ is *not* a deterministic quantity. Recall that random variables are functions from a sample space Ω to $\mathbb{R}$. If X maps elements of Ω to $\mathbb{R}$, then so does Y since $Y(\omega) = g(X(\omega))$. This means that Y is also a random variable. In this section we explain how to characterize the distribution of Y when the distribution of X is known.

If X is discrete, then it is straightforward to compute the pmf of $g(X)$ from the pmf of X,

$$p_Y(y) = \mathrm{P}(Y = y) \tag{2.69}$$

$$= \mathrm{P}(g(X) = y) \tag{2.70}$$

$$= \sum_{\{x \mid g(x)=y\}} p_X(x). \tag{2.71}$$

If X is continuous, the procedure is more subtle. We first compute the cdf of Y by applying the definition,

$$F_Y(y) = \mathrm{P}(Y \leq y) \tag{2.72}$$

$$= \mathrm{P}(g(X) \leq y) \tag{2.73}$$

$$= \int_{\{x \mid g(x)\leq y\}} f_X(x)\,\mathrm{d}x, \tag{2.74}$$

where the last equality obviously only holds if X has a pdf. We can then obtain the pdf of Y from its cdf if it is differentiable. This idea can be used to prove a useful result about Gaussian random variables.

Lemma 2.5.1 (Gaussian random variable)**.** *If X is a Gaussian random variable with mean μ and standard deviation σ, then*

$$U := \frac{X - \mu}{\sigma} \tag{2.75}$$

is a standard Gaussian random variable.

Proof. We apply (2.74) to obtain

$$F_U(u) = \mathrm{P}\left(\frac{X-\mu}{\sigma} \leq u\right) \tag{2.76}$$

$$= \int_{(x-\mu)/\sigma \leq u} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, \mathrm{d}x \tag{2.77}$$

$$= \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} \, \mathrm{d}w \qquad \text{by the change of variables } w = \frac{x-\mu}{\sigma}. \tag{2.78}$$

Differentiating with respect to u yields

$$f_U(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}, \tag{2.79}$$

so U is indeed a standard Gaussian random variable. □

2.6 Generating random variables

Simulation is a fundamental tool in probabilistic modeling. Simulating the outcome of a model requires sampling from the random variables included in it. The most widespread strategy for generating samples from a random variable decouples the process into two steps:

1. Generating samples uniformly from the unit interval $[0, 1]$.
2. Transforming the uniform samples so that they have the desired distribution.

Here we focus on the second step, assuming that we have access to a random-number generator that produces independent samples following a uniform distribution in $[0, 1]$. The construction of good uniform random generators is an important problem, which is beyond the scope of these notes.

2.6.1 Sampling from a discrete distribution

Let X be a discrete random variable with pmf p_X and U a uniform random variable in $[0, 1]$. Our aim is to transform a sample from U so that it is distributed according to p_X. We denote the values that have nonzero probability under p_X by x_1, x_2, ...

For a fixed i, assume that we assign all samples of U within an interval of length $p_X(x_i)$ to x_i. Then the probability that a given sample from U is assigned to x_i is exactly $p_X(x_i)$!

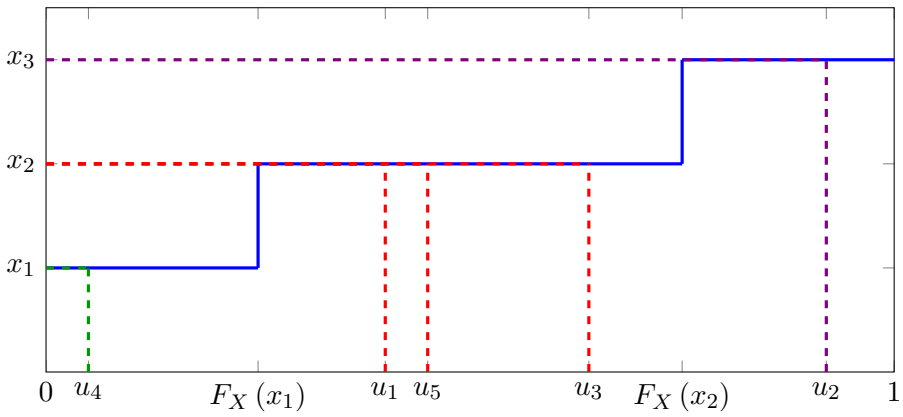


Figure 2.15: Illustration of the method to generate samples from an arbitrary discrete distribution described in Section 2.6.1. The cdf of a discrete random variable is shown in blue. The samples u_4 and u_2 from a uniform distribution are mapped to x_1 and x_3 respectively, whereas u_1, u_3 and u_5 are mapped to x_3.

Very conveniently, the unit interval can be partitioned into intervals of length $p_X(x_i)$. We can consequently generate X by sampling from U and setting

$$X = \begin{cases} x_1 & \text{if } 0 \leq U \leq p_X(x_1), \\ x_2 & \text{if } p_X(x_1) \leq U \leq p_X(x_1) + p_X(x_2), \\ \dots & \\ x_i & \text{if } \sum_{j=1}^{i-1} p_X(x_j) \leq U \leq \sum_{j=1}^{i} p_X(x_j), \\ \dots & \end{cases} \tag{2.80}$$

Recall that the cdf of a discrete random variable equals

$$F_X(x) = \mathrm{P}(X \leq x) \tag{2.81}$$

$$= \sum_{x_i \leq x} p_X(x_i), \tag{2.82}$$

so our algorithm boils down to obtaining a sample u from U and then outputting the x_i such that $F_X(x_{i-1}) \leq u \leq F_X(x_i)$. This is illustrated in Figure 2.15.

2.6.2 Inverse-transform sampling

Inverse-transform sampling makes it possible to sample from an arbitrary distribution with a known cdf by applying a deterministic transformation to uniform samples. Intuitively, we can interpret it as a generalization of the method in Section 2.6.1 to continuous distributions.

Algorithm 2.6.1 (Inverse-transform sampling)**.** *Let X be a continuous random variable with cdf F_X and U a random variable that is uniformly distributed in $[0, 1]$ and independent of X.*

1. *Obtain a sample u of U.*
2. *Set $x := F_X^{-1}(u)$.*

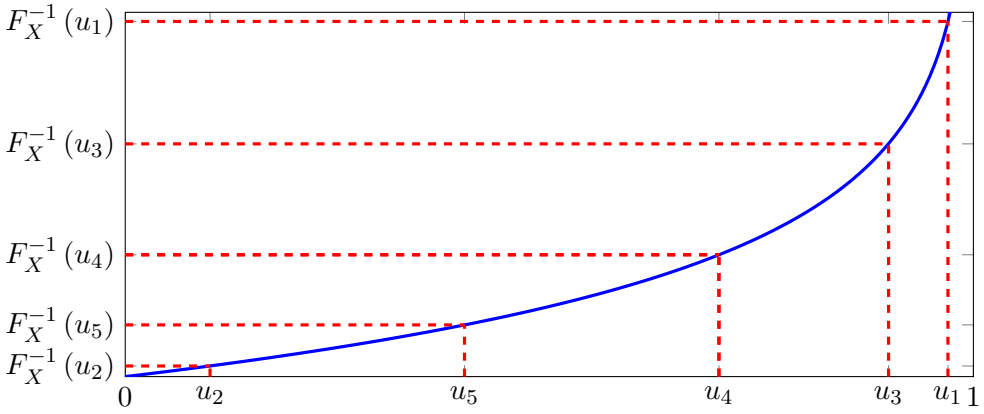


Figure 2.16: Samples from an exponential distribution with parameter $\lambda = 1$ obtained by inverse-transform sampling as described in Example 2.6.4. The samples $u_1, \ldots, u_5$ are generated from a uniform distribution.

The careful reader will point out that F_X may not be invertible at every point. To avoid this problem we define the generalized inverse of the cdf as

$$F_X^{-1}(u) := \min_x \{F_X(x) = u\}. \tag{2.83}$$

The function is well defined because all cdfs are non-decreasing, so F_X is equal to a constant c in any interval $[x_1, x_2]$ where it is not invertible.

We now prove that Algorithm 2.6.1 works.

Theorem 2.6.2 (Inverse-transform sampling works)**.** *The distribution of $Y = F_X^{-1}(U)$ is the same as the distribution of X.*

Proof. We just need to show that the cdf of Y is equal to F_X. We have

$$\begin{aligned} F_Y(y) &= \mathrm{P}(Y \leq y) && (2.84)\\ &= \mathrm{P}\left(F_X^{-1}(U) \leq y\right) && (2.85)\\ &= \mathrm{P}(U \leq F_X(y)) && (2.86)\\ &= \int_{u=0}^{F_X(y)} \mathrm{d}u && (2.87)\\ &= F_X(y), && (2.88) \end{aligned}$$

where in step (2.86) we have to take into account that we are using the generalized inverse of the cdf. This is resolved by the following lemma proved in Section 2.7.4.

Lemma 2.6.3. *The events $\left\{F_X^{-1}(U) \leq y\right\}$ and $\{U \leq F_X(y)\}$ are equivalent.*

□

Example 2.6.4 (Sampling from an exponential distribution)**.** Let X be an exponential random variable with parameter λ. Its cdf $F_X(x) := 1 - e^{-\lambda x}$ is invertible in $[0, \infty]$. Its inverse equals

$$F_X^{-1}(u) = \frac{1}{\lambda} \log\left(\frac{1}{1-u}\right). \tag{2.89}$$

$F_X^{-1}(U)$ is an exponential random variable with parameter λ by Theorem 2.6.2. Figure 2.16 shows how the samples of U are transformed into samples of X.

△

2.7 Proofs

2.7.1 Proof of Lemma 2.2.9

For any fixed constants c_1 and c_2

$$\lim_{n\to\infty} \frac{n - c_1}{n - c_2} = 1, \tag{2.90}$$

so that

$$\lim_{n\to\infty} \frac{n!}{(n-k)!\,(n-\lambda)^k} = \frac{n}{n-\lambda} \cdot \frac{n-1}{n-\lambda} \cdots \frac{n-k+1}{n-\lambda} = 1. \tag{2.91}$$

The result follows from the following basic calculus identity:

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}. \tag{2.92}$$

2.7.2 Proof of Lemma 2.3.2

To establish (2.31)

$$\lim_{x\to-\infty} F_X(x) = 1 - \lim_{x\to-\infty} \mathrm{P}(X > x) \tag{2.93}$$

$$= 1 - \mathrm{P}(X > 0) - \lim_{n\to\infty} \sum_{i=0}^{n} \mathrm{P}(-i \geq X > -(i+1)) \tag{2.94}$$

$$= 1 - \mathrm{P}\left(\lim_{n\to\infty} \{X > 0\} \cup \cup_{i=0}^{n} \{-i \geq X > -(i+1)\}\right) \tag{2.95}$$

$$= 1 - \mathrm{P}(\Omega) = 0. \tag{2.96}$$

The proof of (2.32) follows from this result. Let $Y = -X$, then

$$\lim_{x\to\infty} F_X(x) = \lim_{x\to\infty} \mathrm{P}(X \leq x) \tag{2.97}$$

$$= 1 - \lim_{x\to\infty} \mathrm{P}(X > x) \tag{2.98}$$

$$= 1 - \lim_{x\to-\infty} \mathrm{P}(-X < x) \tag{2.99}$$

$$= 1 - \lim_{x\to-\infty} F_Y(x) = 1 \quad \text{by (2.32)}. \tag{2.100}$$

Finally, (2.33) holds because $\{X \leq a\} \subseteq \{X \leq b\}$.

2.7.3 Proof of Lemma 2.3.11

The result is a consequence of the following lemma.

Lemma 2.7.1.

$$\int_{-\infty}^{\infty} e^{-t^2}\, dt = \sqrt{\pi}. \tag{2.101}$$

Proof. Let us define

$$I = \int_{-\infty}^{\infty} e^{-x^2} \mathrm{d}x. \tag{2.102}$$

Now taking the square and changing to polar coordinates,

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} e^{-x^2} \mathrm{d}x \int_{-\infty}^{\infty} e^{-y^2} \mathrm{d}y && (2.103)\\ &= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} e^{-\left(x^2+y^2\right)} \mathrm{d}x \mathrm{d}y && (2.104)\\ &= \int_{\theta=0}^{2\pi} \int_{r=-\infty}^{\infty} r e^{-\left(r^2\right)} \mathrm{d}\theta \mathrm{d}r && (2.105)\\ &= \pi e^{-\left(r^2\right)}]_0^{\infty} = \pi. && (2.106) \end{aligned}$$

□

To complete the proof we use the change of variables $t = (x-\mu)/\sqrt{2}\sigma$.

2.7.4 Proof of Lemma 2.6.3

$\underline{\left\{F_X^{-1}(U) \leq y\right\} \text{ implies } \{U \leq F_X(y)\}}$

Assume that $U > F_X(y)$, then for all x, such that $F_X(x) = U$, $x > y$ because the cdf is nondecreasing. In particular $\min_x \{F_X(x) = U\} > y$.

$\underline{\{U \leq F_X(y)\} \text{ implies } \left\{F_X^{-1}(U) \leq y\right\}}$

Assume that $\min_x \{F_X(x) = U\} > y$, then $U > F_X(y)$ because the cdf is nondecreasing. The inequality is strict because $U = F_X(y)$ would imply that y belongs to $\{F_X(x) = U\}$, which cannot be the case as we are assuming that it is smaller than the minimum of that set.

Chapter 3

Multivariate Random Variables

Probabilistic models usually include multiple uncertain numerical quantities. In this chapter we describe how to specify random variables to represent such quantities and their interactions. In some occasions, it will make sense to group these random variables as **random vectors**, which we write using uppercase letters with an arrow on top: $\vec{X}$. Realizations of these random vectors are denoted with lowercase letters: $\vec{x}$.

3.1 Discrete random variables

Recall that discrete random variables are numerical quantities that take either finite or countably infinite values. In this section we explain how to manipulate multiple discrete random variables that share a common probability space.

3.1.1 Joint probability mass function

If several discrete random variables are defined on the same probability space, we specify their probabilistic behavior through their **joint probability mass function**, which is the probability that each variable takes a particular value.

Definition 3.1.1 (Joint probability mass function)**.** *Let $X : \Omega \to R_X$ and $Y : \Omega \to R_Y$ be discrete random variables (R_X and R_Y are discrete sets) on the same probability space $(\Omega, \mathcal{F}, \mathrm{P})$. The joint pmf of X and Y is defined as*

$$p_{X,Y}(x, y) := \mathrm{P}(X = x, Y = y) \quad . \tag{3.1}$$

In words, $p_{X,Y}(x, y)$ is the probability of X and Y being equal to x and y respectively.

Similarly, the joint pmf of a discrete random vector of dimension n

$$\vec{X} := \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_n \end{bmatrix} \tag{3.2}$$

with entries $X_i : \Omega \to R_{X_i}$ ($R_1, \ldots, R_n$ are all discrete sets) belonging to the same probability space is defined as

$$p_{\vec{X}}(\vec{x}) := \mathrm{P}(X_1 = \vec{x}_1, X_2 = \vec{x}_2, \ldots, X_n = \vec{x}_n). \tag{3.3}$$

As in the case of the pmf of a single random variable, the joint pmf is a valid probability measure if we consider a probability space where the sample space is $R_X \times R_Y$[1] (or $R_{X_1} \times R_{X_2} \cdots \times R_{X_n}$ in the case of a random vector) and the σ-algebra is just the power set of the sample space. This implies that the joint pmf *completely* characterizes the random variables or the random vector, we don't need to worry about the underlying probability space.

By the definition of probability measure, the joint pmf must be nonnegative and its sum over all its possible arguments must equal one,

$$p_{X,Y}(x,y) \geq 0 \quad \text{for any } x \in R_X, y \in R_Y, \tag{3.4}$$

$$\sum_{x \in R_X} \sum_{y \in R_Y} p_{X,Y}(x,y) = 1. \tag{3.5}$$

By the Law of Total Probability, the joint pmf allows us to obtain the probability of X and Y belonging to any set $\mathcal{S} \subseteq R_X \times R_Y$,

$$\begin{aligned} \mathrm{P}((X,Y) \in \mathcal{S}) &= \mathrm{P}\left(\cup_{(x,y)\in\mathcal{S}} \{X = x, Y = y\}\right) \quad \text{(union of disjoint events)} && (3.6)\\ &= \sum_{(x,y)\in\mathcal{S}} \mathrm{P}(X = x, Y = y) && (3.7)\\ &= \sum_{(x,y)\in\mathcal{S}} p_{X,Y}(x,y). && (3.8) \end{aligned}$$

These properties also hold for random vectors (and groups of more than two random variables). For any random vector $\vec{X}$,

$$p_{\vec{X}}(\vec{x}) \geq 0, \tag{3.9}$$

$$\sum_{\vec{x}_1 \in R_1} \sum_{\vec{x}_2 \in R_2} \cdots \sum_{\vec{x}_n \in R_n} p_{\vec{X}}(\vec{x}) = 1. \tag{3.10}$$

The probability that $\vec{X}$ belongs to a discrete set $\mathcal{S} \subseteq \mathbb{R}^n$ is given by

$$\mathrm{P}\left(\vec{X} \in \mathcal{S}\right) = \sum_{\vec{x}\in\mathcal{S}} p_{\vec{X}}(\vec{x}). \tag{3.11}$$

3.1.2 Marginalization

Assume we have access to the joint pmf of several random variables in a certain probability space, but we are only interested in the behavior of one of them. To compute the value of its pmf for a particular value, we fix that value and sum over the remaining random variables. Indeed, by the Law of Total Probability

$$\begin{aligned} p_X(x) &= \mathrm{P}(X = x) && (3.12)\\ &= \mathrm{P}\left(\cup_{y\in R_Y} \{X = x, Y = y\}\right) \quad \text{(union of disjoint events)} && (3.13)\\ &= \sum_{y\in R_Y} \mathrm{P}(X = x, Y = y) && (3.14)\\ &= \sum_{y\in R_Y} p_{X,Y}(x,y). && (3.15) \end{aligned}$$

[1] This is the Cartesian product of the two sets, defined in Section A.2, which contains all possible pairs (x, y) where $x \in R_X$ and $y \in R_y$.

When the joint pmf involves more than two random variables the argument is exactly the same. This is called **marginalizing** over the other random variables. In this context, the pmf of a single random variable is called its **marginal pmf**. Table 3.1 shows an example of a joint pmf and the corresponding marginal pmfs.

If we are interested in computing the joint pmf of several entries in a random vector, instead of just one, the marginalization process is essentially the same. The pmf is again obtained by summing over the rest of the entries. Let $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$ be a subset of $m < n$ entries of an n-dimensional random vector $\vec{X}$ and $\vec{X}_{\mathcal{I}}$ the corresponding random subvector. To compute the joint pmf of $\vec{X}_{\mathcal{I}}$ we sum over all the entries that are not in $\mathcal{I}$, which we denote by $\{j_1, j_2, \ldots, j_{n-m}\} := \{1, 2, \ldots, n\} / \mathcal{I}$

$$p_{\vec{X}_{\mathcal{I}}}(\vec{x}_{\mathcal{I}}) = \sum_{\vec{x}_{j_1} \in R_{j_1}} \sum_{\vec{x}_{j_2} \in R_{j_2}} \cdots \sum_{\vec{x}_{j_{n-m}} \in R_{j_{n-m}}} p_{\vec{X}}(\vec{x}). \tag{3.16}$$

3.1.3 Conditional distributions

Conditional probabilities allow us to update our uncertainty about the quantities in a probabilistic model when new information is revealed. The conditional distribution of a random variable specifies the behavior of the random variable when we assume that other random variables in the probability space take a fixed value.

Definition 3.1.2 (Conditional probability mass function)**.** *The conditional probability mass function of Y given X, where X and Y are discrete random variables defined on the same probability space, is given by*

$$p_{Y|X}(y|x) = \mathrm{P}(Y = y | X = x) \tag{3.17}$$

$$= \frac{p_{X,Y}(x, y)}{p_X(x)} \qquad \text{if } p_X(x) > 0 \tag{3.18}$$

and is undefined otherwise.

The conditional pmf $p_{X|Y}(\cdot|y)$ characterizes our uncertainty about X conditioned on the event $\{Y = y\}$. This object is a valid pmf of X, so that if R_X is the range of X

$$\sum_{x \in R_X} p_{X|Y}(x|y) = 1 \tag{3.19}$$

for any y for which it is well defined. However, it is *not* a pmf for Y. In particular, there is no reason for $\sum_{y \in R_Y} p_{X|Y}(x|y)$ to add up to one!

We now define the joint conditional pmf of several random variables (equivalently of a subvector of a random vector) given other random variables (or entries of the random vector).

Definition 3.1.3 (Conditional pmf)**.** *The conditional pmf of a discrete random subvector $\vec{X}_{\mathcal{I}}$, $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$, given another subvector $\vec{X}_{\mathcal{J}}$ is*

$$p_{\vec{X}_{\mathcal{I}}|\vec{X}_{\mathcal{J}}}(\vec{x}_{\mathcal{I}}|\vec{x}_{\mathcal{J}}) := \frac{p_{\vec{X}}(\vec{x})}{p_{\vec{X}_{\mathcal{J}}}(\vec{x}_{\mathcal{J}})}, \tag{3.20}$$

where $\{j_1, j_2, \ldots, j_{n-m}\} := \{1, 2, \ldots, n\} / \mathcal{I}$.

R

L \ $p_{L,R}$	0	1
0	$\frac{14}{20}$	$\frac{1}{20}$
1	$\frac{2}{20}$	$\frac{3}{20}$

p_L
$\frac{15}{20}$
$\frac{5}{20}$

$p_{L\|R}(\cdot\|0)$
$\frac{7}{8}$
$\frac{1}{8}$

$p_{L\|R}(\cdot\|1)$
$\frac{1}{4}$
$\frac{3}{4}$

p_R	$\frac{16}{20}$	$\frac{4}{20}$

$p_{R\|L}(\cdot\|0)$	$\frac{14}{15}$	$\frac{1}{15}$

$p_{R\|L}(\cdot\|1)$	$\frac{2}{5}$	$\frac{3}{5}$

Table 3.1: Joint, marginal and conditional pmfs of the random variables L and R defined in Example 3.1.5.

The conditional pmfs $p_{Y|X}(\cdot|x)$ and $p_{\vec{X}_{\mathcal{I}}|\vec{X}_{\mathcal{J}}}(\cdot|\vec{x}_{\mathcal{J}})$ are valid pmfs in the probability space where $X = x$ or $\vec{X}_{\mathcal{J}} = \vec{x}_{\mathcal{J}}$ respectively. For instance, they must be nonnegative and add up to one.

From the definition of conditional pmfs we derive a chain rule for discrete random variables and vectors.

Lemma 3.1.4 (Chain rule for discrete random variables and vectors)**.**

$$p_{X,Y}(x,y) = p_X(x)\, p_{Y|X}(y|x), \tag{3.21}$$

$$p_{\vec{X}}(\vec{x}) = p_{X_1}(\vec{x}_1)\, p_{X_2|X_1}(\vec{x}_2|\vec{x}_1) \dots p_{X_n|X_1,\dots,X_{n-1}}(\vec{x}_n|\vec{x}_1,\dots,\vec{x}_{n-1}) \tag{3.22}$$

$$= \prod_{i=1}^{n} p_{X_i|\vec{X}_{\{1,\dots,i-1\}}}\left(\vec{x}_i|\vec{x}_{\{1,\dots,i-1\}}\right), \tag{3.23}$$

where the order of indices in the random vector is arbitrary (any order works).

The following example illustrates the definitions of marginal and conditional pmfs.

Example 3.1.5 (Flights and rains (continued))**.** Within the probability space described in Example 1.2.1 we define a random variable

$$L = \begin{cases} 1 & \text{if plane is late,} \\ 0 & \text{otherwise,} \end{cases} \tag{3.24}$$

to represent whether the plane is late or not. Similarly,

$$R = \begin{cases} 1 & \text{it rains,} \\ 0 & \text{otherwise,} \end{cases} \tag{3.25}$$

represents whether it rains or not. Equivalently, these random variables are just the indicators $R = 1_{\text{rain}}$ and $L = 1_{\text{late}}$. Table 3.1 shows the joint, marginal and conditional pmfs of L and R.

△

3.2 Continuous random variables

Continuous random variables allow us to model continuous quantities without having to worry about discretization. In exchange, the mathematical tools to manipulate them are somewhat more complicated than in the discrete case.

3.2.1 Joint cdf and joint pdf

As in the case of univariate continuous random variables, we characterize the behavior of several continuous random variables defined on the same probability space through the probability that they belong to Borel sets (or equivalently unions of intervals). In this case we are considering multidimensional Borel sets, which are Cartesian products of one-dimensional Borel sets. Multidimensional Borel sets can be represented as unions of multidimensional intervals or hyperrectangles (defined as Cartesian products of one-dimensional intervals). The **joint cdf** compiles the probability that the random variables belong to the Cartesian product of intervals of the form $(-\infty, r]$ for every $r \in \mathbb{R}$.

Definition 3.2.1 (Joint cumulative distribution function)**.** *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $X, Y : \Omega \to \mathbb{R}$ random variables. The **joint cdf** of X and Y is defined as*

$$F_{X,Y}(x, y) := \mathrm{P}(X \leq x, Y \leq y). \tag{3.26}$$

In words, $F_{X,Y}(x, y)$ is the probability of X and Y being smaller than x and y respectively.

Let $\vec{X} : \Omega \to \mathbb{R}^n$ be a random vector of dimension n on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. The joint cdf of $\vec{X}$ is defined as

$$F_{\vec{X}}(\vec{x}) := \mathrm{P}\left(\vec{X}_1 \leq \vec{x}_1, \vec{X}_2 \leq \vec{x}_2, \ldots, \vec{X}_n \leq \vec{x}_n\right). \tag{3.27}$$

In words, $F_{\vec{X}}(\vec{x})$ is the probability that $\vec{X}_i \leq \vec{x}_i$ for all $i = 1, 2, \ldots, n$.

We now record some properties of the joint cdf.

Lemma 3.2.2 (Properties of the joint cdf)**.**

$$\lim_{x \to -\infty} F_{X,Y}(x, y) = 0, \tag{3.28}$$

$$\lim_{y \to -\infty} F_{X,Y}(x, y) = 0, \tag{3.29}$$

$$\lim_{x \to \infty, y \to \infty} F_{X,Y}(x, y) = 1, \tag{3.30}$$

$$F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2) \quad \text{if } x_2 \geq x_1, y_2 \geq y_1, \quad \text{i.e. } F_{X,Y} \text{ is nondecreasing.} \tag{3.31}$$

Proof. The proof follows along the same lines as that of Lemma 2.3.2. □

The joint cdf completely specifies the behavior of the corresponding random variables. Indeed, we can decompose any Borel set into a union of disjoint n-dimensional intervals and compute their probability by evaluating the joint cdf. Let us illustrate this for the bivariate case:

$$\begin{aligned} \mathrm{P}\left(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2\right) &= \mathrm{P}\left(\{X \leq x_2, Y \leq y_2\} \cap \{X > x_1\} \cap \{Y > y_1\}\right) && (3.32)\\ &= \mathrm{P}\left(X \leq x_2, Y \leq y_2\right) - \mathrm{P}\left(X \leq x_1, Y \leq y_2\right) && (3.33)\\ &\quad - \mathrm{P}\left(X \leq x_2, Y \leq y_1\right) + \mathrm{P}\left(X \leq x_1, Y \leq y_1\right) && (3.34)\\ &= F_{X,Y}\left(x_2, y_2\right) - F_{X,Y}\left(x_1, y_2\right) - F_{X,Y}\left(x_2, y_1\right) + F_{X,Y}\left(x_1, y_1\right). \end{aligned}$$

This means that, as in the univariate case, to define a random vector or a group of random variables all we need to do is define their joint cdf. We don't have to worry about the underlying probability space.

If the joint cdf is differentiable, we can differentiate it to obtain the **joint probability density function** of X and Y. As in the case of univariate random variables, this is often a more convenient way of specifying the joint distribution.

Definition 3.2.3 (Joint probability density function)**.** *If the joint cdf of two random variables X, Y is differentiable, then their joint pdf is defined as*

$$f_{X,Y}\left(x, y\right) := \frac{\partial^2 F_{X,Y}\left(x, y\right)}{\partial x \partial y}. \tag{3.35}$$

If the joint cdf of a random vector $\vec{X}$ is differentiable, then its joint pdf is defined as

$$f_{\vec{X}}\left(\vec{x}\right) := \frac{\partial^n F_{\vec{X}}\left(\vec{x}\right)}{\partial \vec{x}_1 \, \partial \vec{x}_2 \cdots \partial \vec{x}_n}. \tag{3.36}$$

The joint pdf should be understood as an n-dimensional density, *not* as a probability (for instance, it can be larger than one). In the two-dimensional case,

$$\lim_{\Delta_x \to 0, \Delta_y \to 0} \mathrm{P}\left(x \leq X \leq x + \Delta_x, y \leq Y \leq y + \Delta_y\right) = f_{X,Y}\left(x, y\right) \Delta_x \Delta_y. \tag{3.37}$$

Due to the monotonicity of joint cdfs in every variable, joint pmfs are always nonnegative.

The joint pdf of X and Y allows us to compute the probability of any Borel set $\mathcal{S} \subseteq \mathbb{R}^2$ by integrating over $\mathcal{S}$

$$\mathrm{P}\left((X, Y) \in \mathcal{S}\right) = \int_{(x,y) \in \mathcal{S}} f_{X,Y}\left(x, y\right) \, \mathrm{d}x \, \mathrm{d}y. \tag{3.38}$$

Similarly, the joint pdf of an n-dimensional random vector $\vec{X}$ allows to compute the probability that $\vec{X}$ belongs to a set Borel set $\mathcal{S} \subseteq \mathbb{R}^n$,

$$\mathrm{P}\left(\vec{X} \in \mathcal{S}\right) = \int_{\vec{x} \in \mathcal{S}} f_{\vec{X}}\left(\vec{x}\right) \, \mathrm{d}\vec{x}. \tag{3.39}$$

In particular, if we integrate a joint pdf over the whole space $\mathbb{R}^n$, then it must integrate to one by the Law of Total Probability.

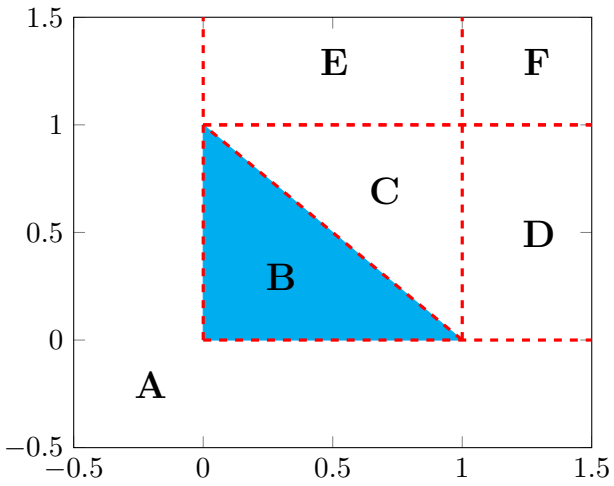


Figure 3.1: Triangle lake in Example 3.2.12.

Example 3.2.4 (Triangle lake)**.** A biologist is tracking an otter that lives in a lake. She decides to model the location of the otter probabilistically. The lake happens to be triangular as shown in Figure 3.1, so that we can represent it by the set

$$\text{Lake} := \{\vec{x} \mid \vec{x}_1 \geq 0, \vec{x}_2 \geq 0, \vec{x}_1 + \vec{x}_2 \leq 1\}. \tag{3.40}$$

The biologist has no idea where the otter is, so she models the position as a random vector $\vec{X}$ which is uniformly distributed over the lake. In other words, the joint pdf of $\vec{X}$ is constant,

$$f_{\vec{X}}(\vec{x}) = \begin{cases} c & \text{if } \vec{x} \in \text{Lake}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.41}$$

To find the normalizing constant c we use the fact that to be a valid joint pdf $f_{\vec{X}}$ should integrate to 1.

$$\int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{\infty} c \, \mathrm{d}x_1 \, \mathrm{d}x_2 = \int_{x_2=0}^{1} \int_{x_1=0}^{1-x_2} c \, \mathrm{d}x_1 \, \mathrm{d}x_2 \tag{3.42}$$

$$= c \int_{x_2=0}^{1} (1 - x_2) \, \mathrm{d}x_2 \tag{3.43}$$

$$= \frac{c}{2} = 1, \tag{3.44}$$

so $c = 2$.

We now compute the cdf of $\vec{X}$. $F_{\vec{X}}(\vec{x})$ represents the probability that the otter is southwest of the point $\vec{x}$. Computing the joint cdf requires dividing the range into the sets shown in Figure 3.1 and integrating the joint pdf. If $\vec{x} \in A$ then $F_{\vec{X}}(\vec{x}) = 0$ because $\mathrm{P}\left(\vec{X} \leq \vec{x}\right) = 0$. If $(\vec{x}) \in B$,

$$F_{\vec{X}}(\vec{x}) = \int_{u=0}^{\vec{x}_2} \int_{v=0}^{\vec{x}_1} 2 \, \mathrm{d}v \, \mathrm{d}u = 2\vec{x}_1\vec{x}_2. \tag{3.45}$$

If $\vec{x} \in C$,

$$F_{\vec{X}}(\vec{x}) = \int_{u=0}^{1-\vec{x}_1} \int_{v=0}^{\vec{x}_1} 2\,\mathrm{d}v\,\mathrm{d}u + \int_{u=1-\vec{x}_1}^{\vec{x}_2} \int_{v=0}^{1-u} 2\,\mathrm{d}v\,\mathrm{d}u = 2\vec{x}_1 + 2\vec{x}_2 - \vec{x}_2^2 - \vec{x}_1^2 - 1. \tag{3.46}$$

If $\vec{x} \in D$,

$$F_{\vec{X}}(\vec{x}) = \mathrm{P}\left(\vec{X}_1 \leq \vec{x}_1, \vec{X}_2 \leq \vec{x}_2\right) = \mathrm{P}\left(\vec{X}_1 \leq 1, \vec{X}_2 \leq \vec{x}_2\right) = F_{\vec{X}}(1, \vec{x}_2) = 2\vec{x}_2 - \vec{x}_2^2, \tag{3.47}$$

where the last step follows from (3.46). Exchanging $\vec{x}_1$ and $\vec{x}_2$, we obtain $F_{\vec{X}}(\vec{x}) = 2\vec{x}_1 - \vec{x}_1^2$ for $\vec{x} \in E$ by the same reasoning. Finally, for $\vec{x} \in F$ $F_{\vec{X}}(\vec{x}) = 1$ because $\mathrm{P}\left(\vec{X}_1 \leq x_1, \vec{X}_2 \leq x_2\right) = 1$. Putting everything together,

$$F_{\vec{X}}(\vec{x}) = \begin{cases} 0 & \text{if } \vec{x}_1 < 0 \text{ or } \vec{x}_2 < 0, \\ 2\vec{x}_1\vec{x}_2, & \text{if } \vec{x}_1 \geq 0, \vec{x}_2 \geq 0, \vec{x}_1 + \vec{x}_2 \leq 1, \\ 2\vec{x}_1 + 2\vec{x}_2 - \vec{x}_2^2 - \vec{x}_1^2 - 1, & \text{if } \vec{x}_1 \leq 1, \vec{x}_2 \leq 1, \vec{x}_1 + \vec{x}_2 \geq 1, \\ 2\vec{x}_2 - \vec{x}_2^2, & \text{if } \vec{x}_1 \geq 1, 0 \leq \vec{x}_2 \leq 1, \\ 2\vec{x}_1 - \vec{x}_1^2, & \text{if } 0 \leq \vec{x}_1 \leq 1, \vec{x}_2 \geq 1, \\ 1, & \text{if } \vec{x}_1 \geq 1, \vec{x}_2 \geq 1. \end{cases} \tag{3.48}$$

△

3.2.2 Marginalization

We now discuss how to characterize the marginal distributions of individual random variables from a joint cdf or a joint pdf. Consider the joint cdf $F_{X,Y}(x, y)$. When $x \to \infty$ the limit of $F_{X,Y}(x, y)$ is by definition the probability of Y being smaller than y, which is precisely the marginal cdf of Y. More formally,

$$\lim_{x\to\infty} F_{X,Y}(x, y) = \lim_{n\to\infty} \mathrm{P}\left(\cup_{i=1}^{n} \{X \leq i, Y \leq y\}\right) \tag{3.49}$$

$$= \mathrm{P}\left(\lim_{n\to\infty} \{X \leq n, Y \leq y\}\right) \tag{3.50}$$

$$= \mathrm{P}(Y \leq y) \tag{3.51}$$

$$= F_Y(y). \tag{3.52}$$

If the random variables have a joint pdf, we can also compute the marginal cdf by integrating over x

$$F_Y(y) = \mathrm{P}(Y \leq y) \tag{3.53}$$

$$= \int_{u=-\infty}^{y} \int_{x=-\infty}^{\infty} f_{X,Y}(x, u)\,\mathrm{d}x\,\mathrm{d}y. \tag{3.54}$$

Differentiating the latter equation with respect to y, we obtain the marginal pdf of Y

$$f_Y(y) = \int_{x=-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}x. \tag{3.55}$$

Similarly, the marginal pdf of a subvector $\vec{X}_{\mathcal{I}}$ of a random vector $\vec{X}$ indexed by $\mathcal{I} := \{i_1, i_2, \ldots, i_m\}$ is obtained by integrating over the rest of the components $\{j_1, j_2, \ldots, j_{n-m}\} := \{1, 2, \ldots, n\} / \mathcal{I}$,

$$f_{\vec{X}_{\mathcal{I}}}(\vec{x}_{\mathcal{I}}) = \int_{\vec{x}_{j_1}} \int_{\vec{x}_{j_2}} \cdots \int_{\vec{x}_{j_{n-m}}} f_{\vec{X}}(\vec{x}) \, \mathrm{d}\vec{x}_{j_1} \, \mathrm{d}\vec{x}_{j_2} \cdots \mathrm{d}\vec{x}_{j_{n-m}}. \tag{3.56}$$

Example 3.2.5 (Triangle lake (continued))**.** The biologist is interested in the probability that the otter is south of x_1. This information is encoded in the cdf of the random vector, we just need to take the limit when $x_2 \to \infty$ to marginalize over x_2.

$$F_{X_1}(x_1) = \begin{cases} 0 & \text{if } x_1 < 0, \\ 2x_1 - x_1^2 & \text{if } 0 \leq x_1 \leq 1, \\ 1 & \text{if } x_1 \geq 1. \end{cases} \tag{3.57}$$

To obtain the marginal pdf of X_1, which represents the latitude of the otter's position, we differentiate the marginal cdf

$$f_{X_1}(x_1) = \frac{\mathrm{d}F_{X_1}(x_1)}{\mathrm{d}x_1} = \begin{cases} 2(1 - x_1) & \text{if } 0 \leq x_1 \leq 1, \\ 0, & \text{otherwise.} \end{cases}. \tag{3.58}$$

Alternatively, we could have integrated the joint uniform pdf over x_2 (we encourage you to check that the result is the same).

△

3.2.3 Conditional distributions

In this section we discuss how to obtain the conditional distribution of a random variable given information about other random variables in the probability space. To begin with, we consider the case of two random variables. As in the case of univariate distributions, we can define the joint cdf and pdf of two random variables given events of the form $\{(X, Y) \in \mathcal{S}\}$ for any Borel set in $\mathbb{R}^2$ by applying the definition of conditional probability.

Definition 3.2.6 (Joint conditional cdf and pdf given an event)**.** *Let X, Y be random variables with joint pdf $f_{X,Y}$ and let $\mathcal{S} \subseteq \mathbb{R}^2$ be any Borel set with nonzero probability, the conditional cdf and pdf of X and Y given the event $(X, Y) \in \mathcal{S}$ is defined as*

$$F_{X,Y|(X,Y)\in\mathcal{S}}(x, y) := \mathrm{P}(X \leq x, Y \leq y | (X, Y) \in \mathcal{S}) \tag{3.59}$$

$$= \frac{\mathrm{P}(X \leq x, Y \leq y, (X, Y) \in \mathcal{S})}{\mathrm{P}((X, Y) \in \mathcal{S})} \tag{3.60}$$

$$= \frac{\int_{u \leq x, v \leq y, (u,v) \in \mathcal{S}} f_{X,Y}(u, v) \; du \, dv}{\int_{(u,v)\in\mathcal{S}} f_{X,Y}(u, v) \; du \, dv}, \tag{3.61}$$

$$f_{X,Y|(X,Y)\in\mathcal{S}}(x, y) := \frac{\partial^2 F_{X,Y|(X,Y)\in\mathcal{S}}(x, y)}{\partial x \partial y}. \tag{3.62}$$

This definition only holds for events with nonzero probability. However, events of the form $\{X = x\}$ have probability equal to zero because the random variable is continuous. Indeed, the

range of X is uncountable, so the probability of almost every event $\{X = x\}$ must be zero, as otherwise the probability their union would be unbounded.

How can we characterize our uncertainty about Y given $X = x$ then? We define a **conditional pdf** that captures what we are trying to do in the limit and then integrate it to obtain a conditional cdf.

Definition 3.2.7 (Conditional pdf and cdf)**.** *If $F_{X,Y}$ is differentiable, then the conditional pdf of Y given X is defined as*

$$f_{Y|X}(y|x) := \frac{f_{X,Y}(x,y)}{f_X(x)} \quad \text{if } f_X(x) > 0 \tag{3.63}$$

and is undefined otherwise.

The conditional cdf of Y given X is defined as

$$F_{Y|X}(y|x) := \int_{u=-\infty}^{y} f_{Y|X}(u|x)\, du \quad \text{if } f_X(x) > 0 \tag{3.64}$$

and is undefined otherwise.

We now justify this definition, beyond the analogy with (3.18). Assume that $f_X(x) > 0$. Let us write the definition of the conditional pdf in terms of limits. We have

$$f_X(x) = \lim_{\Delta_x \to 0} \frac{\mathrm{P}(x \le X \le x + \Delta_x)}{\Delta_x}, \tag{3.65}$$

$$f_{X,Y}(x,y) = \lim_{\Delta_x \to 0} \frac{1}{\Delta_x} \frac{\partial \mathrm{P}(x \le X \le x + \Delta_x, Y \le y)}{\partial y}. \tag{3.66}$$

This implies

$$\frac{f_{X,Y}(x,y)}{f_X(x)} = \lim_{\Delta_x \to 0, \Delta_y \to 0} \frac{1}{\mathrm{P}(x \le X \le x + \Delta_x)} \frac{\partial \mathrm{P}(x \le X \le x + \Delta_x, Y \le y)}{\partial y}. \tag{3.67}$$

We can now write the conditional cdf as

$$F_{Y|X}(y|x) = \int_{u=-\infty}^{y} \lim_{\Delta_x \to 0, \Delta_y \to 0} \frac{1}{\mathrm{P}(x \le X \le x + \Delta_x)} \frac{\partial \mathrm{P}(x \le X \le x + \Delta_x, Y \le u)}{\partial y}\, \mathrm{d}u \tag{3.68}$$

$$= \lim_{\Delta_x \to 0} \frac{1}{\mathrm{P}(x \le X \le x + \Delta_x)} \int_{u=-\infty}^{y} \frac{\partial \mathrm{P}(x \le X \le x + \Delta_x, Y \le u)}{\partial y}\, \mathrm{d}u \tag{3.69}$$

$$= \lim_{\Delta_x \to 0} \frac{\mathrm{P}(x \le X \le x + \Delta_x, Y \le y)}{\mathrm{P}(x \le X \le x + \Delta_x)} \tag{3.70}$$

$$= \lim_{\Delta_x \to 0} \mathrm{P}(Y \le y | x \le X \le x + \Delta_x). \tag{3.71}$$

We can therefore interpret the conditional cdf as the limit of the cdf of Y at y conditioned on X belonging to an interval around x when the width of the interval tends to zero.

Remark 3.2.8. *Interchanging limits and integrals as in (3.69) is not necessarily justified in general. In this case it is, as long as the integral converges and the quantities involved are bounded.*

An immediate consequence of Definition 3.2.7 is the chain rule for continuous random variables.

Lemma 3.2.9 (Chain rule for continuous random variables)**.**

$$f_{X,Y}(x,y) = f_X(x)\, f_{Y|X}(y|x). \tag{3.72}$$

Applying the same ideas as in the bivariate case, we define the conditional distribution of a subvector given the rest of the random vector.

Definition 3.2.10 (Conditional pdf)**.** *The conditional pdf of a random subvector $\vec{X}_{\mathcal{I}}$, $\mathcal{I} \subseteq \{1,2,\ldots,n\}$, given the subvector $\vec{X}_{\{1,\ldots,n\}/\mathcal{I}}$ is*

$$f_{\vec{X}_{\mathcal{I}}|\vec{X}_{\{1,\ldots,n\}/\mathcal{I}}}\left(\vec{x}_{\mathcal{I}}|\vec{x}_{\{1,\ldots,n\}/\mathcal{I}}\right) := \frac{f_{\vec{X}}(\vec{x})}{f_{\vec{X}_{\{1,\ldots,n\}/\mathcal{I}}}\left(\vec{x}_{\{1,\ldots,n\}/\mathcal{I}}\right)}. \tag{3.73}$$

It is often useful to represent the joint pdf of a random vector by factoring it into conditional pdfs using the chain rule for random vectors.

Lemma 3.2.11 (Chain rule for random vectors)**.** *The joint pdf of a random vector $\vec{X}$ can be decomposed into*

$$f_{\vec{X}}(\vec{x}) = f_{\vec{X}_1}(\vec{x}_1)\, f_{\vec{X}_2|\vec{X}_1}(\vec{x}_2|\vec{x}_1) \ldots f_{\vec{X}_n|\vec{X}_1,\ldots,\vec{X}_{n-1}}(\vec{x}_n|\vec{x}_1,\ldots,\vec{x}_{n-1}) \tag{3.74}$$

$$= \prod_{i=1}^{n} f_{\vec{X}_i|\vec{X}_{\{1,\ldots,i-1\}}}\left(\vec{x}_i|\vec{x}_{\{1,\ldots,i-1\}}\right). \tag{3.75}$$

Note that the order is arbitrary, you can reorder the components of the vector in any way you like.

Proof. The result follows from applying the definition of conditional pdf recursively. □

Example 3.2.12 (Triangle lake (continued))**.** The biologist spots the otter from the shore of the lake. She is standing on the west side of the lake at a latitude of $x_1 = 0.75$ looking east and the otter is right in front of her. The otter is consequently also at a latitude of $x_1 = 0.75$, but she cannot tell at what distance. The distribution of the location of the otter given its latitude X_1 is characterized by the conditional pdf of the longitude X_2 given X_1,

$$f_{X_2|X_1}(x_2|x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)} \tag{3.76}$$

$$= \frac{1}{1-x_1}, \quad 0 \leq x_2 \leq 1 - x_1. \tag{3.77}$$

The biologist is interested in the probability that the otter is closer than x_2 to her. This probability is given by the conditional cdf

$$F_{X_2|X_1}(x_2|x_1) = \int_{-\infty}^{x_2} f_{X_2|X_1}(u|x_1)\, \mathrm{d}u \tag{3.78}$$

$$= \frac{x_2}{1-x_1}. \tag{3.79}$$

The probability that the otter is less than x_2 away is $4x_2$ for $0 \leq x_2 \leq 1/4$.

△

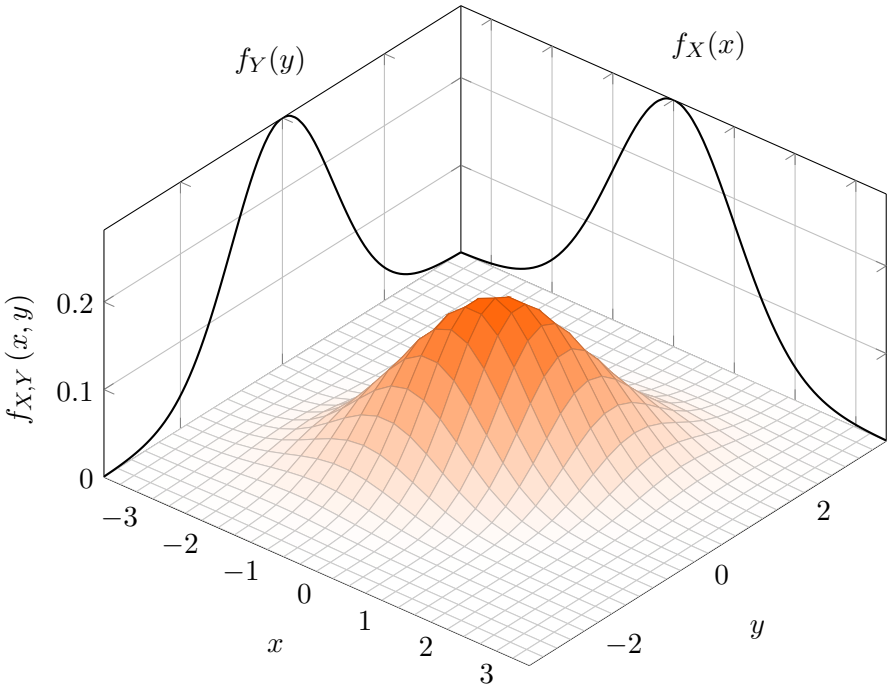


Figure 3.2: Joint pdf of a bivariate Gaussian random variable (X, Y) together with the marginal pdfs of X and Y.

3.2.4 Gaussian random vectors

Gaussian random vectors are a multidimensional generalization of Gaussian random variables. They are parametrized by a vector and a matrix that correspond to their mean and covariance matrix (we define these quantities for general multivariate random variables in Chapter 4).

Definition 3.2.13 (Gaussian random vector)**.** *A Gaussian random vector $\vec{X}$ is a random vector with joint pdf*

$$f_{\vec{X}}\left(\vec{x}\right) = \frac{1}{\sqrt{\left(2\pi\right)^n \left|\Sigma\right|}} \exp\left(-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^T \Sigma^{-1}\left(\vec{x}-\vec{\mu}\right)\right) \tag{3.80}$$

where the mean vector $\vec{\mu} \in \mathbb{R}^n$ and the covariance matrix Σ, which is symmetric and positive definite, parametrize the distribution. A Gaussian distribution with mean $\vec{\mu}$ and covariance matrix Σ is usually denoted by $\mathcal{N}\left(\vec{\mu}, \Sigma\right)$.

A fundamental property of Gaussian random vectors is that performing linear transformations on them always yields vectors with joint distributions that are also Gaussian. We will not prove this result formally, but the proof is similar to Lemma 2.5.1 (in fact this is a multidimensional generalization of that result).

Theorem 3.2.14 (Linear transformations of Gaussian random vectors are Gaussian)**.** *Let $\vec{X}$ be a Gaussian random vector of dimension n with mean $\vec{\mu}$ and covariance matrix Σ. For any matrix $A \in \mathbb{R}^{m\times n}$ and $\vec{b} \in \mathbb{R}^m$ $\vec{Y} = A\vec{X} + \vec{b}$ is a Gaussian random vector with mean $A\vec{\mu} + \vec{b}$ and covariance matrix $A\Sigma A^T$.*

A corollary of this result is that the joint pdf of a subvector of a Gaussian random vector is also a Gaussian vector.

Corollary 3.2.15 (Marginals of Gaussian random vectors are Gaussian)**.** *The joint pdf of any subvector of a Gaussian random vector is Gaussian. Without loss of generality, assume that the subvector $\vec{X}$ consists of the first m entries of the Gaussian random vector,*

$$\vec{Z} := \begin{bmatrix} \vec{X} \\ \vec{Y} \end{bmatrix}, \qquad \text{with mean} \quad \vec{\mu} := \begin{bmatrix} \mu_{\vec{X}} \\ \mu_{\vec{Y}} \end{bmatrix} \tag{3.81}$$

and covariance matrix

$$\Sigma_{\vec{Z}} = \begin{bmatrix} \Sigma_{\vec{X}} & \Sigma_{\vec{X}\vec{Y}} \\ \Sigma_{\vec{X}\vec{Y}}^T & \Sigma_{\vec{Y}} \end{bmatrix}. \tag{3.82}$$

Then $\vec{X}$ is a Gaussian random vector with mean $\mu_{\vec{X}}$ and covariance matrix $\Sigma_{\vec{X}}$.

Proof. Note that

$$\vec{X} = \begin{bmatrix} I_m & 0_{m\times n-m} \\ 0_{n-m\times m} & 0_{n-m\times n-m} \end{bmatrix} \begin{bmatrix} \vec{X} \\ \vec{Y} \end{bmatrix} = \begin{bmatrix} I_m & 0_{m\times n-m} \\ 0_{n-m\times m} & 0_{n-m\times n-m} \end{bmatrix} \vec{Z}, \tag{3.83}$$

where $I \in \mathbb{R}^{m\times m}$ is an identity matrix and $0_{c\times d}$ represents a matrix of zeros of dimensions $c \times d$. The result then follows from Theorem 3.2.14. □

Figure 3.2 shows the joint pdf of a bivariate Gaussian random variable along with its marginal pdfs.

3.3 Joint distributions of discrete and continuous variables

Probabilistic models often include both discrete and continuous random variables. However, the joint pmf or pdf of a discrete and a continuous random variable is not well defined. In order to specify the joint distribution in such cases we use their marginal and conditional pmfs and pdfs.

Assume that we have a continuous random variable C and a discrete random variable D with range R_D. We define the conditional cdf and pdf of C given D as follows.

Definition 3.3.1 (Conditional cdf and pdf of a continuous random variable given a discrete random variable)**.** *Let C and D be a continuous and a discrete random variable defined on the same probability space. Then, the conditional cdf and pdf of C given D are of the form*

$$F_{C|D}(c|d) := \mathrm{P}(C \leq c|d), \tag{3.84}$$

$$f_{C|D}(c|d) := \frac{\mathrm{d}F_{C|D}(c|d)}{\mathrm{d}c}. \tag{3.85}$$

We obtain the marginal cdf and pdf of C from the conditional cdfs and pdfs by computing a weighted sum.

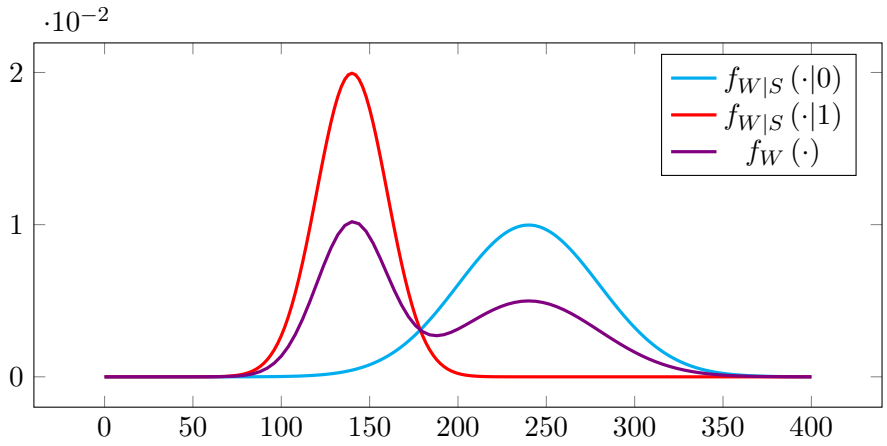


Figure 3.3: Conditional and marginal distributions of the weight of the bears W in Example 3.3.3.

Lemma 3.3.2. *Let $F_{C|D}$ and $f_{C|D}$ be the conditional cdf and pdf of a continuous random variable C given a discrete random variable D. Then,*

$$F_C(c) = \sum_{d \in R_D} p_D(d) F_{C|D}(c|d), \tag{3.86}$$

$$f_C(c) = \sum_{d \in R_D} p_D(d) f_{C|D}(c|d). \tag{3.87}$$

Proof. The events $\{D = d\}$ are a partition of the whole probability space (one of them must happen and they are all disjoint), so

$$F_C(c) = \mathrm{P}(C \leq c) \tag{3.88}$$

$$= \sum_{d \in R_D} \mathrm{P}(D = d)\, \mathrm{P}(C \leq c|d) \quad \text{by the Law of Total Probability} \tag{3.89}$$

$$= \sum_{d \in R_D} p_D(d) F_{C|D}(c|d). \tag{3.90}$$

Now, (3.87) follows by differentiating. □

Combining a discrete marginal pmf with a continuous conditional distribution allows us to define **mixture models** where the data is drawn from a continuous distribution whose parameters are chosen from a discrete set. If a Gaussian is used as the continuous distribution, this yields a Gaussian mixture model. Fitting Gaussian mixture models is a popular technique for clustering data.

Example 3.3.3 (Grizzlies in Yellowstone)**.** A scientist is gathering data on the bears in Yellowstone. It turns out that the weight of the males is well modeled by a Gaussian random variable with mean 240 kg and standard variation 40 kg, whereas the weight of the females is well modeled by a Gaussian with mean 140 kg and standard deviation 20 kg. There are about the same number of females and males.

The distribution of the weights of all the grizzlies can consequently be modeled by a Gaussian mixture that includes a continuous random variable W to represent the weight and a discrete random variable S to represent the sex of the bears. S is Bernoulli with parameter $1/2$, W given $S = 0$ (male) is $\mathcal{N}(240, 1600)$ and W given $S = 1$ (female) is $\mathcal{N}(140, 400)$. By (3.87) the pdf of W is consequently of the form

$$f_W(w) = \sum_{s=0}^{1} p_S(s) f_{W|S}(w|s) \tag{3.91}$$

$$= \frac{1}{2\sqrt{2\pi}} \left(\frac{e^{-\frac{(w-240)^2}{3200}}}{40} + \frac{e^{-\frac{(w-140)^2}{800}}}{20} \right). \tag{3.92}$$

Figure 3.3 shows the conditional and marginal distributions of W.

△

Defining the conditional pmf of a discrete random variable D given a continuous random variable C is challenging because the probability of the event $\{C = c\}$ is zero. We follow the same approach as in Definition 3.2.7 and define the conditional pmf as a limit.

Definition 3.3.4 (Conditional pmf of a discrete random variable given a continuous random variable)**.** *Let C and D be a continuous and a discrete random variable defined on the same probability space. Then, the conditional pmf of D given C is defined as*

$$p_{D|C}(d|c) := \lim_{\Delta \to 0} \frac{\mathrm{P}(D = d, c \leq C \leq c + \Delta)}{\mathrm{P}(c \leq C \leq c + \Delta)}. \tag{3.93}$$

Analogously to Lemma 3.3.2, we obtain the marginal pmf of D from the conditional pmfs by computing a weighted sum.

Lemma 3.3.5. *Let $p_{D|C}$ be the conditional pmf of a discrete random variable D given a continuous random variable C. Then,*

$$p_D(d) = \int_{c=-\infty}^{\infty} f_C(c) p_{D|C}(d|c) \, dc. \tag{3.94}$$

Proof. We will not give a formal proof but rather an intuitive argument that can be made rigorous. If we take a grid of values for c which are on a grid $\ldots, c_{-1}, c_0, c_1, \ldots$ of width Δ, then

$$p_D(d) = \sum_{i=-\infty}^{\infty} \mathrm{P}(D = d, c_i \leq C \leq c_i + \Delta) \tag{3.95}$$

by the Law of Total probability. Taking the limit as $\Delta \to 0$ the sum becomes an integral and we have

$$p_D(d) = \int_{c=-\infty}^{\infty} \lim_{\Delta \to 0} \frac{\mathrm{P}(D = d, c \leq C \leq c + \Delta)}{\Delta} \mathrm{d}c \tag{3.96}$$

$$= \int_{c=-\infty}^{\infty} \lim_{\Delta \to 0} \frac{\mathrm{P}(c \leq C \leq c + \Delta)}{\Delta} \cdot \frac{\mathrm{P}(D = d, c \leq C \leq c + \Delta)}{\mathrm{P}(c \leq C \leq c + \Delta)} \mathrm{d}c \tag{3.97}$$

$$= \int_{c=-\infty}^{\infty} f_C(c) p_{D|C}(d|c) \, \mathrm{d}c. \tag{3.98}$$

since $f_C(c) = \lim_{\Delta \to 0} \frac{\mathrm{P}(c \leq C \leq c + \Delta)}{\Delta}$. □

Combining continuous marginal distributions with discrete conditional distributions is particularly useful in Bayesian statistical models, as illustrated in the following example (see Chapter 10 for more information). The continuous distribution is used to quantify our uncertainty about the parameter of a discrete distribution.

Example 3.3.6 (Bayesian coin flip)**.** Your uncle bets you ten dollars that a coin flip will turn out heads. You suspect that the coin is biased, but you are not sure to what extent. To model this uncertainty you represent the bias as a continuous random variable B with the following pdf:

$$f_B(b) = 2b \quad \text{for } b \in [0,1] \ . \tag{3.99}$$

You can now compute the probability that the coin lands on heads denoted by X using Lemma 3.3.5. Conditioned on the bias B, the result of the coin flip is Bernoulli with parameter B.

$$\begin{aligned} p_X(1) &= \int_{b=-\infty}^{\infty} f_B(b)\, p_{X|B}(1|b)\ \mathrm{d}b & (3.100)\\ &= \int_{b=0}^{1} 2b^2\, \mathrm{d}b & (3.101)\\ &= \frac{2}{3}. & (3.102) \end{aligned}$$

According to your model the probability that the coin lands heads is $2/3$. △

The following lemma provides an analogue to the chain rule for jointly distributed continuous and discrete random variables.

Lemma 3.3.7 (Chain rule for jointly distributed continuous and discrete random variables)**.** *Let C be a continuous random variable with conditional pdf $f_{C|D}$ and D a discrete random variable with conditional pmf $p_{D|C}$. Then,*

$$p_D(d)\, f_{C|D}(c|d) = f_C(c)\, p_{D|C}(d|c)\,. \tag{3.103}$$

Proof. Applying the definitions,

$$\begin{aligned} p_D(d)\, f_{C|D}(c|d) &= \lim_{\Delta\to 0} \mathrm{P}(D=d)\, \frac{\mathrm{P}(c\le C\le c+\Delta|D=d)}{\Delta} & (3.104)\\ &= \lim_{\Delta\to 0} \frac{\mathrm{P}(D=d, c\le C\le c+\Delta)}{\Delta} & (3.105)\\ &= \lim_{\Delta\to 0} \frac{\mathrm{P}(c\le C\le c+\Delta)}{\Delta} \cdot \frac{\mathrm{P}(D=d, c\le C\le c+\Delta)}{\mathrm{P}(c\le C\le c+\Delta)} & (3.106)\\ &= f_C(c)\, p_{D|C}(d|c)\,. & (3.107) \end{aligned}$$

□

Example 3.3.8 (Grizzlies in Yellowstone (continued))**.** The scientist observes a bear with her binoculars. From their size she estimates that its weight is 180 kg. What is the probability that the bear is male?

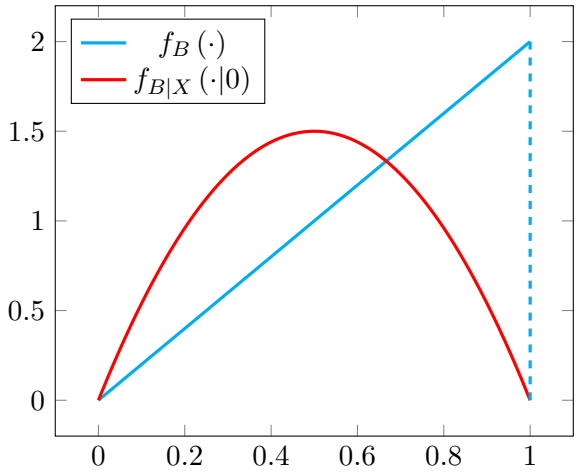


Figure 3.4: Conditional and marginal distributions of the bias of the coin flip in Example 3.3.9.

We apply Lemma 3.3.7 to compute

$$p_{S|W}(0|180) = \frac{p_S(0)\, f_{W|S}(180|0)}{f_W(180)} \tag{3.108}$$

$$= \frac{\frac{1}{40}\exp\left(-\frac{60^2}{3200}\right)}{\frac{1}{40}\exp\left(-\frac{60^2}{3200}\right) + \frac{1}{20}\exp\left(-\frac{40^2}{800}\right)} \tag{3.109}$$

$$= 0.545. \tag{3.110}$$

According to the probabilistic model, the probability that it's a male is 0.545.

△

Example 3.3.9 (Bayesian coin flip (continued))**.** The coin lands on tails. You decide to recompute the distribution of the bias conditioned on this information. By Lemma 3.3.7

$$f_{B|X}(b|0) = \frac{f_B(b)\, p_{X|B}(0|b)}{p_X(0)} \tag{3.111}$$

$$= \frac{2b(1-b)}{1/3} \tag{3.112}$$

$$= 6b(1-b). \tag{3.113}$$

Conditioned on the outcome, the pdf of the bias is now centered instead of concentrated near one as before, as shown in Figure 3.4.

△

3.4 Independence

In this section we define independence and conditional independence for random variables and vectors.

3.4.1 Definition

When knowledge about a random variable X does not affect our uncertainty about another random variable Y, we say that X and Y are **independent**. Formally, this is reflected by the marginal and conditional cdf and the conditional pmf or pdf which must be equal, i.e.

$$F_Y(y) = F_{Y|X}(y|x) \tag{3.114}$$

and

$$p_Y(y) = p_{Y|X}(y|x) \quad \text{or} \quad f_Y(y) = f_{Y|X}(y|x), \tag{3.115}$$

depending on whether the variable is discrete or continuous, for any x and any y for which the conditional distributions are well defined. Equivalently, the joint cdf and the conditional pmf or pdf factors into the marginals.

Definition 3.4.1 (Independent random variables)**.** *Two random variables X and Y are independent if and only if*

$$F_{X,Y}(x,y) = F_X(x) F_Y(y), \quad \text{for all } (x,y) \in \mathbb{R}^2. \tag{3.116}$$

If the variables are discrete, the following condition is equivalent

$$p_{X,Y}(x,y) = p_X(x) p_Y(y), \quad \text{for all } x \in R_X, y \in R_Y. \tag{3.117}$$

If the variables are continuous have joint and marginal pdfs, the following condition is equivalent

$$f_{X,Y}(x,y) = f_X(x) f_Y(y), \quad \text{for all } (x,y) \in \mathbb{R}^2. \tag{3.118}$$

We now extend the definition to account for several random variables (or equivalently several entries in a random vector) that do not provide information about each other.

Definition 3.4.2 (Independent random variables)**.** *The n entries X_1, X_2, ..., X_n in a random vector $\vec{X}$ are independent if and only if*

$$F_{\vec{X}}(\vec{x}) = \prod_{i=1}^{n} F_{X_i}(\vec{x}_i), \tag{3.119}$$

which is equivalent to

$$p_{\vec{X}}(\vec{x}) = \prod_{i=1}^{n} p_{X_i}(\vec{x}_i) \tag{3.120}$$

for discrete vectors and

$$f_{\vec{X}}(\vec{x}) = \prod_{i=1}^{n} f_{X_i}(\vec{x}_i) \tag{3.121}$$

for continuous vectors, if the joint pdf exists.

The following example shows that pairwise independence does *not imply* independence.

Example 3.4.3 (Pairwise independence does not imply joint independence)**.** Let X_1 and X_2 be the outcomes of independent unbiased coin flips. Let X_3 be the indicator of the event $\{X_1 \text{ and } X_2 \text{ have the same outcome}\}$,

$$X_3 = \begin{cases} 1 & \text{if } X_1 = X_2, \\ 0 & \text{if } X_1 \neq X_2. \end{cases} \tag{3.122}$$

The pmf of X_3 is

$$p_{X_3}(1) = p_{X_1,X_2}(1,1) + p_{X_1,X_2}(0,0) = \frac{1}{2}, \tag{3.123}$$

$$p_{X_3}(0) = p_{X_1,X_2}(0,1) + p_{X_1,X_2}(1,0) = \frac{1}{2}. \tag{3.124}$$

X_1 and X_2 are independent by assumption. X_1 and X_3 are independent because

$$p_{X_1,X_3}(0,0) = p_{X_1,X_2}(0,1) = \frac{1}{4} = p_{X_1}(0)\, p_{X_3}(0), \tag{3.125}$$

$$p_{X_1,X_3}(1,0) = p_{X_1,X_2}(1,0) = \frac{1}{4} = p_{X_1}(1)\, p_{X_3}(0), \tag{3.126}$$

$$p_{X_1,X_3}(0,1) = p_{X_1,X_2}(0,0) = \frac{1}{4} = p_{X_1}(0)\, p_{X_3}(1), \tag{3.127}$$

$$p_{X_1,X_3}(1,1) = p_{X_1,X_2}(1,1) = \frac{1}{4} = p_{X_1}(1)\, p_{X_3}(1). \tag{3.128}$$

X_2 and X_3 are independent too (the reasoning is the same).

However, are X_1, X_2 and X_3 all independent?

$$p_{X_1,X_2,X_3}(1,1,1) = \mathrm{P}(X_1 = 1, X_2 = 1) = \frac{1}{4} \neq p_{X_1}(1)\, p_{X_2}(1)\, p_{X_3}(1) = \frac{1}{8}. \tag{3.129}$$

They are not, which makes sense since X_3 is a function of X_1 and X_2. △

Conditional independence indicates that two random variables do not depend on each other, as long as an additional random variable is known.

Definition 3.4.4 (Conditionally independent random variables)**.** *Two random variables X and Y are independent with respect to another random variable Z if and only if*

$$F_{X,Y\,|\,Z}(x,y\,|\,z) = F_{X\,|\,Z}(x\,|\,z)\, F_{Y\,|\,Z}(y\,|\,z), \quad \textit{for all } (x,y) \in \mathbb{R}^2, \tag{3.130}$$

and any z for which the conditional cdfs are well defined. If the variables are discrete, the following condition is equivalent

$$p_{X,Y\,|\,Z}(x,y\,|\,z) = p_{X\,|\,Z}(x\,|\,z)\, p_{Y\,|\,Z}(y\,|\,z), \quad \textit{for all } x \in R_X, y \in R_Y, \tag{3.131}$$

and any z for which the conditional pmfs are well defined. If the variables are continuous have joint and marginal pdfs, the following condition is equivalent

$$f_{X,Y\,|\,Z}(x,y\,|\,z) = f_{X\,|\,Z}(x\,|\,z)\, f_{Y\,|\,Z}(y\,|\,z), \quad \textit{for all } (x,y) \in \mathbb{R}^2, \tag{3.132}$$

and any z for which the conditional pmfs are well defined.

The definition can be extended to condition on several random variables.

Definition 3.4.5 (Conditionally independent random variables)**.** *The components of a subvector $\vec{X}_{\mathcal{I}}$, $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$ are conditionally independent given another subvector $\vec{X}_{\mathcal{J}}$, $\mathcal{J} \subseteq \{1, 2, \ldots, n\}$, if and only if*

$$F_{\vec{X}_{\mathcal{I}}|\vec{X}_{\mathcal{J}}}\left(\vec{x}_{\mathcal{I}}|\vec{x}_{\mathcal{J}}\right) = \prod_{i\in\mathcal{I}} F_{X_i|\vec{X}_{\mathcal{J}}}\left(\vec{x}_i|\vec{x}_{\mathcal{J}}\right), \tag{3.133}$$

which is equivalent to

$$p_{\vec{X}_{\mathcal{I}}|\vec{X}_{\mathcal{J}}}\left(\vec{x}_{\mathcal{I}}|\vec{x}_{\mathcal{J}}\right) = \prod_{i\in\mathcal{I}} p_{X_i|\vec{X}_{\mathcal{J}}}\left(\vec{x}_i|\vec{x}_{\mathcal{J}}\right) \tag{3.134}$$

for discrete vectors and

$$f_{\vec{X}_{\mathcal{I}}|\vec{X}_{\mathcal{J}}}\left(\vec{x}_{\mathcal{I}}|\vec{x}_{\mathcal{J}}\right) = \prod_{i\in\mathcal{I}} f_{X_i|\vec{X}_{\mathcal{J}}}\left(\vec{x}_i|\vec{x}_{\mathcal{J}}\right) \tag{3.135}$$

for continuous vectors if the conditional joint pdf exists.

As established in Examples 1.3.5 and 1.3.6, independence does **not** imply conditional independence or vice versa.

3.4.2 Variable dependence in probabilistic modeling

A fundamental consideration when designing a probabilistic model is the dependence between the different variables, i.e. what variables are independent or conditional independent from each other. Although it may sound surprising, if the number of variables is large, introducing some independence assumptions may be necessary to make the model tractable, even if we know that all the variables are dependent. To illustrate this, consider a model for the US presidential election where there are 50 random variables, each representing a state. If the variables only take two possible values (representing what candidate wins that state), the joint pmf of their distribution has $2^{50} - 1 \geq 10^{15}$ degrees of freedom. We wouldn't be able to store the pmf with all the computer memory in the world! In contrast, if we assume that all the variables are independent, then the distribution only has 50 free parameters. Of course, this is not necessarily a good idea because failing to represent dependencies may severely affect the prediction accuracy of a model, as illustrated in Example 3.5.1 below. Striking a balance between tractability and accuracy is a crucial challenge in probabilistic modeling.

We now illustrate how the dependence structure of the random variables in a probabilistic model can be exploited to reduce the number of parameters describing the distribution through an appropriate factorization of their joint pmf or pdf. Consider three Bernoulli random variables A, B and C. In general, we need $7 = 2^3 - 1$ parameters to describe the pmf. However, if B and C are conditionally independent given A we can perform the following factorization

$$p_{A,B,C} = p_A \, p_{B\,|\,A} \, p_{C\,|\,A} \tag{3.136}$$

which only depends on five parameters (one for p_A and two each for $p_{B\,|\,A}$ and $p_{C\,|\,B}$). It is important to note that there are many other possible factorizations that do not exploit the dependence assumptions, such as for example

$$p_{A,B,C} = p_B \, p_{A\,|\,B} \, p_{C\,|\,A,B}. \tag{3.137}$$

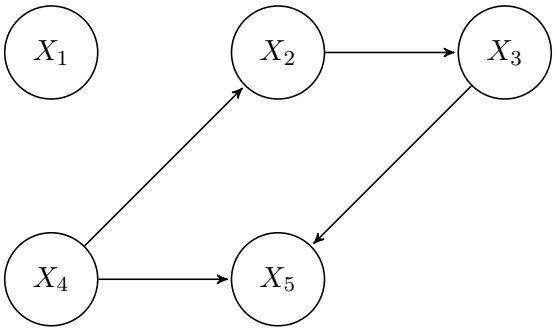


Figure 3.5: Example of a directed acyclic graphic representing a probabilistic model.

For large probabilistic models it is crucial to find factorizations that reduce the number of parameters as much as possible.

3.4.3 Graphical models

Graphical models are a tool for characterizing the dependence structure of probabilistic models. In this section we give a brief description of **directed** graphical models, which are also called Bayesian networks. Undirected graphical models, known as Markov random fields, are out of the scope of these notes. We refer the interested reader to more advanced texts in probabilistic modeling and machine learning for a more in-depth treatment of graphical models.

Directed acyclic graphs, known as DAGs, can be interpreted as diagrams representing a factorization of the joint pmf or pdf of a probabilistic model. In order to specify a valid factorization, the graphs are constrained to not have any cycles (hence the term acyclic). Each node in the DAG represents a random variable. The edges between the nodes indicate the dependence between the variables. The factorization corresponding to a DAG contains:

- The marginal pmf or pdf of the variables corresponding to all nodes with no incoming edges.
- The conditional pmf or pdf of the remaining random variables given their *parents*. A is a parent of B if there is a directed edge from (the node assigned to) A to (the node assigned to) B.

To be concrete, consider the DAG in Figure 3.5. For simplicity we denote each node using the corresponding random variable and assume that they are all discrete. Nodes X_1 and X_4 have no parents, so the factorization of the joint pmf includes their marginal pmfs. Node X_2 only descends from X_4 so we include $p_{X_2 \,|\, X_4}$. Node X_3 descends from X_2 so we include $p_{X_3 \,|\, X_2}$. Finally, node X_5 descends from X_3 and X_4 so we include $p_{X_5 \,|\, X_3, X_4}$. The factorization is of the form

$$p_{X_1,X_2,X_3,X_4,X_5} = p_{X_1} \, p_{X_4} \, p_{X_2 \,|\, X_4} \, p_{X_3 \,|\, X_2} \, p_{X_5 \,|\, X_3,X_4}. \tag{3.138}$$

This factorization reveals some dependence assumptions. By the chain rule another valid factorization of the joint pmf is

$$p_{X_1,X_2,X_3,X_4,X_5} = p_{X_1} \, p_{X_4 \,|\, X_1} \, p_{X_2 \,|\, X_1,X_4} \, p_{X_3 \,|\, X_1,X_2,X_4} \, p_{X_5 \,|\, X_1,X_2,X_3,X_4}. \tag{3.139}$$

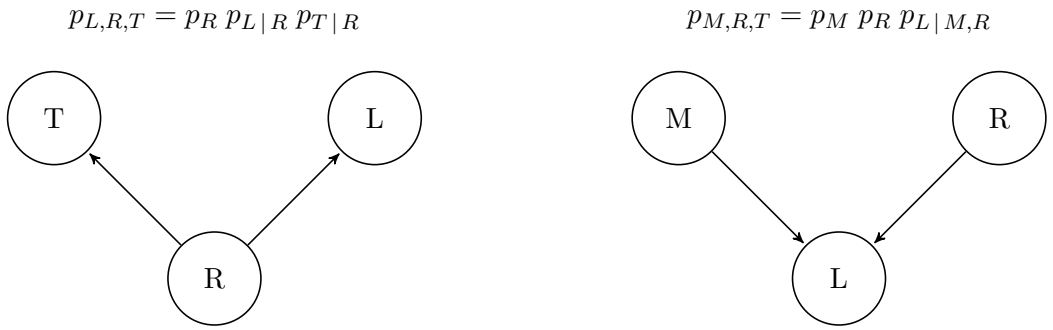


Figure 3.6: Directed graphical models corresponding to the variables in Examples 1.3.5 and 1.3.6.

Comparing both expressions, we see that X_1 and all the other variables are independent, since $p_{X_4\,|\,X_1} = p_{X_4}$, $p_{X_2\,|\,X_1,X_4} = p_{X_2\,|\,X_4}$ and so on. In addition, X_3 is conditionally independent of X_4 given X_2 since $p_{X_3\,|\,X_2,X_4} = p_{X_3\,|\,X_2}$. These dependence assumptions can be read directly from the graph, using the following property.

Theorem 3.4.6 (Local Markov property)**.** *The factorization of the joint pmf or pdf represented by a DAG satisfies the local Markov property: each variable is conditionally independent of its non-descendants given all its parent variables. In particular, if it has no parents, it is independent of its non-descendants. To be clear, B is a non-descendant of A if there is no directed path from A to B.*

Proof. Let X_i be an arbitrary variable. We denote by X_N the set of non-descendants of X_i, by X_P the set of parents and by X_D the set of descendants. The factorization represented by the graphical model is of the form

$$p_{X_1,\ldots,X_n} = p_{X_N}\, p_{X_P|X_N}\, p_{X_i|X_P}\, p_{X_D|X_i}. \tag{3.140}$$

By the chain rule another valid factorization is

$$p_{X_1,\ldots,X_n} = p_{X_N}\, p_{X_P|X_N}\, p_{X_i|X_P,X_N}\, p_{X_D|X_i,X_P,X_N}. \tag{3.141}$$

Comparing both expressions we conclude that $p_{X_i|X_P,X_N} = p_{X_i|X_P}$ so X_i is conditionally independent of X_N given X_P. □

We illustrate these ideas by showing the DAGs for Examples 1.3.5 and 1.3.6.

Example 3.4.7 (Graphical model for Example 1.3.5)**.** We model the different events in Example 1.3.5 using indicator random variables. T represents whether a taxi is available ($T = 1$) or not ($T = 0$), L whether the plane is late ($L = 1$) or not ($L = 0$), and R whether it rains ($R = 1$) or not ($R = 0$). In the example, T and L are conditionally independent given R. We can represent the corresponding factorization using the graph on the left of Figure 3.6.

△

Example 3.4.8 (Graphical model for Example 1.3.6)**.** We model the different events in Example 1.3.6 using indicator random variables. M represents whether a mechanical problem occurs

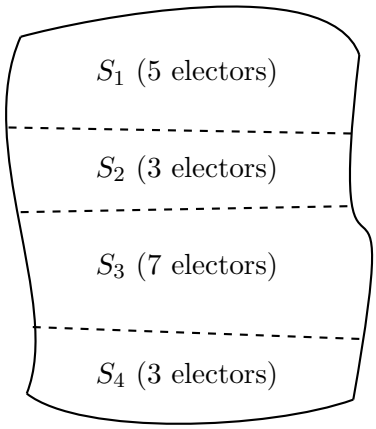


Figure 3.7: Fictitious country considered in Example 3.4.9.

($M = 1$) or not ($M = 0$) and L and R are the same as in Example 3.4.7. In the example, M and R are independent, but L depends on both of them. We can represent the corresponding factorization using the graph on the right of Figure 3.6.

△

The following example that introduces an important class of graphical models called Markov chains, which we will discuss at length in Chapter 7.

Example 3.4.9 (Election)**.** In the country shown in Figure 3.7 the presidential election follows the same system as in the United States. Citizens cast ballots for *electors* in the Electoral College. Each state is entitled to a number of electors (in the US this is usually the same as the members of Congress). In every state, the electors are pledged to the candidate that wins the state. Our goal is to model the election probabilistically. We assume that there are only two candidates A and B. Each state is represented by a random variable S_i, $1 \leq i \leq 4$,

$$S_i = \begin{cases} 1 & \text{if candidate A wins state } i, \\ -1 & \text{if candidate B wins state } i. \end{cases} \tag{3.142}$$

An important decision to make is what independence assumptions to assume about the model. Figure 3.8 shows three different options. If we model each state as independent, then we only need to estimate a single parameter for each state. However, the model may not be accurate, as the outcome in states with similar demographics is bound to be related. Another option is to estimate the full joint pmf. The problem is that it may be quite challenging to compute the parameters. We can estimate the marginal pmfs of the individual states using poll data, but conditional probabilities are more difficult to estimate. In addition, for larger models it is not tractable to consider fully dependent models (for instance in the case of the US election, as mentioned previously). A reasonable compromise could be to model the states that are not adjacent as conditionally independent given the states between them. For example, we assume that the outcome of states 1 and 3 are only related through state 2. The corresponding graphical model, depicted on the right of Figure 3.8, is called a Markov chain. It corresponds

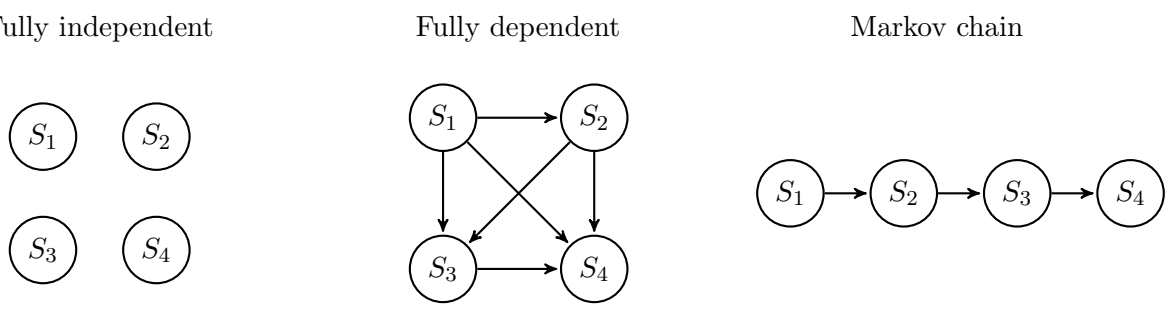


Figure 3.8: Graphical models capturing different assumptions about the distribution of the random variables considered in Example 3.4.9.

to a factorization of the form

$$p_{S_1,S_2,S_3,S_4} = p_{S_1}\, p_{S_2\,|\,S_1}\, p_{S_3\,|\,S_2}\, p_{S_4\,|\,S_3}. \tag{3.143}$$

Under this model we only need to worry about estimating pairwise conditional probabilities, as opposed to the full joint pmf. We discuss Markov chains at length in Chapter 7.

△

We conclude the section with an example involving continuous variables.

Example 3.4.10 (Desert)**.** Dani and Felix are traveling through the desert in Arizona. They become concerned that their car might break down and decide to build a probabilistic model to evaluate the risk. They model the time until the car breaks down as an exponential random variable T with a parameter that depends on the state of the motor M and the state of the road R. These three quantities are represented by random variables in the same probability space.

Unfortunately they have no idea what the state of the motor is so they assume that it is uniform between 0 (no problem with the motor) and 1 (the motor is almost dead). Similarly, they have no information about the road, so they also assume that its state is a uniform random variable between 0 (no problem with the road) and 1 (the road is terrible). In addition, they assume that the states of the road and the car are independent and that the parameter of the exponential random variable that represents the time in hours until there is a breakdown is equal to $M + R$. The corresponding graphical model is shown in Figure 3.9

To find the joint distribution of the random variables, we apply the chain rule to obtain,

$$\begin{aligned} f_{M,R,T}\,(m,r,t) &= f_M\,(m)\, f_{R|M}\,(r|m)\, f_{T|M,R}\,(t|m,r) && (3.144)\\ &= f_M\,(m)\, f_R\,(r)\, f_{T|M,R}\,(t|m,r) \quad \text{(by independence of } M \text{ and } R) && (3.145)\\ &= \begin{cases} (m+r)\, e^{-(m+r)t} & \text{for } t \geq 0,\ 0 \leq m \leq 1,\ 0 \leq r \leq 1,\\ 0 & \text{otherwise.} \end{cases} && (3.146) \end{aligned}$$

Note that we start with M and R because we know their marginal distribution, whereas we only know the conditional distribution of T given M and R.

After 15 minutes, the car breaks down. The road seems OK, about a 0.2 in the scale they defined for the value of R, so they naturally wonder about the state of the motor. Given their

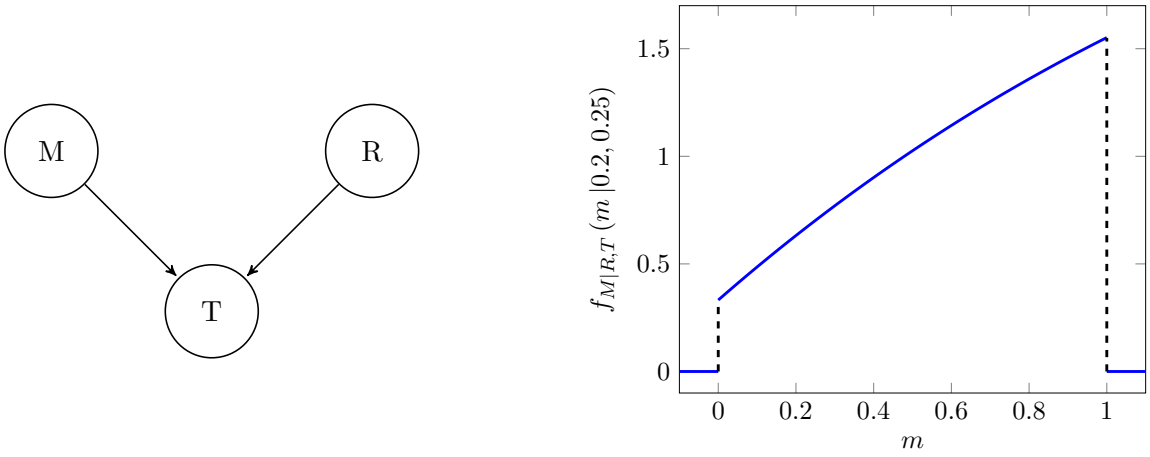


Figure 3.9: The left image is a graphical model representing the random variables in Example 3.4.10. The right plot shows the conditional pdf of M given $T = 0.25$ and $R = 0.2$.

probabilistic model, their uncertainty about the motor given all of this information is captured by the conditional distribution of M given T and R.

To compute the conditional pdf, we first need to compute the joint marginal distribution of T and R by marginalizing over M. In order to simplify the computations, we use the following simple lemma.

Lemma 3.4.11. *For any constant $c > 0$,*

$$\int_0^1 e^{-cx}\,dx = \frac{1 - e^{-c}}{c}, \tag{3.147}$$

$$\int_0^1 xe^{-cx}\,dx = \frac{1 - (1 + c)\,e^{-c}}{c^2}. \tag{3.148}$$

Proof. Equation (3.147) is obtained using the antiderivative of the exponential function (itself), whereas integrating by parts yields (3.148). △

We have

$$f_{R,T}(r, t) = \int_{m=0}^{1} f_{M,R,T}(m, r, t)\,\mathrm{d}m \tag{3.149}$$

$$= e^{-tr}\left(\int_{m=0}^{1} me^{-tm}\,\mathrm{d}m + r\int_{m=0}^{1} e^{-tm}\,\mathrm{d}m\right) \tag{3.150}$$

$$= e^{-tr}\left(\frac{1 - (1 + t)\,e^{-t}}{t^2} + \frac{r\left(1 - e^{-t}\right)}{t}\right) \quad \text{by (3.147) and (3.148)} \tag{3.151}$$

$$= \frac{e^{-tr}}{t^2}\left(1 + tr - e^{-t}\left(1 + t + tr\right)\right), \tag{3.152}$$

for $t \geq 0,\ 0 \leq r \leq 1$.

The conditional pdf of M given T and R is

$$f_{M|R,T}(m|r,t) = \frac{f_{M,R,T}(m,r,t)}{f_{R,T}(r,t)} \tag{3.153}$$

$$= \frac{(m+r)\,e^{-(m+r)t}}{\frac{e^{-tr}}{t^2}\left(1+tr-e^{-t}\left(1+t+tr\right)\right)} \tag{3.154}$$

$$= \frac{(m+r)\,t^2 e^{-tm}}{1+tr-e^{-t}\left(1+t+tr\right)}, \tag{3.155}$$

for $t \geq 0,\ 0 \leq m \leq 1,\ 0 \leq r \leq 1$. Plugging in the observed values, the conditional pdf is equal to

$$f_{M|R,T}(m|0.2,0.25) = \frac{(m+0.2)\,0.25^2 e^{-0.25m}}{1+0.25\cdot 0.2-e^{-0.25}\left(1+0.25+0.25\cdot 0.2\right)} \tag{3.156}$$

$$= 1.66\,(m+0.2)\,e^{-0.25m}. \tag{3.157}$$

for $0 \leq m \leq 1$ and to zero otherwise. The pdf is plotted in Figure 3.9. According to the model, it seems quite likely that the state of the motor was not good. △

3.5 Functions of several random variables

The pmf of a random variable $Y := g(X_1, \ldots, X_n)$ defined as a function $g : \mathbb{R}^n \to \mathbb{R}$ of several discrete random variables $X_1, \ldots, X_n$ is given by

$$p_Y(y) = \sum_{y=g(x_1,\ldots,x_n)} p_{X_1,\ldots,X_n}(x_1,\ldots,x_n). \tag{3.158}$$

This follows directly from (3.11). In words, the probability that $g(X_1, \ldots, X_n) = y$ is the sum of the joint pmf over all possible values such that $y = g(x_1, \ldots, x_n)$.

Example 3.5.1 (Election)**.** In Example 3.4.9 we discussed several possible models for a presidential election for a country with four states. Imagine that you are trying to predict the result of the election using poll data from individual states. The goal is to predict the outcome of the election, represented by the random variable

$$O := \begin{cases} 1 & \text{if } \sum_{i=1}^{4} n_i\, S_i > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.159}$$

where n_i denotes the number of electors in state i (notice that the sum can never be zero).

From analyzing the poll data you conclude that the probability that candidate A wins each of the states is 0.15. If you assume that all the states are independent, this is enough to characterize the joint pmf. Table 3.2 lists the probability of all possible outcomes for this model. By (3.158) we only need to add up the outcomes for which $O = 1$. Under the full-independence assumption, the probability that candidate A wins is 6%.

You are not satisfied by the result because you suspect that the outcomes in different states are highly dependent. From past elections, you determine that the conditional probability of a

S_1	S_2	S_3	S_4	O	Prob. (indep.)	Prob. (Markov)
-1	-1	-1	-1	0	0.5220	0.6203
-1	-1	-1	1	0	0.0921	0.0687
-1	-1	1	-1	0	0.0921	0.0431
-1	-1	1	1	1	0.0163	0.0332
-1	1	-1	-1	0	0.0921	0.0431
-1	1	-1	1	0	0.0163	0.0048
-1	1	1	-1	1	0.0163	0.0208
-1	1	1	1	1	0.0029	0.0160
1	-1	-1	-1	0	0.0921	0.0687
1	-1	-1	1	0	0.0163	0.0077
1	-1	1	-1	1	0.0163	0.0048
1	-1	1	1	1	0.0029	0.0037
1	1	-1	-1	0	0.0163	0.0332
1	1	-1	1	1	0.0029	0.0037
1	1	1	-1	1	0.0029	0.0160
1	1	1	1	1	0.0005	0.0123

Table 3.2: Table of auxiliary values for Example 3.5.1.

candidate winning a state if they win an adjacent state is indeed very high. You incorporate your estimate of the conditional probabilities into a Markov-chain model described by (3.143):

$$p_{S_1}(1) = 0.15, \tag{3.160}$$

$$p_{S_{i+1}\,|\,S_i}(1\,|\,1) = 0.435, \quad 2 \leq i \leq 4, \tag{3.161}$$

$$p_{S_{i+1}\,|\,S_i}(-1\,|\,-1) = 0.900 \quad 2 \leq i \leq 4. \tag{3.162}$$

This means that if candidate B wins a state, they are very likely to win the adjacent one. If candidate A wins a state, their chance to win an adjacent state is significantly higher than if they don't (but still lower than candidate B). Under this model the marginal probability that candidate A wins each state is still 0.15. Table 3.2 lists the probability of all possible outcomes. The probability that candidate A wins is now 11%, almost double the probability than that obtained under the fully-independent model. This illustrates the danger of not accounting for dependencies between states, which for example may have been one of the reasons why many forecasts severely underestimated Donald Trump's chances in the 2016 election.

△

Section 2.5 explains how to derive the distribution of functions of univariate random variables by first computing their cdf and then differentiating it to obtain their pdf. This directly extends to multivariable random functions. Let X, Y be random variables defined on the same probability

space, and let $U = g(X,Y)$ and $V = h(X,Y)$ for two arbitrary functions $g, h : \mathbb{R}^2 \to \mathbb{R}$. Then,

$$F_{U,V}(u,v) = \mathrm{P}(U \leq u, V \leq v) \tag{3.163}$$
$$= \mathrm{P}(g(X,Y) \leq u, h(X,Y) \leq v) \tag{3.164}$$
$$= \int_{\{(x,y) \mid g(x,y) \leq u, h(x,y) \leq v\}} f_{X,Y}(x,y)\, \mathrm{d}x\, \mathrm{d}y, \tag{3.165}$$

where the last equality only holds if the joint pdf of X and Y exists. The joint pdf can then be obtained by differentiation.

Theorem 3.5.2 (Pdf of the sum of two independent random variables)**.** *The pdf of $Z = X + Y$, where X and Y are independent random variables is equal to the* ***convolution*** *of their respective pdfs f_X and f_Y,*

$$f_Z(z) = \int_{u=-\infty}^{\infty} f_X(z-u)\, f_Y(u)\ du. \tag{3.166}$$

Proof. First we derive the cdf of Z

$$F_Z(z) = \mathrm{P}(X + Y \leq z) \tag{3.167}$$
$$= \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{z-y} f_X(x)\, f_Y(y)\, \mathrm{d}x\, \mathrm{d}y \tag{3.168}$$
$$= \int_{y=-\infty}^{\infty} F_X(z-y)\, f_Y(y)\, \mathrm{d}y. \tag{3.169}$$

Note that the joint pdf of X and Y is the product of the marginal pdfs because the random variables are independent. We now differentiate the cdf to obtain the pdf. Note that this requires an interchange of a limit operator with a differentiation operator and another interchange of an integral operator with a differentiation operator, which are justified because the functions involved are bounded and integrable.

$$f_Z(z) = \frac{\mathrm{d}}{\mathrm{d}z} \lim_{u \to \infty} \int_{y=-u}^{u} F_X(z-y)\, f_Y(y)\, \mathrm{d}y \tag{3.170}$$
$$= \lim_{u \to \infty} \frac{\mathrm{d}}{\mathrm{d}z} \int_{y=-u}^{u} F_X(z-y)\, f_Y(y)\, \mathrm{d}y \tag{3.171}$$
$$= \lim_{u \to \infty} \int_{y=-u}^{u} \frac{\mathrm{d}}{\mathrm{d}z} F_X(z-y)\, f_Y(y)\, \mathrm{d}y \tag{3.172}$$
$$= \lim_{u \to \infty} \int_{y=-u}^{u} f_X(z-y)\, f_Y(y)\, \mathrm{d}y. \tag{3.173}$$

□

Example 3.5.3 (Coffee beans)**.** A company that makes coffee buys beans from two small local producers in Colombia and Vietnam. The amount of beans they can buy from each producer varies depending on the weather. The company models these quantities C and V as independent random variables (assuming that the weather in Colombia is independent from the weather in Vietnam) which have uniform distributions in $[0,1]$ and $[0,2]$ (the unit is tons) respectively.

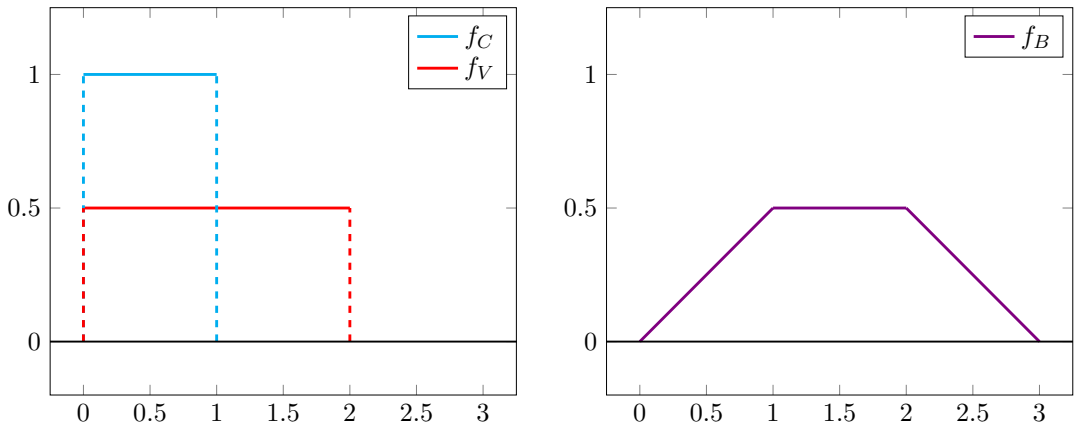


Figure 3.10: Probability density functions in Example 3.5.3.

We now compute the pdf of the total amount of coffee beans $B := E+V$ applying Theorem 3.5.2,

$$f_B(b) = \int_{u=-\infty}^{\infty} f_C(b-u)\, f_V(u)\, \mathrm{d}u \tag{3.174}$$

$$= \frac{1}{2}\int_{u=0}^{2} f_C(b-u)\, \mathrm{d}u \tag{3.175}$$

$$= \begin{cases} \frac{1}{2}\int_{u=0}^{b} \mathrm{d}u = \frac{b}{2} & \text{if } b \leq 1 \\ \frac{1}{2}\int_{u=b-1}^{b} \mathrm{d}u = \frac{1}{2} & \text{if } 1 \leq b \leq 2 \\ \frac{1}{2}\int_{u=b-1}^{2} \mathrm{d}u = \frac{3-b}{2} & \text{if } 2 \leq b \leq 3. \end{cases} \tag{3.176}$$

The pdf of B is shown in Figure 3.10.

△

3.6 Generating multivariate random variables

In Section 2.6 we consider the problem of generating independent samples from an arbitrary univariate distribution. Assuming that a procedure to achieve this is available, we can use it to sample from an arbitrary multivariate distribution by generating samples from the appropriate conditional distributions.

Algorithm 3.6.1 (Sampling from a multivariate distribution)**.** *Let X_1, X_2, ..., X_n be random variables belonging to the same probability space. To generate samples from their joint distribution we sequentially sample from their conditional distributions:*

1. *Obtain a sample x_1 of X_1.*
2. *For $i = 2, 3, \ldots, n$, obtain a sample x_i of X_i given the event $\{X_1 = x_1, \ldots, X_{i-1} = x_{i-1}\}$ by sampling from $F_{X_i|X_1,\ldots,X_{i-1}}(\cdot|x_1, \ldots, x_{i-1})$.*

The chain rule implies that the output $x_1, \ldots, x_n$ of this procedure are samples from the joint distribution of the random variables. The following example considers the problem of sampling from a mixture of exponential random variables.

Example 3.6.2 (Mixture of exponentials)**.** Let B be a Bernoulli random variable with parameter p and X an exponential random variable with parameter 1 if $B = 0$ and 2 if $B = 1$. Assume that we have access to two independent samples u_1 and u_2 from a uniform distribution in $[0, 1]$. To obtain samples from B and X:

1. We set $b := 1$ if $u_1 \leq p$ and $b := 0$ otherwise. This ensures that b is a Bernoulli sample with the right parameter.
2. Then, we set

$$x := \frac{1}{\lambda} \log \left(\frac{1}{1 - u_2} \right) \tag{3.177}$$

where $\lambda := 1$ if $b = 0$ and $\lambda := 2$ if $b = 1$. By Example 2.6.4 x is distributed as an exponential with parameter λ.

△

3.7 Rejection sampling

We end the chapter by describing rejection sampling, also known as the accept-reject method, an alternative procedure for sampling from univariate distributions. The reason we have deferred it to this chapter is that analyzing this technique requires an understanding of multivariate random variables. Before presenting the method, we motivate it using discrete random variables.

3.7.1 Rejection sampling for discrete random variables

Our goal is to simulate a random variable Y using samples from another random variable X. To simplify the exposition, we assume that their pmfs p_X and p_Y have nonzero values in the set $\{1, 2, \ldots, n\}$ (generalizing to other discrete sets is straightforward). The idea behind rejection sampling is that we can choose a subset of the samples of X in a way that reshapes its distribution. When we obtain a sample of X we decide whether to accept it or reject with a certain probability. The probability depends on the value of the sample x, if $p_X(x)$ is much larger than $p_Y(x)$ we should probably reject it most of the time (but not always!). For each $x \in \{1, 2, \ldots, n\}$ we define the probability of accepting the sample by a_x.

We are interested in the distribution of only the accepted samples. Mathematically, the pmf of the accepted samples is equal to the conditional pmf of X, conditioned on the event that the sample is accepted,

$$p_{X \,|\, \text{Accepted}} \left(x \,|\, \text{Accepted} \right) = \frac{p_X \left(x \right) \mathrm{P} \left(\text{Accepted} \,|\, X = x \right)}{\sum_{i=1}^{n} p_X \left(i \right) \mathrm{P} \left(\text{Accepted} \,|\, X = i \right)} \quad \text{by Bayes' rule} \tag{3.178}$$

$$= \frac{p_X \left(x \right) a_x}{\sum_{i=1}^{n} p_X \left(i \right) a_i}. \tag{3.179}$$

We would like to fix the accept probabilities so that for all $x \in \{1, 2, \ldots, n\}$

$$p_{X \,|\, \text{Accepted}} \left(x \,|\, \text{Accepted} \right) = p_Y(x). \tag{3.180}$$

This can be achieved by fixing

$$a_x := \frac{p_Y(x)}{c p_X(x)}, \qquad x \in \{1, \ldots, n\}, \tag{3.181}$$

for any constant c. However, this will not yield a valid probability for any arbitrary c, because a_i could be larger than one! To avoid this issue, we need

$$c \geq \max_{x \in \{1,\ldots,n\}} \frac{p_Y(x)}{p_X(x)}, \quad \text{for all } x \in \{1, \ldots, n\}. \tag{3.182}$$

Finally, we can use a uniform random variable U between 0 and 1 to accept or reject, accepting each sample x if $U \leq a_x$. You might be wondering why we can't just generate Y directly from U. That would be indeed work and is much simpler; here we are just presenting the discrete case as a pedagogical introduction to the continuous case.

Algorithm 3.7.1 (Rejection sampling)**.** *Let X and Y be random variables with pmfs p_X and p_Y such that*

$$c \geq \max_{x \in \{1,\ldots,n\}} \frac{p_Y(x)}{p_X(x)} \tag{3.183}$$

for all x such that $p_Y(x)$ is nonzero, and U a random variable that is uniformly distributed in $[0, 1]$ and independent of X.

1. *Obtain a sample y of X.*
2. *Obtain a sample u of U.*
3. *Declare y to be a sample of Y if*

$$u \leq \frac{p_Y(y)}{c\, p_X(y)}. \tag{3.184}$$

3.7.2 Rejection sampling for continuous random variables

Here we show that the idea presented in the previous section can be applied in the continuous case. The goal is to obtain samples according to a target pdf f_Y by choosing samples obtained according to a different pdf f_X. As in the discrete case, we need

$$f_Y(y) \leq c\, f_X(y) \tag{3.185}$$

for all y, where c is a fixed positive constant. In words, the pdf of Y must be bounded by a scaled version of the pdf of X.

Algorithm 3.7.2 (Rejection sampling)**.** *Let X be a random variable with pdf f_X and U a random variable that is uniformly distributed in $[0, 1]$ and independent of X. We assume that (3.185) holds.*

1. *Obtain a sample y of X.*
2. *Obtain a sample u of U.*

3. Declare y to be a sample of Y if

$$u \leq \frac{f_Y(y)}{c\, f_X(y)}. \tag{3.186}$$

The following theorem establishes that the samples obtained by rejection sampling have the desired distribution.

Theorem 3.7.3 (Rejection sampling works)**.** *If assumption* (3.185) *holds, then the samples produced by rejection sampling are distributed according to f_Y.*

Proof. Let Z denote the random variable produced by rejection sampling. The cdf of Z is equal to

$$F_Z(y) = \mathrm{P}\left(X \leq y \,|\, U \leq \frac{f_Y(X)}{c\, f_X(X)}\right) \tag{3.187}$$

$$= \frac{\mathrm{P}\left(X \leq y,\, U \leq \frac{f_Y(X)}{c\, f_X(X)}\right)}{\mathrm{P}\left(U \leq \frac{f_Y(X)}{c\, f_X(X)}\right)}. \tag{3.188}$$

To compute the numerator we integrate the joint pdf of U and X over the region of interest

$$\mathrm{P}\left(X \leq y,\, U \leq \frac{f_Y(X)}{c\, f_X(X)}\right) = \int_{x=-\infty}^{y} \int_{u=0}^{\frac{f_Y(x)}{c\, f_X(x)}} f_X(x)\, \mathrm{d}u\, \mathrm{d}x \tag{3.189}$$

$$= \int_{x=-\infty}^{y} \frac{f_Y(x)}{c\, f_X(x)} f_X(x)\, \mathrm{d}x \tag{3.190}$$

$$= \frac{1}{c} \int_{x=-\infty}^{y} f_Y(x)\, \mathrm{d}x \tag{3.191}$$

$$= \frac{1}{c} F_Y(y). \tag{3.192}$$

The denominator is obtained in a similar way

$$\mathrm{P}\left(U \leq \frac{f_Y(X)}{c\, f_X(X)}\right) = \int_{x=-\infty}^{\infty} \int_{u=0}^{\frac{f_Y(x)}{c\, f_X(x)}} f_X(x)\, \mathrm{d}u\, \mathrm{d}x \tag{3.193}$$

$$= \int_{x=-\infty}^{\infty} \frac{f_Y(x)}{c\, f_X(x)} f_X(x)\, \mathrm{d}x \tag{3.194}$$

$$= \frac{1}{c} \int_{x=-\infty}^{\infty} f_Y(x)\, \mathrm{d}x \tag{3.195}$$

$$= \frac{1}{c}. \tag{3.196}$$

We conclude that

$$F_Z(y) = F_Y(y), \tag{3.197}$$

so the method produces samples from the distribution of Y. □

We now illustrate the method by applying it to produce a Gaussian random variable from an exponential and a uniform random variable.

Example 3.7.4 (Generating a Gaussian random variable)**.** In Example 2.6.4 we learned how to generate an exponential random variables using samples from a uniform distribution. In this example we will use samples from an exponential distribution to generate a standard Gaussian random variable applying rejection sampling.

The following lemma shows that we can generate a standard Gaussian random variable Y by:

1. Generating a random variable H with pdf
$$f_H(h) := \begin{cases} \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{h^2}{2}\right) & \text{if } h \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.198}$$
2. Generating a random variable S which is equal to 1 or -1 with probability 1/2, for example by applying the method described in Section 2.6.1.
3. Setting $Y := SH$.

Lemma 3.7.5. *Let H be a continuous random variable with pdf given by (3.198) and S a discrete random variable which equals 1 with probability 1/2 and -1 with probability 1/2. The random variable of $Y := SH$ is a standard Gaussian.*

Proof. The conditional pdf of Y given S is given by
$$f_{Y|S}(y|1) = \begin{cases} f_H(y) & \text{if } y \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.199}$$
$$f_{Y|S}(y|-1) = \begin{cases} f_H(-y) & \text{if } y < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.200}$$
By Lemma 3.3.5 we have
$$f_Y(y) = p_S(1) f_{Y|S}(y|1) + p_S(-1) f_{Y|S}(y|-1) \tag{3.201}$$
$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right). \tag{3.202}$$
△

The reason why we reduce the problem to generating H is that its pdf is only nonzero on the positive axis, which allows us to bound it with the exponential pdf of an exponential random variable X with parameter 1. If we set $c := \sqrt{2e/\pi}$ then $f_H(x) \leq c f_X(x)$ for all x, as illustrated in Figure 3.11. Indeed,
$$\frac{f_H(x)}{f_X(x)} = \frac{\frac{2}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)}{\exp(-x)} \tag{3.203}$$
$$= \sqrt{\frac{2e}{\pi}} \exp\left(\frac{-(x-1)^2}{2}\right) \tag{3.204}$$
$$\leq \sqrt{\frac{2e}{\pi}}. \tag{3.205}$$

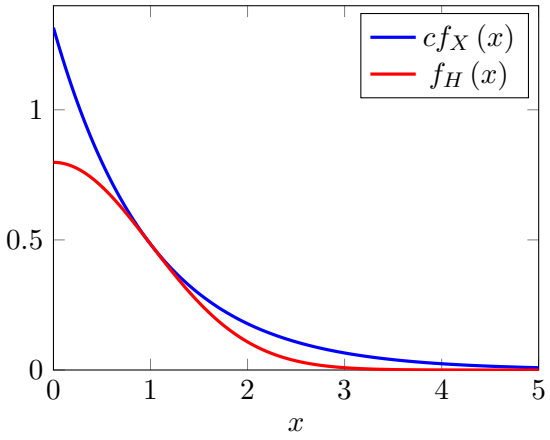


Figure 3.11: Bound on the pdf of the target distribution in Example 3.7.4.

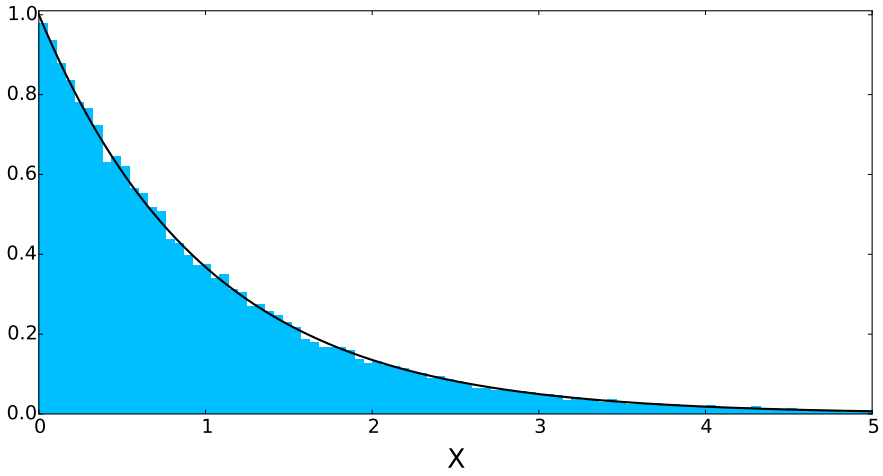
We can now apply rejection sampling to generate H. The steps are

1. Obtain a sample x from an exponential random variable X with parameter one
2. Obtain a sample u from U, which is uniformly distributed in $[0, 1]$.
3. Accept x as a sample of H if

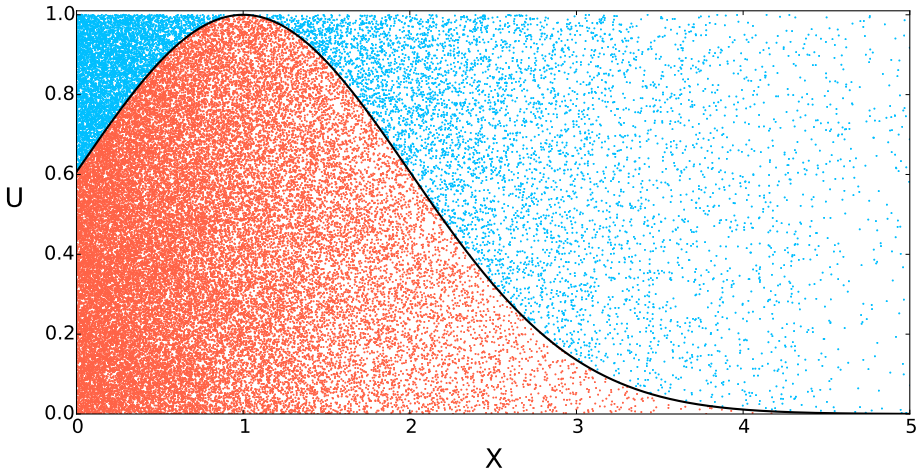
$$u \leq \exp\left(\frac{-(x-1)^2}{2}\right). \tag{3.206}$$

This procedure is illustrated in Figure 3.12. The rejection mechanism ensures that the accepted samples have the right distribution. △

Histogram of 50,000 iid samples from X (f_X is shown in black)



Scatterplot of samples from X and samples from U (accepted samples are colored red, $\exp\left(-(x-1)^2/2\right)$ is shown in black)



Histogram of accepted samples (f_H is shown in black)

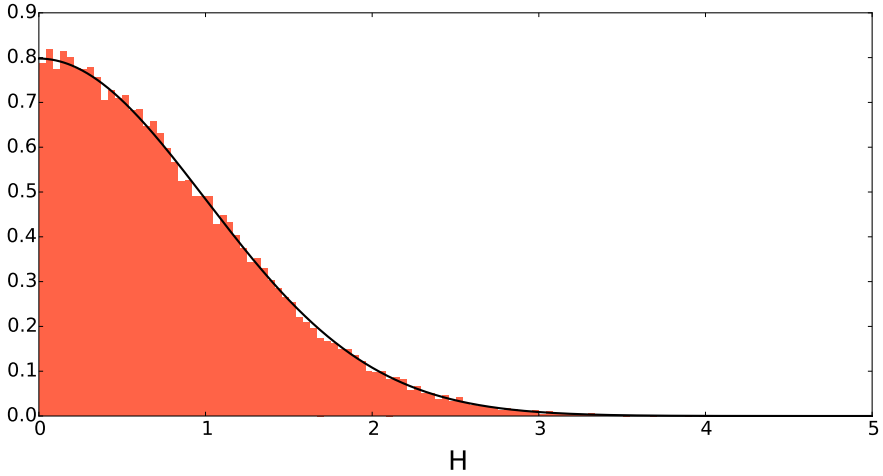


Figure 3.12: Illustration of how to generate 50,000 samples from the random variable H defined in Example 3.7.4 via rejection sampling.

Chapter 4

Expectation

In this section we introduce some quantities that describe the behavior of random variables very succinctly. The mean is the value around which the distribution of a random variable is centered. The variance quantifies the extent to which a random variable fluctuates around the mean. The covariance of two random variables indicates whether they tend to deviate from their means in a similar way. In multiple dimensions, the covariance matrix of a random vector encodes its variance in every possible direction. These quantities do not completely characterize the distribution of a random variable or vector, but they provide a useful summary of their behavior with just a few numbers.

4.1 Expectation operator

The expectation operator allows us to define the mean, variance and covariance rigorously. It maps a function of a random variable or of several random variables to an average weighted by the corresponding pmf or pdf.

Definition 4.1.1 (Expectation for discrete random variables)**.** *Let X be a discrete random variable with range R. The expected value of a function $g(X)$, $g : \mathbb{R} \to \mathbb{R}$, of X is*

$$\mathrm{E}\left(g\left(X\right)\right) := \sum_{x \in R} g\left(x\right) p_X\left(x\right). \tag{4.1}$$

Similarly, if X, Y are both discrete random variables with ranges R_X and R_Y then the expected value of a function $g(X, Y)$, $g : \mathbb{R}^2 \to \mathbb{R}$, of X and Y is

$$\mathrm{E}\left(g\left(X,Y\right)\right) := \sum_{x \in R_X} \sum_{x \in R_Y} g\left(x,y\right) p_{X,Y}\left(x,y\right). \tag{4.2}$$

If $\vec{X}$ is an n-dimensional discrete random vector, the expected value of a function $g(\vec{X})$, $g : \mathbb{R}^n \to \mathbb{R}$, of $\vec{X}$ is

$$\mathrm{E}\left(g(\vec{X})\right) := \sum_{\vec{x}_1} \sum_{\vec{x}_2} \cdots \sum_{\vec{x}_n} g\left(\vec{x}\right) p_{\vec{X}}\left(\vec{x}\right). \tag{4.3}$$

Definition 4.1.2 (Expectation for continuous random variables)**.** *Let X be a continuous random variable. The expected value of a function $g(X)$, $g:\mathbb{R}\to\mathbb{R}$, of X is*

$$\mathrm{E}\left(g\left(X\right)\right) := \int_{x=-\infty}^{\infty} g\left(x\right) f_X\left(x\right)\, dx. \tag{4.4}$$

Similarly, if X, Y are both continuous random variables then the expected value of a function $g(X,Y)$, $g:\mathbb{R}^2\to\mathbb{R}$, of X and Y is

$$\mathrm{E}\left(g\left(X,Y\right)\right) := \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} g\left(x,y\right) f_{X,Y}\left(x,y\right)\, dx\, dy. \tag{4.5}$$

If $\vec{X}$ is an n-dimensional random vector, the expected value of a function $g(X)$, $g:\mathbb{R}^n\to\mathbb{R}$, of $\vec{X}$ is

$$\mathrm{E}\left(g(\vec{X})\right) := \int_{x_1=-\infty}^{\infty}\int_{x_2=-\infty}^{\infty}\cdots\int_{x_n=-\infty}^{\infty} g\left(\vec{x}\right) f_{\vec{X}}\left(\vec{x}\right)\, dx_1\, dx_2 \ldots\, dx_n \tag{4.6}$$

In the case of quantities that depend on both continuous and discrete random variables, the product of the marginal and conditional distributions plays the role of the joint pdf or pmf.

Definition 4.1.3 (Expectation with respect to continuous and discrete random variables)**.** *If C is a continuous random variable and D a discrete random variable with range R_D defined on the same probability space, the expected value of a function $g(C,D)$ of C and D is*

$$\mathrm{E}\left(g\left(C,D\right)\right) := \int_{c=-\infty}^{\infty}\sum_{d\in R_D} g\left(c,d\right) f_C\left(c\right) p_{D|C}\left(d|c\right)\, dc \tag{4.7}$$

$$= \sum_{d\in R_D}\int_{c=-\infty}^{\infty} g\left(c,d\right) p_D\left(d\right) f_{C|D}\left(c|d\right)\, dc. \tag{4.8}$$

The expected value of a certain quantity may be infinite or not even exist if the corresponding sum or integral tends towards infinity or has an undefined value. This is illustrated by Examples 4.1.4 and 4.2.2 below.

Example 4.1.4 (St Petersburg paradox)**.** A casino offers you the following game. You will flip an unbiased coin until it lands on heads and the casino will pay you 2^k dollars where k is the number of flips. How much are you willing to pay in order to play?

Let us compute the expected gain. If the flips are independent, the total number of flips X is a geometric random variable, so $p_X(k) = 1/2^k$. The gain is 2^X which means that

$$\mathrm{E}\left(\text{Gain}\right) = \sum_{k=1}^{\infty} 2^k\cdot\frac{1}{2^k} = \infty. \tag{4.9}$$

The expected gain is infinite, but since you only get to play once, the amount of money that you are willing to pay is probably bounded. This is known as the St Petersburg paradox.

△

A fundamental property of the expectation operator is that it is linear.

Theorem 4.1.5 (Linearity of expectation)**.** *For any constant $a \in \mathbb{R}$, any function $g : \mathbb{R} \to \mathbb{R}$ and any continuous or discrete random variable X*

$$\mathrm{E}\left(a\, g\left(X\right)\right) = a\, \mathrm{E}\left(g\left(X\right)\right). \tag{4.10}$$

For any constants $a, b \in \mathbb{R}$, any functions $g_1, g_2 : \mathbb{R}^n \to \mathbb{R}$ and any continuous or discrete random variables X and Y

$$\mathrm{E}\left(a\, g_1\left(X,Y\right) + b\, g_2\left(X,Y\right)\right) = a\, \mathrm{E}\left(g_1\left(X,Y\right)\right) + b\, \mathrm{E}\left(g_2\left(X,Y\right)\right). \tag{4.11}$$

Proof. The theorem follows immediately from the linearity of sums and integrals. □

Linearity of expectation makes it very easy to compute the expectation of linear functions of random variables. In contrast, computing the joint pdf or pmf is usually much more complicated.

Example 4.1.6 (Coffee beans (continued from Example 3.5.3))**.** Let us compute the expected total amount of beans that can be bought. C is uniform in $[0, 1]$, so $\mathrm{E}\left(C\right) = 1/2$. V is uniform in $[0, 2]$, so $\mathrm{E}\left(V\right) = 1$. By linearity of expectation

$$\begin{aligned} \mathrm{E}\left(C + V\right) &= \mathrm{E}\left(C\right) + \mathrm{E}\left(V\right) && (4.12)\\ &= 1.5 \text{ tons}. && (4.13) \end{aligned}$$

Note that this holds even if the two quantities are *not* independent.

△

If two random variables are independent, then the expectation of the product factors into a product of expectations.

Theorem 4.1.7 (Expectation of functions of independent random variables)**.** *If X, Y are independent random variables defined on the same probability space, and $g, h : \mathbb{R} \to \mathbb{R}$ are univariate real-valued functions, then*

$$\mathrm{E}\left(g\left(X\right) h\left(Y\right)\right) = \mathrm{E}\left(g\left(X\right)\right) \mathrm{E}\left(h\left(Y\right)\right). \tag{4.14}$$

Proof. We prove the result for continuous random variables, but the proof for discrete random variables is essentially the same.

$$\begin{aligned} \mathrm{E}\left(g\left(X\right) h\left(Y\right)\right) &= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} g\left(x\right) h\left(y\right) f_{X,Y}\left(x,y\right)\, \mathrm{d}x\, \mathrm{d}y && (4.15)\\ &= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} g\left(x\right) h\left(y\right) f_X\left(x\right) f_Y\left(y\right)\, \mathrm{d}x\, \mathrm{d}y \quad \text{by independence} && (4.16)\\ &= \mathrm{E}\left(g\left(X\right)\right) \mathrm{E}\left(h\left(Y\right)\right). && (4.17) \end{aligned}$$

□

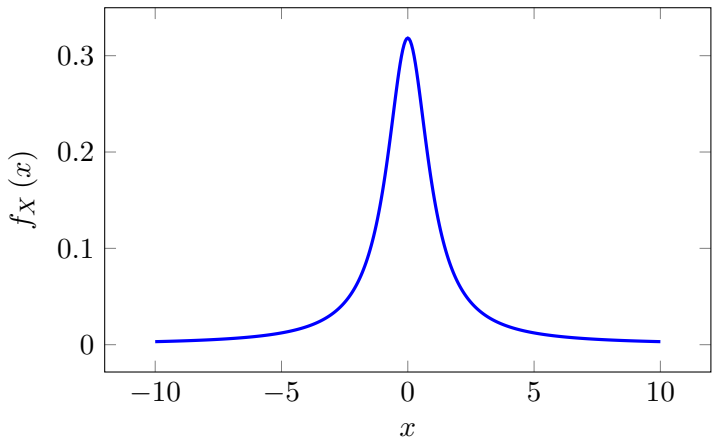


Figure 4.1: Probability density function of a Cauchy random variable.

4.2 Mean and variance

4.2.1 Mean

The mean of a random variable is equal to its expected value.

Definition 4.2.1 (Mean)**.** *The mean or first moment of X is the expected value of X:* $\mathrm{E}(X)$.

Table 4.1 lists the means of some important random variables. The derivations can be found in Section 4.5.1. As illustrated by Figure 4.3, the mean is the center of mass of the pmf or the pdf of the corresponding random variable.

If the distribution of a random variable is very *heavy tailed*, which means that the probability of the random variable taking large values decays slowly, its mean may be infinite. This is the case of the random variable representing the gain in Example 4.1.4. The following example shows that the mean may not exist if the value of the corresponding sum or integral is not well defined.

Example 4.2.2 (Cauchy random variable)**.** The pdf of the Cauchy random variable, which is shown in Figure 4.1, is given by

$$f_X(x) = \frac{1}{\pi(1+x^2)}. \tag{4.18}$$

By the definition of expected value,

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)}\,\mathrm{d}x = \int_{0}^{\infty} \frac{x}{\pi(1+x^2)}\,\mathrm{d}x - \int_{0}^{\infty} \frac{x}{\pi(1+x^2)}\,\mathrm{d}x. \tag{4.19}$$

Now, by the change of variables $t = x^2$,

$$\int_{0}^{\infty} \frac{x}{\pi(1+x^2)}\,\mathrm{d}x = \int_{0}^{\infty} \frac{1}{2\pi(1+t)}\mathrm{d}t = \lim_{t\to\infty} \frac{\log(1+t)}{2\pi} = \infty, \tag{4.20}$$

so $\mathrm{E}(X)$ does not exist, as it is the difference of two limits that tend to infinity.

△

The mean of a random vector is defined as the vector formed by the means of its components.

Definition 4.2.3 (Mean of a random vector)**.** *The mean of a random vector $\vec{X}$ is*

$$\mathrm{E}(\vec{X}) := \begin{bmatrix} \mathrm{E}\left(\vec{X}_1\right) \\ \mathrm{E}\left(\vec{X}_2\right) \\ \cdots \\ \mathrm{E}\left(\vec{X}_n\right) \end{bmatrix}. \tag{4.21}$$

As in the univariate case, the mean can be interpreted as the value around which the distribution of the random vector is centered.

It follows immediately from the linearity of the expectation operator in one dimension that the mean operator is linear.

Theorem 4.2.4 (Mean of linear transformation of a random vector)**.** *For any random vector $\vec{X}$ of dimension n, any matrix $A \in \mathbb{R}^{m \times n}$ and $\vec{b} \in \mathbb{R}^m$*

$$\mathrm{E}\left(A\vec{X} + \vec{b}\right) = A\,\mathrm{E}(\vec{X}) + \vec{b}. \tag{4.22}$$

Proof.

$$\mathrm{E}\left(A\vec{X} + \vec{b}\right) = \begin{bmatrix} \mathrm{E}\left(\sum_{i=1}^n A_{1i}\vec{X}_i + b_1\right) \\ \mathrm{E}\left(\sum_{i=1}^n A_{2i}\vec{X}_i + b_2\right) \\ \cdots \\ \mathrm{E}\left(\sum_{i=1}^n A_{mi}\vec{X}_i + b_n\right) \end{bmatrix} \tag{4.23}$$

$$= \begin{bmatrix} \sum_{i=1}^n A_{1i}\mathrm{E}\left(\vec{X}_i\right) + b_1 \\ \sum_{i=1}^n A_{2i}\mathrm{E}\left(\vec{X}_i\right) + b_2 \\ \cdots \\ \sum_{i=1}^n A_{mi}\mathrm{E}\left(\vec{X}_i\right) + b_n \end{bmatrix} \quad \text{by linearity of expectation} \tag{4.24}$$

$$= A\,\mathrm{E}(\vec{X}) + \vec{b}. \tag{4.25}$$

□

4.2.2 Median

The mean is often interpreted as representing a *typical* value taken by the random variable. However, the probability of a random variable being equal to its mean may be zero! For instance, a Bernoulli random variable cannot equal 0.5. In addition, the mean can be severely distorted by a small subset of extreme values, as illustrated by Example 4.2.6 below. The median is an alternative characterization of a *typical* value taken by the random variable, which is designed to be more robust to such situations. It is defined as the midpoint of the pmf or pdf of the random variable. If the random variable is continuous, the probability that it is either larger or smaller than the median is equal to 1/2.

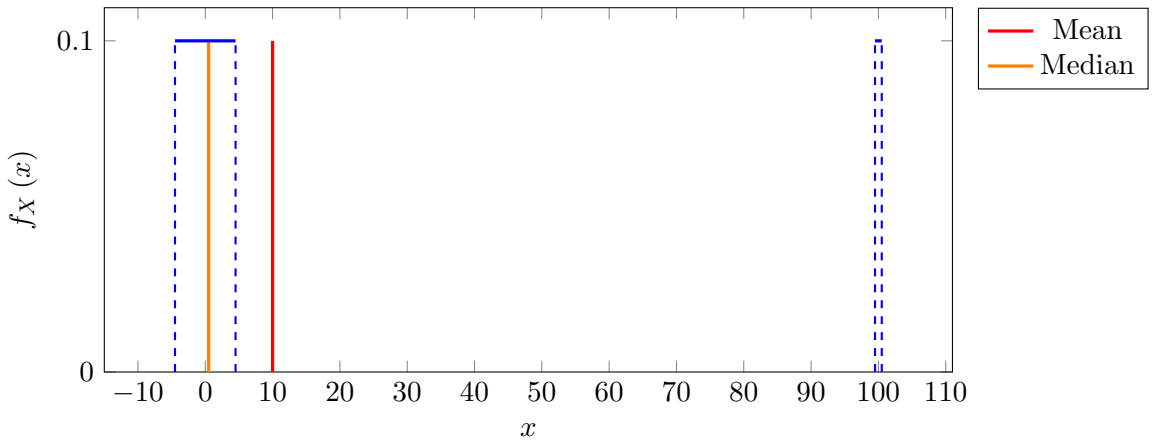


Figure 4.2: Uniform pdf in $[-4.5, 4.5] \cup [99.5, 100.5]$. The mean is 10 and the median is 0.5.

Definition 4.2.5 (Median)**.** *The median of a discrete random variable X is a number m such that*

$$P(X \leq m) \geq \frac{1}{2} \quad \textit{and} \quad P(X \geq m) \geq \frac{1}{2}. \tag{4.26}$$

The median of a continuous random variable X is a number m such that

$$F_X(m) = \int_{-\infty}^{m} f_X(x)\ dx = \frac{1}{2}. \tag{4.27}$$

The following example illustrates the robustness of the median to the presence of a small subset of extreme values with nonzero probability.

Example 4.2.6 (Mean vs median)**.** Consider a uniform random variable X with support $[-4.5, 4.5] \cup [99.5, 100.5]$. The mean of X equals

$$\begin{aligned} \mathrm{E}(X) &= \int_{x=-4.5}^{4.5} x f_X(x)\ \mathrm{d}x + \int_{x=99.5}^{100.5} x f_X(x)\ \mathrm{d}x && (4.28)\\ &= \frac{1}{10}\frac{100.5^2 - 99.5^2}{2} && (4.29)\\ &= 10. && (4.30) \end{aligned}$$

The cdf of X between -4.5 and 4.5 is equal to

$$\begin{aligned} F_X(m) &= \int_{-4.5}^{m} f_X(x)\ \mathrm{d}x && (4.31)\\ &= \frac{m + 4.5}{10}. && (4.32) \end{aligned}$$

Setting this equal to 1/2 allows to compute the median which is equal to 0.5. Figure 4.2 shows the pdf of X and the location of the median and the mean. The median provides a more realistic measure of the center of the distribution.

△

Random variable	Parameters	Mean	Variance
Bernoulli	p	p	$p\,(1-p)$
Geometric	p	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Binomial	n, p	np	$np\,(1-p)$
Poisson	λ	λ	λ
Uniform	a, b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gaussian	μ, σ	μ	σ^2

Table 4.1: Means and variance of common random variables, derived in Section 4.5.1 of the appendix.

4.2.3 Variance and standard deviation

The expected value of the square of a random variable is sometimes used to quantify the *energy* of the random variable.

Definition 4.2.7 (Second moment)**.** *The mean square or second moment of a random variable X is the expected value of X^2:* $\mathrm{E}\left(X^2\right)$.

The definition generalizes to higher moments, defined as $\mathrm{E}\left(X^p\right)$ for integers larger than two. The mean square of the difference between the random variable and its mean is called the variance of the random value. It quantifies the variation of the random variable around its mean and is also referred to as the second *centered* moment of the distribution. The square root of this quantity is the standard deviation of the random variable.

Definition 4.2.8 (Variance and standard deviation)**.** *The variance of X is the mean square deviation from the mean*

$$\mathrm{Var}\left(X\right) := \mathrm{E}\left(\left(X - \mathrm{E}\left(X\right)\right)^2\right) \tag{4.33}$$

$$= \mathrm{E}\left(X^2\right) - \mathrm{E}^2\left(X\right). \tag{4.34}$$

The standard deviation σ_X of X is

$$\sigma_X := \sqrt{\mathrm{Var}\left(X\right)}. \tag{4.35}$$

We have compiled the variances of some important random variables in Table 4.1. The derivations can be found in Section 4.5.1. In Figure 4.3 we plot the pmfs and pdfs of these random variables and display the range of values that fall within one standard deviation of the mean.

The variance operator is not linear, but it is straightforward to determine the variance of a linear function of a random variable.

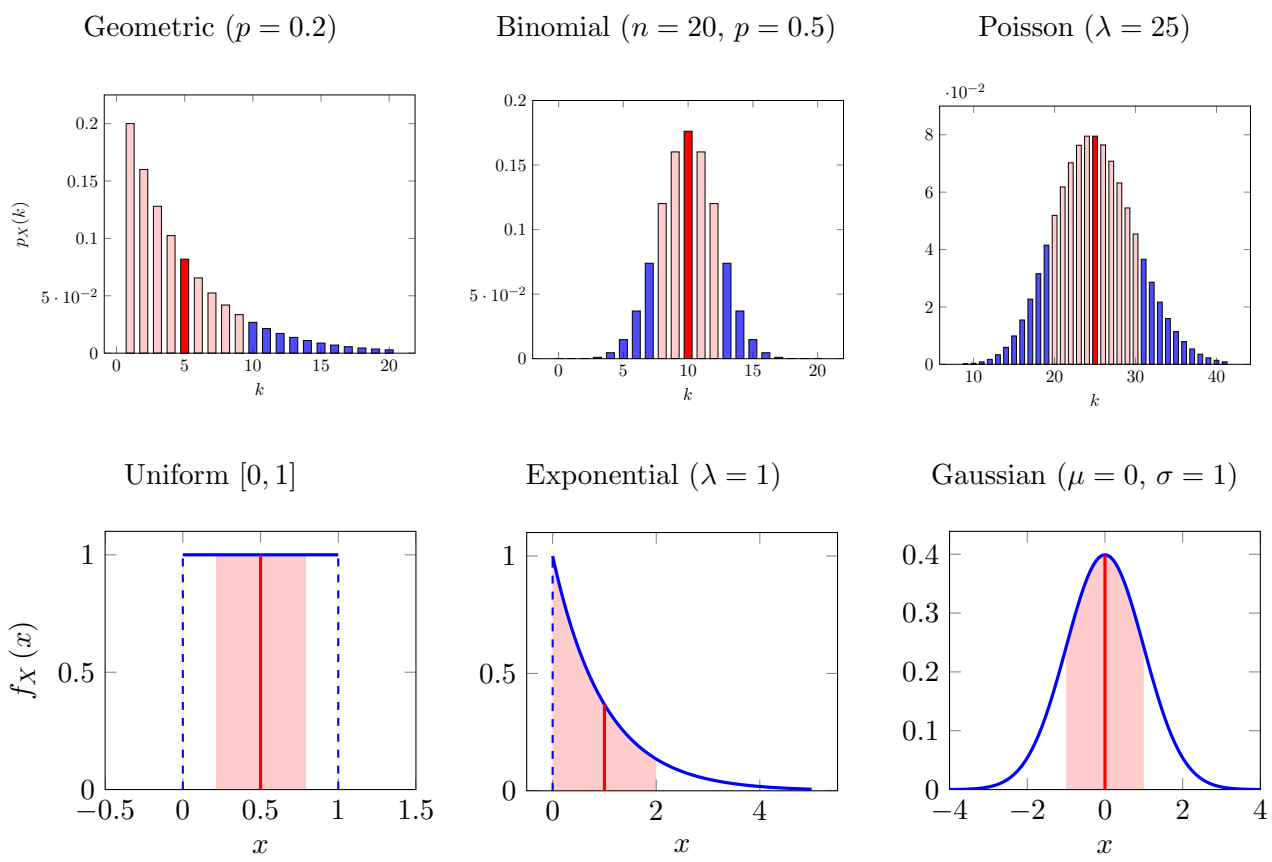


Figure 4.3: Pmfs of discrete random variables (top row) and pdfs of continuous random variables (bottom row). The mean of the random variable is marked in red. Values that are within one standard deviation of the mean are marked in pink.

Lemma 4.2.9 (Variance of linear functions)**.** *For any constants a and b*

$$\operatorname{Var}(a\,X + b) = a^2 \operatorname{Var}(X)\,. \tag{4.36}$$

Proof.

$$\begin{aligned}
\operatorname{Var}(a\,X + b) &= \mathrm{E}\left((a\,X + b - \mathrm{E}(a\,X + b))^2\right) && (4.37)\\
&= \mathrm{E}\left((a\,X + b - a\mathrm{E}(X) - b)^2\right) && (4.38)\\
&= a^2\,\mathrm{E}\left((X - \mathrm{E}(X))^2\right) && (4.39)\\
&= a^2 \operatorname{Var}(X)\,. && (4.40)
\end{aligned}$$

□

This result makes sense: If we change the center of the random variable by adding a constant, then the variance is not affected because the variance only measures the deviation from the mean. If we multiply a random variable by a constant, the standard deviation is scaled by the same factor.

4.2.4 Bounding probabilities using the mean and variance

In this section we introduce two inequalities that allow to characterize the behavior of a random valuable to some extent just from knowing its mean and variance. The first is the Markov inequality, which quantifies the intuitive idea that if a random variable is nonnegative and small then the probability that it takes large values must be small.

Theorem 4.2.10 (Markov's inequality)**.** *Let X be a nonnegative random variable. For any positive constant $a > 0$,*

$$\mathrm{P}\left(X \geq a\right) \leq \frac{\mathrm{E}\left(X\right)}{a}. \tag{4.41}$$

Proof. Consider the indicator variable $1_{X \geq a}$. We have

$$X - a\, 1_{X \geq a} \geq 0. \tag{4.42}$$

In particular its expectation is nonnegative (as it is the sum or integral of a nonnegative quantity over the positive real line). By linearity of expectation and the fact that $1_{X \geq a}$ is a Bernoulli random variable with expectation $\mathrm{P}\left(X \geq a\right)$ we have

$$\mathrm{E}\left(X\right) \geq a\, \mathrm{E}\left(1_{X \geq a}\right) = a\, \mathrm{P}\left(X \geq a\right). \tag{4.43}$$

□

Example 4.2.11 (Age of students)**.** You hear that the mean age of NYU students is 20 years, but you know quite a few students that are older than 30. You decide to apply Markov's inequality to bound the fraction of students above 30 by modeling age as a nonnegative random variable A.

$$\mathrm{P}(A \geq 30) \leq \frac{\mathrm{E}\left(A\right)}{30} = \frac{2}{3}. \tag{4.44}$$

At most two thirds of the students are over 30.

△

As illustrated Example 4.2.11, Markov's inequality can be rather loose. The reason is that it barely uses any information about the distribution of the random variable.

Chebyshev's inequality controls the deviation of the random variable from its mean. Intuitively, if the variance (and hence the standard deviation) is small, then the probability that the random variable is far from its mean must be low.

Theorem 4.2.12 (Chebyshev's inequality)**.** *For any positive constant $a > 0$ and any random variable X with bounded variance,*

$$\mathrm{P}\left(\left|X - \mathrm{E}\left(X\right)\right| \geq a\right) \leq \frac{\mathrm{Var}\left(X\right)}{a^2}. \tag{4.45}$$

Proof. Applying Markov's inequality to the random variable $Y = \left(X - \mathrm{E}\left(X\right)\right)^2$ yields the result. □

An interesting corollary to Chebyshev's inequality shows that if the variance of a random variable is zero, then the random variable is a constant or, to be precise, the probability that it deviates from its mean is zero.

Corollary 4.2.13. *If* $\mathrm{Var}(X) = 0$ *then* $\mathrm{P}(X \neq \mathrm{E}(X)) = 0$.

Proof. Take any $\epsilon > 0$, by Chebyshev's inequality

$$\mathrm{P}(|X - \mathrm{E}(X)| \geq \epsilon) \leq \frac{\mathrm{Var}(X)}{\epsilon^2} = 0. \tag{4.46}$$

□

Example 4.2.14 (Age of students (continued))**.** You are not very satisfied with your bound on the number of students above 30. You find out that the standard deviation of student age is actually just 3 years. Applying Chebyshev's inequality, this implies that

$$\mathrm{P}(A \geq 30) \leq \mathrm{P}(|A - \mathrm{E}(A)| \geq 10) \tag{4.47}$$

$$\leq \frac{\mathrm{Var}(A)}{100} = \frac{9}{100}. \tag{4.48}$$

So actually at least 91% of the students are under 30 (and above 10).

△

4.3 Covariance

4.3.1 Covariance of two random variables

The covariance of two random variables describes their joint behavior. It is the expected value of the product between the difference of the random variables and their respective means. Intuitively, it measures to what extent the random variables fluctuate together.

Definition 4.3.1 (Covariance)**.** *The covariance of* X *and* Y *is*

$$\mathrm{Cov}(X, Y) := \mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))) \tag{4.49}$$

$$= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y). \tag{4.50}$$

If $\mathrm{Cov}(X, Y) = 0$*,* X *and* Y *are* ***uncorrelated****.*

Figure 4.4 shows samples from bivariate Gaussian distributions with different covariances. If the covariance is zero, then the joint pdf has a spherical form. If the covariance is positive and large, then the joint pdf becomes skewed so that the two variables tend to have similar values. If the covariance is large and negative, then the two variables will tend to have similar values with opposite sign.

The variance of the sum of two random variables can be expressed in terms of their individual variances and their covariance. As a result, their fluctuations reinforce each other if the covariance is positive and cancel each other if it is negative.

Theorem 4.3.2 (Variance of the sum of two random variables)**.**

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y). \tag{4.51}$$

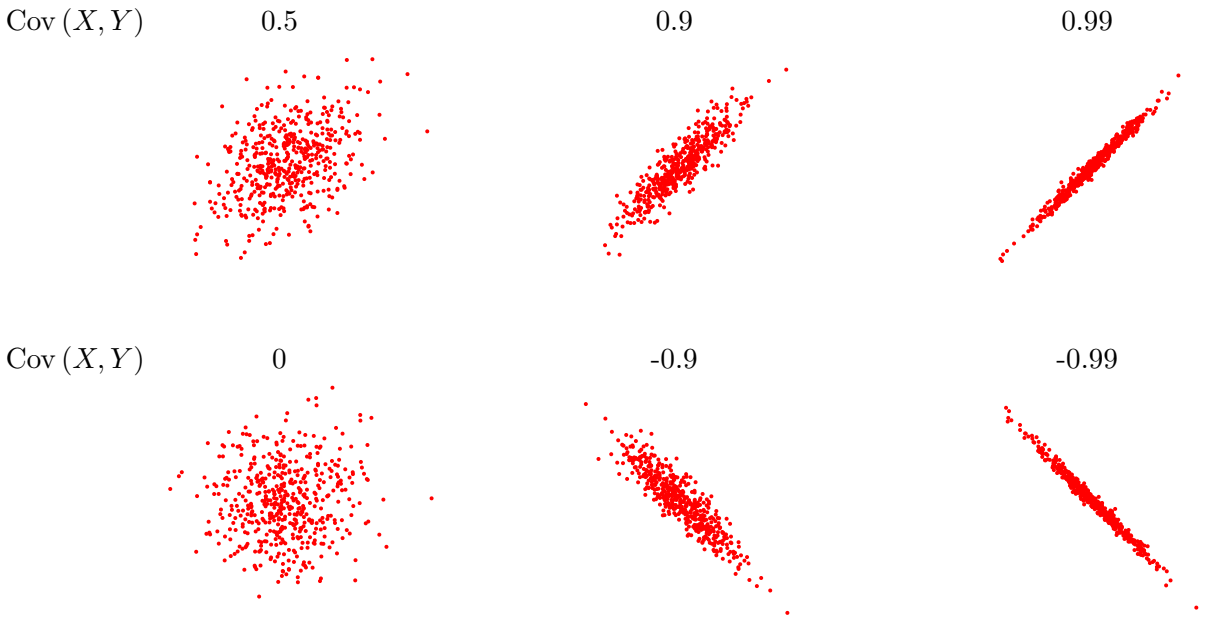


Figure 4.4: Samples from 2D Gaussian vectors (X, Y), where X and Y are standard Gaussian random variables with zero mean and unit variance, for different values of the covariance between X and Y.

Proof.

$$\begin{aligned}\operatorname{Var}(X+Y) &= \mathrm{E}\left((X+Y-\mathrm{E}(X+Y))^2\right) &(4.52)\\ &= \mathrm{E}\left((X-\mathrm{E}(X))^2\right)+\mathrm{E}\left((Y-\mathrm{E}(Y))^2\right)+2\mathrm{E}\left((X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\right)\\ &= \operatorname{Var}(X)+\operatorname{Var}(Y)+2\operatorname{Cov}(X,Y). &(4.53)\end{aligned}$$

□

An immediate consequence is that if two random variables are uncorrelated, then the variance of their sum equals the sum of their variances.

Corollary 4.3.3. *If X and Y are uncorrelated, then*

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X)+\operatorname{Var}(Y). \tag{4.54}$$

The following lemma and example show that independence implies uncorrelation, but uncorrelation does not always imply independence.

Lemma 4.3.4 (Independence implies uncorrelation)**.** *If two random variables are independent, then they are uncorrelated.*

Proof. By Theorem 4.1.7, if X and Y are independent

$$\operatorname{Cov}(X,Y) = \mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y) = \mathrm{E}(X)\mathrm{E}(Y)-\mathrm{E}(X)\mathrm{E}(Y) = 0. \tag{4.55}$$

□

Example 4.3.5 (Uncorrelation does not imply independence)**.** Let X and Y be two independent Bernoulli random variables with parameter $1/2$. Consider the random variables

$$U = X + Y, \tag{4.56}$$
$$V = X - Y. \tag{4.57}$$

Note that

$$p_U(0) = \mathrm{P}(X = 0, Y = 0) = \frac{1}{4}, \tag{4.58}$$
$$p_V(0) = \mathrm{P}(X = 1, Y = 1) + \mathrm{P}(X = 0, Y = 0) = \frac{1}{2}, \tag{4.59}$$
$$p_{U,V}(0,0) = \mathrm{P}(X = 0, Y = 0) = \frac{1}{4} \neq p_U(0)\, p_V(0) = \frac{1}{8}, \tag{4.60}$$

so U and V are not independent. However, they are uncorrelated as

$$\mathrm{Cov}(U, V) = \mathrm{E}(UV) - \mathrm{E}(U)\,\mathrm{E}(V) \tag{4.61}$$
$$= \mathrm{E}((X + Y)(X - Y)) - \mathrm{E}(X + Y)\,\mathrm{E}(X - Y) \tag{4.62}$$
$$= \mathrm{E}\left(X^2\right) - \mathrm{E}\left(Y^2\right) - \mathrm{E}^2(X) + \mathrm{E}^2(Y) = 0. \tag{4.63}$$

The final equality holds because X and Y have the same distribution.

△

4.3.2 Correlation coefficient

The covariance does not take into account the magnitude of the variances of the random variables involved. The Pearson correlation coefficient is obtained by normalizing the covariance using the standard deviations of both variables.

Definition 4.3.6 (Pearson correlation coefficient)**.** *The Pearson correlation coefficient of two random variables X and Y is*

$$\rho_{X,Y} := \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{4.64}$$

The correlation coefficient between X and Y is equal to the covariance between X/σ_X and Y/σ_Y. Figure 4.5 compares samples of bivariate Gaussian random variables that have the same correlation coefficient, but different covariance and vice versa.

Although it might not be immediately obvious, the magnitude of the correlation coefficient is bounded by one because the covariance of two random variables cannot exceed the product of their standard deviations. A useful interpretation of the correlation coefficient is that it quantifies to what extent X and Y are linearly related. In fact, if it is equal to 1 or -1 then one of the variables is a linear function of the other! All of this follows from the Cauchy-Schwarz inequality. The proof is in Section 4.5.3.

Theorem 4.3.7 (Cauchy-Schwarz inequality)**.** *For any random variables X and Y defined on the same probability space*

$$|\mathrm{E}(XY)| \leq \sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)}. \tag{4.65}$$

$\sigma_Y = 1$, $\mathrm{Cov}(X,Y) = 0.9$, $\rho_{X,Y} = 0.9$ | $\sigma_Y = 3$, $\mathrm{Cov}(X,Y) = 0.9$, $\rho_{X,Y} = 0.3$ | $\sigma_Y = 3$, $\mathrm{Cov}(X,Y) = 2.7$, $\rho_{X,Y} = 0.9$

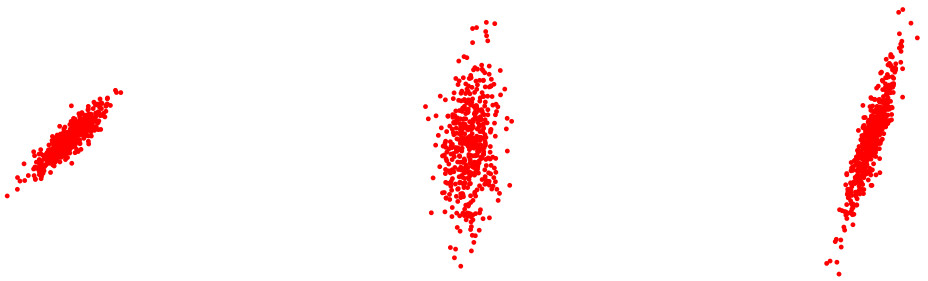


Figure 4.5: Samples from 2D Gaussian vectors (X, Y), where X is a standard Gaussian random variables with zero mean and unit variance, for different values of the standard deviation σ_Y of Y (which is mean zero) and of the covariance between X and Y.

Assume $\mathrm{E}(X^2) \neq 0$,

$$\mathrm{E}(XY) = \sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)} \iff Y = \sqrt{\frac{\mathrm{E}(Y^2)}{\mathrm{E}(X^2)}}X, \tag{4.66}$$

$$\mathrm{E}(XY) = -\sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)} \iff Y = -\sqrt{\frac{\mathrm{E}(Y^2)}{\mathrm{E}(X^2)}}X. \tag{4.67}$$

Corollary 4.3.8. *For any random variables X and Y,*

$$\mathrm{Cov}(X,Y) \leq \sigma_X \sigma_Y. \tag{4.68}$$

Equivalently, the Pearson correlation coefficient satisfies

$$|\rho_{X,Y}| \leq 1, \tag{4.69}$$

with equality if and only if there is a linear relationship between X and Y

$$|\rho_{X,Y}| = 1 \iff Y = cX + d. \tag{4.70}$$

where

$$c := \begin{cases} \frac{\sigma_Y}{\sigma_X} & \text{if } \rho_{X,Y} = 1, \\ -\frac{\sigma_Y}{\sigma_X} & \text{if } \rho_{X,Y} = -1, \end{cases} \qquad d := \mathrm{E}(Y) - c\,\mathrm{E}(X). \tag{4.71}$$

Proof. Let

$$U := X - \mathrm{E}(X), \tag{4.72}$$

$$V := Y - \mathrm{E}(Y). \tag{4.73}$$

From the definition of the variance and the correlation coefficient,

$$\mathrm{E}(U^2) = \mathrm{Var}(X), \tag{4.74}$$

$$\mathrm{E}(V^2) = \mathrm{Var}(Y) \tag{4.75}$$

$$\rho_{X,Y} = \frac{\mathrm{E}(UV)}{\sqrt{\mathrm{E}(U^2)\,\mathrm{E}(V^2)}}. \tag{4.76}$$

The result now follows from applying Theorem 4.3.7 to U and V. □

4.3.3 Covariance matrix of a random vector

The covariance matrix of a random vector captures the interaction between the components of the vector. It contains the variance of each component in the diagonal and the covariances between different components in the off diagonals.

Definition 4.3.9. *The covariance matrix of a random vector $\vec{X}$ is defined as*

$$\Sigma_{\vec{X}} := \begin{bmatrix} \operatorname{Var}\left(\vec{X}_1\right) & \operatorname{Cov}\left(\vec{X}_1, \vec{X}_2\right) & \cdots & \operatorname{Cov}\left(\vec{X}_1, \vec{X}_n\right) \\ \operatorname{Cov}\left(\vec{X}_2, \vec{X}_1\right) & \operatorname{Var}\left(\vec{X}_2\right) & \cdots & \operatorname{Cov}\left(\vec{X}_2, \vec{X}_n\right) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}\left(\vec{X}_n, \vec{X}_1\right) & \operatorname{Cov}\left(\vec{X}_n, \vec{X}_2\right) & \cdots & \operatorname{Var}\left(\vec{X}_n\right) \end{bmatrix} \tag{4.77}$$

$$= \mathrm{E}\left(\vec{X}\vec{X}^T\right) - \mathrm{E}(\vec{X})\mathrm{E}(\vec{X})^T. \tag{4.78}$$

Note that if all the entries of a vector are uncorrelated, then its covariance matrix is diagonal.

From Theorem 4.2.4 we obtain a simple expression for the covariance matrix of the linear transformation of a random vector.

Theorem 4.3.10 (Covariance matrix after a linear transformation)**.** *Let $\vec{X}$ be a random vector of dimension n with covariance matrix Σ. For any matrix $A \in \mathbb{R}^{m\times n}$ and $\vec{b} \in \mathbb{R}^m$,*

$$\Sigma_{A\vec{X}+\vec{b}} = A\Sigma_{\vec{X}}A^T. \tag{4.79}$$

Proof.

$$\Sigma_{A\vec{X}+\vec{b}} = \mathrm{E}\left(\left(A\vec{X}+\vec{b}\right)\left(A\vec{X}+\vec{b}\right)^T\right) - \mathrm{E}\left(A\vec{X}+\vec{b}\right)\mathrm{E}\left(A\vec{X}+\vec{b}\right)^T \tag{4.80}$$

$$= A\,\mathrm{E}\left(\vec{X}\vec{X}^T\right)A^T + \vec{b}\,\mathrm{E}\left(\vec{X}\right)^T A^T + A\,\mathrm{E}(\vec{X})\vec{b}^T + \vec{b}\vec{b}^T$$
$$- A\,\mathrm{E}(\vec{X})\mathrm{E}(\vec{X})^T A^T - A\,\mathrm{E}(\vec{X})\vec{b}^T - \vec{b}\,\mathrm{E}(\vec{X})^T A^T - \vec{b}\vec{b}^T \tag{4.81}$$

$$= A\left(\mathrm{E}\left(\vec{X}\vec{X}^T\right) - \mathrm{E}(\vec{X})\mathrm{E}(\vec{X})^T\right)A^T \tag{4.82}$$

$$= A\Sigma_{\vec{X}}A^T. \tag{4.83}$$

□

An immediate corollary of this result is that we can easily decode the variance of the random vector *in any direction* from the covariance matrix. Mathematically, the variance of the random vector in the direction of a unit vector $\vec{v}$ is equal to the variance of its projection onto $\vec{v}$.

Corollary 4.3.11. *Let $\vec{v}$ be a unit vector,*

$$\operatorname{Var}\left(\vec{v}^T\vec{X}\right) = \vec{v}^T\Sigma_{\vec{X}}\vec{v}. \tag{4.84}$$

Consider the eigendecomposition of the covariance matrix of an n-dimensional random vector $\vec{X}$

$$\Sigma_{\vec{X}} = U \Lambda U^T \tag{4.85}$$

$$= \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n \end{bmatrix}^T, \tag{4.86}$$

where the eigenvalues are ordered $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Covariance matrices are symmetric by definition, so by Theorem B.7.1 the eigenvectors $\vec{u}_1$, $\vec{u}_2$, $\ldots$, $\vec{u}_n$ can be chosen to be orthogonal. These eigenvectors and the eigenvalues completely characterize the variance of the random vector in different directions. The theorem is a direct consequence of Corollary 4.3.11 and Theorem B.7.2.

Theorem 4.3.12. *Let $\vec{X}$ be a random vector of dimension n with covariance matrix $\Sigma_{\vec{X}}$. The eigendecomposition of $\Sigma_{\vec{X}}$ given by (4.86) satisfies*

$$\lambda_1 = \max_{||\vec{v}||_2=1} \text{Var}\left(\vec{v}^T \vec{X}\right), \tag{4.87}$$

$$\vec{u}_1 = \arg \max_{||\vec{v}||_2=1} \text{Var}\left(\vec{v}^T \vec{X}\right), \tag{4.88}$$

$$\lambda_k = \max_{||\vec{v}||_2=1, \vec{v} \perp \vec{u}_1, \ldots, \vec{u}_{k-1}} \text{Var}\left(\vec{v}^T \vec{X}\right), \tag{4.89}$$

$$\vec{u}_k = \arg \max_{||\vec{v}||_2=1, \vec{v} \perp \vec{u}_1, \ldots, \vec{u}_{k-1}} \text{Var}\left(\vec{v}^T \vec{X}\right). \tag{4.90}$$

In words, $\vec{u}_1$ is the *direction of maximum variance*. The eigenvector $\vec{u}_2$ corresponding to the second largest eigenvalue λ_2 is the direction of maximum variation that is orthogonal to $\vec{u}_1$. In general, the eigenvector $\vec{u}_k$ corresponding to the kth largest eigenvalue λ_k reveals the direction of maximum variation that is orthogonal to $\vec{u}_1$, $\vec{u}_2$, $\ldots$, $\vec{u}_{k-1}$. Finally, $\vec{u_n}$ is the direction of minimum variance. Figure 4.6 illustrates this with an example, where $n = 2$. As we discuss in Chapter 8, principal component analysis– a popular method for unsupervised learning and dimensionality reduction– applies the same principle to determine the directions of variation of a data set.

To conclude the section, we describe an algorithm to transform samples from an uncorrelated random vector so that they have a prescribed covariance matrix. The process of transforming uncorrelated samples for this purpose is called *coloring* because uncorrelated samples are usually described as being *white* noise. As we will see in the next section, coloring allows to simulate Gaussian random vectors.

Algorithm 4.3.13 (Coloring uncorrelated samples)**.** *Let $\vec{x}$ be a realization from an n-dimensional random vector with covariance matrix I. To generate samples with covariance matrix Σ, we:*

1. *Compute the eigendecomposition $\Sigma = U \Lambda U^T$.*

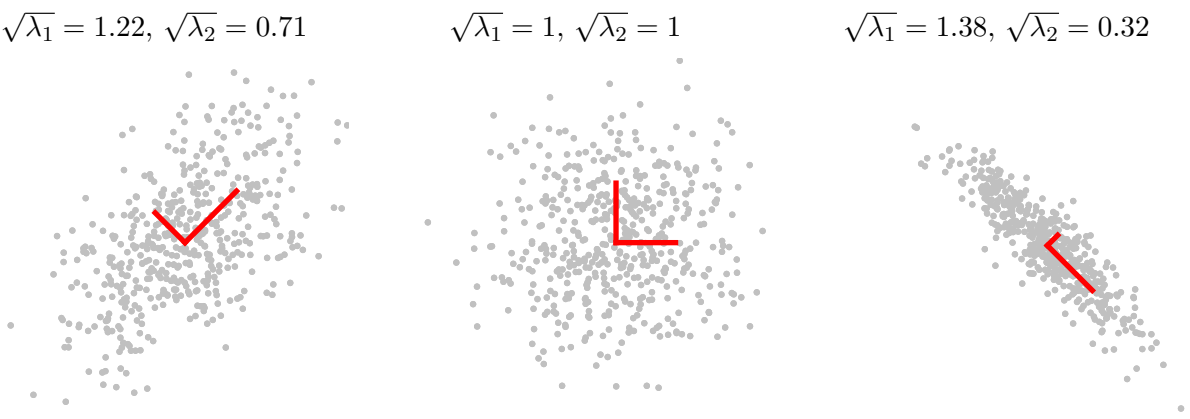


Figure 4.6: Samples from bivariate Gaussian random vectors with different covariance matrices are shown in gray. The eigenvectors of the covariance matrices are plotted in red. Each is scaled by the square roof of the corresponding eigenvalue λ_1 or λ_2.

2. *Set $\vec{y} := U\sqrt{\Lambda}\vec{x}$, where $\sqrt{\Lambda}$ is a diagonal matrix containing the square roots of the eigenvalues of Σ,*

$$\sqrt{\Lambda} := \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{bmatrix}. \tag{4.91}$$

By Theorem 4.3.10 the covariance matrix of $Y := U\sqrt{\Lambda}\vec{x}$ indeed equals Σ.

$$\Sigma_{\vec{Y}} = U\sqrt{\Lambda}\Sigma_{\vec{X}}\sqrt{\Lambda}^T U^T \tag{4.92}$$
$$= U\sqrt{\Lambda} I \sqrt{\Lambda}^T U^T \tag{4.93}$$
$$= \Sigma. \tag{4.94}$$

Figure 4.7 illustrates the two steps of coloring in 2D: First the samples are stretched according to the eigenvalues of Σ and then they are rotated to align them with the corresponding eigenvectors.

4.3.4 Gaussian random vectors

We have mostly used Gaussian vectors to visualize the different properties of the covariance operator. As opposed to other random vectors, Gaussian random vectors are completely determined by their mean and their covariance matrix. An important consequence, is that if the entries of a Gaussian random vector are uncorrelated then they are also mutually independent.

Lemma 4.3.14 (Uncorrelation implies mutual independence for Gaussian random vectors)**.** *If all the components of a Gaussian random vector $\vec{X}$ are uncorrelated, this implies that they are mutually independent.*

Proof. The parameter Σ of the joint pdf of a Gaussian random vector is its covariance matrix (one can verify this by applying the definition of covariance and integrating). If all the components

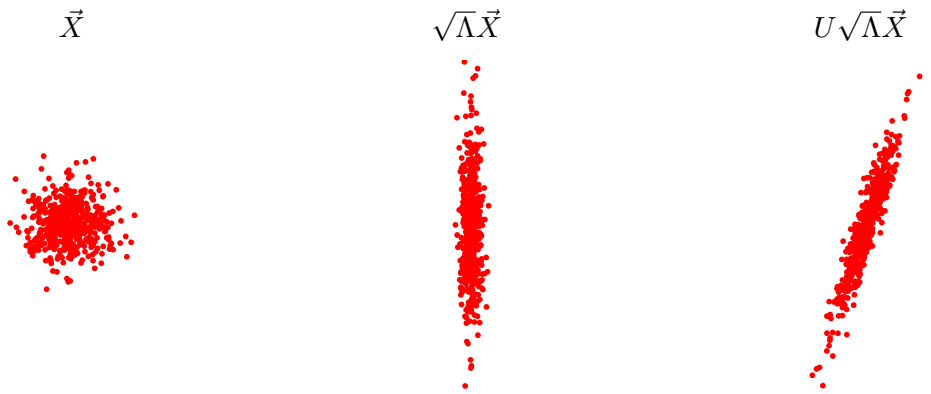


Figure 4.7: When we color two-dimensional uncorrelated samples (left), first the diagonal matrix $\sqrt{\Lambda}$ *stretches* them differently along different directions according to the eigenvalues of the desired covariance matrix (center) and then U rotates them so that they are aligned with the correspondent eigenvectors (right).

are uncorrelated then

$$\Sigma_{\vec{X}} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}, \tag{4.95}$$

where σ_i is the standard deviation of the ith component. Now, the inverse of this diagonal matrix is just

$$\Sigma_{\vec{X}}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{bmatrix}, \tag{4.96}$$

and its determinant is $|\Sigma| = \prod_{i=1}^{n} \sigma_i^2$ so that

$$f_{\vec{X}}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})\right) \tag{4.97}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)\sigma_i}} \exp\left(-\frac{(\vec{x}_i - \mu_i)^2}{2\sigma_i^2}\right) \tag{4.98}$$

$$= \prod_{i=1}^{n} f_{\vec{X}_i}(\vec{x}_i)\,. \tag{4.99}$$

Since the joint pdf factors into a product of the marginals, the components are all mutually independent. □

The following algorithm generates samples from a Gaussian random vector with an arbitrary mean and covariance matrix by coloring (and centering) a vector of independent samples from a standard Gaussian distribution.

Algorithm 4.3.15 (Generating a Gaussian random vector)**.** *To sample from an n-dimensional Gaussian random vector with mean $\vec{\mu}$ and covariance matrix Σ, we:*

1. *Generate a vector $\vec{x}$ containing n independent standard Gaussian samples.*
2. *Compute the eigendecomposition $\Sigma = U \Lambda U^T$.*
3. *Set $\vec{y} := U\sqrt{\Lambda}\vec{x} + \vec{\mu}$, where $\sqrt{\Lambda}$ is defined by (8.20).*

The algorithm just centers and colors the random vector $\vec{Y} := U\sqrt{\Lambda}\vec{X} + \vec{\mu}$. By linearity of expectation its mean is

$$\mathrm{E}(\vec{Y}) = U\sqrt{\Lambda}\mathrm{E}(\vec{X}) + \vec{\mu} \tag{4.100}$$
$$= \vec{\mu} \tag{4.101}$$

since the mean of $\vec{X}$ is zero. The same argument used in equation (4.94) shows that the covariance matrix of $\vec{X}$ is Σ. Since coloring and centering are linear operations, by Theorem 3.2.14 $\vec{Y}$ is Gaussian with the desired mean and covariance matrix. For example, in Figure 4.7 the generated samples are Gaussian. For non-Gaussian random vectors, coloring will modify the covariance matrix, but not necessarily preserve the distribution.

4.4 Conditional expectation

Conditional expectation is a useful tool for manipulating random variables. Unfortunately, it can be somewhat confusing (as we see below it's a random variable not an expectation!). Consider a function g of two random variables X and Y. The expectation of g conditioned on the event $X = x$ for any fixed value x can be computed using the conditional pmf or pdf of Y given X.

$$\mathrm{E}\left(g\left(X, Y\right) | X = x\right) = \sum_{y \in R} g\left(x, y\right) p_{Y|X}\left(y|x\right), \tag{4.102}$$

if Y is discrete and has range R, whereas

$$\mathrm{E}\left(g\left(X, Y\right) | X = x\right) = \int_{y=-\infty}^{\infty} g\left(x, y\right) f_{Y|X}\left(y|x\right) \, \mathrm{d}y, \tag{4.103}$$

if Y is continuous.

Note that $\mathrm{E}\left(g\left(X, Y\right) | X = x\right)$ can actually be interpreted as a *function of* x since it maps every value of x to a real number. This allows to define the conditional expectation of $g\left(X, Y\right)$ given X as follows.

Definition 4.4.1 (Conditional expectation)**.** *The conditional expectation of $g\left(X, Y\right)$ given X is*

$$\mathrm{E}\left(g\left(X, Y\right) | X\right) := h\left(X\right), \tag{4.104}$$

where

$$h\left(x\right) := \mathrm{E}\left(g\left(X, Y\right) | X = x\right). \tag{4.105}$$

Beware the confusing definition, the conditional expectation is actually a random variable!

One of the main uses of conditional expectation is applying iterated expectation for computing expected values. The idea is that the expected value of a certain quantity can be expressed as the expectation of the conditional expectation of the quantity.

Theorem 4.4.2 (Iterated expectation)**.** *For any random variables X and Y and any function $g : \mathbb{R}^2 \to \mathbb{R}$*

$$\mathrm{E}\left(g\left(X,Y\right)\right) = \mathrm{E}\left(\mathrm{E}\left(g\left(X,Y\right)|X\right)\right). \tag{4.106}$$

Proof. We prove the result for continuous random variables, the proof for discrete random variables, and for quantities that depend on both continuous and discrete random variables, is almost identical. To make the explanation clearer, we define

$$h\left(x\right) := \mathrm{E}\left(g\left(X,Y\right)|X=x\right) \tag{4.107}$$

$$= \int_{y=-\infty}^{\infty} g\left(x,y\right) f_{Y|X}\left(y|x\right)\,\mathrm{d}y. \tag{4.108}$$

Now,

$$\mathrm{E}\left(\mathrm{E}\left(g\left(X,Y\right)|X\right)\right) = \mathrm{E}\left(h\left(X\right)\right) \tag{4.109}$$

$$= \int_{x=-\infty}^{\infty} h\left(x\right) f_X\left(x\right)\,\mathrm{d}x \tag{4.110}$$

$$= \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} f_X\left(x\right) f_{Y|X}\left(y|x\right) g\left(x,y\right)\,\mathrm{d}y\,\mathrm{d}x \tag{4.111}$$

$$= \mathrm{E}\left(g\left(X,Y\right)\right). \tag{4.112}$$

□

Iterated expectation allows to obtain the expectation of quantities that depend on several quantities very easily if we have access to the marginal and conditional distributions. We illustrate this with several examples taken from the previous chapters.

Example 4.4.3 (Desert (continued from Example 3.4.10))**.** Let us compute the mean time at which the car breaks down, i.e. the mean of T. By iterated expectation

$$\mathrm{E}\left(T\right) = \mathrm{E}\left(\mathrm{E}\left(T|M,R\right)\right) \tag{4.113}$$

$$= \mathrm{E}\left(\frac{1}{M+R}\right) \quad \text{because } T \text{ is exponential when conditioned on } M \text{ and } R \tag{4.114}$$

$$= \int_0^1\int_0^1 \frac{1}{m+r}\,\mathrm{d}m\,\mathrm{d}r \tag{4.115}$$

$$= \int_0^1 \log\left(r+1\right) - \log\left(r\right)\,\mathrm{d}r \tag{4.116}$$

$$= \log 4 \approx 1.39 \qquad \text{integrating by parts.} \tag{4.117}$$

△

Example 4.4.4 (Grizzlies in Yellowstone (continued from Example 3.3.3))**.** Let us compute the mean weight of a bear in Yosemite. By iterated expectation

$$\begin{aligned} \mathrm{E}(W) &= \mathrm{E}(\mathrm{E}(W|S)) && (4.118)\\ &= \frac{\mathrm{E}(W|S=0)+\mathrm{E}(W|S=1)}{2} && (4.119)\\ &= 180 \text{ kg}. && (4.120) \end{aligned}$$

△

Example 4.4.5 (Bayesian coin flip (continued from Example 3.3.6)**.** Let us compute the mean of the coin-flip outcome X. By iterated expectation

$$\begin{aligned} \mathrm{E}(X) &= \mathrm{E}(\mathrm{E}(X|B)) && (4.121)\\ &= \mathrm{E}(B) \quad \text{because } X \text{ is Bernoulli when conditioned on } B && (4.122)\\ &= \int_0^1 2b^2 \,\mathrm{d}b && (4.123)\\ &= \frac{2}{3}. && (4.124) \end{aligned}$$

△

4.5 Proofs

4.5.1 Derivation of means and variances in Table 4.1

Bernoulli

$$\begin{aligned} \mathrm{E}(X) &= p_X(1) = p, && (4.125)\\ \mathrm{E}\left(X^2\right) &= p_X(1), && (4.126)\\ \mathrm{Var}(X) &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2(X) = p(1-p). && (4.127) \end{aligned}$$

Geometric

To compute the mean of a geometric random variable, we need to deal with a geometric series. By Lemma 4.5.3 in Section 4.5.2 below we have:

$$\begin{aligned} \mathrm{E}(X) &= \sum_{k=1}^{\infty} k\, p_X(k) && (4.128)\\ &= \sum_{k=1}^{\infty} k\, p\,(1-p)^{k-1} && (4.129)\\ &= \frac{p}{1-p}\sum_{k=1}^{\infty} k\,(1-p)^k = \frac{1}{p}. && (4.130) \end{aligned}$$

To compute the mean square value we apply Lemma 4.5.4 in the same section:

$$\mathrm{E}\left(X^2\right) = \sum_{k=1}^{\infty} k^2 \, p_X(k) \tag{4.131}$$

$$= \sum_{k=1}^{\infty} k^2 \, p \, (1-p)^{k-1} \tag{4.132}$$

$$= \frac{p}{1-p} \sum_{k=1}^{\infty} k^2 \, (1-p)^k \tag{4.133}$$

$$= \frac{2-p}{p^2}. \tag{4.134}$$

Binomial

As shown in Example 2.2.6, we can express a binomial random variable with parameters n and p as the sum of n independent Bernoulli random variables $B_1, B_2, \ldots$ with parameter p

$$X = \sum_{i=1}^{n} B_i. \tag{4.135}$$

Since the mean of the Bernoulli random variables is p, by linearity of expectation

$$\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(B_i) = np. \tag{4.136}$$

Note that $\mathrm{E}\left(B_i^2\right) = p$ and $\mathrm{E}(B_i B_j) = p^2$ by independence, so

$$\mathrm{E}\left(X^2\right) = \mathrm{E}\left(\sum_{i=1}^{n} \sum_{j=1}^{n} B_i B_j\right) \tag{4.137}$$

$$= \sum_{i=1}^{n} \mathrm{E}\left(B_i^2\right) + 2 \sum_{i=1}^{n-1} \sum_{i=j+1}^{n} \mathrm{E}(B_i B_j) = np + n(n-1)p^2. \tag{4.138}$$

Poisson

From calculus we have

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}, \tag{4.139}$$

which is the Taylor series expansion of the exponential function. This implies

$$\mathrm{E}(X) = \sum_{k=1}^{\infty} k p_X(k) \tag{4.140}$$

$$= \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-1)!} \tag{4.141}$$

$$= e^{-\lambda} \sum_{m=0}^{\infty} \frac{\lambda^{m+1}}{m!} = \lambda, \tag{4.142}$$

and

$$\mathrm{E}\left(X^2\right) = \sum_{k=1}^{\infty} k^2 p_X\left(k\right) \tag{4.143}$$

$$= \sum_{k=1}^{\infty} \frac{k\lambda^k e^{-\lambda}}{(k-1)!} \tag{4.144}$$

$$= e^{-\lambda}\left(\sum_{k=1}^{\infty} \frac{(k-1)\,\lambda^k}{(k-1)!} + \frac{k\lambda^k}{(k-1)!}\right) \tag{4.145}$$

$$= e^{-\lambda}\left(\sum_{m=1}^{\infty} \frac{\lambda^{m+2}}{m!} + \sum_{m=1}^{\infty} \frac{\lambda^{m+1}}{m!}\right) = \lambda^2 + \lambda. \tag{4.146}$$

Uniform

We apply the definition of expected value for continuous random variables to obtain

$$\mathrm{E}\left(X\right) = \int_{-\infty}^{\infty} x f_X\left(x\right) \mathrm{d}x = \int_a^b \frac{x}{b-a} \mathrm{d}x \tag{4.147}$$

$$= \frac{b^2 - a^2}{2\left(b-a\right)} = \frac{a+b}{2}. \tag{4.148}$$

Similarly,

$$\mathrm{E}\left(X^2\right) = \int_a^b \frac{x^2}{b-a} \mathrm{d}x \tag{4.149}$$

$$= \frac{b^3 - a^3}{3\left(b-a\right)} \tag{4.150}$$

$$= \frac{a^2 + ab + b^2}{3}. \tag{4.151}$$

Exponential

Applying integration by parts,

$$\mathrm{E}\left(X\right) = \int_{-\infty}^{\infty} x f_X\left(x\right) \mathrm{d}x \tag{4.152}$$

$$= \int_0^{\infty} x\lambda e^{-\lambda x} \mathrm{d}x \tag{4.153}$$

$$= x e^{-\lambda x}\rfloor_0^{\infty} + \int_0^{\infty} e^{-\lambda x} \mathrm{d}x = \frac{1}{\lambda}. \tag{4.154}$$

Similarly,

$$\mathrm{E}\left(X^2\right) = \int_0^{\infty} x^2 \lambda e^{-\lambda x} \mathrm{d}x \tag{4.155}$$

$$= x^2 e^{-\lambda x}\rfloor_0^{\infty} + 2\int_0^{\infty} x e^{-\lambda x} \mathrm{d}x = \frac{2}{\lambda^2}. \tag{4.156}$$

Gaussian

We apply the change of variables $t = (x - \mu)/\sigma$.

$$\begin{aligned}
\mathrm{E}(X) &= \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x && (4.157)\\
&= \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x && (4.158)\\
&= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t e^{-\frac{t^2}{2}}\,\mathrm{d}t + \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\,\mathrm{d}t && (4.159)\\
&= \mu, && (4.160)
\end{aligned}$$

where the last step follows from the fact that the integral of a bounded odd function over a symmetric interval is zero.

Applying the change of variables $t = (x - \mu)/\sigma$ and integrating by parts, we obtain that

$$\begin{aligned}
\mathrm{E}\left(X^2\right) &= \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x && (4.161)\\
&= \int_{-\infty}^{\infty} \frac{x^2}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x && (4.162)\\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 e^{-\frac{t^2}{2}}\,\mathrm{d}t + \frac{2\mu\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t e^{-\frac{t^2}{2}}\,\mathrm{d}t + \frac{\mu^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\,\mathrm{d}t && (4.163)\\
&= \frac{\sigma^2}{\sqrt{2\pi}} \left(t^2 e^{-\frac{t^2}{2}}]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\,\mathrm{d}t \right) + \mu^2 && (4.164)\\
&= \sigma^2 + \mu^2. && (4.165)
\end{aligned}$$

4.5.2 Geometric series

Lemma 4.5.1. *For any $\alpha \neq 0$ and any integers n_1 and n_2*

$$\sum_{k=n_1}^{n_2} \alpha^k = \frac{\alpha^{n_1} - \alpha^{n_2+1}}{1-\alpha}. \tag{4.166}$$

Corollary 4.5.2. *If $0 < \alpha < 1$*

$$\sum_{k=0}^{\infty} \alpha^k = \frac{\alpha}{1-\alpha}. \tag{4.167}$$

Proof. We just multiply the sum by the factor $(1-\alpha)/(1-\alpha)$ which obviously equals one,

$$\begin{aligned}
\alpha^{n_1} + \alpha^{n_1+1} + \cdots + \alpha^{n_2-1} + \alpha^{n_2} &= \frac{1-\alpha}{1-\alpha} \left(\alpha^{n_1} + \alpha^{n_1+1} + \cdots + \alpha^{n_2-1} + \alpha^{n_2} \right) && (4.168)\\
&= \frac{\alpha^{n_1} - \alpha^{n_1+1} + \alpha^{n_1+1} + \cdots - \alpha^{n_2} + \alpha^{n_2} - \alpha^{n_2+1}}{1-\alpha}\\
&= \frac{\alpha^{n_1} - \alpha^{n_2+1}}{1-\alpha} && (4.169)
\end{aligned}$$

□

Lemma 4.5.3. *For* $0 < \alpha < 1$

$$\sum_{k=1}^{\infty} k\,\alpha^k = \frac{\alpha}{(1-\alpha)^2}. \tag{4.170}$$

Proof. By Corollary 4.5.2,

$$\sum_{k=0}^{\infty} \alpha^k = \frac{1}{1-\alpha}. \tag{4.171}$$

Since the left limit converges, we can differentiate on both sides to obtain

$$\sum_{k=0}^{\infty} k\alpha^{k-1} = \frac{1}{(1-\alpha)^2}. \tag{4.172}$$

□

Lemma 4.5.4. *For* $0 < \alpha < 1$

$$\sum_{k=1}^{\infty} k^2\,\alpha^k = \frac{\alpha\,(1+\alpha)}{(1-\alpha)^3}. \tag{4.173}$$

Proof. By Lemma 4.5.3,

$$\sum_{k=1}^{\infty} k\,\alpha^k = \frac{\alpha}{(1-\alpha)^2}. \tag{4.174}$$

Since the left limit converges, we can differentiate on both sides to obtain

$$\sum_{k=1}^{\infty} k^2\alpha^{k-1} = \frac{1+\alpha}{(1-\alpha)^3}. \tag{4.175}$$

□

4.5.3 Proof of Theorem 4.3.7

If $\mathrm{E}\left(X^2\right) = 0$ then $X = 0$ by Corollary 4.2.13 $X = 0$ with probability one, which implies $\mathrm{E}\left(XY\right) = 0$ and consequently that equality holds in (4.65). The same is true if $\mathrm{E}\left(Y^2\right) = 0$.

Now assume that $\mathrm{E}\left(X^2\right) \neq 0$ and $\mathrm{E}\left(Y^2\right) \neq 0$. Let us define the constants $a = \sqrt{\mathrm{E}\left(Y^2\right)}$ and $b = \sqrt{\mathrm{E}\left(X^2\right)}$. By linearity of expectation,

$$\begin{aligned}
\mathrm{E}\left((aX + bY)^2\right) &= a^2\mathrm{E}\left(X^2\right) + b^2\mathrm{E}\left(Y^2\right) + 2\,a\,b\,\mathrm{E}\left(XY\right) && (4.176)\\
&= 2\left(\mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right) + \sqrt{\mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right)}\mathrm{E}\left(XY\right)\right), && (4.177)\\
\mathrm{E}\left((aX - bY)^2\right) &= a^2\mathrm{E}\left(X^2\right) + b^2\mathrm{E}\left(Y^2\right) - 2\,a\,b\,\mathrm{E}\left(XY\right) && (4.178)\\
&= 2\left(\mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right) - \sqrt{\mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right)}\mathrm{E}\left(XY\right)\right). && (4.179)
\end{aligned}$$

The expectation of a nonnegative quantity is nonzero because the integral or sum of a nonnegative quantity is nonnegative. Consequently, the left-hand side of (4.176) and (4.178) is nonnegative, so (B.117) and (B.118) are both nonnegative, which implies (4.65).

Let us prove (B.21) by proving both implications.

($\Rightarrow$). Assume $\mathrm{E}(XY) = -\sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)}$. Then (B.117) equals zero, so

$$\mathrm{E}\left(\left(\sqrt{\mathrm{E}(X^2)}X + \sqrt{\mathrm{E}(X^2)}Y\right)^2\right) = 0, \tag{4.180}$$

which by Corollary 4.2.13 means that $\sqrt{\mathrm{E}(Y^2)}X = -\sqrt{\mathrm{E}(X^2)}Y$ with probability one.

($\Leftarrow$). Assume $Y = -\frac{\mathrm{E}(Y^2)}{\mathrm{E}(X^2)}X$. Then one can easily check that (B.117) equals zero, which implies $\mathrm{E}(XY) = -\sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)}$.

The proof of (B.22) is almost identical (using (4.176) instead of (B.117)).

Chapter 5

Random Processes

Random processes, also known as stochastic processes, are used to model uncertain quantities that evolve in time: the trajectory of a particle, the price of oil, the temperature in New York, the national debt of the United States, etc. In these notes we introduce a mathematical framework that makes it possible to reason probabilistically about such quantities.

5.1 Definition

We denote random processes using a tilde over an upper case letter $\widetilde{X}$. This is **not** standard notation, but we want to emphasize the difference with random variables and random vectors. Formally, a random process $\widetilde{X}$ is a function that maps elements in a sample space Ω to real-valued functions.

Definition 5.1.1 (Random process)**.** *Given a probability space $(\Omega, \mathcal{F}, \mathrm{P})$, a random process $\widetilde{X}$ is a function that maps each element ω in the sample space Ω to a function $\widetilde{X}(\omega, \cdot) : \mathcal{T} \to \mathbb{R}$, where $\mathcal{T}$ is a discrete or continuous set.*

There are two possible interpretations for $\widetilde{X}(\omega, t)$:

- If we fix ω, then $\widetilde{X}(\omega, t)$ is a *deterministic function* of t known as a **realization** of the random process.
- If we fix t then $\widetilde{X}(\omega, t)$ is a *random variable*, which we usually just denote by $\widetilde{X}(t)$.

We can consequently interpret $\widetilde{X}$ as an infinite collection of random variables indexed by t. The set of possible values that the random variable $\widetilde{X}(t)$ can take for fixed t is called the **state space** of the random process. Random processes can be classified according to the indexing variable or to their state space.

- If the indexing variable t is defined on $\mathbb{R}$, or on a semi-infinite interval (t_0, ∞) for some $t_0 \in \mathbb{R}$, then $\widetilde{X}$ is a **continuous-time** random process.
- If the indexing variable t is defined on a discrete set, usually the integers or the natural numbers, then $\widetilde{X}$ is a **discrete-time** random process. In such cases we often use a different letter from t, such as i, as an indexing variable.

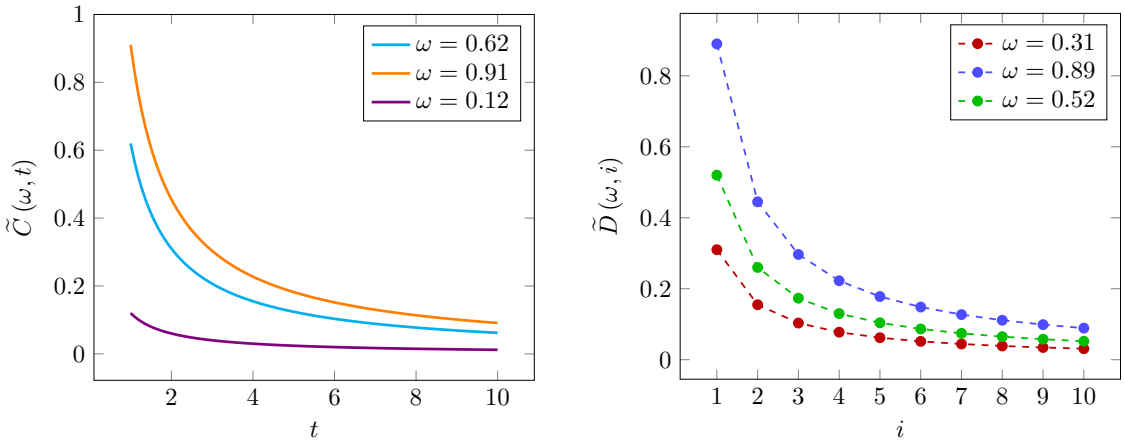


Figure 5.1: Realizations of the continuous-time (left) and discrete-time (right) random process defined in Example 5.1.2.

- If $\widetilde{X}(t)$ is a discrete random variable for all t, then $\widetilde{X}$ is a **discrete-state** random process. If the discrete random variable takes a finite number of values that is the same for all t, then $\widetilde{X}$ is a **finite-state** random process.
- If $\widetilde{X}(t)$ is a continuous random variable for all t, then $\widetilde{X}$ is a **continuous-state** random process.

Note that there are continuous-state discrete-time random processes and discrete-state continuous-time random processes. Any combination is possible.

The underlying probability space $(\Omega, \mathcal{F}, \mathrm{P})$ mentioned in the definition completely determines the stochastic behavior of the random process. In principle we can specify random processes by defining (1) a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and (2) a mapping that assigns a function to each element of Ω, as illustrated in the following example. This way of specifying random processes is only tractable for very simple cases.

Example 5.1.2 (Puddle)**.** Bob asks Mary to model a puddle probabilistically. When the puddle is formed, it contains an amount of water that is distributed uniformly between 0 and 1 gallon. As time passes, the water evaporates. After a time interval t the water that is left is t times less than the initial quantity.

Mary models the water in the puddle as a continuous-state continuous-time random process $\widetilde{C}$. The underlying sample space is $(0, 1)$, the σ algebra is the corresponding Borel σ algebra (all possible countable unions of intervals in $(0, 1)$) and the probability measure is the uniform probability measure on $(0, 1)$. For a particular element in the sample space $\omega \in (0, 1)$

$$\widetilde{C}(\omega, t) := \frac{\omega}{t}, \quad t \in [1, \infty), \tag{5.1}$$

where the unit of t is days in this example. Figure 6.1 shows different realizations of the random process. Each realization is a deterministic function on $[1, \infty)$.

Bob points out that he only cares what the state of the puddle is each day, as opposed to at any time t. Mary decides to simplify the model by using a continuous-state discrete-time random

process $\widetilde{D}$. The underlying probability space is exactly the same as before, but the time index is now discrete. For a particular element in the sample space $\omega \in (0, 1)$

$$\widetilde{D}(\omega, i) := \frac{\omega}{i}, \quad i = 1, 2, \ldots \tag{5.2}$$

Figure 6.1 shows different realizations of the continuous random process. Note that each realization is just a deterministic discrete sequence.

△

Recall that the value of the random process at a specific time is a random variable. We can therefore characterize the behavior of the process at that time by computing the distribution of the corresponding random variable. Similarly, we can consider the joint distribution of the process sampled at n fixed times. This is given by the nth-order distribution of the random process.

Definition 5.1.3 (nth-order distribution)**.** *The nth-order distribution of a random process $\widetilde{X}$ is the joint distribution of the random variables $\widetilde{X}(t_1)$, $\widetilde{X}(t_2)$, ..., $\widetilde{X}(t_n)$ for any n samples $\{t_1, t_2, \ldots, t_n\}$ of the time index t.*

Example 5.1.4 (Puddle (continued))**.** The first-order cdf of $\widetilde{C}(t)$ in Example 5.1.2 is

$$F_{\widetilde{C}(t)}(x) := \mathrm{P}\left(\widetilde{C}(t) \leq x\right) \tag{5.3}$$

$$= \mathrm{P}(\omega \leq t\, x) \tag{5.4}$$

$$= \begin{cases} \int_{u=0}^{t\,x} \mathrm{d}u = t\, x & \text{if } 0 \leq x \leq \frac{1}{t}\ , \\ 1 & \text{if } x > \frac{1}{t}\ , \\ 0 & \text{if } x < 0. \end{cases} \tag{5.5}$$

We obtain the first-order pdf by differentiating.

$$f_{\widetilde{C}(t)}(x) = \begin{cases} t & \text{if } 0 \leq x \leq \frac{1}{t}\ , \\ 0 & \text{otherwise.} \end{cases} \tag{5.6}$$

△

If the nth order distribution of a random process is shift-invariant, then the process is said to be strictly or strongly stationary.

Definition 5.1.5 (Strictly/strongly stationary process)**.** *A process is stationary in a strict or strong sense if for any $n \geq 0$ if we select n samples t_1, t_2, ..., t_n and any displacement τ the random variables $\widetilde{X}(t_1)$, $\widetilde{X}(t_2)$, ..., $\widetilde{X}(t_n)$ have the same joint distribution as $\widetilde{X}(t_1 + \tau)$, $\widetilde{X}(t_2 + \tau)$, ..., $\widetilde{X}(t_n + \tau)$.*

The random processes in Example 5.1.2 are clearly not strictly stationary because their first-order pdf and pmf are not the same at every point. An important example of strictly stationary processes are independent identically-distributed sequences, presented in Section 5.3.

As in the case of random variables and random vectors, defining the underlying probability space in order to specify a random process is usually not very practical, except for very simple

cases like the one in Example 5.1.2. The reason is that it is challenging to come up with a probability space that gives rise to a given n-th order distribution of interest. Fortunately, we can also specify a random process by directly specifying its n-th order distribution *for all values of* $n = 1, 2, \ldots$ This completely characterizes the random process. Most of the random processes described in this chapter, e.g. independent identically-distributed sequences, Markov chains, Poisson processes and Gaussian processes, are specified in this way.

Finally, random processes can also be specified by expressing them as functions of other random processes. A function $\widetilde{Y} := g(\widetilde{X})$ of a random process $\widetilde{X}$ is also a random process, as it maps any element ω in the sample space Ω to a function $\widetilde{Y}(\omega, \cdot) := g(\widetilde{X}(\omega, \cdot))$. In Section 5.6 we define random walks in this way.

5.2 Mean and autocovariance functions

As in the case of random variables and random vectors, the expectation operator allows to derive quantities that summarize the behavior of the random process. The mean of the random vector is the mean of $\widetilde{X}(t)$ at any fixed time t.

Definition 5.2.1 (Mean)**.** *The mean of a random process is the function*

$$\mu_{\widetilde{X}}(t) := \mathrm{E}\left(\widetilde{X}(t)\right). \tag{5.7}$$

Note that the mean is a *deterministic* function of t. The autocovariance of a random process is another deterministic function that is equal to the covariance of $\widetilde{X}(t_1)$ and $\widetilde{X}(t_2)$ for any two points t_1 and t_2. If we set $t_1 := t_2$, then the autocovariance equals the variance at t_1.

Definition 5.2.2 (Autocovariance)**.** *The autocovariance of a random process is the function*

$$R_{\widetilde{X}}(t_1, t_2) := \mathrm{Cov}\left(\widetilde{X}(t_1), \widetilde{X}(t_2)\right). \tag{5.8}$$

In particular,

$$R_{\widetilde{X}}(t, t) := \mathrm{Var}\left(\widetilde{X}(t)\right). \tag{5.9}$$

Intuitively, the autocovariance quantifies the correlation between the process at two different time points. If this correlation only depends on the separation between the two points, then the process is said to be wide-sense stationary.

Definition 5.2.3 (Wide-sense/weakly stationary process)**.** *A process is stationary in a wide or weak sense if its mean is constant*

$$\mu_{\widetilde{X}}(t) := \mu \tag{5.10}$$

and its autocovariance function is shift invariant*, i.e.*

$$R_{\widetilde{X}}(t_1, t_2) := R_{\widetilde{X}}(t_1 + \tau, t_2 + \tau) \tag{5.11}$$

for any t_1 and t_2 and any shift τ. For weakly stationary processes, the autocovariance is usually expressed as a function of the difference between the two time points,

$$R_{\widetilde{X}}(s) := R_{\widetilde{X}}(t, t + s) \quad \text{for any } t. \tag{5.12}$$

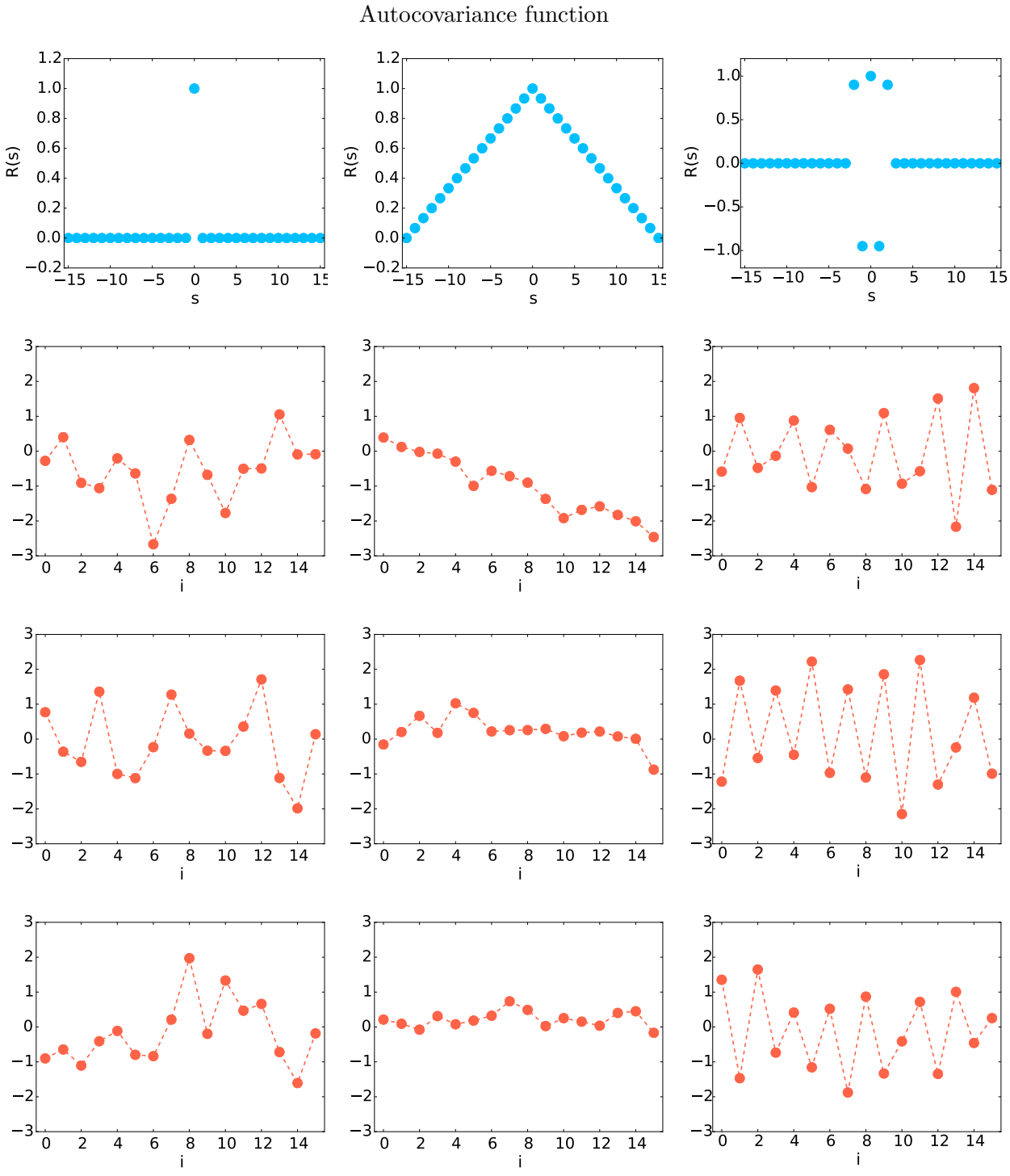


Figure 5.2: Realizations (bottom three rows) of Gaussian processes with zero mean and the autocovariance functions shown on the top row.

Note that any strictly stationary process is necessarily weakly stationary because its first and second-order distributions are shift invariant.

Figure 5.2 shows several stationary random processes with different autocovariance functions. If the autocovariance function is zero everywhere except at the origin, then the values of the random processes at different points are uncorrelated. This results in erratic fluctuations. When the autocovariance at neighboring times is high, the trajectory random process becomes smoother. The autocorrelation can also induce more structured behavior, as in the right column of the figure. In that example $\widetilde{X}(i)$ is negatively correlated with its two neighbors $\widetilde{X}(i-1)$ and $\widetilde{X}(i+1)$, but positively correlated with $\widetilde{X}(i-2)$ and $\widetilde{X}(i+2)$. This results in rapid periodic fluctuations.

5.3 Independent identically-distributed sequences

An independent identically-distributed (iid) sequence $\widetilde{X}$ is a discrete-time random process where $\widetilde{X}(i)$ has the same distribution for any fixed i and $\widetilde{X}(i_1)$, $\widetilde{X}(i_2)$, $\ldots$, $\widetilde{X}(i_n)$ are mutually independent for any n fixed indices and any $n \geq 2$. If $\widetilde{X}(i)$ is a discrete random variable (or equivalently the state space of the random process is discrete), then we denote the pmf associated to the distribution of each entry by $p_{\widetilde{X}}$. This pdf completely characterizes the random process, since for any n indices $i_1, i_2, \ldots, i_n$ and any n:

$$p_{\widetilde{X}(i_1),\widetilde{X}(i_2),\ldots,\widetilde{X}(i_n)}(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) = \prod_{i=1}^{n} p_{\widetilde{X}}(x_i). \tag{5.13}$$

Note that the distribution that does not vary if we shift every index by the same amount, so the process is strictly stationary.

Similarly, if $\widetilde{X}(i)$ is a continuous random variable, then we denote the pdf associated to the distribution by $f_{\widetilde{X}}$. For any n indices $i_1, i_2, \ldots, i_n$ and any n we have

$$f_{\widetilde{X}(i_1),\widetilde{X}(i_2),\ldots,\widetilde{X}(i_n)}(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) = \prod_{i=1}^{n} f_{\widetilde{X}}(x_i). \tag{5.14}$$

Figure 5.3 shows several realizations from iid sequences which follow a uniform and a geometric distribution.

The mean of an iid random sequence is constant and equal to the mean of its associated distribution, which we denote by μ,

$$\mu_{\widetilde{X}}(i) := \mathrm{E}\left(\widetilde{X}(i)\right) \tag{5.15}$$

$$= \mu. \tag{5.16}$$

Let us denote the variance of the distribution associated to the iid sequence by σ^2. The autocovariance function is given by

$$R_{\widetilde{X}}(i,j) := \mathrm{E}\left(\widetilde{X}(i)\,\widetilde{X}(j)\right) - \mathrm{E}\left(\widetilde{X}(i)\right)\mathrm{E}\left(\widetilde{X}(j)\right) \tag{5.17}$$

$$= \begin{cases} \sigma^2, \\ 0. \end{cases} \tag{5.18}$$

This is not surprising, $\widetilde{X}(i)$ and $\widetilde{X}(j)$ are independent for all $i \neq j$, so they are also uncorrelated.

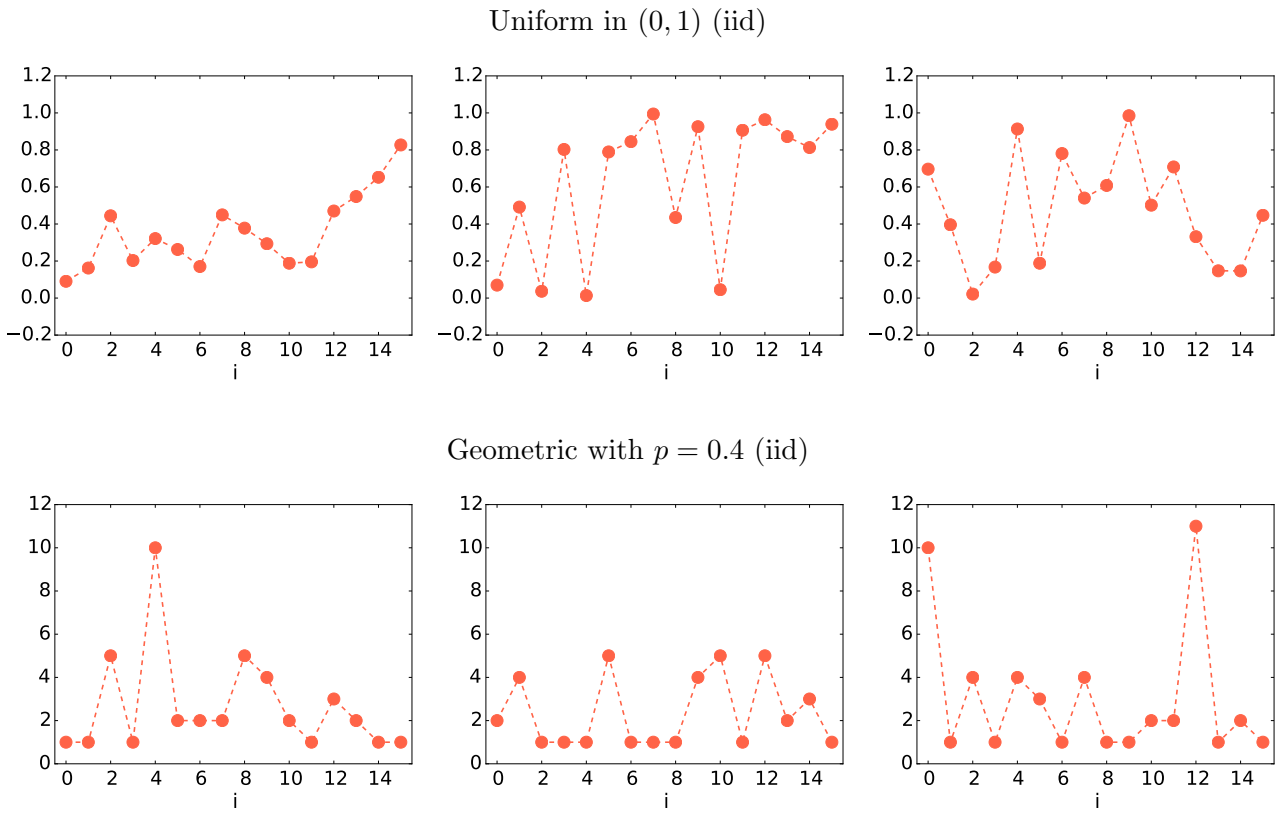


Figure 5.3: Realizations of an iid uniform sequence in $(0, 1)$ (first row) and an iid geometric sequence with parameter $p = 0.4$ (second row).

5.4 Gaussian process

A random process $\widetilde{X}$ is Gaussian if any set of samples is a Gaussian random vector. A Gaussian process $\widetilde{X}$ is fully characterized by its mean function $\mu_{\widetilde{X}}$ and its autocovariance function $R_{\widetilde{X}}$. For all t_1, t_2, ..., t_n and any $n \geq 1$, the random vector

$$\vec{X} := \begin{bmatrix} \widetilde{X}(t_1) \\ \widetilde{X}(t_2) \\ \cdots \\ \widetilde{X}(t_n) \end{bmatrix} \tag{5.19}$$

is a Gaussian random vector with mean

$$\vec{\mu}_{\vec{X}} := \begin{bmatrix} \mu_{\widetilde{X}}(t_1) \\ \mu_{\widetilde{X}}(t_2) \\ \cdots \\ \mu_{\widetilde{X}}(t_n) \end{bmatrix} \tag{5.20}$$

and covariance matrix

$$\Sigma_{\vec{X}} := \begin{bmatrix} R_{\widetilde{X}}(t_1,t_1) & R_{\widetilde{X}}(t_1,t_2) & \cdots & R_{\widetilde{X}}(t_1,t_n) \\ R_{\widetilde{X}}(t_1,t_2) & R_{\widetilde{X}}(t_2,t_2) & \cdots & R_{\widetilde{X}}(t_2,t_n) \\ \vdots & \vdots & \ddots & \vdots \\ R_{\widetilde{X}}(t_2,t_n) & R_{\widetilde{X}}(t_2,t_n) & \cdots & R_{\widetilde{X}}(t_n,t_n) \end{bmatrix} \tag{5.21}$$

Figure 5.2 shows realizations of several discrete Gaussian processes with different autocovariance functions. Sampling from a Gaussian random process boils down to sampling a Gaussian random vector with the appropriate mean and covariance matrix.

Algorithm 5.4.1 (Generating a Gaussian random process)**.** *To sample from an Gaussian random process with mean function $\mu_{\widetilde{X}}$ and autocovariance function $\Sigma_{\widetilde{X}}$ at n points $t_1, \ldots, t_n$ we:*

1. *Compute the mean vector $\vec{\mu}_{\vec{X}}$ given by (5.20) and the covariance matrix $\Sigma_{\vec{X}}$ given by (5.21).*
2. *Generate n independent samples from a standard Gaussian.*
3. *Color the samples according to $\Sigma_{\vec{X}}$ and center them around $\vec{\mu}_{\vec{X}}$, as described in Algorithm 4.3.15.*

5.5 Poisson process

In Example 2.2.8 we motivate the definition of Poisson random variable by deriving the distribution of the number of events that occur in a fixed time interval under the following conditions:

1. Each event occurs independently from every other event.
2. Events occur uniformly.
3. Events occur at a rate of λ events per time interval.

We now assume that these conditions hold in the semi-infinite interval $[0, \infty)$ and define a random process $\widetilde{N}$ that counts the events. To be clear $\widetilde{N}(t)$ is the number of events that happen between 0 and t.

By the same reasoning as in Example 2.2.8, the distribution of the random variable $\widetilde{N}(t_2) - \widetilde{N}(t_1)$, which represents the number of events that occur between t_1 and t_2, is a Poisson random variable with parameter $\lambda(t_2 - t_1)$. This holds for any t_1 and t_2. In addition the random variables $\widetilde{N}(t_2) - \widetilde{N}(t_1)$ and $\widetilde{N}(t_4) - \widetilde{N}(t_3)$ are independent as along as the intervals $[t_1, t_2]$ and (t_3, t_4) do not overlap by Condition 1. A Poisson process is a discrete-state continuous random process that satisfies these two properties.

Poisson processes are often used to model events such as earthquakes, telephone calls, decay of radioactive particles, neural spikes, etc. Figure 2.6 shows an example of a real scenario where the number of calls received at a call center is well approximated as a Poisson process (as long as we only consider a few hours). Note that here we are using the word *event* to mean *something that happens*, such as the arrival of an email, instead of a set within a sample space, which is the meaning that it usually has elsewhere in these notes.

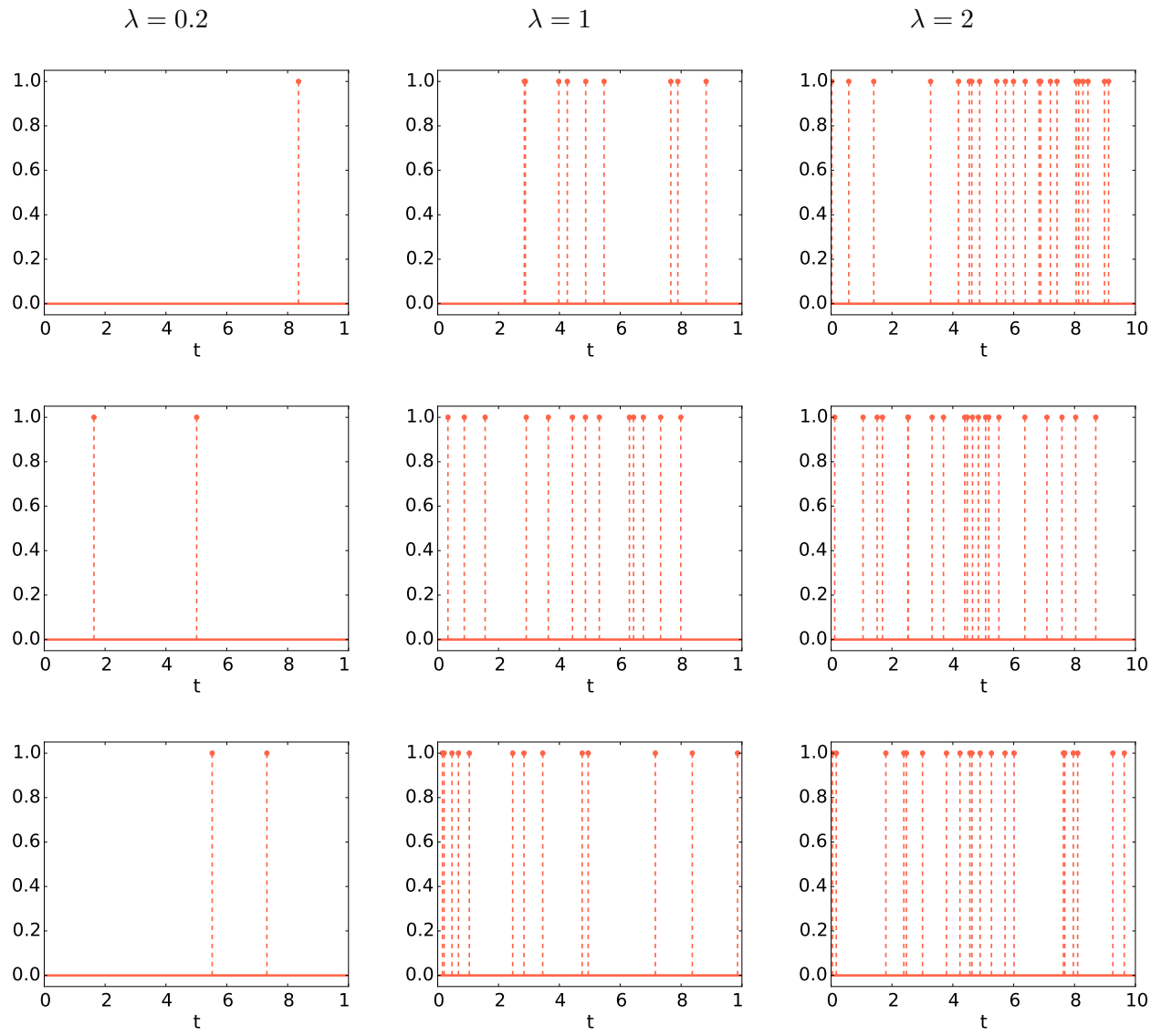


Figure 5.4: Events corresponding to the realizations of a Poisson process $\widetilde{N}$ for different values of the parameter λ. $\widetilde{N}(t)$ equals the number of events up to time t.

Definition 5.5.1 (Poisson process)**.** *A Poisson process with parameter λ is a discrete-state continuous random process $\widetilde{N}$ such that*

1. $\widetilde{N}(0) = 0$.
2. *For any $t_1 < t_2 < t_3 < t_4$ $\widetilde{N}(t_2) - \widetilde{N}(t_1)$ is a Poisson random variable with parameter $\lambda(t_2 - t_1)$.*
3. *For any $t_1 < t_2 < t_3 < t_4$ the random variables $\widetilde{N}(t_2) - \widetilde{N}(t_1)$ and $\widetilde{N}(t_4) - \widetilde{N}(t_3)$ are independent.*

We now check that the random process is well defined, by proving that we can derive the joint pmf of $\widetilde{N}$ at any n points $t_1 < t_2 < \ldots < t_n$ for any $n \geq 0$. To alleviate notation let $p\left(\tilde{\lambda}, x\right)$ be the value of the pmf of a Poisson random variable with parameter $\tilde{\lambda}$ at x, i.e.

$$p\left(\tilde{\lambda}, x\right) := \frac{\tilde{\lambda}^x \, e^{-\tilde{\lambda}}}{x!}. \tag{5.22}$$

We have

$$\begin{aligned}
&p_{\widetilde{N}(t_1),\ldots,\widetilde{N}(t_n)}\left(x_1, \ldots, x_n\right) && (5.23)\\
&= \mathrm{P}\left(\widetilde{N}\left(t_1\right) = x_1, \ldots, \widetilde{N}\left(t_n\right) = x_n\right) && (5.24)\\
&= \mathrm{P}\left(\widetilde{N}\left(t_1\right) = x_1, \widetilde{N}\left(t_2\right) - \widetilde{N}\left(t_1\right) = x_2 - x_1, \ldots, \widetilde{N}\left(t_n\right) - \widetilde{N}\left(t_{n-1}\right) = x_n - x_{n-1}\right) && (5.25)\\
&= \mathrm{P}\left(\widetilde{N}\left(t_1\right) = x_1\right) \mathrm{P}\left(\widetilde{N}\left(t_2\right) - \widetilde{N}\left(t_1\right) = x_2 - x_1\right) \ldots \mathrm{P}\left(\widetilde{N}\left(t_n\right) - \widetilde{N}\left(t_{n-1}\right) = x_n - x_{n-1}\right)\\
&= p\left(\lambda t_1, x_1\right) p\left(\lambda\left(t_2 - t_1\right), x_2 - x_1\right) \ldots p\left(\lambda\left(t_n - t_{n-1}\right), x_n - x_{n-1}\right). && (5.26)
\end{aligned}$$

In words, we have expressed the event that $\widetilde{N}\left(t_i\right) = x_i$ for $1 \leq i \leq n$ in terms of the random variables $\widetilde{N}\left(t_1\right)$ and $\widetilde{N}\left(t_i\right) - \widetilde{N}\left(t_{i-1}\right)$, $2 \leq i \leq n$, which are independent Poisson random variables with parameters λt_1 and $\lambda\left(t_i - t_{i-1}\right)$ respectively.

Figure 5.4 shows several sequences of events corresponding to the realizations of a Poisson process $\widetilde{N}$ for different values of the parameter λ ($\widetilde{N}\left(t\right)$ equals the number of events up to time t). Interestingly, the interarrival time of the events, i.e. the time between contiguous events, always has the same distribution: it is an exponential random variable.

Lemma 5.5.2 (Interarrival times of a Poisson process are exponential)**.** *Let T denote the time between two contiguous events in a Poisson process with parameter λ. T is an exponential random variable with parameter λ.*

The proof is in Section 5.7.1 of the appendix. Figure 2.11 shows that the interarrival times of telephone calls at a call center are indeed well modeled as exponential.

Lemma 5.5.2 suggests that to simulate a Poisson process all we need to do is sample from an exponential distribution.

Algorithm 5.5.3 (Generating a Poisson random process)**.** *To sample from a Poisson random process with parameter λ we:*

1. *Generate independent samples from an exponential random variable with parameter λ t_1, t_2, t_3, ...*
2. *Set the events of the Poisson process to occur at t_1, $t_1 + t_2$, $t_1 + t_2 + t_3$, ...*

Figure 5.4 was generated in this way. To confirm that the algorithm allows to sample from a Poisson process, we would have to prove that the resulting process satisfies the conditions in Definition 5.5.1. This is indeed the case, but we omit the proof.

The following lemma, which derives the mean and autocovariance functions of a Poisson process is proved in Section 5.7.2.

Lemma 5.5.4 (Mean and autocovariance of a Poisson process)**.** *The mean and autocovariance of a Poisson process equal*

$$\mathrm{E}\left(\widetilde{X}\left(t\right)\right) = \lambda t, \tag{5.27}$$

$$R_{\widetilde{X}}\left(t_1, t_2\right) = \lambda \min\left\{t_1, t_2\right\}. \tag{5.28}$$

The mean of the Poisson process is not constant and its autocovariance is not shift-invariant, so the process is neither strictly nor wide-sense stationary.

Example 5.5.5 (Earthquakes)**.** The number of earthquakes with intensity at least 3 on the Richter scale occurring in the San Francisco peninsula is modeled using a Poisson process with parameter 0.3 earthquakes/year. What is the probability that there are no earthquakes in the next ten years and then at least one earthquake over the following twenty years?

We define a Poisson process $\widetilde{X}$ with parameter 0.3 to model the problem. The number of earthquakes in the next 10 years, i.e. $\widetilde{X}\left(10\right)$, is a Poisson random variable with parameter $0.3 \cdot 10 = 3$. The earthquakes in the following 20 years, $\widetilde{X}\left(30\right) - \widetilde{X}\left(10\right)$, are Poisson with parameter $0.3 \cdot 20 = 6$. The two random variables are independent because the intervals do not overlap.

$$\mathrm{P}\left(\widetilde{X}\left(10\right) = 0, \widetilde{X}\left(30\right) \geq 1\right) = \mathrm{P}\left(\widetilde{X}\left(10\right) = 0, \widetilde{X}\left(30\right) - \widetilde{X}\left(10\right) \geq 1\right) \tag{5.29}$$

$$= \mathrm{P}\left(\widetilde{X}\left(10\right) = 0\right) \mathrm{P}\left(\widetilde{X}\left(30\right) - \widetilde{X}\left(10\right) \geq 1\right) \tag{5.30}$$

$$= \mathrm{P}\left(\widetilde{X}\left(10\right) = 0\right) \left(1 - \mathrm{P}\left(\widetilde{X}\left(30\right) - \widetilde{X}\left(10\right) = 0\right)\right) \tag{5.31}$$

$$= e^{-3}\left(1 - e^{-6}\right) = 4.97\,10^{-2}. \tag{5.32}$$

The probability is 4.97%.

△

5.6 Random walk

A random walk is a discrete-time random process that models a sequence of steps in random directions. To specify a random walk formally, we first define an iid sequence of steps $\widetilde{S}$ such that

$$\widetilde{S}\left(i\right) = \begin{cases} +1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2}. \end{cases} \tag{5.33}$$

We define a random walk $\widetilde{X}$ as the discrete-state discrete-time random process

$$\widetilde{X}\left(i\right) := \begin{cases} 0 & \text{for } i = 0, \\ \sum_{j=1}^{i} \widetilde{S}\left(j\right) & \text{for } i = 1, 2, \ldots \end{cases} \tag{5.34}$$

We have specified $\widetilde{X}$ as a function of an iid sequence, so it is well defined. Figure 5.5 shows several realizations of the random walk.

$\widetilde{X}$ is symmetric (there is the same probability of taking a positive step and a negative step) and begins at the origin. It is easy to define variations where the walk is non-symmetric and begins

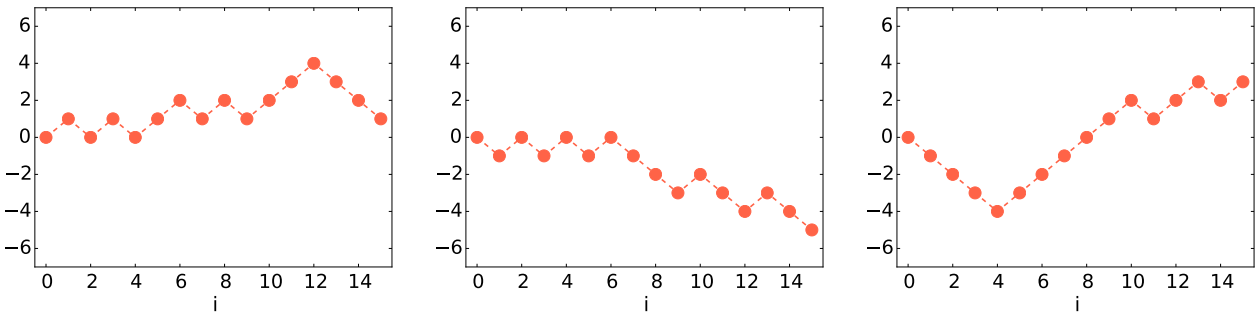


Figure 5.5: Realizations of the random walk defined in Section 5.5.

at another point. Generalizations to higher dimensional spaces– for instance to model random processes on a 2D surface– are also possible.

We derive the first-order pmf of the random walk in the following lemma, proved in Section 5.7.3 of the appendix.

Lemma 5.6.1 (First-order pmf of a random walk)**.** *The first-order pmf of the random walk* $\widetilde{X}$ *is*

$$p_{\widetilde{X}(i)}(x) = \begin{cases} \binom{i}{\frac{i+x}{2}} \frac{1}{2^i} & \text{if } i+x \text{ is even and } -i \leq x \leq i \\ 0 & \text{otherwise.} \end{cases} \tag{5.35}$$

The first-order distribution of the random walk is clearly time-dependent, so the random process is not strictly stationary. By the following lemma, the mean of the random walk is constant (it equals zero). The autocovariance, however, is not shift invariant, so the process is not weakly stationary either.

Lemma 5.6.2 (Mean and autocovariance of a random walk)**.** *The mean and autocovariance of the random walk* $\widetilde{X}$ *are*

$$\mu_{\widetilde{X}}(i) = 0, \tag{5.36}$$

$$R_{\widetilde{X}}(i,j) = \min\{i,j\}. \tag{5.37}$$

Proof.

$$\mu_{\widetilde{X}}(i) := \mathrm{E}\left(\widetilde{X}(i)\right) \tag{5.38}$$

$$= \mathrm{E}\left(\sum_{j=1}^{i} \widetilde{S}(j)\right) \tag{5.39}$$

$$= \sum_{j=1}^{i} \mathrm{E}\left(\widetilde{S}(j)\right) \qquad \text{by linearity of expectation} \tag{5.40}$$

$$= 0. \tag{5.41}$$

$$R_{\widetilde{X}}(i,j) := \mathrm{E}\left(\widetilde{X}(i)\,\widetilde{X}(j)\right) - \mathrm{E}\left(\widetilde{X}(i)\right)\mathrm{E}\left(\widetilde{X}(j)\right) \tag{5.42}$$

$$= \mathrm{E}\left(\sum_{k=1}^{i}\sum_{l=1}^{j}\widetilde{S}(k)\,\widetilde{S}(l)\right) \tag{5.43}$$

$$= \mathrm{E}\left(\sum_{k=1}^{\min\{i,j\}}\widetilde{S}(k)^2 + \sum_{k=1}^{i}\sum_{\substack{l=1\\ l\neq k}}^{j}\widetilde{S}(k)\,\widetilde{S}(l)\right) \tag{5.44}$$

$$= \sum_{k=1}^{\min\{i,j\}} 1 + \sum_{k=1}^{i}\sum_{\substack{l=1\\ l\neq k}}^{j}\mathrm{E}\left(\widetilde{S}(k)\right)\mathrm{E}\left(\widetilde{S}(l)\right) \tag{5.45}$$

$$= \min\{i,j\}, \tag{5.46}$$

where (5.45) follows from linearity of expectation and independence. □

The variance of $\widetilde{X}$ at i equals $R_{\widetilde{X}}(i,i) = i$ which means that the standard deviation of the random walk scales as $\sqrt{i}$.

Example 5.6.3 (Gambler)**.** A gambler is playing the following game. A fair coin is flipped sequentially. Every time the result is heads the gambler wins a dollar, every time it lands on tails she loses a dollar. We can model the amount of money earned (or lost) by the gambler as a random walk, as long as the flips are independent. This allows us to estimate that the expected gain equals zero or that the probability that the gambler is up 6 dollars or more after the first 10 flips is

$$\mathrm{P}(\text{gambler is up \$6 or more}) = p_{\widetilde{X}(10)}(6) + p_{\widetilde{X}(10)}(8) + p_{\widetilde{X}(10)}(10) \tag{5.47}$$

$$= \binom{10}{8}\frac{1}{2^{10}} + \binom{10}{9}\frac{1}{2^{10}} + \frac{1}{2^{10}} \tag{5.48}$$

$$= 5.47\,10^{-2}. \tag{5.49}$$

△

5.7 Proofs

5.7.1 Proof of Lemma 5.5.2

We begin by deriving the cdf of T,

$$F_T(t) := \mathrm{P}(T \leq t) \tag{5.50}$$

$$= 1 - \mathrm{P}(T > t) \tag{5.51}$$

$$= 1 - \mathrm{P}(\text{no events in an interval of length } t) \tag{5.52}$$

$$= 1 - e^{-\lambda t} \tag{5.53}$$

because the number of points in an interval of length t follows a Poisson distribution with parameter λt. Differentiating we conclude that

$$f_T(t) = \lambda e^{-\lambda t}. \tag{5.54}$$

5.7.2 Proof of Lemma 5.5.4

By definition the number of events between 0 and t is distributed as a Poisson random variables with parameter $\lambda\, t$ and hence its mean is equal to $\lambda\, t$.

The autocovariance equals

$$R_{\widetilde{X}}\left(t_1, t_2\right) := \mathrm{E}\left(\widetilde{X}\left(t_1\right)\widetilde{X}\left(t_2\right)\right) - \mathrm{E}\left(\widetilde{X}\left(t_1\right)\right)\mathrm{E}\left(\widetilde{X}\left(t_2\right)\right) \tag{5.55}$$

$$= \mathrm{E}\left(\widetilde{X}\left(t_1\right)\widetilde{X}\left(t_2\right)\right) - \lambda^2 t_1 t_2. \tag{5.56}$$

By assumption $\widetilde{X}\left(t_1\right)$ and $\widetilde{X}\left(t_2\right) - \widetilde{X}\left(t_1\right)$ are independent so that

$$\mathrm{E}\left(\widetilde{X}\left(t_1\right)\widetilde{X}\left(t_2\right)\right) = \mathrm{E}\left(\widetilde{X}\left(t_1\right)\left(\widetilde{X}\left(t_2\right) - \widetilde{X}\left(t_1\right)\right) + \widetilde{X}\left(t_1\right)^2\right) \tag{5.57}$$

$$= \mathrm{E}\left(\widetilde{X}\left(t_1\right)\right)\mathrm{E}\left(\widetilde{X}\left(t_2\right) - \widetilde{X}\left(t_1\right)\right) + \mathrm{E}\left(\widetilde{X}\left(t_1\right)^2\right) \tag{5.58}$$

$$= \lambda^2 t_1\left(t_2 - t_1\right) + \lambda t_1 + \lambda^2 t_1^2 \tag{5.59}$$

$$= \lambda^2 t_1 t_2 + \lambda t_1. \tag{5.60}$$

5.7.3 Proof of Lemma 5.6.1

Let us define the number of positive steps S_+ that the random walk takes. Given the assumptions on $\widetilde{S}$, this is a binomial random variable with parameters i and $1/2$. The number of negative steps is $S_- := i - S_+$. In order for $\widetilde{X}\left(i\right)$ to equal x we need for the net number of steps to equal x, which implies

$$x = S_+ - S_- \tag{5.61}$$

$$= 2S_+ - i. \tag{5.62}$$

This means that S_+ must equal $\frac{i+x}{2}$. We conclude that

$$p_{\widetilde{X}(i)}\left(i\right) = \mathrm{P}\left(\sum_{j=0}^{i}\widetilde{S}\left(i\right) = x\right) \tag{5.63}$$

$$= \binom{i}{\frac{i+x}{2}}\frac{1}{2^i} \qquad \text{if } \frac{i+x}{2} \text{ is an integer between 0 and } i. \tag{5.64}$$

Chapter 6

Convergence of Random Processes

In this chapter we study the convergence of discrete random processes. This allows to characterize two phenomena that are fundamental in statistical estimation and probabilistic modeling: the law of large numbers and the central limit theorem.

6.1 Types of convergence

Let us quickly recall the concept of convergence for a deterministic sequence of real numbers x_1, x_2, ... We have

$$\lim_{i \to \infty} x_i = x \tag{6.1}$$

if x_i is arbitrarily close to x as the index i grows. More formally, the sequence converges to x if for any $\epsilon > 0$ there is an index i_0 such that for all indices i greater than i_0 we have $|x_i - x| < \epsilon$. Recall that any realization of a discrete-time random process $\widetilde{X}(\omega, i)$ where we fix the outcome ω is a deterministic sequence. Establishing convergence of such realizations to a fixed number can therefore be achieved by computing the corresponding limit. However, if we consider the random process itself instead of a realization and we want to determine whether it eventually converges to a random variable X, then deterministic convergence no longer makes sense. In this section we describe several alternative definitions of convergence, which allow to extend this concept to random quantities.

6.1.1 Convergence with probability one

Consider a discrete random process $\widetilde{X}$ and a random variable X defined on the same probability space. If we fix an element ω of the sample space Ω, then $\widetilde{X}(i, \omega)$ is a deterministic sequence and $X(\omega)$ is a constant. It is consequently possible to verify whether $\widetilde{X}(i, \omega)$ converges deterministically to $X(\omega)$ as $i \to \infty$ *for that particular value of* ω. In fact, we can ask: what is the probability that this happens? To be precise, this would be the probability that if we draw ω we have

$$\lim_{i \to \infty} \widetilde{X}(i, \omega) = X(\omega). \tag{6.2}$$

If this probability equals one then we say that $\widetilde{X}(i)$ converges to X with probability one.

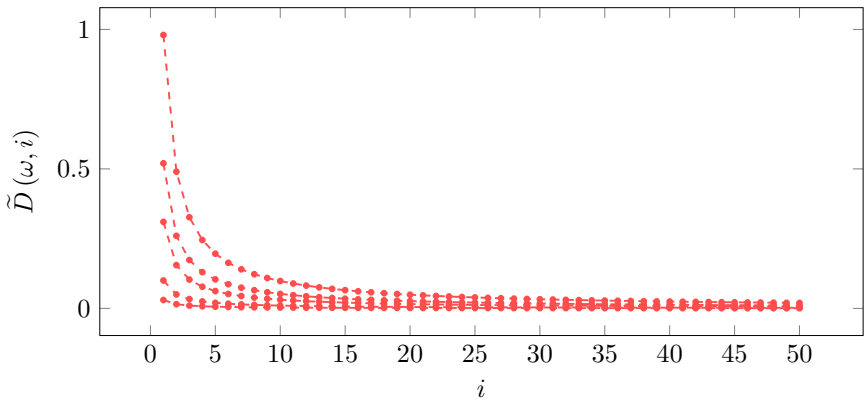


Figure 6.1: Convergence to zero of the discrete random process $\widetilde{D}$ defined in Example 5.1.2.

Definition 6.1.1 (Convergence with probability one)**.** *A discrete random vector $\widetilde{X}$ converges with probability one to a random variable X belonging to the same probability space $(\Omega, \mathcal{F}, \mathrm{P})$ if*

$$\mathrm{P}\left(\left\{\omega \mid \omega \in \Omega, \quad \lim_{i\to\infty} \widetilde{X}(\omega, i) = X(\omega)\right\}\right) = 1. \tag{6.3}$$

Recall that in general the sample space Ω is very difficult to define and manipulate explicitly, except for very simple cases.

Example 6.1.2 (Puddle (continued from Example 5.1.2))**.** Let us consider the discrete random process $\widetilde{D}$ defined in Example 5.1.2. If we fix $\omega \in (0, 1)$

$$\lim_{i\to\infty} \widetilde{D}(\omega, i) = \lim_{i\to\infty} \frac{\omega}{i} \tag{6.4}$$

$$= 0. \tag{6.5}$$

It turns out the realizations tend to zero for all possible values of ω in the sample space. This implies that $\widetilde{D}$ converges to zero with probability one.

△

6.1.2 Convergence in mean square and in probability

To verify convergence with probability one we fix the outcome ω and check whether the corresponding realizations of the random process converge deterministically. An alternative viewpoint is to fix the indexing variable i and consider how close the random variable $\widetilde{X}(i)$ is to another random variable X as we increase i.

A possible measure of the distance between two random variables is the mean square of their difference. If $\mathrm{E}((X - Y)^2) = 0$ then $X = Y$ with probability one by Chebyshev's inequality. The mean square deviation between $\widetilde{X}(i)$ and X is a deterministic quantity (a number), so we can evaluate its convergence as $i \to \infty$. If it converges to zero then we say that the random sequence converges in mean square.

Definition 6.1.3 (Convergence in mean square)**.** *A discrete random process $\widetilde{X}$ converges in mean square to a random variable X belonging to the same probability space if*

$$\lim_{i\to\infty} \mathrm{E}\left(\left(X - \widetilde{X}\left(i\right)\right)^2\right) = 0. \tag{6.6}$$

Alternatively, we can consider the probability that $\widetilde{X}\left(i\right)$ is separated from X by a certain fixed $\epsilon > 0$. If for any ϵ, no matter how small, this probability converges to zero as $i \to \infty$ then we say that the random sequence converges in probability.

Definition 6.1.4 (Convergence in probability)**.** *A discrete random process $\widetilde{X}$ converges in probability to another random variable X belonging to the same probability space if for any $\epsilon > 0$*

$$\lim_{i\to\infty} \mathrm{P}\left(\left|X - \widetilde{X}\left(i\right)\right| > \epsilon\right) = 0. \tag{6.7}$$

Note that as in the case of convergence in mean square, the limit in this definition is deterministic, as it is a limit of probabilities, which are just real numbers.

As a direct consequence of Markov's inequality, convergence in mean square implies convergence in probability.

Theorem 6.1.5. *Convergence in mean square implies convergence in probability.*

Proof. We have

$$\lim_{i\to\infty} \mathrm{P}\left(\left|X - \widetilde{X}\left(i\right)\right| > \epsilon\right) = \lim_{i\to\infty} \mathrm{P}\left(\left(X - \widetilde{X}\left(i\right)\right)^2 > \epsilon^2\right) \tag{6.8}$$

$$\leq \lim_{i\to\infty} \frac{\mathrm{E}\left(\left(X - \widetilde{X}\left(i\right)\right)^2\right)}{\epsilon^2} \quad \text{by Markov's inequality} \tag{6.9}$$

$$= 0, \tag{6.10}$$

if the sequence converges in mean square. □

It turns out that convergence with probability one also implies convergence in probability. Convergence in probability one does not imply convergence in mean square or vice versa. The difference between these three types of convergence is not very important for the purposes of this course.

6.1.3 Convergence in distribution

In some cases, a random process $\tilde{X}$ does not converge to the value of any random variable, but the cdf of $\tilde{X}\left(i\right)$ converges pointwise to the cdf of another random variable X. In that case, the actual values of $\tilde{X}\left(i\right)$ and X are *not* necessarily close, but in the limit they have the same *distribution*. In this case, we say that $\widetilde{X}$ converges in distribution to X.

Definition 6.1.6 (Convergence in distribution)**.** *A random process $\widetilde{X}$ converges in distribution to a random variable X belonging to the same probability space if*

$$\lim_{i\to\infty} F_{\widetilde{X}(i)}\left(x\right) = F_X\left(x\right) \tag{6.11}$$

for all $x \in \mathbb{R}$ where F_X is continuous.

Note that convergence in distribution is a much weaker notion than convergence with probability one, in mean square or in probability. If a discrete random process $\widetilde{X}$ converges to a random variable X in distribution, this only means that as i becomes large the distribution of $\widetilde{X}(i)$ tends to the distribution of X, not that *the values* of the two random variables are close. However, convergence in probability (and hence convergence with probability one or in mean square) does imply convergence in distribution.

Example 6.1.7 (Binomial converges to Poisson)**.** Let us define a discrete random process $\widetilde{X}(i)$ such that the distribution of $\widetilde{X}(i)$ is binomial with parameters i and $p := \lambda/i$. $\widetilde{X}(i)$ and $\widetilde{X}(j)$ are independent for $i \neq j$, which completely characterizes the n-order distributions of the process for all $n > 1$. Consider a Poisson random variable X with parameter λ that is independent of $\tilde{X}(i)$ for all i. Do you expect the values of X and $\tilde{X}(i)$ to be close as $i \to \infty$?

No! In fact even $\tilde{X}(i)$ and $\tilde{X}(i+1)$ will not be close in general. However, $\tilde{X}$ converges in distribution to X, as established in Example 2.2.8:

$$\lim_{i\to\infty} p_{\widetilde{X}(i)}(x) = \lim_{i\to\infty} \binom{i}{x} p^x (1-p)^{(i-x)} \tag{6.12}$$

$$= \frac{\lambda^x e^{-\lambda}}{x!} \tag{6.13}$$

$$= p_X(x). \tag{6.14}$$

△

6.2 Law of large numbers

Let us define the average of a discrete random process.

Definition 6.2.1 (Moving average)**.** *The moving or running average $\widetilde{A}$ of a discrete random process $\widetilde{X}$, defined for $i = 1, 2, \ldots$ (i.e. 1 is the starting point), is equal to*

$$\widetilde{A}(i) := \frac{1}{i} \sum_{j=1}^{i} \widetilde{X}(j). \tag{6.15}$$

Consider an iid sequence. A very natural interpretation for the moving average is that it is a real-time estimate of the mean. In fact, in statistical terms the moving average is the sample mean of the process up to time i (the sample mean is defined in Chapter 8). The law of large numbers establishes that the average does indeed converge to the mean of the iid sequence.

Theorem 6.2.2 (Weak law of large numbers)**.** *Let $\widetilde{X}$ be an iid discrete random process with mean $\mu_{\widetilde{X}} := \mu$ such that the variance of $\widetilde{X}(i)$ σ^2 is bounded. Then the average $\widetilde{A}$ of $\widetilde{X}$ converges in mean square to μ.*

Proof. First, we establish that the mean of $\widetilde{A}(i)$ is constant and equal to μ,

$$\mathrm{E}\left(\widetilde{A}(i)\right) = \mathrm{E}\left(\frac{1}{i}\sum_{j=1}^{i}\widetilde{X}(j)\right) \tag{6.16}$$

$$= \frac{1}{i}\sum_{j=1}^{i}\mathrm{E}\left(\widetilde{X}(j)\right) \tag{6.17}$$

$$= \mu. \tag{6.18}$$

Due to the independence assumption, the variance scales linearly in i. Recall that for independent random variables the variance of the sum equals the sum of the variances,

$$\mathrm{Var}\left(\widetilde{A}(i)\right) = \mathrm{Var}\left(\frac{1}{i}\sum_{j=1}^{i}\widetilde{X}(j)\right) \tag{6.19}$$

$$= \frac{1}{i^2}\sum_{j=1}^{i}\mathrm{Var}\left(\widetilde{X}(j)\right) \tag{6.20}$$

$$= \frac{\sigma^2}{i}. \tag{6.21}$$

We conclude that

$$\lim_{i\to\infty}\mathrm{E}\left(\left(\widetilde{A}(i)-\mu\right)^2\right) = \lim_{i\to\infty}\mathrm{E}\left(\left(\widetilde{A}(i)-\mathrm{E}\left(\widetilde{A}(i)\right)\right)^2\right) \quad \text{by (6.18)} \tag{6.22}$$

$$= \lim_{i\to\infty}\mathrm{Var}\left(\widetilde{A}(i)\right) \tag{6.23}$$

$$= \lim_{i\to\infty}\frac{\sigma^2}{i} \quad \text{by (6.21)} \tag{6.24}$$

$$= 0. \tag{6.25}$$

□

By Theorem 6.1.5 the average also converges to the mean of the iid sequence in probability. In fact, one can also prove convergence with probability one under the same assumptions. This result is known as the strong law of large numbers, but the proof is beyond the scope of these notes. We refer the interested reader to more advanced texts in probability theory.

Figure 6.2 shows averages of realizations of several iid sequences. When the iid sequence is Gaussian or geometric we observe convergence to the mean of the distribution, however when the sequence is Cauchy the moving average diverges. The reason is that, as shown in Example 4.2.2, the Cauchy distribution does not have a well defined mean! Intuitively, extreme values have non-negligeable probability under the Cauchy distribution so from time to time the iid sequence takes values with very large magnitudes and this prevents the moving average from converging.

6.3 Central limit theorem

In the previous section we established that the moving average of a sequence of iid random variables converges to the mean of their distribution (as long as the mean is well defined and

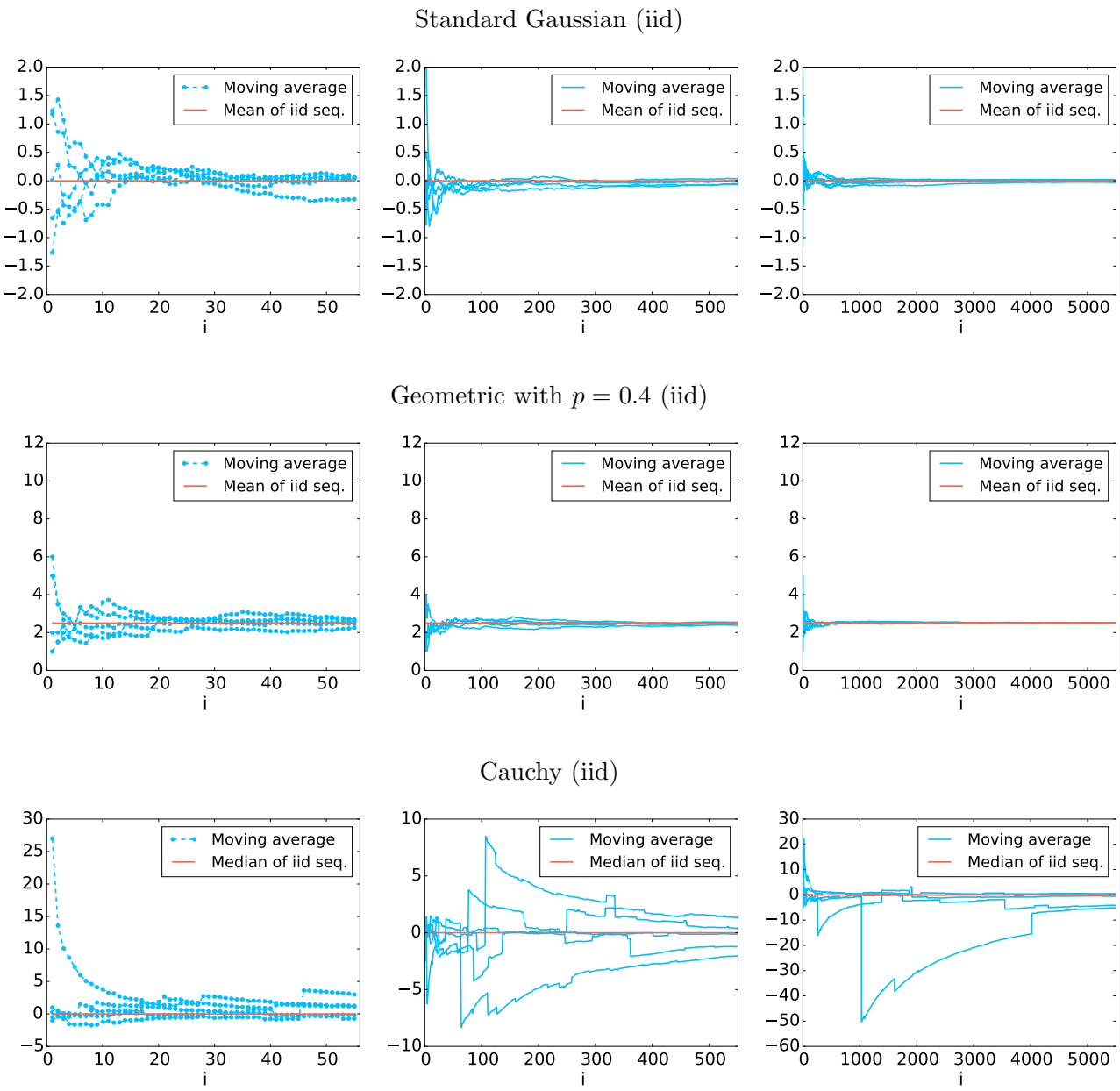


Figure 6.2: Realization of the moving average of an iid standard Gaussian sequence (top), an iid geometric sequence with parameter $p = 0.4$ (center) and an iid Cauchy sequence (bottom).

the variance is finite). In this section, we characterize the *distribution* of the average $\widetilde{A}(i)$ as i increases. It turns out that $\widetilde{A}$ converges to a Gaussian random variable in distribution, which is very useful in statistics as we will see later on.

This result, known as the central limit theorem, justifies the use of Gaussian distributions to model data that are the result of many different independent factors. For example, the distribution of height or weight of people in a certain population often has a Gaussian shape– as illustrated by Figure 2.13– because the height and weight of a person depends on many different factors that are roughly independent. In many signal-processing applications noise is well modeled as having a Gaussian distribution for the same reason.

Theorem 6.3.1 (Central limit theorem)**.** *Let $\widetilde{X}$ be an iid discrete random process with mean $\mu_{\widetilde{X}} := \mu$ such that the variance of $\widetilde{X}(i)$ σ^2 is bounded. The random process $\sqrt{n}(\widetilde{A} - \mu)$, which corresponds to the centered and scaled moving average of $\widetilde{X}$, converges in distribution to a Gaussian random variable with mean 0 and variance σ^2.*

Proof. The proof of this remarkable result is beyond the scope of these notes. It can be found in any advanced text on probability theory. However, we would still like to provide some intuition as to why the theorem holds. Theorem 3.5.2 establishes that the pdf of the sum of two independent random variables is equal to the convolutions of their individual pdfs. The same holds for discrete random variables: the pmf of the sum is equal to the convolution of the pmfs, as long as the random variables are independent.

If each of the entries of the iid sequence has pdf f, then the pdf of the sum of the first i elements can be obtained by convolving f with itself i times

$$f_{\sum_{j=1}^{1} \widetilde{X}(j)}(x) = (f * f * \cdots * f)(x). \tag{6.26}$$

If the sequence has a discrete state and each of the entries has pmf p, the pmf of the sum of the first i elements can be obtained by convolving p with itself i times

$$p_{\sum_{j=1}^{1} \widetilde{X}(j)}(x) = (p * p * \cdots * p)(x). \tag{6.27}$$

Normalizing by i just results in scaling the result of the convolution, so the pmf or pdf of the moving mean $\widetilde{A}$ is the result of repeated convolutions of a fixed function. These convolutions have a smoothing effect, which eventually transforms the pmf/pdf into a Gaussian! We show this numerically in Figure 6.3 for two very different distributions: a uniform distribution and a very irregular one. Both converge to Gaussian-like shapes after just 3 or 4 convolutions. The central limit theorem makes this precise, establishing that the shape of the pmf or pdf becomes Gaussian asymptotically. □

In statistics the central limit theorem is often invoked to justify treating averages as if they have a Gaussian distribution. The idea is that for large enough n $\sqrt{n}(\widetilde{A} - \mu)$ is approximately Gaussian with mean 0 and variance σ^2, which implies that $\widetilde{A}$ is approximately Gaussian with mean μ and variance σ^2/n. It's important to remember that we have *not* established this rigorously. The rate of convergence will depend on the particular distribution of the entries of the iid sequence.

In practice convergence is usually very fast. Figure 6.4 shows the empirical distribution of the moving average of an exponential and a geometric iid sequence. In both cases the approximation obtained by the central limit theory is very accurate even for an average of 100 samples. The

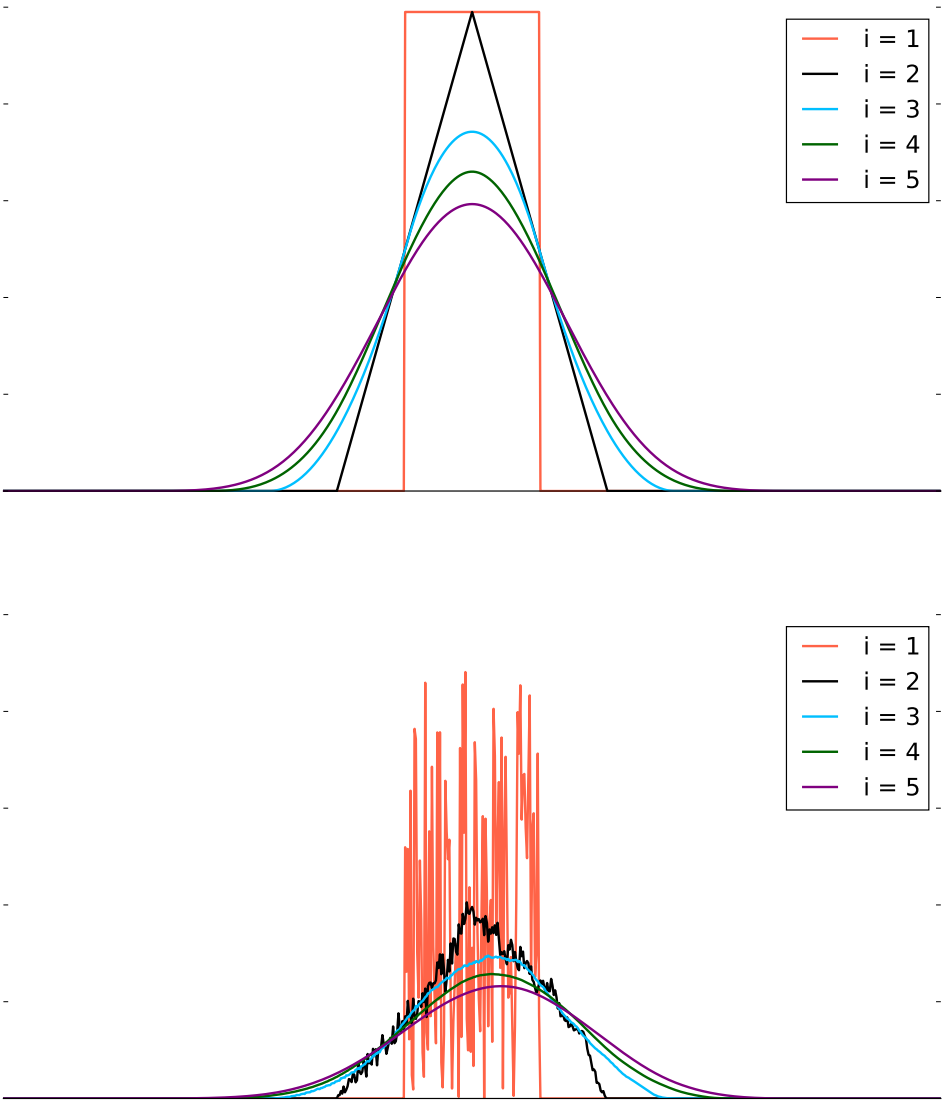


Figure 6.3: Result of convolving two different distributions with themselves several times. The shapes quickly become Gaussian-like.

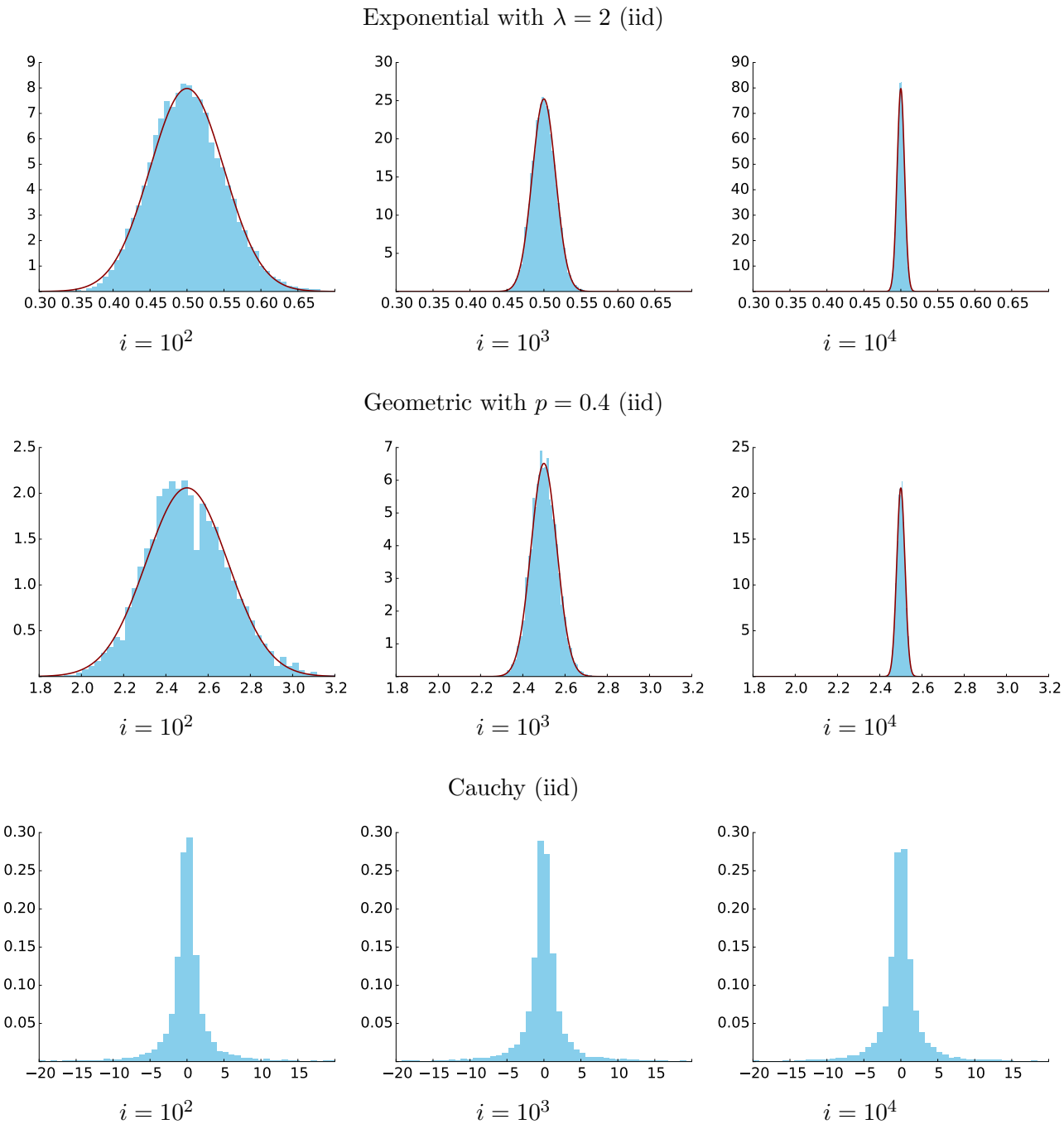


Figure 6.4: Empirical distribution of the moving average of an iid standard Gaussian sequence (top), an iid geometric sequence with parameter $p = 0.4$ (center) and an iid Cauchy sequence (bottom). The empirical distribution is computed from 10^4 samples in all cases. For the two first rows the estimate provided by the central limit theorem is plotted in red.

figure also shows that for a Cauchy iid sequence, the distribution of the moving average does not become Gaussian, which does not contradict the central limit theorem as the distribution does not have a well defined mean. To close this section we derive a useful approximation to the binomial distribution using the central limit theorem.

Example 6.3.2 (Gaussian approximation to the binomial distribution)**.** Let X have a binomial distribution with parameters n and p, such that n is large. Computing the probability that X is in a certain interval requires summing its pmf over all the values in that interval. Alternatively, we can obtain a quick approximation using the fact that for large n the distribution of a binomial random variable is approximately Gaussian. Indeed, we can write X as the sum of n independent Bernoulli random variables with parameter p,

$$X = \sum_{i=1}^{n} B_i. \tag{6.28}$$

The mean of B_i is p and its variance is $p\left(1-p\right)$. By the central limit theorem $\frac{1}{n}X$ is approximately Gaussian with mean p and variance $p\left(1-p\right)/n$. Equivalently, by Lemma 2.5.1, X is approximately Gaussian with mean np and variance $np\left(1-p\right)$.

Assume that a basketball player makes each shot she takes with probability $p = 0.4$. If we assume that each shot is independent, what is the probability that she makes more than 420 shots out of 1000? We can model the shots made as a binomial X with parameters 1000 and 0.4. The exact answer is

$$\begin{aligned} \mathrm{P}\left(X \geq 420\right) &= \sum_{x=420}^{1000} p_X\left(x\right) && (6.29)\\ &= \sum_{x=420}^{1000} \binom{1000}{x} 0.4^x 0.6^{(n-x)} && (6.30)\\ &= 10.4\,10^{-2}. && (6.31) \end{aligned}$$

If we apply the Gaussian approximation, by Lemma 2.5.1 X being larger than 420 is the same as a standard Gaussian U being larger than $\frac{420-\mu}{\sigma}$ where μ and σ are the mean and standard deviation of X, equal to $np = 400$ and $\sqrt{np(1-p)} = 15.5$ respectively.

$$\begin{aligned} \mathrm{P}\left(X \geq 420\right) &\approx \mathrm{P}\left(\sqrt{np\left(1-p\right)}U + np \geq 420\right) && (6.32)\\ &= \mathrm{P}\left(U \geq 1.29\right) && (6.33)\\ &= 1 - \Phi\left(1.29\right) && (6.34)\\ &= 9.85\,10^{-2}. && (6.35) \end{aligned}$$

△

6.4 Monte Carlo simulation

Simulation is a powerful tool in probability and statistics. Probabilistic models are often too complex for us to derive closed-form solutions of the distribution or expectation of quantities of interest, as we do in homework problems.

As an example, imagine that you set up a probabilistic model to determine the probability of winning a game of solitaire. If the cards are well shuffled, this probability equals

$$\mathrm{P}\left(\text{Win}\right) = \frac{\text{Number of permutations that lead to a win}}{\text{Total number}}. \tag{6.36}$$

The problem is that characterizing what permutations lead to a win is very difficult without actually playing out the game to see the outcome. Doing this for every possible permutation is computationally intractable, since there are $52! \approx 8\,10^{67}$ of them. However, there is a simple way to approximate the probability of interest: simulating a large number of games and recording what fraction result in wins. The game of solitaire was precisely what inspired Stanislaw Ulam to propose simulation-based methods, known as the Monte Carlo method (a code name, inspired by the Monte Carlo Casino in Monaco), in the context of nuclear-weapon research in the 1940s:

The first thoughts and attempts I made to practice (the Monte Carlo Method) were suggested by a question which occurred to me in 1946 as I was convalescing from an illness and playing solitaires. The question was what are the chances that a Canfield solitaire laid out with 52 cards will come out successfully? After spending a lot of time trying to estimate them by pure combinatorial calculations, I wondered whether a more practical method than "abstract thinking" might not be to lay it out say one hundred times and simply observe and count the number of successful plays.

This was already possible to envisage with the beginning of the new era of fast computers, and I immediately thought of problems of neutron diffusion and other questions of mathematical physics, and more generally how to change processes described by certain differential equations into an equivalent form interpretable as a succession of random operations. Later, I described the idea to John von Neumann, and we began to plan actual calculations.[1]

Monte Carlo methods use simulation to estimate quantities that are challenging to compute exactly. In this section, we consider the problem of approximating the probability of an event $\mathcal{E}$, as in the *game of solitaire* example.

Algorithm 6.4.1 (Monte Carlo approximation)**.** *To approximate the probability of an event $\mathcal{E}$, we:*

1. *Generate n independent samples from the indicator function $1_{\mathcal{E}}$ associated to the event: I_1, I_2, $\ldots$, I_n.*
2. *Compute the average of the n samples*

$$\widetilde{A}\left(n\right) := \frac{1}{n}\sum_{i=1}^{n} I_i \tag{6.37}$$

which is the estimate for the probability of $\mathcal{E}$

The probability of interest can be interpreted as the expectation of the indicator function $1_{\mathcal{E}}$ associated to the event,

$$\mathrm{E}\left(1_{\mathcal{E}}\right) = \mathrm{P}\left(\mathcal{E}\right). \tag{6.38}$$

By the law of large numbers, the estimate $\widetilde{A}$ converges to the true probability as $n \to \infty$. The following example illustrates the power of this simple technique.

[1] http://en.wikipedia.org/wiki/Monte_Carlo_method#History

Game outcomes			Rank			Probability
1-2	1-3	2-3	R_1	R_2	R_3	
1	1	2	1	2	3	1/6
1	1	3	1	3	2	1/6
1	3	2	1	1	1	1/12
1	3	3	2	3	1	1/12
2	1	2	2	1	3	1/6
2	1	3	1	1	1	1/6
2	3	2	3	1	2	1/12
2	3	3	3	2	1	1/12

Probability mass function

	R_1	R_2	R_3
1	7/12	1/2	5/12
2	1/4	1/4	1/4
3	1/6	1/4	1/3

Table 6.1: The table on the left shows all possible outcomes in a league of three teams ($m = 3$), the resulting ranks for each team and the corresponding probability. The table on the right shows the pmf of the ranks of each of the teams.

Example 6.4.2 (Basketball league)**.** In an intramural basketball league m teams play each other once every season. The teams are ordered according to their past results: team 1 being the best and team m the worst. We model the probability that team i beats team j, for $1 \leq i < j \leq m$ as

$$\mathrm{P}\left(\text{team } j \text{ beats team } i\right) := \frac{1}{j-i+1}. \tag{6.39}$$

The best team beats the second with probability 1/2 and the third with probability 2/3, the second beats the third with probability 1/2, the fourth with probability 2/3 and the fifth with probability 3/4, and so on. We assume that the outcomes of the different games are independent.

At the end of the season, after every team has played with every other team, the teams are ranked according to their number of wins. If several teams have the same number of wins, then they share the same rank. For example, if two teams have the most wins, they both have rank 1, and the next team has rank 3. The goal is to compute the distribution of the final rank of each team in the league, which we model as the random variables $R_1, R_2, \ldots, R_m$. We have all the information to compute the joint pmf of these random variables by applying the law of total probability. As shown in Table 6.1 for $m = 3$, all we need to do is enumerate all the possible outcomes of the games and sum the probabilities of the outcomes that result in a particular rank.

Unfortunately, the number of possible outcomes grows dramatically with m. The number of games equals $m\left(m-1\right)/2$, so the possible outcomes are $2^{m(m-1)/2}$. When there are just 10 teams, this is larger than 10^{13}. Computing the exact distribution of the final ranks for leagues that are not very small is therefore very computationally demanding. Fortunately, Algorithm 6.4.1 offers a more tractable alternative: We can sample a large number of seasons n by simulating each game as a Bernoulli random variable with a parameter given by equation (6.39) and approximate the pmf using the fraction of times that each team ends up in each position. Simulating a whole season only requires sampling $m\left(m-1\right)/2$ games, which can be done very fast.

Table 6.2 illustrates the Monte Carlo approach for $m = 3$. The approximation is quite coarse if we only use $n = 10$ simulated seasons, but becomes very accurate when $n = 2,000$. Figure 6.5

Game outcomes			Rank		
1-2	1-3	2-3	R_1	R_2	R_3
1	3	2	1	1	1
1	1	3	1	3	2
2	1	2	2	1	3
2	3	2	3	1	2
2	1	3	1	1	1
1	1	2	1	2	3
2	1	3	1	1	1
2	3	2	3	1	2
1	1	2	1	2	3
2	3	2	3	1	2

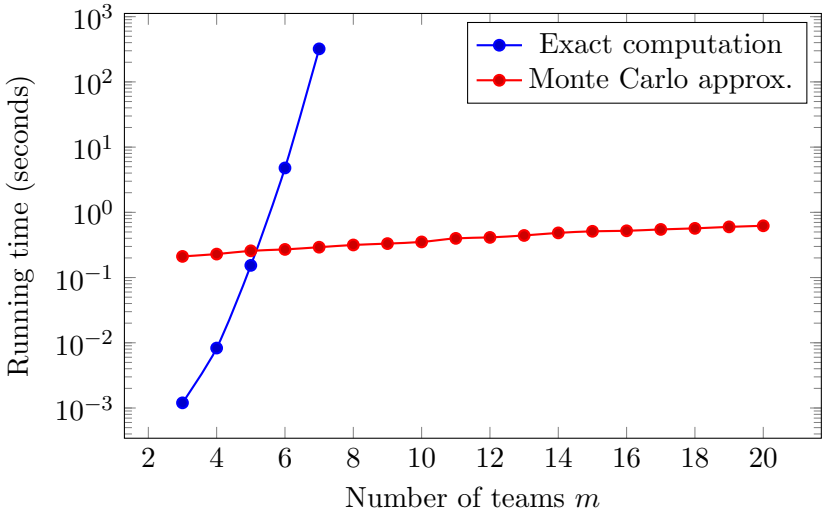
Estimated pmf ($n = 10$)

	R_1	R_2	R_3
1	0.6 (0.583)	0.7 (0.5)	0.3 (0.417)
2	0.1 (0.25)	0.2 (0.25)	0.4 (0.25)
3	0.3 (0.167)	0.1 (0.25)	0.3 (0.333)

Estimated pmf ($n = 2,000$)

	R_1	R_2	R_3
1	0.582 (0.583)	0.496 (0.5)	0.417 (0.417)
2	0.248 (0.25)	0.261 (0.25)	0.244 (0.25)
3	0.171 (0.167)	0.245 (0.25)	0.339 (0.333)

Table 6.2: The table on the left shows 10 simulated outcomes of a league of three teams ($m = 3$) and the resulting ranks. The tables on the right show the estimated pmf obtained by Monte Carlo simulation from the simulated outcomes on the left (top) and from 2,000 simulated outcomes (bottom). The exact values are included in brackets for comparison.



Monte Carlo error

m	Average error
3	$9.28\,10^{-3}$
4	$12.7\,10^{-3}$
5	$7.95\,10^{-3}$
6	$7.12\,10^{-3}$
7	$7.12\,10^{-3}$

Figure 6.5: The graph on the left shows the time needed to obtain the exact pmf of the final ranks in Example 6.4.2 and to approximate them by Monte Carlo approximation using 2,000 simulated league outcomes. The table on the right shows the average error per entry of the Monte Carlo approximation.

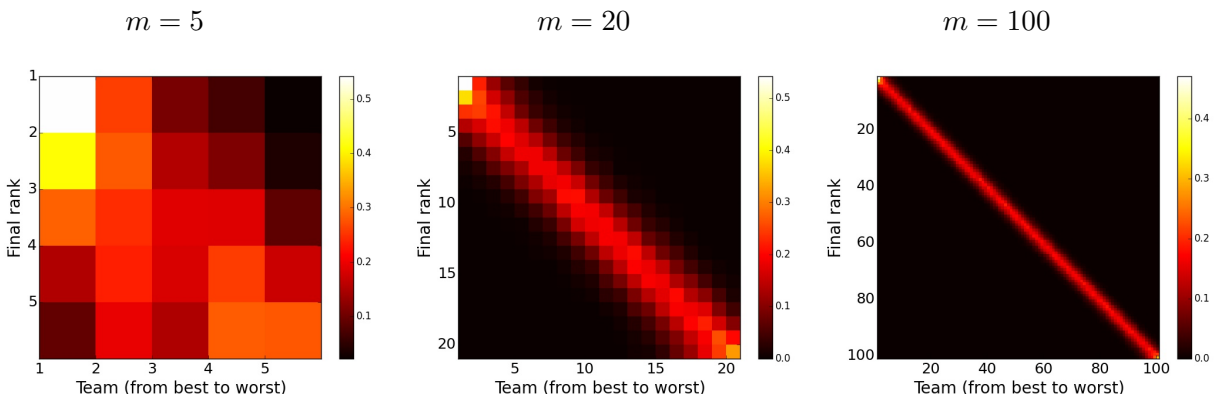


Figure 6.6: Approximate pmf of the final ranks in Example 6.4.2 using 2,000 simulated league outcomes.

shows the running time needed to compute the exact pmf and to approximate it with the Monte Carlo approach for different numbers of teams. When the number of teams is very small the exact computation is very fast, but the running time increases exponentially with m as expected, so that for 7 teams the computation already takes 5 and a half minutes. In contrast, the Monte Carlo approximation is dramatically faster. For $m = 20$ it just takes half a second. Figure 6.6 shows the approximate pmf of the final ranks for 5, 20 and 100 teams. Higher ranks have higher probabilities because when two teams are tied they are awarded the higher rank.

△

Chapter 7

Markov Chains

The Markov property is satisfied by any random process for which *the future is conditionally independent from the past given the present.*

Definition 7.0.1 (Markov property)**.** *A random process satisfies the Markov property if* $\widetilde{X}(t_{i+1})$ *is conditionally independent of* $\widetilde{X}(t_1), \ldots, \widetilde{X}(t_{i-1})$ *given* $\widetilde{X}(t_i)$ *for any* $t_1 < t_2 < \ldots < t_i < t_{i+1}$*. If the state space of the random process is discrete, then for any* x_1*,* x_2*, ...,* x_{i+1}

$$p_{\widetilde{X}(t_{n+1})|\widetilde{X}(t_1),\widetilde{X}(t_2),\ldots,\widetilde{X}(t_i)}(x_{n+1}|x_1, x_2, \ldots, x_n) = p_{\widetilde{X}(t_{i+1})|\widetilde{X}(t_i)}(x_{i+1}|x_i). \tag{7.1}$$

If the state space of the random process is continuous (and the distribution has a joint pdf),

$$f_{\widetilde{X}(t_{i+1})|\widetilde{X}(t_1),\widetilde{X}(t_2),\ldots,\widetilde{X}(t_i)}(x_{i+1}|x_1, x_2, \ldots, x_i) = f_{\widetilde{X}(t_{i+1})|\widetilde{X}(t_i)}(x_{i+1}|x_i). \tag{7.2}$$

Figure 7.1 shows the directed graphical model that corresponds to the dependence assumptions implied by the Markov property. Any iid sequence satisfies the Markov property, since all conditional pmfs or pdfs are just equal to the marginals (in this case there would be no edges in the directed acyclic graph of Figure 7.1). The random walk also satisfies the property, since once we fix where the walk is at a certain time i the path that it took before i has no influence in its next steps.

Lemma 7.0.2. *The random walk satisfies the Markov property.*

Proof. Let $\widetilde{X}$ denote the random walk defined in Section 5.6. Conditioned on $\widetilde{X}(j) = x_i$ for $j \leq i$, $\widetilde{X}(i+1)$ equals $x_i + \widetilde{S}(i+1)$. This does not depend on $x_1, \ldots, x_{i-1}$, which implies (7.1). □

7.1 Time-homogeneous discrete-time Markov chains

A Markov chain is a random process that satisfies the Markov property. Here we consider **discrete-time** Markov chains with a **finite state space**, which means that the process can only take a finite number of values at any given time point. To specify a Markov chain, we only need to define the pmf of the random process at its starting point (which we will assume is at $i = 0$) and its transition probabilities. This follows from the Markov property, since for any

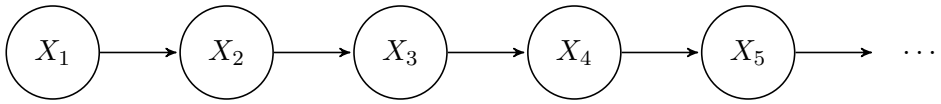


Figure 7.1: Directed graphical model describing the dependence assumptions implied by the Markov property.

$n \geq 0$

$$p_{\widetilde{X}(0),\widetilde{X}(1),\ldots,\widetilde{X}(n)}\left(x_0, x_1, \ldots, x_n\right) := \prod_{i=0}^{n} p_{\widetilde{X}(i)|\widetilde{X}(0),\ldots,\widetilde{X}(i-1)}\left(x_i | x_0, \ldots, x_{i-1}\right) \tag{7.3}$$

$$= \prod_{i=0}^{n} p_{\widetilde{X}(i)|\widetilde{X}(i-1)}\left(x_i | x_{i-1}\right). \tag{7.4}$$

If these transition probabilities are the same at every time step (i.e. they are constant and do not depend on i), then the Markov chain is said to be **time homogeneous**. In this case, we can store the probability of each possible transition in an $s \times s$ matrix $T_{\widetilde{X}}$, where s is the number of states.

$$\left(T_{\widetilde{X}}\right)_{jk} := p_{\widetilde{X}(i+1)|\widetilde{X}(i)}\left(x_j | x_k\right). \tag{7.5}$$

In this chapter we focus on time-homogeneous finite-state Markov chains. The transition probabilities of these chains can be visualized using a state diagram, which shows each state and the probability of every possible transition. See Figure 7.2 below for an example. The state diagram should not be confused with the directed acyclic graph (DAG) that represents the dependence structure of the model, illustrated in Figure 7.1. In the state diagram, each node corresponds to a state and the edges to transition probabilities between states, whereas the DAG just indicates the dependence structure of the random process in time and is usually the same for all Markov chains.

To simplify notation we define an s-dimensional vector $\vec{p}_{\widetilde{X}(i)}$ called the **state vector**, which contains the marginal pmf of the Markov chain at each time i,

$$\vec{p}_{\widetilde{X}(i)} := \begin{bmatrix} p_{\widetilde{X}(i)}\left(x_1\right) \\ p_{\widetilde{X}(i)}\left(x_2\right) \\ \cdots \\ p_{\widetilde{X}(i)}\left(x_s\right) \end{bmatrix}. \tag{7.6}$$

Each entry in the state vector contains the probability that the Markov chain is in that particular state at time i. It is *not* the value of the Markov chain, which is a random variable.

The initial state space $\vec{p}_{\widetilde{X}(0)}$ and the transition matrix $T_{\widetilde{X}}$ suffice to completely specify a time-homogeneous finite-state Markov chain. Indeed, we can compute the joint distribution of the chain at any n time points $i_1, i_2, \ldots, i_n$ for any $n \geq 1$ from $\vec{p}_{\widetilde{X}(0)}$ and $T_{\widetilde{X}}$ by applying (7.4) and marginalizing over any times that we are not interested in. We illustrate this in the following example.

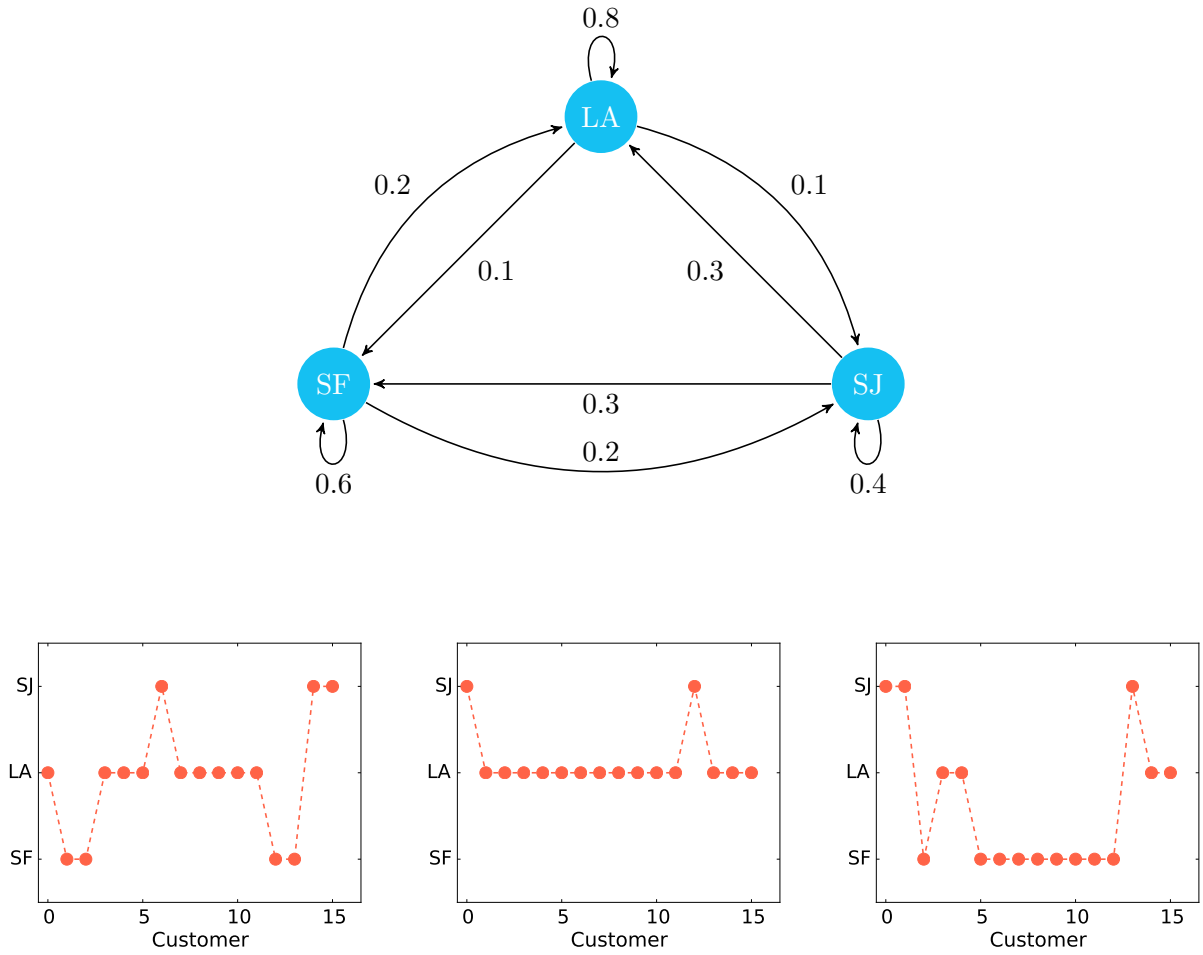


Figure 7.2: State diagram of the Markov chain described in Example (7.1.1) (top). Each arrow shows the probability of a transition between the two states. Below we show three realizations of the Markov chain.

Example 7.1.1 (Car rental)**.** A car-rental company hires you to model the location of their cars. The company operates in Los Angeles, San Francisco and San Jose. Customers regularly take a car in a city and drop it off in another. It would be very useful for the company to be able to compute how likely it is for a car to end up in a given city. You decide to model the location of the car as a Markov chain, where each time step corresponds to a new customer taking the car. The company allocates new cars evenly between the three cities. The transition probabilities, obtained from past data, are given by

$$
\begin{array}{c}
\begin{array}{ccc} \text{San Francisco} & \text{Los Angeles} & \text{San Jose} \end{array} \\
\begin{pmatrix} 0.6 & 0.1 & 0.3 \\ 0.2 & 0.8 & 0.3 \\ 0.2 & 0.1 & 0.4 \end{pmatrix}
\begin{array}{c} \text{San Francisco} \\ \text{Los Angeles} \\ \text{San Jose} \end{array}
\end{array}
$$

To be clear, the probability that a customer moves the car from San Francisco to LA is 0.2, the

probability that the car stays in San Francisco is 0.6, and so on.

The initial state vector and the transition matrix of the Markov chain are

$$\vec{p}_{\widetilde{X}(0)} := \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}, \qquad T_{\widetilde{X}} := \begin{bmatrix} 0.6 & 0.1 & 0.3 \\ 0.2 & 0.8 & 0.3 \\ 0.2 & 0.1 & 0.4 \end{bmatrix}. \tag{7.7}$$

State 1 is assigned to *San Francisco*, state 2 to *Los Angeles* and state 3 to *San Jose*. Figure 7.2 shows a state diagram of the Markov chain. Figure 7.2 shows some realizations of the Markov chain.

The company wants to find out the probability that the car starts in San Francisco, but is in San Jose right after the second customer. This is given by

$$\begin{aligned} p_{\widetilde{X}(0),\widetilde{X}(2)}(1,3) &= \sum_{i=1}^{3} p_{\widetilde{X}(0),\widetilde{X}(1),\widetilde{X}(2)}(1,i,3) && (7.8)\\ &= \sum_{i=1}^{3} p_{\widetilde{X}(0)}(1)\, p_{\widetilde{X}(1)|\widetilde{X}(0)}(i|1)\, p_{\widetilde{X}(2)|\widetilde{X}(1)}(3|i) && (7.9)\\ &= \left(\vec{p}_{\widetilde{X}(0)}\right)_1 \sum_{i=1}^{3} \left(T_{\widetilde{X}}\right)_{i1} \left(T_{\widetilde{X}}\right)_{3i} && (7.10)\\ &= \frac{0.6 \cdot 0.2 + 0.2 \cdot 0.1 + 0.2 \cdot 0.4}{3} \approx 7.33\, 10^{-2}. && (7.11) \end{aligned}$$

The probability is 7.33%.

△

The following lemma provides a simple expression for the state vector at time i $\vec{p}_{\widetilde{X}(i)}$ in terms of $T_{\widetilde{X}}$ and the previous state vector.

Lemma 7.1.2 (State vector and transition matrix)**.** *For a Markov chain $\widetilde{X}$ with transition matrix $T_{\widetilde{X}}$*

$$\vec{p}_{\widetilde{X}(i)} = T_{\widetilde{X}}\, \vec{p}_{\widetilde{X}(i-1)}. \tag{7.12}$$

If the Markov chain starts at time 0 then

$$\vec{p}_{\widetilde{X}(i)} = T_{\widetilde{X}}^{i}\, \vec{p}_{\widetilde{X}(0)}, \tag{7.13}$$

where $T_{\widetilde{X}}^{i}$ denotes multiplying i times by matrix $T_{\widetilde{X}}$.

Proof. The proof follows directly from the definitions,

$$\vec{p}_{\widetilde{X}(i)} := \begin{bmatrix} p_{\widetilde{X}(i)}(x_1) \\ p_{\widetilde{X}(i)}(x_2) \\ \cdots \\ p_{\widetilde{X}(i)}(x_s) \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{s} p_{\widetilde{X}(i-1)}(x_j)\, p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_1|x_j) \\ \sum_{j=1}^{s} p_{\widetilde{X}(i-1)}(x_j)\, p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_2|x_j) \\ \cdots \\ \sum_{j=1}^{s} p_{\widetilde{X}(i-1)}(x_j)\, p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_s|x_j) \end{bmatrix} \tag{7.14}$$

$$= \begin{bmatrix} p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_1|x_1) & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_1|x_2) & \cdots & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_1|x_s) \\ p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_2|x_1) & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_2|x_2) & \cdots & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_2|x_s) \\ & & \cdots & \\ p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_s|x_1) & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_s|x_2) & \cdots & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_s|x_s) \end{bmatrix} \begin{bmatrix} p_{\widetilde{X}(i-1)}(x_1) \\ p_{\widetilde{X}(i-1)}(x_2) \\ \cdots \\ p_{\widetilde{X}(i-1)}(x_s) \end{bmatrix}$$

$$= T_{\widetilde{X}}\, \vec{p}_{\widetilde{X}(i-1)} \tag{7.15}$$

Equation (7.13) is obtained by applying (7.12) i times and taking into account the Markov property. □

Example 7.1.3 (Car rental (continued))**.** The company wants to estimate the distribution of locations right after the 5th customer has used a car. Applying Lemma 7.1.2 we obtain

$$\vec{p}_{\widetilde{X}(5)} = T_{\widetilde{X}}^{5}\, \vec{p}_{\widetilde{X}(0)} \tag{7.16}$$

$$= \begin{bmatrix} 0.281 \\ 0.534 \\ 0.185 \end{bmatrix}. \tag{7.17}$$

The model estimates that after 5 customers more than half of the cars are in Los Angeles.

△

7.2 Recurrence

The states of a Markov chain can be classified depending on whether the Markov chain is guaranteed to always return to them or whether it may eventually stop visiting those states.

Definition 7.2.1 (Recurrent and transient states)**.** *Let $\widetilde{X}$ be a time-homogeneous finite-state Markov chain. We consider a particular state x. If*

$$\mathrm{P}\left(\widetilde{X}(j) = s \text{ for some } j > i \mid \widetilde{X}(i) = s\right) = 1 \tag{7.18}$$

*then the state is **recurrent**. In words, given that the Markov chain is at x, the probability that it returns to x is one. In contrast, if*

$$\mathrm{P}\left(\widetilde{X}(j) \neq s \text{ for all } j > i \mid \widetilde{X}(i) = s\right) > 0 \tag{7.19}$$

*the state is **transient**. Given that the Markov chain is at x, there is nonzero probability that it will never return.*

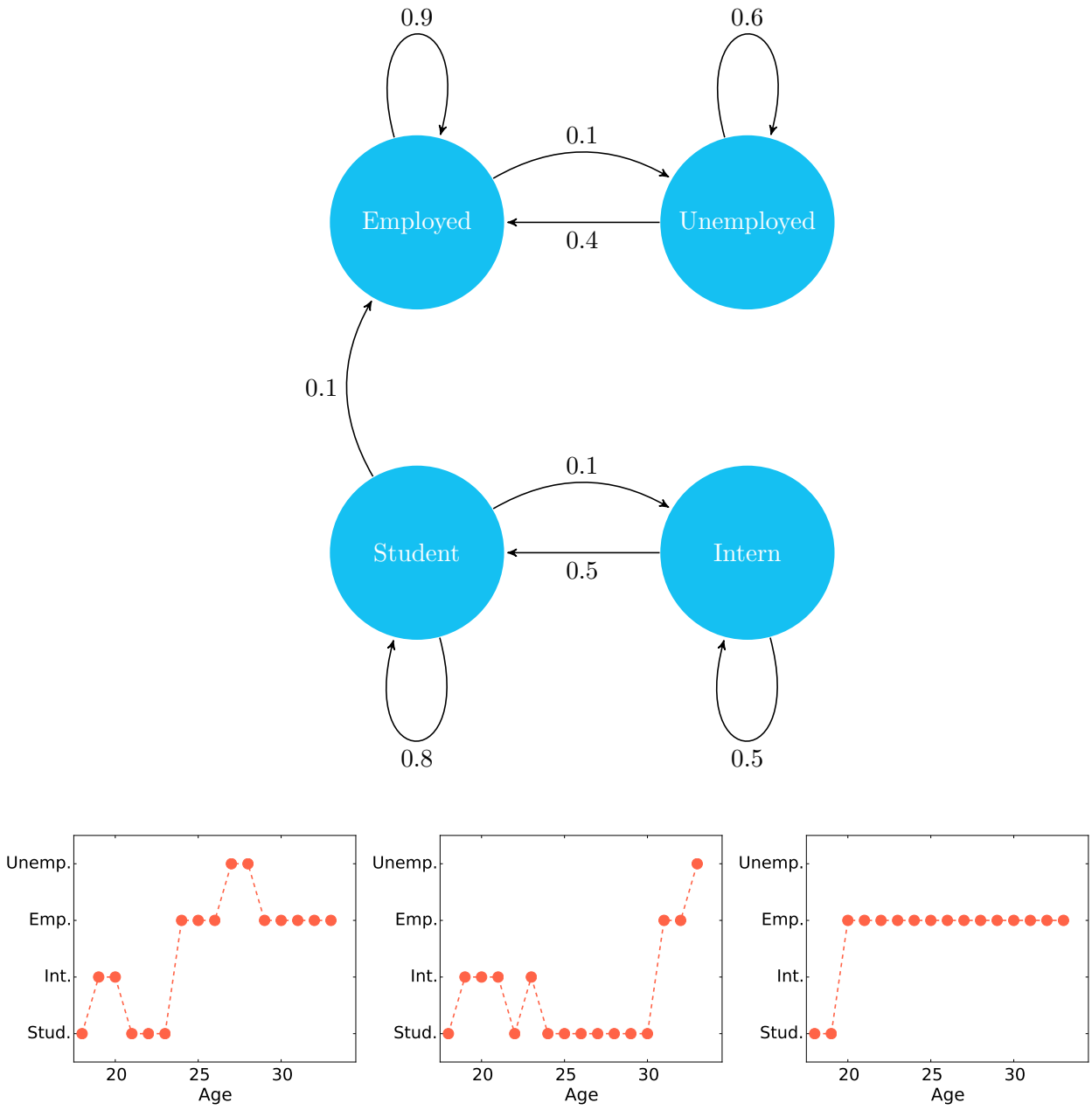


Figure 7.3: State diagram of the Markov chain described in Example (7.2.2) (top). Below we show three realizations of the Markov chain.

The following example illustrates the difference between recurrent and transient states.

Example 7.2.2 (Employment dynamics)**.** A researcher is interested in modeling the employment dynamics of young people using a Markov chain.

She determines that at age 18 a person is either a student with probability 0.9 or an intern with probability 0.1. After that she estimates the following transition probabilities:

$$\begin{array}{cccc} \text{Student} & \text{Intern} & \text{Employed} & \text{Unemployed} & \\ \left(\begin{array}{cccc} 0.8 & 0.5 & 0 & 0 \\ 0.1 & 0.5 & 0 & 0 \\ 0.1 & 0 & 0.9 & 0.4 \\ 0 & 0 & 0.1 & 0.6 \end{array}\right) & \begin{array}{c} \text{Student} \\ \text{Intern} \\ \text{Employed} \\ \text{Unemployed} \end{array} \end{array}$$

The Markov assumption is obviously not completely precise, someone who has been a student for longer is probably less likely to remain a student, but such Markov models are easier to fit (we only need to estimate the transition probabilities) and often yield useful insights.

The initial state vector and the transition matrix of the Markov chain are

$$\vec{p}_{\widetilde{X}(0)} := \begin{bmatrix} 0.9 \\ 0.1 \\ 0 \\ 0 \end{bmatrix}, \qquad T_{\widetilde{X}} := \begin{bmatrix} 0.8 & 0.5 & 0 & 0 \\ 0.1 & 0.5 & 0 & 0 \\ 0.1 & 0 & 0.9 & 0.4 \\ 0 & 0 & 0.1 & 0.6 \end{bmatrix}. \tag{7.20}$$

Figure 7.3 shows the state diagram and some realizations of the Markov chain.

States 1 (student) and 2 (intern) are transient states. Note that the probability that the Markov chain returns to those states after visiting state 3 (employed) is zero, so

$$\begin{aligned} \mathrm{P}\left(\widetilde{X}\left(j\right) \neq 1 \text{ for all } j > i \mid \widetilde{X}\left(i\right) = 1\right) &\geq \mathrm{P}\left(\widetilde{X}\left(i+1\right) = 3 \mid \widetilde{X}\left(i\right) = 1\right) && (7.21) \\ &= 0.1 > 0, && (7.22) \\ \mathrm{P}\left(\widetilde{X}\left(j\right) \neq 2 \text{ for all } j > i \mid \widetilde{X}\left(i\right) = 2\right) &\geq \mathrm{P}\left(\widetilde{X}\left(i+2\right) = 3 \mid \widetilde{X}\left(i\right) = 2\right) && (7.23) \\ &= 0.5 \cdot 0.1 > 0. && (7.24) \end{aligned}$$

In contrast, states 3 and 4 (unemployed) are recurrent. We prove this for state 3 (the argument for state 4 is exactly the same):

$$\begin{aligned} &\mathrm{P}\left(\widetilde{X}\left(j\right) \neq 3 \text{ for all } j > i \mid \widetilde{X}\left(i\right) = 3\right) && (7.25) \\ &= \mathrm{P}\left(\widetilde{X}\left(j\right) = 4 \text{ for all } j > i \mid \widetilde{X}\left(i\right) = 3\right) && (7.26) \\ &= \lim_{k\to\infty} \mathrm{P}\left(\widetilde{X}\left(i+1\right) = 4 \mid \widetilde{X}\left(i\right) = 3\right) \prod_{j=1}^{k} \mathrm{P}\left(\widetilde{X}\left(i+j+1\right) = 4 \mid \widetilde{X}\left(i+j\right) = 4\right) && (7.27) \\ &= \lim_{k\to\infty} 0.1 \cdot 0.6^k && (7.28) \\ &= 0. && (7.29) \end{aligned}$$

△

In this example, it is not possible to reach the states *student* and *intern* from the states *employed* or *unemployed*. Markov chains for which there is a possible transition between any two states (even if it is not direct) are called irreducible.

Definition 7.2.3 (Irreducible Markov chain)**.** *A time-homogeneous finite-state Markov chain is irreducible if for any state x, the probability of reaching every other state $y \neq x$ in a finite number of steps is nonzero, i.e. there exists $m \geq 0$ such that*

$$\mathrm{P}\left(\widetilde{X}\left(i+m\right)=y \mid \widetilde{X}\left(i\right)=x\right)>0. \tag{7.30}$$

One can easily check that the Markov chain in Example 7.1.1 is irreducible, whereas the one in Example 7.2.2. An important result is that all states in an irreducible Markov chain are recurrent.

Theorem 7.2.4 (Irreducible Markov chains)**.** *All states in an irreducible Markov chain are recurrent.*

Proof. In any finite-state Markov chain there must be at least one state that is recurrent. If all the states are transient there is a nonzero probability that it leaves all of the states forever, which is not possible. Without loss of generality let us assume that state x is recurrent. We now provide a sketch of a proof that another arbitrary state y must also be recurrent. To alleviate notation let

$$p_{x,x} := \mathrm{P}\left(\widetilde{X}\left(j\right)=x \text{ for some } j>i \mid \widetilde{X}\left(i\right)=x\right), \tag{7.31}$$

$$p_{x,y} := \mathrm{P}\left(\widetilde{X}\left(j\right)=y \text{ for some } j>i \mid \widetilde{X}\left(i\right)=x\right), \tag{7.32}$$

$$p_{y,x} := \mathrm{P}\left(\widetilde{X}\left(k\right)=x \text{ for some } j>i \mid \widetilde{X}\left(i\right)=y\right). \tag{7.33}$$

The chain is irreducible so there is a nonzero probability $p_m > 0$ of reaching y from x in at most m steps for some $m > 0$. The probability that the chain goes from x to y and never goes back to x is consequently at least $p_m\left(1-p_{y,x}\right)$. However, x is recurrent, so this probability must be zero! Since $p_m > 0$ this implies $p_{y,x} = 1$.

Consider the following event:

1. $\widetilde{X}$ goes from y to x.
2. $\widetilde{X}$ does not return to y in m steps after reaching x.
3. $\widetilde{X}$ eventually reaches x again at a time $m' > m$.

The probability of this event is equal to $p_{y,x}\left(1-p_m\right)p_{x,x} = 1-p_m$ (recall that x is recurrent so $p_{x,x} = 1$). Now imagine that steps 2 and 3 repeat k times, i.e. that $\widetilde{X}$ fails to go from x to y in m steps k times. The probability of this event is $p_{y,x}\left(1-p_m\right)^k p_{x,x}^k = \left(1-p_m\right)^k$. Taking $k \to \infty$ this is equal to zero for any m so the probability that $\widetilde{X}$ does not eventually return to x must be zero (this can be made rigorous, but the details are beyond the scope of these notes). □

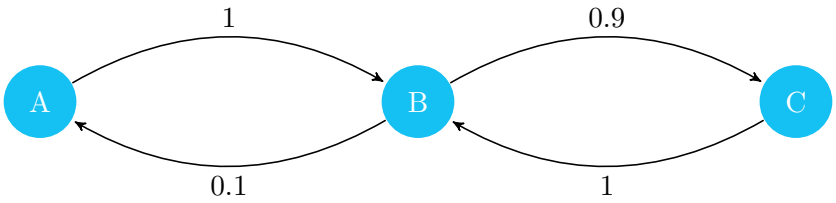


Figure 7.4: State diagram of a Markov chain where states the states have period two.

7.3 Periodicity

Another important consideration is whether the Markov chain always visits a given state at regular intervals. If this is the case, then the state has a period greater than one.

Definition 7.3.1 (Period of a state)**.** *Let $\widetilde{X}$ be a time-homogeneous finite-state Markov chain and x a state of the Markov chain. The period m of x is the largest integer such that it is only possible to return to x in a number of steps that is a multiple of m, i.e. we can only return in km steps with nonzero probability where k is a positive integer.*

Figure 7.4 shows a Markov chain where the states have a period equal to two. Aperiodic Markov chains do not contain states with periods greater than one.

Definition 7.3.2 (Aperiodic Markov chain)**.** *A time-homogeneous finite-state Markov chain $\widetilde{X}$ is aperiodic if all states have period equal to one.*

The Markov chains in Examples 7.1.1 and 7.2.2 are both aperiodic.

7.4 Convergence

In this section we study under what conditions a finite-state time-homogeneous Markov chain $\widetilde{X}$ converges in distribution. If a Markov chain converges in distribution, then its state vector $\vec{p}_{\widetilde{X}(i)}$, which contains the first order pmf of $\widetilde{X}$, converges to a fixed vector $\vec{p}_\infty$,

$$\vec{p}_\infty := \lim_{i\to\infty} \vec{p}_{\widetilde{X}(i)}. \tag{7.34}$$

In that case the probability of the Markov chain being in each state eventually tends to a fixed value (which does *not* imply that the Markov chain will stay at a given state!).

By Lemma 7.1.2 we can express (7.34) in terms of the initial state vector and the transition matrix of the Markov chain

$$\vec{p}_\infty = \lim_{i\to\infty} T_{\widetilde{X}}^i \, \vec{p}_{\widetilde{X}(0)}. \tag{7.35}$$

Computing this limit analytically for a particular $T_{\widetilde{X}}$ and $\vec{p}_{\widetilde{X}(0)}$ may seem challenging at first sight. However, it is often possible to leverage the eigendecomposition of the transition matrix (if it exists) to find $\vec{p}_\infty$. This is illustrated in the following example.

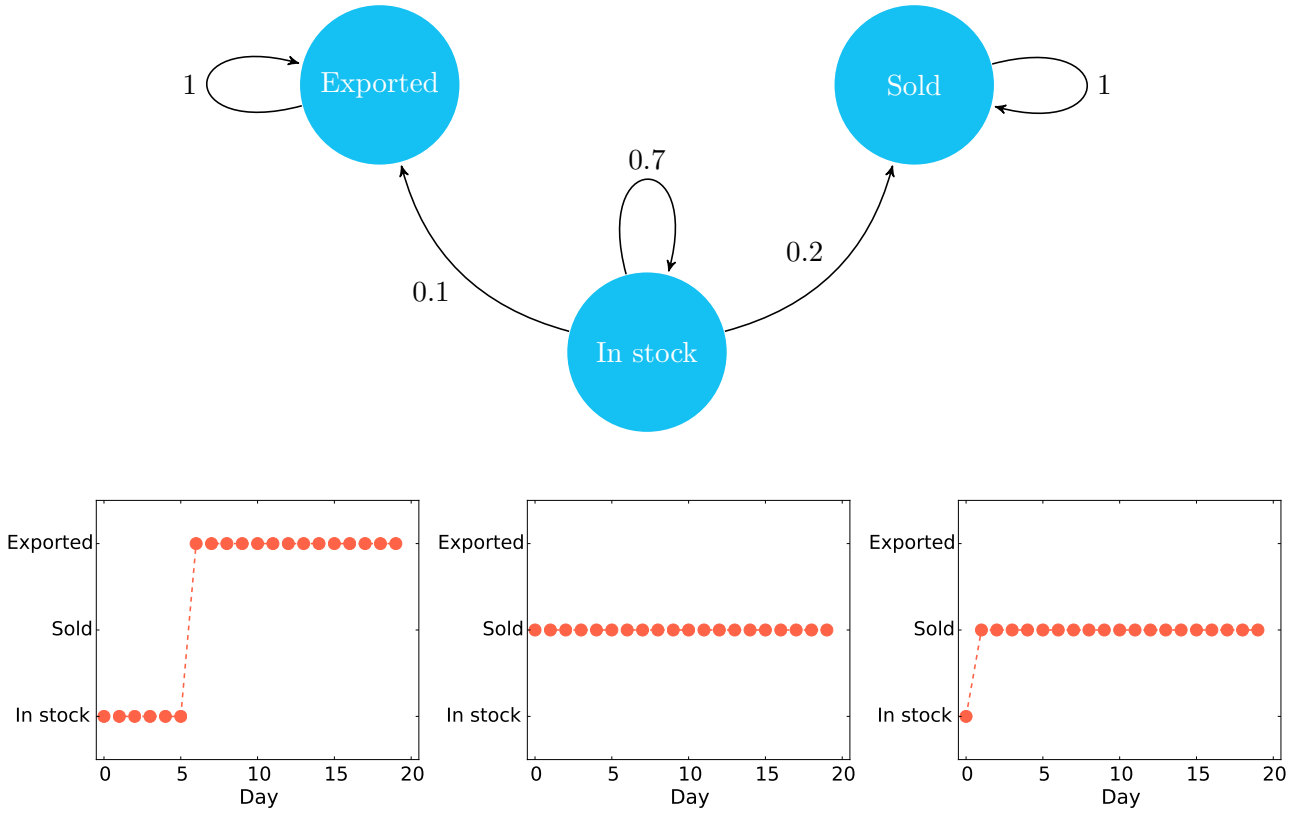


Figure 7.5: State diagram of the Markov chain described in Example (7.4.1) (top). Below we show three realizations of the Markov chain.

Example 7.4.1 (Mobile phones)**.** A company that makes mobile phones wants to model the sales of a new model they have just released. At the moment 90% of the phones are in stock, 10% have been sold locally and none have been exported. Based on past data, the company determines that each day a phone is sold with probability 0.2 and exported with probability 0.1. The initial state vector and the transition matrix of the Markov chain are

$$\vec{a} := \begin{bmatrix} 0.9 \\ 0.1 \\ 0 \end{bmatrix}, \qquad T_{\widetilde{X}} = \begin{bmatrix} 0.7 & 0 & 0 \\ 0.2 & 1 & 0 \\ 0.1 & 0 & 1 \end{bmatrix}. \tag{7.36}$$

We have used $\vec{a}$ to denote $\vec{p}_{\widetilde{X}(0)}$ because later we will consider other possible initial state vectors. Figure 7.6 shows the state diagram and some realizations of the Markov chain.

The company is interested in the fate of the new model. In particular, it would like to compute what fraction of mobile phones will end up exported and what fraction will be sold locally. This

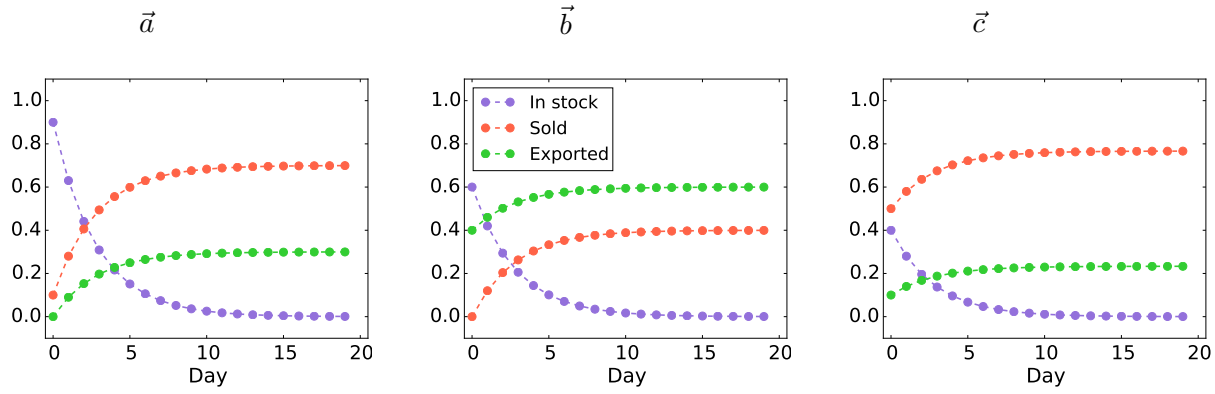


Figure 7.6: Evolution of the state vector of the Markov chain in Example (7.4.1) for different values of the initial state vector $\vec{p}_{\widetilde{X}(0)}$.

is equivalent to computing

$$\lim_{i\to\infty} \vec{p}_{\widetilde{X}(i)} = \lim_{i\to\infty} T_{\widetilde{X}}^{i}\, \vec{p}_{\widetilde{X}(0)} \tag{7.37}$$

$$= \lim_{i\to\infty} T_{\widetilde{X}}^{i}\, \vec{a}. \tag{7.38}$$

The transition matrix $T_{\widetilde{X}}$ has three eigenvectors

$$\vec{q}_1 := \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \qquad \vec{q}_2 := \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \qquad \vec{q}_3 := \begin{bmatrix} 0.80 \\ -0.53 \\ -0.27 \end{bmatrix}. \tag{7.39}$$

The corresponding eigenvalues are $\lambda_1 := 1$, $\lambda_2 := 1$ and $\lambda_3 := 0.7$. We gather the eigenvectors and eigenvalues into two matrices

$$Q := \begin{bmatrix} \vec{q}_1 & \vec{q}_2 & \vec{q}_3 \end{bmatrix}, \qquad \Lambda := \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}, \tag{7.40}$$

so that the eigendecomposition of $T_{\widetilde{X}}$ is

$$T_{\widetilde{X}} := Q \Lambda Q^{-1}. \tag{7.41}$$

It will be useful to express the initial state vector $\vec{a}$ in terms of the different eigenvectors. This is achieved by computing

$$Q^{-1} \vec{p}_{\widetilde{X}(0)} = \begin{bmatrix} 0.3 \\ 0.7 \\ 1.122 \end{bmatrix}, \tag{7.42}$$

so that

$$\vec{a} = 0.3\, \vec{q}_1 + 0.7\, \vec{q}_2 + 1.122\, \vec{q}_3. \tag{7.43}$$

We conclude that

$$\lim_{i\to\infty} T_{\widetilde{X}}^i \vec{a} = \lim_{i\to\infty} T_{\widetilde{X}}^i \left(0.3\,\vec{q}_1 + 0.7\,\vec{q}_2 + 1.122\,\vec{q}_3\right) \tag{7.44}$$

$$= \lim_{i\to\infty} 0.3\,T_{\widetilde{X}}^i \vec{q}_1 + 0.7\,T_{\widetilde{X}}^i \vec{q}_2 + 1.122\,T_{\widetilde{X}}^i \vec{q}_3 \tag{7.45}$$

$$= \lim_{i\to\infty} 0.3\,\lambda_1^i \vec{q}_1 + 0.7\,\lambda_2^i \vec{q}_2 + 1.122\,\lambda_3^i \vec{q}_3 \tag{7.46}$$

$$= \lim_{i\to\infty} 0.3\,\vec{q}_1 + 0.7\,\vec{q}_2 + 1.122\,0.5^{\,i}\,\vec{q}_3 \tag{7.47}$$

$$= 0.3\,\vec{q}_1 + 0.7\,\vec{q}_2 \tag{7.48}$$

$$= \begin{bmatrix} 0 \\ 0.7 \\ 0.3 \end{bmatrix}. \tag{7.49}$$

This means that eventually the probability that each phone has been sold locally is 0.7 and the probability that it has been exported is 0.3. The left graph in Figure 7.6 shows the evolution of the state vector. As predicted, it eventually converges to the vector in equation (7.49).

In general, because of the special structure of the two eigenvectors with eigenvalues equal to one in this example, we have

$$\lim_{i\to\infty} T_{\widetilde{X}}^i \vec{p}_{\widetilde{X}(0)} = \begin{bmatrix} 0 \\ \left(Q^{-1}\vec{p}_{\widetilde{X}(0)}\right)_2 \\ \left(Q^{-1}\vec{p}_{\widetilde{X}(0)}\right)_1 \end{bmatrix}. \tag{7.50}$$

This is illustrated in Figure 7.6 where you can see the evolution of the state vector if it is initialized to these other two distributions:

$$\vec{b} := \begin{bmatrix} 0.6 \\ 0 \\ 0.4 \end{bmatrix}, \qquad Q^{-1}\vec{b} = \begin{bmatrix} 0.6 \\ 0.4 \\ 0.75 \end{bmatrix}, \tag{7.51}$$

$$\vec{c} := \begin{bmatrix} 0.4 \\ 0.5 \\ 0.1 \end{bmatrix}, \qquad Q^{-1}\vec{c} = \begin{bmatrix} 0.23 \\ 0.77 \\ 0.50 \end{bmatrix}. \tag{7.52}$$

△

The transition matrix of the Markov chain in Example 7.4.1 has two eigenvectors with eigenvalue equal to one. If we set the initial state vector to equal either of these eigenvectors (note that we must make sure to normalize them so that the state vector contains a valid pmf) then

$$T_{\widetilde{X}}\,\vec{p}_{\widetilde{X}(0)} = \vec{p}_{\widetilde{X}(0)}, \tag{7.53}$$

so that

$$\vec{p}_{\widetilde{X}(i)} = T_{\widetilde{X}}^i \vec{p}_{\widetilde{X}(0)} \tag{7.54}$$

$$= \vec{p}_{\widetilde{X}(0)} \tag{7.55}$$

for all i. In particular,

$$\lim_{i\to\infty} \vec{p}_{\widetilde{X}(i)} = \vec{p}_{\widetilde{X}(0)}, \tag{7.56}$$

so $\vec{X}$ converges to a random variable with pmf $\vec{p}_{\widetilde{X}(0)}$ in distribution. A distribution that satisfies (7.56) is called a *stationary* distribution of the Markov chain.

Definition 7.4.2 (Stationary distribution)**.** *Let $\widetilde{X}$ be a finite-state time-homogeneous Markov chain and let $\vec{p}_{stat}$ be a state vector containing a valid pmf over the possible states of $\widetilde{X}$. If $\vec{p}_{stat}$ is an eigenvector associated to an eigenvalue equal to one, so that*

$$T_{\widetilde{X}}\, \vec{p}_{stat} = \vec{p}_{stat}, \tag{7.57}$$

then the distribution corresponding to $\vec{p}_{stat}$ is a stationary or steady-state distribution of $\widetilde{X}$.

Establishing whether a distribution is stationary by checking whether (7.57) holds may be challenging computationally if the state space is very large. We now derive an alternative condition that implies stationarity. Let us first define reversibility of Markov chains.

Definition 7.4.3 (Reversibility)**.** *Let $\widetilde{X}$ be a finite-state time-homogeneous Markov chain with s states and transition matrix $T_{\widetilde{X}}$. Assume that $\widetilde{X}\,(i)$ is distributed according to the state vector $\vec{p} \in \mathbb{R}^s$. If*

$$\mathrm{P}\left(\widetilde{X}\,(i) = x_j, \widetilde{X}\,(i+1) = x_k\right) = \mathrm{P}\left(\widetilde{X}\,(i) = x_k, \widetilde{X}\,(i+1) = x_j\right), \quad \text{for all } 1 \leq j,k \leq s, \tag{7.58}$$

then we say that $\widetilde{X}$ is reversible with respect to $\vec{p}$. This is equivalent to the detailed-balance condition

$$\left(T_{\widetilde{X}}\right)_{kj} \vec{p}_j = \left(T_{\widetilde{X}}\right)_{jk} \vec{p}_k, \quad \text{for all } 1 \leq j,k \leq s. \tag{7.59}$$

As proved in the following theorem, reversibility implies stationarity, but the converse does not hold. A Markov chain is not necessarily reversible with respect to a stationary distribution (and often will not be). The detailed-balance condition therefore only provides a sufficient condition for stationarity.

Theorem 7.4.4 (Reversibility implies stationarity)**.** *If a time-homogeneous Markov chain $\widetilde{X}$ is reversible with respect to a distribution p_X, then p_X is a stationary distribution of $\widetilde{X}$.*

Proof. Let $\vec{p}$ be the state vector containing p_X. By assumption $T_{\widetilde{X}}$ and $\vec{p}$ satisfy (7.59), so for $1 \leq j \leq s$

$$\left(T_{\widetilde{X}}\vec{p}\right)_j = \sum_{k=1}^{s} \left(T_{\widetilde{X}}\right)_{jk} \vec{p}_k \tag{7.60}$$

$$= \sum_{k=1}^{s} \left(T_{\widetilde{X}}\right)_{kj} \vec{p}_j \tag{7.61}$$

$$= \vec{p}_j \sum_{k=1}^{s} \left(T_{\widetilde{X}}\right)_{kj} \tag{7.62}$$

$$= \vec{p}_j. \tag{7.63}$$

$$\vec{p}_{\widetilde{X}(0)} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \qquad \vec{p}_{\widetilde{X}(0)} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad \vec{p}_{\widetilde{X}(0)} = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.7 \end{bmatrix}$$

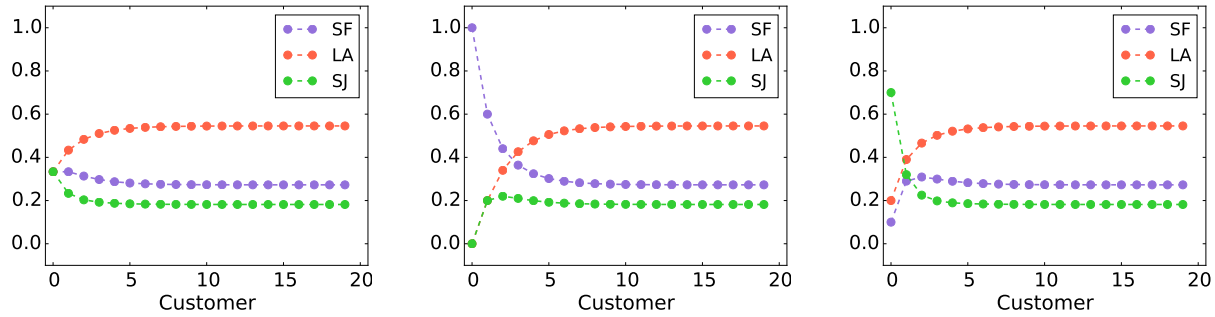


Figure 7.7: Evolution of the state vector of the Markov chain in Example (7.4.7).

The last step follows from the fact that the columns of a valid transition matrix must add to one (the chain always has to go somewhere). □

In Example 7.4.1 the Markov chain has two stationary distributions. It turns out that this is not possible for irreducible Markov chains.

Theorem 7.4.5. *Irreducible Markov chains have a single stationary distribution.*

Proof. This follows from the Perron-Frobenius theorem, which states that the transition matrix of an irreducible Markov chain has a single eigenvector with eigenvalue equal to one and nonnegative entries. □

If in addition, the Markov chain is aperiodic, then it is guaranteed to converge in distribution to a random variable with its stationary distribution *for any initial state vector*. Such Markov chains are called **ergodic**.

Theorem 7.4.6 (Convergence of Markov chains)**.** *If a discrete-time time-homogeneous Markov chain $\widetilde{X}$ is irreducible and aperiodic its state vector converges to the stationary distribution $\vec{p}_{stat}$ of $\widetilde{X}$ for any initial state vector $\vec{p}_{\widetilde{X}(0)}$. This implies that $\widetilde{X}$ converges in distribution to a random variable with pmf given by $\vec{p}_{stat}$.*

The proof of this result is beyond the scope of these notes.

Example 7.4.7 (Car rental (continued))**.** The Markov chain in the car rental example is irreducible and aperiodic. We will now check that it indeed converges in distribution. Its transition matrix has the following eigenvectors

$$\vec{q}_1 := \begin{bmatrix} 0.273 \\ 0.545 \\ 0.182 \end{bmatrix}, \qquad \vec{q}_2 := \begin{bmatrix} -0.577 \\ 0.789 \\ -0.211 \end{bmatrix}, \qquad \vec{q}_3 := \begin{bmatrix} -0.577 \\ -0.211 \\ 0.789 \end{bmatrix}. \tag{7.64}$$

The corresponding eigenvalues are $\lambda_1 := 1$, $\lambda_2 := 0.573$ and $\lambda_3 := 0.227$. As predicted by Theorem 7.4.5 the Markov chain has a single stationary distribution.

For any initial state vector, the component that is collinear with $\vec{q}_1$ will be preserved by the transitions of the Markov chain, but the other two components will become negligible after a while. The chain consequently converges in distribution to a random variable with pmf $\vec{q}_1$ (note that $\vec{q}_1$ has been normalized to be a valid pmf), as predicted by Theorem 7.4.6. This is illustrated in Figure 7.7. No matter how the company allocates the new cars, eventually 27.3% will end up in San Francisco, 54.5% in LA and 18.2% in San Jose. △

7.5 Markov-chain Monte Carlo

The convergence of Markov chains to a stationary distribution is very useful for simulating random variables. Markov-chain Monte Carlo (MCMC) methods generate samples from a target distribution by constructing a Markov chain in such a way that the stationary distribution equals the desired distribution. These techniques are of huge importance in modern statistics and in particular in Bayesian modeling. In this section we describe one of the most popular MCMC methods and illustrate it with a simple example.

The key challenge in MCMC methods is to design an irreducible aperiodic Markov chain for which the target distribution is stationary. The Metropolis-Hastings algorithm uses an auxiliary Markov chain to achieve this.

Algorithm 7.5.1 (Metropolis-Hastings algorithm)**.** *We store the pmf p_X of the target distribution in a vector $\vec{p} \in \mathbb{R}^s$, such that*

$$\vec{p}_j := p_X(x_j), \qquad 1 \leq j \leq s. \tag{7.65}$$

Let T denote the transition matrix of an irreducible Markov chain with the same state space $\{x_1, \ldots, x_s\}$ as $\vec{p}$.
Initialize $\widetilde{X}(0)$ randomly or to a fixed state, then repeat the following steps for $i = 1, 2, 3, \ldots$.

1. *Generate a candidate random variable C from $\widetilde{X}(i-1)$ by using the transition matrix T, i.e.*

$$\mathrm{P}\left(C = k \,|\, \widetilde{X}(i-1) = j\right) = T_{kj}, \qquad 1 \leq j, k \leq s. \tag{7.66}$$

2. *Set*

$$\widetilde{X}(i) := \begin{cases} C & \text{with probability } p_{\text{acc}}\left(\widetilde{X}(i-1), C\right), \\ \widetilde{X}(i-1) & \text{otherwise,} \end{cases} \tag{7.67}$$

where the acceptance probability *is defined as*

$$p_{\text{acc}}(j,k) := \min\left\{\frac{T_{jk}\,\vec{p}_k}{T_{kj}\,\vec{p}_j}, 1\right\} \qquad 1 \leq j, k \leq s. \tag{7.68}$$

It turns out that this algorithm yields a Markov chain that is reversible with respect to the distribution of interest, which ensures that the distribution is stationary.

Theorem 7.5.2. *The pmf in $\vec{p}$ corresponds to a stationary distribution of the Markov chain $\widetilde{X}$ obtained by the Metropolis-Hastings algorithm.*

Proof. We show that the Markov chain $\widetilde{X}$ is reversible with respect to $\vec{p}$, i.e. that

$$\left(T_{\widetilde{X}}\right)_{kj} \vec{p}_j = \left(T_{\widetilde{X}}\right)_{jk} \vec{p}_k, \tag{7.69}$$

holds for all $1 \leq j, k \leq s$. This establishes the result by Theorem 7.4.4. The detailed-balanced condition holds trivially if $j = k$. If $j \neq k$ we have

$$\begin{aligned}
\left(T_{\widetilde{X}}\right)_{kj} &:= \mathrm{P}\left(\widetilde{X}\left(i\right) = k \,|\, \widetilde{X}\left(i-1\right) = j\right) && (7.70)\\
&= \mathrm{P}\left(\widetilde{X}\left(i\right) = C, C = k \,|\, \widetilde{X}\left(i-1\right) = j\right) && (7.71)\\
&= \mathrm{P}\left(\widetilde{X}\left(i\right) = C \,|\, C = k, \widetilde{X}\left(i-1\right) = j\right) P\left(C = k \,|\, \widetilde{X}\left(i-1\right) = j\right) && (7.72)\\
&= p_{\mathrm{acc}}\left(j,k\right) T_{kj} && (7.73)
\end{aligned}$$

and by exactly the same argument $\left(T_{\widetilde{X}}\right)_{jk} = p_{\mathrm{acc}}\left(k,j\right) T_{jk}$. We conclude that

$$\begin{aligned}
\left(T_{\widetilde{X}}\right)_{kj} \vec{p}_j &= p_{\mathrm{acc}}\left(j,k\right) T_{kj}\, \vec{p}_j && (7.74)\\
&= T_{kj}\, \vec{p}_j \min\left\{\frac{T_{jk}\, \vec{p}_k}{T_{kj}\, \vec{p}_j}, 1\right\} && (7.75)\\
&= \min\left\{T_{jk}\, \vec{p}_k, T_{kj}\, \vec{p}_j\right\} && (7.76)\\
&= T_{jk}\, \vec{p}_k \min\left\{1, \frac{T_{kj}\, \vec{p}_j}{T_{jk}\, \vec{p}_k}\right\} && (7.77)\\
&= p_{\mathrm{acc}}\left(k,j\right) T_{jk}\, \vec{p}_k && (7.78)\\
&= \left(T_{\widetilde{X}}\right)_{jk} \vec{p}_k. && (7.79)
\end{aligned}$$

□

The following example is taken from Hastings's seminal paper *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*.

Example 7.5.3 (Generating a Poisson random variable)**.** Our aim is to generate a Poisson random variable X. Note that we don't need to know the normalizing constant in the Poisson pmf, which equals to e^{λ}, as long as we know that it is proportional to

$$p_X\left(x\right) \propto \frac{\lambda^x}{x!} \tag{7.80}$$

The auxiliary Markov chain must be able to reach any possible value of X, i.e. all positive integers. We will use a modified random walk that takes steps upwards and downwards with probability $1/2$, but never goes below 0. Its transition matrix equals

$$T_{kj} := \begin{cases} \frac{1}{2} & \text{if } j = 0 \text{ and } k = 0,\\ \frac{1}{2} & \text{if } k = j+1,\\ \frac{1}{2} & \text{if } j > 0 \text{ and } k = j-1,\\ 0 & \text{otherwise.} \end{cases} \tag{7.81}$$

T is symmetric so the acceptance probability is equal to the ratio of the pmfs:

$$p_{\text{acc}}(j,k) := \min\left\{\frac{T_{jk}\, p_X(k)}{T_{kj}\, p_X(j)}, 1\right\} \tag{7.82}$$

$$= \min\left\{\frac{p_X(k)}{p_X(j)}, 1\right\}. \tag{7.83}$$

To compute the acceptance probability, we only consider transitions that are possible under the random walk. If $j = 0$ and $k = 0$

$$p_{\text{acc}}(j,k) = 1. \tag{7.84}$$

If $k = j + 1$

$$p_{\text{acc}}(j, j+1) = \min\left\{\frac{\frac{\lambda^{j+1}}{(j+1)!}}{\frac{\lambda^j}{j!}}, 1\right\} \tag{7.85}$$

$$= \min\left\{\frac{\lambda}{j+1}, 1\right\}. \tag{7.86}$$

If $k = j - 1$

$$p_{\text{acc}}(j, j-1) = \min\left\{\frac{\frac{\lambda^{j-1}}{(j-1)!}}{\frac{\lambda^j}{j!}}, 1\right\} \tag{7.87}$$

$$= \min\left\{\frac{j}{\lambda}, 1\right\}. \tag{7.88}$$

We now spell out the steps of the Metropolis-Hastings method. To simulate the auxiliary random walk we use a sequence of Bernoulli random variables that indicate whether the random walk is trying to go up or down (or stay at zero). We initialize the chain at $x_0 = 0$. Then, for $i = 1, 2, \ldots$, we

- Generate a sample b from a Bernoulli distribution with parameter $1/2$ and a sample u uniformly distributed in $[0, 1]$.
- If $b = 0$:
  - If $x_{i-1} = 0$, $x_i := 0$.
  - If $x_{i-1} > 0$:
    - If $u < \frac{x_{i-1}}{\lambda}$, $x_i := x_{i-1} - 1$.
    - Otherwise $x_i := x_{i-1}$.
- If $b = 1$:
  - If $u < \frac{\lambda}{x_{i-1}+1}$, $x_i := x_{i-1} + 1$.
  - Otherwise $x_i := x_{i-1}$.

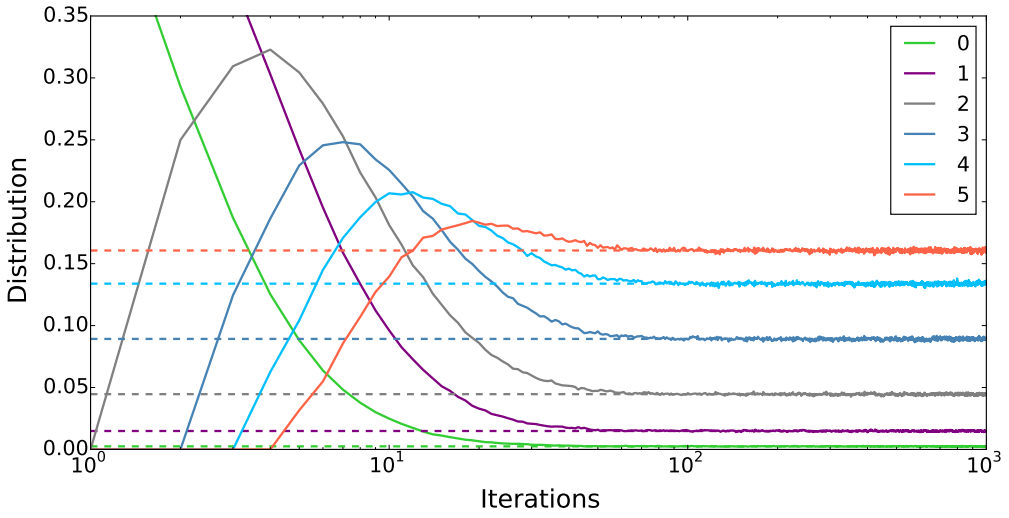


Figure 7.8: Convergence in distribution of the Markov chain constructed in Example 7.8 for $\lambda := 6$. To prevent clutter we only plot the empirical distribution of 6 states, computed by running the Markov chain 10^4 times.

The Markov chain that we have built is irreducible: there is nonzero probability of going from any nonnegative integer to any other nonnegative integer (although it could take a while!). We have not really proved that the chain should converge to the desired distribution, since we have not discussed convergence of Markov chains with infinite state spaces, but Figure 7.8 shows that the method indeed allows to sample from a Poisson distribution with $\lambda := 6$.

△

For the example in Figure 7.8, approximate convergence in distribution occurs after around 100 iterations. This is called the **mixing time** of the Markov chain. To account for it, MCMC methods usually discard the samples from the chain over an initial period known as *burn-in* time.

The careful reader might be wondering about the point of using MCMC methods if we already have access to the desired distribution. It seems much simpler to just apply the method described in Section 2.6.1 instead. However, the Metropolis-Hastings method can be applied to discrete distributions with infinite supports and also to continuous distributions (justifying this is beyond the scope of these notes). Crucially, in contrast with inverse-transform and rejection sampling, Metropolis-Hastings does not require having access to the pmf p_X or pdf f_X of the target distribution, but rather to the ratio $p_X(x)/p_X(y)$ or $f_X(x)/f_X(y)$ for every $x \neq y$. This is very useful when computing conditional distributions within probabilistic models.

Imagine that we have access to the marginal distribution of a continuous random variable A and the conditional distribution of another continuous random variable B given A. Computing the

conditional pdf

$$f_{A|B}(a|b) = \frac{f_A(a) f_{B|A}(b|a)}{\int_{u=-\infty}^{\infty} f_A(u) f_{B|A}(b|u) \, \mathrm{d}u} \tag{7.89}$$

is not necessary feasible due to the integral in the denominator. However, if we apply Metropolis-Hastings to sample from $f_{A|B}$ we don't need to compute the normalizing factor since for any $a_1 \neq a_2$

$$\frac{f_{A|B}(a_1|b)}{f_{A|B}(a_2|b)} = \frac{f_A(a_1) f_{B|A}(b|a_1)}{f_A(a_2) f_{B|A}(b|a_2)}. \tag{7.90}$$

Chapter 8

Descriptive statistics

In this chapter we describe several techniques for visualizing data, as well as for computing quantities that summarize it effectively. Such quantities are known as descriptive statistics. As we will see in the following chapters, these statistics can often be interpreted within a probabilistic framework, but they are also useful when probabilistic assumptions are not warranted. Because of this, we present them from a **deterministic** point of view.

8.1 Histogram

We begin by considering data sets containing one-dimensional data. One of the most natural ways of visualizing 1D data is to plot their histogram. The histogram is obtained by binning the range of the data and counting the number of instances that fall within each bin. The width of the bins is a parameter that can be adjusted to yield higher or lower resolution. If we interpret the data as corresponding to samples from a random variable, then the histogram would be a piecewise constant approximation to their pmf or pdf.

Figure 8.1 shows two histograms computed from temperature data gathered at a weather station in Oxford over 150 years.[1] Each data point represents the maximum temperature recorded in January or August of a particular year. Figure 8.2 shows a histogram of the GDP per capita of all countries in the world in 2014 according to the United Nations.[2]

8.2 Sample mean and variance

Averaging the elements in a one-dimensional data set provides a one-number summary of the data, which is a deterministic counterpart to the mean of a random variable (recall that we are making no probabilistic assumptions in this chapter). This can be extended to multi-dimensional data by averaging over each dimension separately.

Definition 8.2.1 (Sample mean)**.** *Let* $\{x_1, x_2, \ldots, x_n\}$ *be a set of real-valued data. The sample*

[1] The data is available at http://www.metoffice.gov.uk/pub/data/weather/uk/climate/stationdata/oxforddata.txt.

[2] The data is available at http://unstats.un.org/unsd/snaama/selbasicFast.asp.

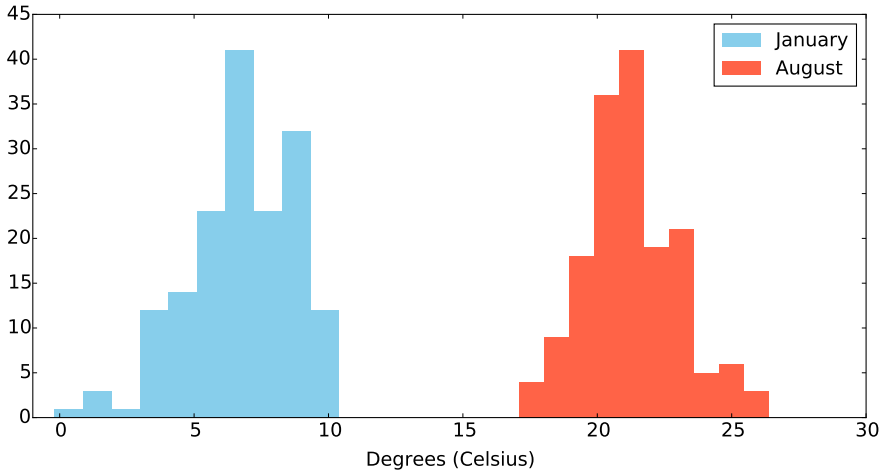


Figure 8.1: Histograms of temperature data taken in a weather station in Oxford over 150 years. Each data point equals the maximum temperature recorded in a certain month in a particular year.

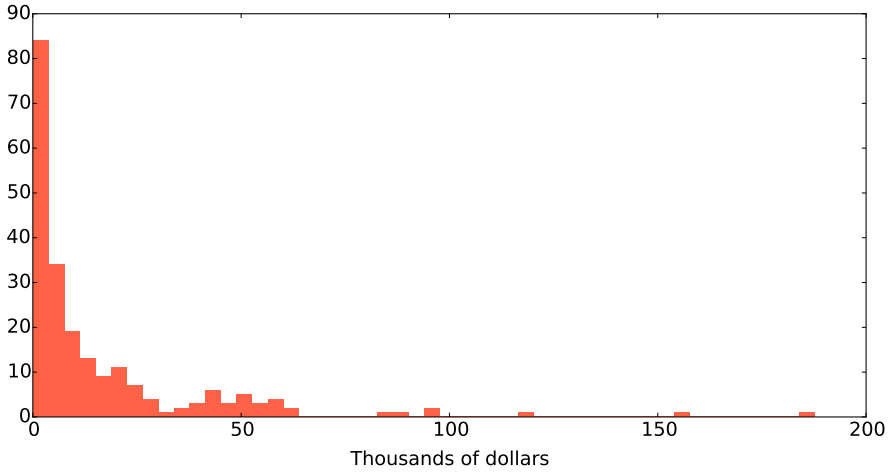


Figure 8.2: Histogram of the GDP per capita of all countries in the world in 2014.

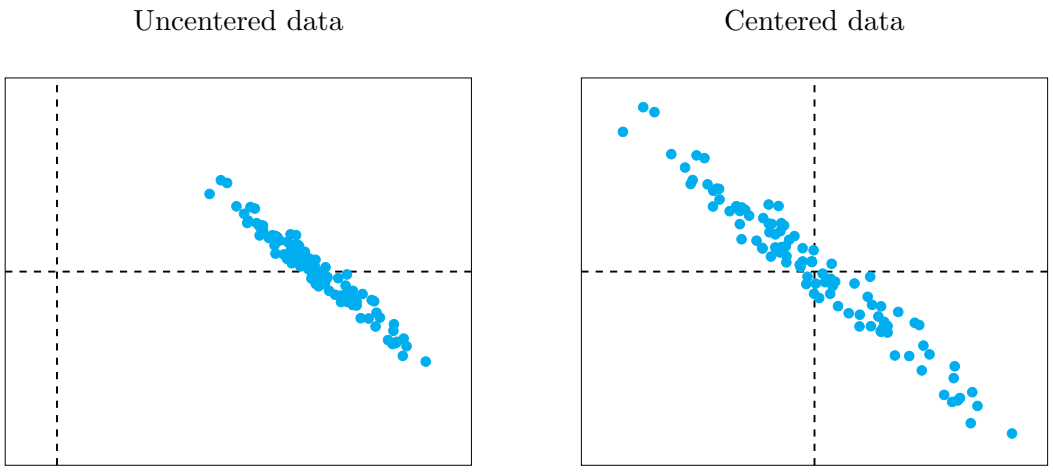


Figure 8.3: Effect of centering a two-dimensional data set. The axes are depicted using dashed lines.

mean of the data is defined as

$$\text{av}\left(x_1, x_2, \ldots, x_n\right) := \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{8.1}$$

Let $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ *be a set of d-dimensional real-valued data vectors. The sample mean is*

$$\text{av}\left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\right) := \frac{1}{n}\sum_{i=1}^{n} \vec{x}_i. \tag{8.2}$$

The sample mean of the data in Figure 8.1 is 6.73 °C in January and 21.3 °C in August. The sample mean of the GDPs per capita in Figure 8.2 is $16,500.

Geometrically, the average, also known as the sample mean, is the center of mass of the data. A common preprocessing step in data analysis is to **center** a set of data by subtracting its sample mean. Figure 8.3 shows an example.

Algorithm 8.2.2 (Centering)**.** *Let* $\vec{x}_1, \ldots, \vec{x}_n$ *be a set of d-dimensional data. To center the data set we:*

1. *Compute the sample mean following Definition 8.2.1.*
2. *Subtract the sample mean from each vector of data. For* $1 \leq i \leq n$

$$\vec{y}_i := \vec{x}_i - \text{av}\left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\right). \tag{8.3}$$

The resulting data set $\vec{y}_1, \ldots, \vec{y}_n$ *has sample mean equal to zero; it is centered at the origin.*

The sample variance is the average of the squared deviations from the sample mean. Geometrically, it quantifies the average variation of the data set around its center. It is a deterministic counterpart to the variance of a random variable.

Definition 8.2.3 (Sample variance and standard deviation)**.** *Let* $\{x_1, x_2, \ldots, x_n\}$ *be a set of real-valued data. The sample variance is defined as*

$$\text{var}\left(x_1, x_2, \ldots, x_n\right) := \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \text{av}\left(x_1, x_2, \ldots, x_n\right)\right)^2 \tag{8.4}$$

The sample standard deviation is the square root of the sample variance

$$\text{std}\left(x_1, x_2, \ldots, x_n\right) := \sqrt{\text{var}\left(x_1, x_2, \ldots, x_n\right)}. \tag{8.5}$$

You might be wondering why the normalizing constant is $1/\left(n-1\right)$ instead of $1/n$. The reason is that this ensures that the expectation of the sample variance equals the true variance when the data are iid (see Lemma 9.2.5). In practice there is not much difference between the two normalizations.

The sample standard deviation of the temperature data in Figure 8.1 is 1.99 °C in January and 1.73 °C in August. The sample standard deviation of the GDP data in Figure 8.2 is $25,300.

8.3 Order statistics

In some cases, a data set is well described by its mean and standard deviation.

In January the temperature in Oxford is around 6.73 °C give or take 2 °C.

This a pretty accurate account of the temperature data from the previous section. However, imagine that someone describes the GDP data set in Figure 8.2 as:

Countries typically have a GDP per capita of about $16 500 give or take $25 300.

This description is pretty terrible. The problem is that most countries have very small GDPs per capita, whereas a few have really large ones and the sample mean and standard deviation don't really convey this information. Order statistics provide an alternative description, which is usually more informative when there are extreme values in the data.

Definition 8.3.1 (Quantiles and percentiles)**.** *Let* $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ *denote the ordered elements of a set of data* $\{x_1, x_2, \ldots, x_n\}$*. The* q *quantile of the data for* $0 < q < 1$ *is* $x_{([q(n+1)])}$*, where* $[q\left(n+1\right)]$ *is the result of rounding* $q\left(n+1\right)$ *to the closest integer. The* $100\,p$ *quantile is known as the* p *percentile.*

The 0.25 *and* 0.75 *quantiles are known as the first and third* ***quartiles****, whereas the* 0.5 *quantile is known as the* ***sample median****. A quarter of the data are smaller than the* 0.25 *quantile, half are smaller (or larger) than the median and three quarters are smaller than the* 0.75 *quartile. If* n *is even, the sample median is usually set to*

$$\frac{x_{(n/2)} + x_{(n/2+1)}}{2}. \tag{8.6}$$

The difference between the third and the first quartile is known as the ***interquartile range*** *(IQR).*

It turns out that for the temperature data set in Figure 8.1 the sample median is 6.80 °C in January and 21.2 °C in August, which is essentially the same as the sample mean. The IQR is

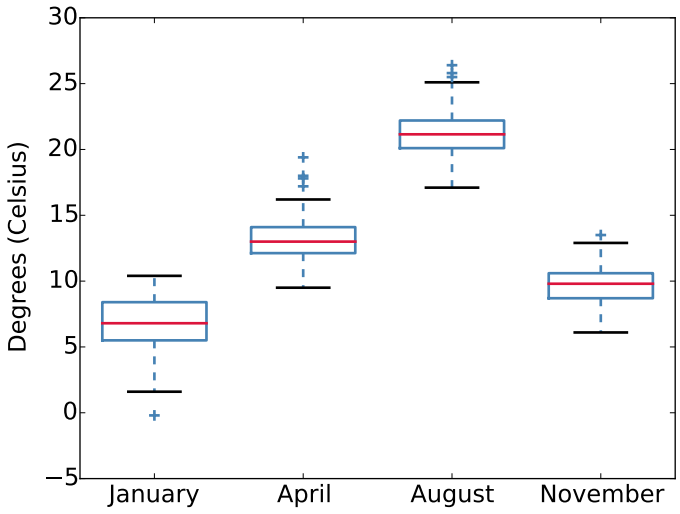


Figure 8.4: Box plots of the Oxford temperature data set used in Figure 8.1. Each box plot corresponds to the maximum temperature in a particular month (January, April, August and November) over the last 150 years.

2.9 °C in January and 2.1 °C in August. This gives a very similar spread around the median, as the sample mean. In this particular example, there does not seem to be an advantage in using order statistics.

For the GDP data set, the median is $6,350. This means that half of the countries have a GDP of less than $6,350. In contrast, 71% of the countries have a GDP per capita lower than the sample mean! The IQR of these data is $18,200. To provide a more complete description of the data set, we can list a **five-number summary** of order statistics: the minimum $x_{(1)}$, the first quartile, the sample median, the third quartile and the maximum $x_{(n)}$. For the GDP data set these are $130, $1,960, $6,350, $20,100, and $188,000 respectively.

We can visualize the main order statistics of a data set by using a **box plot**, which shows the median value of the data enclosed in a box. The bottom and top of the box are the first and third quartiles. This way of visualizing a data set was proposed by the mathematician John Tukey. Tukey's box plot also includes *whiskers*. The lower whisker is a line extending from the bottom of the box to the smallest value within 1.5 IQR of the first quartile. The higher whisker extends from the top of the box to the highest value within 1.5 IQR of the third quartile. Values beyond the whiskers are considered **outliers** and are plotted separately.

Figure 8.4 applies box plots to visualize the temperature data set used in Figure 8.1. Each box plot corresponds to the maximum temperature in a particular month (January, April, August and November) over the last 150 years. The box plots allow us to quickly compare the spread of temperatures in the different months. Figure 8.5 shows a box plot of the GDP data from Figure 8.2. From the box plot it is immediately apparent that most countries have very small GDPs per capita, that the spread between countries increases for larger GDPs per capita and that a small number of countries have very large GDPs per capita.

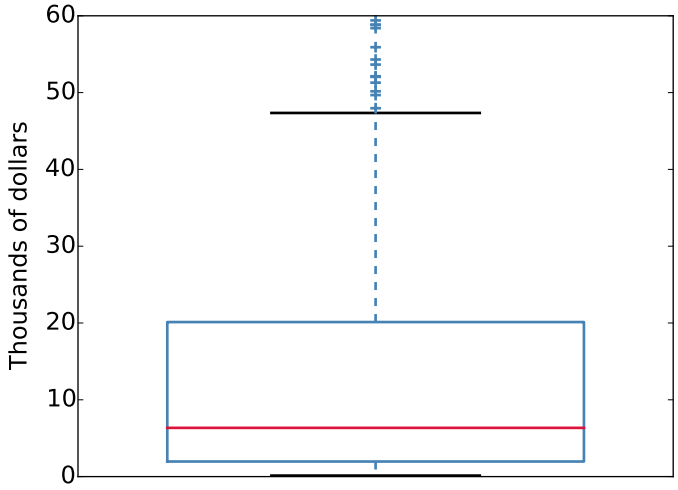


Figure 8.5: Box plot of the GDP per capita of all countries in the world in 2014. Not all of the outliers are shown.

8.4 Sample covariance

In the previous sections we mostly considered data sets consisting of one-dimensional data (except when we discussed the sample mean of a multidimensional data set). In machine-learning lingo, there was only one feature per data point. We now study a multidimensional scenario, where there are several features associated to each data point.

If the dimension of the data set equals to two (i.e. there are two features per data point), we can visualize the data using a **scatter plot**, where each axis represents one of the features. Figure 8.6 shows several scatter plots of temperature data. These data are the same as in Figure 8.1, but we have now arranged them to form two-dimensional data sets. In the plot on the left, one dimension corresponds to the temperature in January and the other dimension to the temperature in August (there is one data point per year). In the plot on the right, one dimension represents the minimum temperature in a particular month and the other dimension represents the maximum temperature in the same month (there is one data point per month). The sample covariance quantifies whether the two features of a two-dimensional data set tend to vary in a similar way on average, just as the covariance quantifies the expected joint variation of two random variables.

Definition 8.4.1 (Sample covariance)**.** *Let* $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ *be a data set where each example consists of a measurement of two different features. The sample covariance is defined as*

$$\text{cov}\left((x_1, y_1), \ldots, (x_n, y_n)\right) := \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \text{av}\left(x_1, \ldots, x_n\right)\right) \left(y_i - \text{av}\left(y_1, \ldots, y_n\right)\right). \tag{8.7}$$

In order to take into account that each individual feature may vary on a different scale, a common preprocessing step is to *normalize* each feature, dividing it by its sample standard deviation.

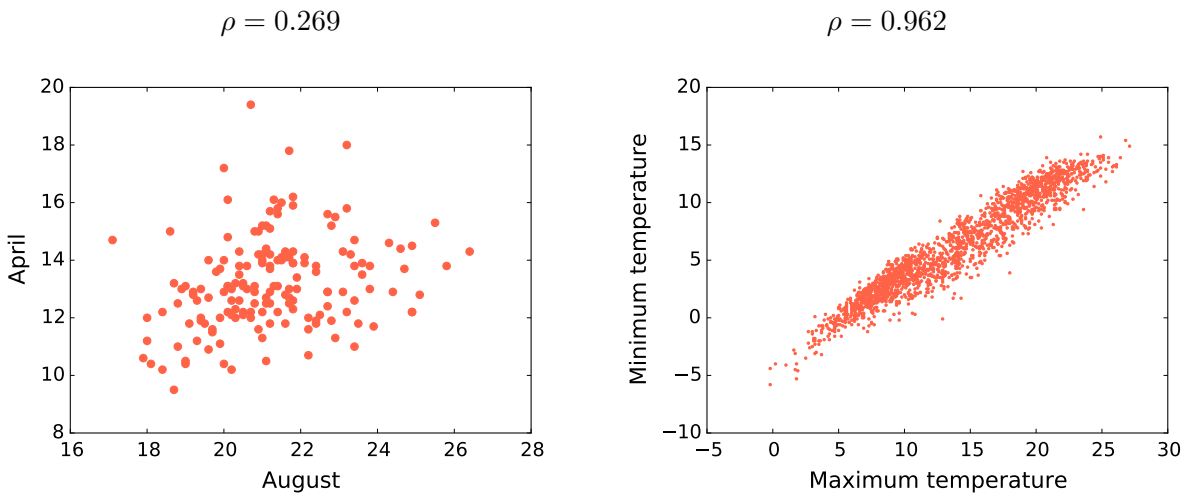


Figure 8.6: Scatterplot of the temperature in January and in August (left) and of the maximum and minimum monthly temperature (right) in Oxford over the last 150 years.

If we normalize before computing the covariance, we obtain the sample correlation coefficient of the two features. One of the advantages of the correlation coefficient is that we don't need to worry about the units in which the features are measured. In contrast, measuring a feature representing distance in inches or miles can severely distort the covariance, if we don't scale the other feature accordingly.

Definition 8.4.2 (Sample correlation coefficient)**.** *Let* $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ *be a data set where each example consists of two features. The sample correlation coefficient is defined as*

$$\rho\left((x_1, y_1), \ldots, (x_n, y_n)\right) := \frac{\operatorname{cov}\left((x_1, y_1), \ldots, (x_n, y_n)\right)}{\operatorname{std}\left(x_1, \ldots, x_n\right) \operatorname{std}\left(y_1, \ldots, y_n\right)}. \tag{8.8}$$

By the Cauchy-Schwarz inequality (Theorem B.2.4), which states that for any vectors $\vec{a}$ and $\vec{b}$

$$-1 \leq \frac{\vec{a}^T \vec{b}}{||a||_2 \, ||b||_2} \leq 1, \tag{8.9}$$

the magnitude of the sample correlation coefficient is bounded by one. If it is equal to 1 or -1, then the two centered data sets are collinear. The Cauchy-Schwarz inequality is related to the Cauchy-Schwarz inequality for random variables (Theorem 4.3.7), but here it applies to deterministic vectors.

Figure 8.6 is annotated with the sample correlation coefficients corresponding to the two plots. Maximum and minimum temperatures within the same month are highly correlated, whereas the maximum temperature in January and August within the same year are only somewhat correlated.

8.5 Sample covariance matrix

8.5.1 Definition

We now consider sets of multidimensional data. In particular, we are interested in analyzing the variation in the data. The sample covariance matrix of a data set contains the pairwise sample covariance between every pair of features.

Definition 8.5.1 (Sample covariance matrix)**.** *Let $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ be a set of d-dimensional real-valued data vectors. The sample covariance matrix of these data is the $d \times d$ matrix*

$$\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right) := \frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{x}_i - \text{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right) \left(\vec{x}_i - \text{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)^T. \tag{8.10}$$

The (i, j) entry of the covariance matrix, where $1 \leq i, j \leq d$, is given by

$$\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right)_{ij} = \begin{cases} \text{var}\left(\left(\vec{x}_1\right)_i, \ldots, \left(\vec{x}_n\right)_i\right) & \text{if } i = j, \\ \text{cov}\left(\left(\left(\vec{x}_1\right)_i, \left(\vec{x}_1\right)_j\right), \ldots, \left(\left(\vec{x}_n\right)_i, \left(\vec{x}_n\right)_j\right)\right) & \text{if } i \neq j. \end{cases} \tag{8.11}$$

In order to characterize the variation of a multidimensional data set around its center, we consider its variation in different directions. The average variation of the data in a certain direction is quantified by the sample variance of the projections of the data onto that direction. Let $\vec{v}$ be a unit-norm vector aligned with a direction of interest, the sample variance of the data set in the direction of $\vec{v}$ is given by

$$\text{var}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{v}^T\vec{x}_i - \text{av}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right)\right)^2 \tag{8.12}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{v}^T\left(\vec{x}_i - \text{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)\right)^2 \tag{8.13}$$

$$= \vec{v}^T \left(\frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{x}_i - \text{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right) \left(\vec{x}_i - \text{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)^T\right) \vec{v}$$

$$= \vec{v}^T \Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right) \vec{v}. \tag{8.14}$$

Using the sample covariance matrix we can express the variation in every direction! This is a deterministic analog of the fact that the covariance matrix of a random vector encodes its variance in every direction.

8.5.2 Principal component analysis

Consider the eigendecomposition of the covariance matrix

$$\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right) = \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n \end{bmatrix}^T. \tag{8.15}$$

By definition, $\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right)$ is symmetric, so its eigenvectors u_1, u_2, ..., u_n are orthogonal. By equation (8.14) and Theorem B.7.2, the eigenvectors and eigenvalues completely characterize the variation of the data in every direction.

$\lambda_1/n = 0.497$	$\lambda_1/n = 0.967$	$\lambda_1/n = 1.820$
$\lambda_2/n = 0.476$	$\lambda_2/n = 0.127$	$\lambda_2/n = 0.021$

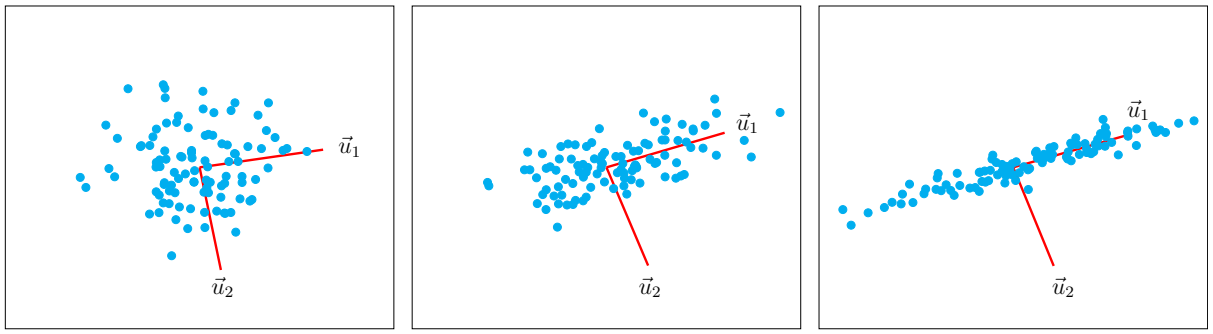


Figure 8.7: PCA of a set consisting of $n = 100$ two-dimensional data points with different configurations.

Theorem 8.5.2. *Let the sample covariance of a set of vectors $\Sigma(\vec{x}_1, \ldots, \vec{x}_n)$ have an eigendecomposition given by (8.15) where the eigenvalues are ordered $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Then,*

$$\lambda_1 = \max_{||\vec{v}||_2=1} \text{var}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right), \tag{8.16}$$

$$\vec{u}_1 = \arg\max_{||\vec{v}||_2=1} \text{var}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right), \tag{8.17}$$

$$\lambda_k = \max_{||\vec{v}||_2=1, \vec{u}\perp\vec{u}_1,\ldots,\vec{u}_{k-1}} \text{var}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right), \tag{8.18}$$

$$\vec{u}_k = \arg\max_{||\vec{v}||_2=1, \vec{u}\perp\vec{u}_1,\ldots,\vec{u}_{k-1}} \text{var}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right). \tag{8.19}$$

This means that $\vec{u}_1$ is the direction of maximum variation. The eigenvector $\vec{u}_2$ corresponding to the second largest eigenvalue λ_2 is the direction of maximum variation that is orthogonal to $\vec{u}_1$. In general, the eigenvector $\vec{u}_k$ corresponding to the kth largest eigenvalue λ_k reveals the direction of maximum variation that is orthogonal to $\vec{u}_1$, $\vec{u}_2$, ..., $\vec{u}_{k-1}$. Finally, $\vec{u}_n$ is the direction of minimum variation.

In data analysis, the eigenvectors of the sample covariance matrix are usually called principal directions. Computing these eigenvectors to quantify the variation of a data set in different directions is called **principal component analysis** (PCA). Figure 8.7 shows the principal directions for several 2D examples.

Figure 8.8 illustrates the importance of centering before applying PCA. Theorem 8.5.2 still holds if the data are not centered. However, the norm of the projection onto a certain direction no longer reflects the variation of the data. In fact, if the data are concentrated around a point that is far from the origin, the first principal direction tends be aligned with that point. This makes sense as projecting onto that direction captures more energy. As a result, the principal directions do not reflect the directions of maximum variation *within* the cloud of data. Centering the data set before applying PCA solves the issue.

The following example explains how to apply principal component analysis to dimensionality reduction. The motivation is that in many cases directions of higher variation are more informative about the structure of the data set.

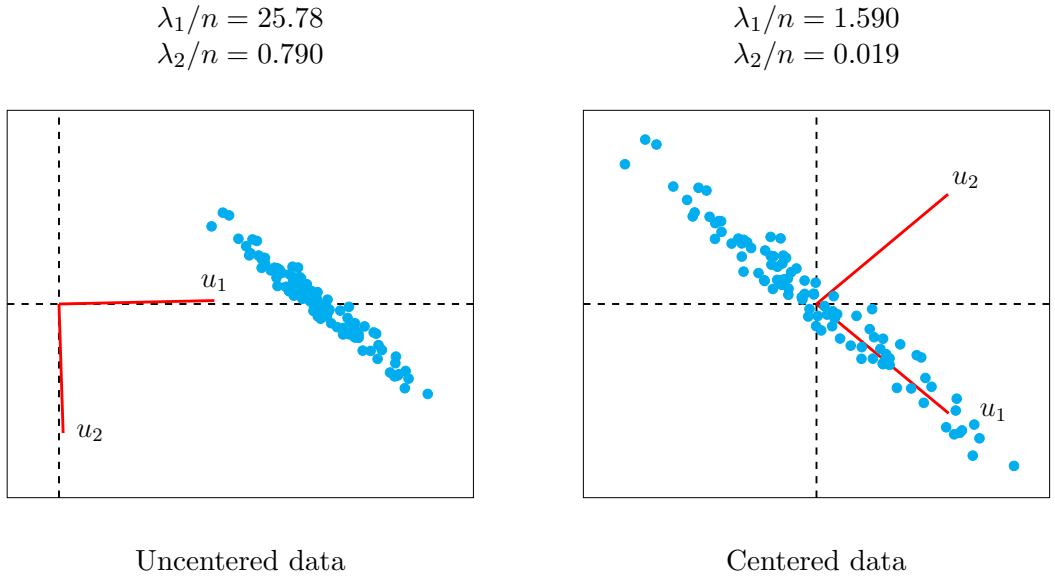


Figure 8.8: PCA applied to $n = 100$ 2D data points. On the left the data are not centered. As a result the dominant principal direction u_1 lies in the direction of the mean of the data and PCA does not reflect the actual structure. Once we center, u_1 becomes aligned with the direction of maximal variation.

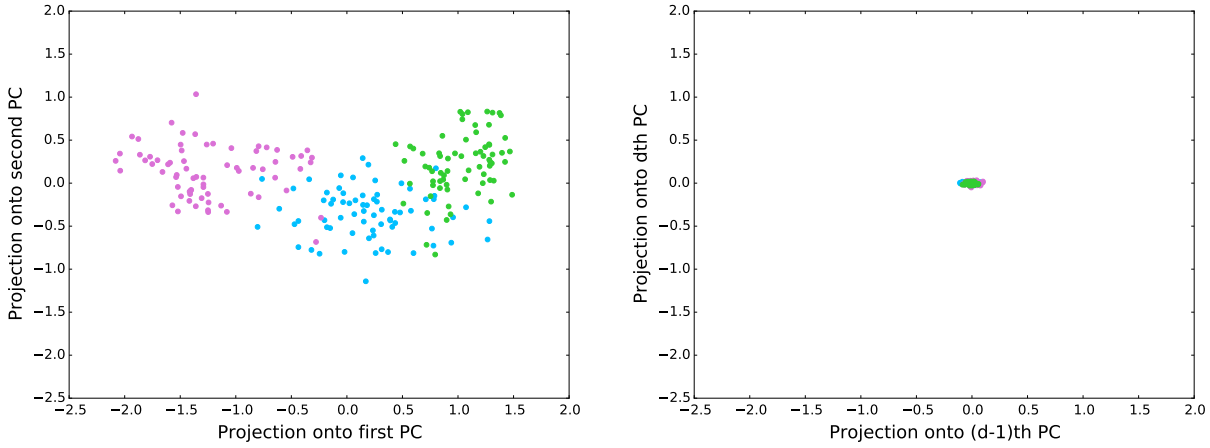


Figure 8.9: Projection of 7-dimensional vectors describing different wheat seeds onto the first two (left) and the last two (right) principal directions of the data set. Each color represents a variety of wheat.

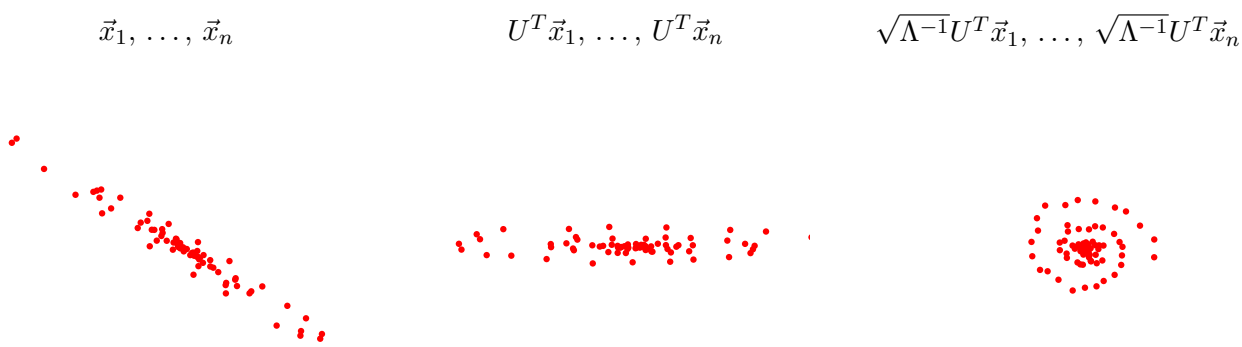


Figure 8.10: Effect of whitening a set of data. The original data are dominated by a linear skew (left). Applying U^T aligns the axes with the eigenvectors of the sample covariance matrix (center). Finally, $\sqrt{\Lambda^{-1}}$ reweights the data along those axes so that they have the same average variation, revealing the nonlinear structure that was obscured by the linear skew (right).

Example 8.5.3 (Dimensionality reduction via PCA)**.** We consider a data set where each data point corresponds to a seed which has seven features: area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient and length of kernel groove. The seeds belong to three different varieties of wheat: Kama, Rosa and Canadian.[3] Our aim is to visualize the data by projecting the data down to two dimensions in a way that preserves as much variation as possible. This can be achieved by projecting each point onto the two first principal dimensions of the data set.

Figure 8.9 shows the projection of the data onto the first two and the last two principal directions. In the latter case, there is almost no discernible variation. The structure of the data is much better conserved by the two first directions, which allow to clearly visualize the difference between the three types of seeds. Note however that projection onto the first principal directions only ensures that we preserve as much variation as possible, but it does not necessarily preserve useful features for tasks such as classification. △

8.5.3 Whitening

Whitening is a useful procedure for preprocessing data that contains nonlinear patterns. The goal is to eliminate the linear skew in the data by rotating and contracting the data along different directions in order to reveal its underlying nonlinear structure. This can be achieved by applying a linear transformation that essentially inverts the sample covariance matrix, so that the result is uncorrelated. The process is known as **whitening**, because random vectors with uncorrelated entries are often referred to as white noise. It is closely related to Algorithm 8.5.4 for coloring random vectors.

Algorithm 8.5.4 (Whitening)**.** *Let $\vec{x}_1, \ldots, \vec{x}_n$ be a set of d-dimensional data, which we assume to be centered and to have a full-rank covariance matrix. To whiten the data set we:*

1. *Compute the eigendecomposition of the sample covariance matrix $\Sigma(\vec{x}_1, \ldots, \vec{x}_n) = U\Lambda U^T$.*

[3]The data can be found at https://archive.ics.uci.edu/ml/datasets/seeds.

2. *Set* $\vec{y}_i := \sqrt{\Lambda}^{-1} U^T \vec{x}_i$, *for* $i = 1, \ldots, n$, *where*

$$\sqrt{\Lambda} := \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{bmatrix}, \tag{8.20}$$

so that $\Sigma(\vec{x}_1, \ldots, \vec{x}_n) = U\sqrt{\Lambda}\sqrt{\Lambda}U^T$.

The whitened data set $\vec{y}_1, \ldots, \vec{y}_n$ has a sample covariance matrix equal to the identity,

$$\Sigma(\vec{y}_1, \ldots, \vec{y}_n) := \frac{1}{n-1} \sum_{i=1}^{n} \vec{y}_i \vec{y}_i^T \tag{8.21}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \sqrt{\Lambda}^{-1} U^T \vec{x}_i \left(\sqrt{\Lambda}^{-1} U^T \vec{x}_i \right)^T \tag{8.22}$$

$$= \sqrt{\Lambda}^{-1} U^T \left(\frac{1}{n-1} \sum_{i=1}^{n} \vec{x}_i \vec{x}_i^T \right) U \sqrt{\Lambda}^{-1} \tag{8.23}$$

$$= \sqrt{\Lambda}^{-1} U^T \Sigma(\vec{x}_1, \ldots, \vec{x}_n) U \sqrt{\Lambda}^{-1} \tag{8.24}$$

$$= \sqrt{\Lambda}^{-1} U^T U \sqrt{\Lambda} \sqrt{\Lambda} U^T U \sqrt{\Lambda}^{-1} \tag{8.25}$$

$$= I. \tag{8.26}$$

Intuitively, whitening first rotates the data and then shrinks or expands it so that the average variation is the same in every direction. As a result, nonlinear patterns become more apparent, as illustrated by Figure 8.10.

Chapter 9

Frequentist Statistics

The goal of statistical analysis is to *extract information* from data by computing **statistics**, which are deterministic functions of the data. In Chapter 8 we describe several statistics from a deterministic and geometric point of view, without making any assumptions about the data-generation process. This makes it very challenging to evaluate the accuracy of the acquired information.

In this chapter we model the data-acquisition process probabilistically. This allows to analyze statistical techniques and derive theoretical guarantees on their performance. The data are interpreted as **realizations** of random variables, vectors or processes (depending on the dimensionality). The information that we want to extract can then be expressed in terms of the joint distribution of these quantities. We consider this distribution to be unknown but **fixed**, taking a **frequentist** perspective. The alternative framework of Bayesian statistics is described in Chapter 10.

9.1 Independent identically-distributed sampling

In this chapter we consider one-dimensional real-valued data, modeled as the realization of an iid sequence. Figure 9.1 depicts the corresponding graphical model. This is a very popular assumption, which holds for controlled experiments, such as randomized trials to test drugs, and can often be a good approximation in other settings. However, in practice it is crucial to evaluate to what extent the independence assumptions of a model actually hold.

The following example shows that measuring a quantity by sampling a subset of individuals randomly from a large population produces data satisfying the iid assumption, as long as we sample with replacement (if the population is large, sampling without replacement will have a negligible effect).

Example 9.1.1 (Sampling from a population)**.** Assume that we are studying a population of m individuals. We are interested in a certain quantity associated to each person, e.g. their cholesterol level, their salary or who they are voting for in an election. There are k possible values for the quantity $\{z_1, z_2, \ldots, z_k\}$, where k can be equal to m or much smaller. We denote by m_j the number of people for whom the quantity is equal to z_j, $1 \leq j \leq k$. In the case of an election with two candidates, k would equal two and m_1 and m_2 would represent the people voting for each of the candidates.

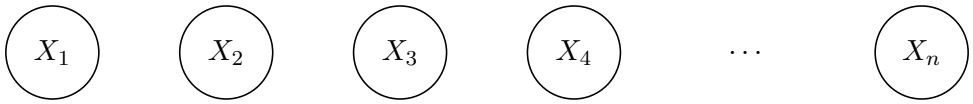


Figure 9.1: Directed graphical model corresponding to an independent sequence. If the sequence is also identically distributed, then X_1, X_2, ..., X_n all have the same distribution.

Let us assume that we select n individuals independently at random with replacement, which means that one individual could be chosen more than once, and record the value of the quantity of interest. Under these assumptions the measurements can be modeled as a random sequence of independent variables $\widetilde{X}$. Since the probability of choosing any individual is the same every time we make a selection, the first-order pmf of the sequence is

$$p_{\widetilde{X}(i)}(z_j) = P(\text{The } i\text{th measurement equals } z_j) \tag{9.1}$$

$$= \frac{\text{People such that the quantity equals } z_j}{\text{Total number of people}} \tag{9.2}$$

$$= \frac{m_j}{m}, \qquad 1 \leq j \leq k, \tag{9.3}$$

for $1 \leq i \leq n$ by the law of total probability. We conclude that the data can be modeled as a realization of an iid sequence. △

9.2 Mean square error

We define an **estimator** as a deterministic function of the available data $x_1, x_2, \ldots, x_n$ which provides an approximation to a quantity associated to the distribution that generates the data

$$y := h(x_1, x_2, \ldots, x_n). \tag{9.4}$$

For example, as we will see, if we want to estimate the expectation of the underlying distribution, a reasonable estimator is the average of the data. Since we are taking a frequentist viewpoint, the quantity of interest is modeled as deterministic (in contrast to the Bayesian viewpoint which would model it as a random variable). For a fixed data set, the estimator is a deterministic function of the data. However, if we model the data as realizations of a sequence of random variables, then the estimator is also a realization of the random variable

$$Y := h(X_1, X_2, \ldots, X_n). \tag{9.5}$$

This allows to evaluate the estimator probabilistically (usually under some assumptions on the underlying distribution). For instance, we can measure the error incurred by the estimator by computing the mean square of the difference between the estimator and the true quantity of interest.

Definition 9.2.1 (Mean square error)**.** *The mean square error (MSE) of an estimator Y that approximates a deterministic quantity $\gamma \in \mathbb{R}$ is*

$$\text{MSE}(Y) := \text{E}\left((Y - \gamma)^2\right). \tag{9.6}$$

The MSE can be decomposed into a **bias** term and a **variance** term. The bias term is the difference between the quantity of interest and the expected value of the estimator. The variance term corresponds to the variation of the estimator around its expected value.

Lemma 9.2.2 (Bias-variance decomposition)**.** *The MSE of an estimator Y that approximates $\gamma \in \mathbb{R}$ satisfies*

$$MSE(Y) = \underbrace{\mathrm{E}\left((Y - \mathrm{E}(Y))^2\right)}_{\text{variance}} + \underbrace{(\mathrm{E}(Y) - \gamma)^2}_{\text{bias}}. \tag{9.7}$$

Proof. The lemma is a direct consequence of linearity of expectation. □

If the bias is zero, then the estimator equals the quantity of interest on average.

Definition 9.2.3 (Unbiased estimator)**.** *An estimator Y that approximates $\gamma \in \mathbb{R}$ is unbiased if its bias is equal to zero, i.e. if and only if*

$$\mathrm{E}(Y) = \gamma. \tag{9.8}$$

An estimator may be unbiased but still incur in a large mean square error due to its variance.

The following lemmas establish that the sample mean and variance are unbiased estimators of the true mean and variance of an iid sequence of random variables.

Lemma 9.2.4 (The sample mean is unbiased)**.** *The sample mean is an unbiased estimator of the mean of an iid sequence of random variables.*

Proof. We consider the sample mean of an iid sequence $\widetilde{X}$ with mean μ,

$$\widetilde{Y}(n) := \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}(i). \tag{9.9}$$

By linearity of expectation

$$\mathrm{E}\left(\widetilde{Y}(n)\right) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left(\widetilde{X}(i)\right) \tag{9.10}$$

$$= \mu. \tag{9.11}$$

□

Lemma 9.2.5 (The sample variance is unbiased)**.** *The sample variance is an unbiased estimator of the variance of an iid sequence of random variables.*

The proof of this result is in Section 9.7.1.

9.3 Consistency

If we are estimating a scalar quantity, the estimate should improve as we gather more data. Ideally the estimate should converge to the true value in the limit when the number of data $n \to \infty$. Estimators that achieve this are said to be consistent.

Definition 9.3.1 (Consistency)**.** *An estimator* $\widetilde{Y}(n) := h\left(\widetilde{X}(1), \widetilde{X}(2), \ldots, \widetilde{X}(n)\right)$ *that approximates* $\gamma \in \mathbb{R}$ *is consistent if it converges to* γ *as* $n \to \infty$ *in mean square, with probability one or in probability.*

The following theorem shows that the mean is consistent.

Theorem 9.3.2 (The sample mean is consistent)**.** *The sample mean is a consistent estimator of the mean of an iid sequence of random variables as long as the variance of the sequence is bounded.*

Proof. We consider the sample mean of an iid sequence $\widetilde{X}$ with mean μ,

$$\widetilde{Y}(n) := \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}(i). \tag{9.12}$$

The estimator is equal to the moving average of the data. As a result it converges to μ in mean square (and with probability one) by the law of large numbers (Theorem 6.2.2), as long as the variance σ^2 of each of the entries in the iid sequence is bounded. □

Example 9.3.3 (Estimating the average height)**.** In this example we illustrate the consistency of the sample mean. Imagine that we want to estimate the mean height in a population. To be concrete we consider a population of $m := 25000$ people. Figure 9.2 shows a histogram of their heights.[1] As explained in Example 9.1.1 if we sample n individuals from this population with replacement, then their heights form an iid sequence $\widetilde{X}$. The mean of this sequence is

$$\mathrm{E}\left(\widetilde{X}(i)\right) := \sum_{j=1}^{m} \mathrm{P}(\text{Person } j \text{ is chosen}) \cdot \text{height of person } j \tag{9.13}$$

$$= \frac{1}{m}\sum_{j=1}^{m} h_j \tag{9.14}$$

$$= \mathrm{av}(h_1, \ldots, h_m) \tag{9.15}$$

for $1 \leq i \leq n$, where $h_1, \ldots, h_m$ are the heights of the people. In addition, the variance is bounded because the heights are finite. By Theorem 9.3.2 the sample mean of the n data should converge to the mean of the iid sequence and hence to the average height over the whole population. Figure 9.3 illustrates this numerically.

△

If the mean of the underlying distribution is not well defined, or its variance is unbounded, then the sample mean is not necessarily a consistent estimator. This is related to the fact that

[1] The data are available here: wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights.

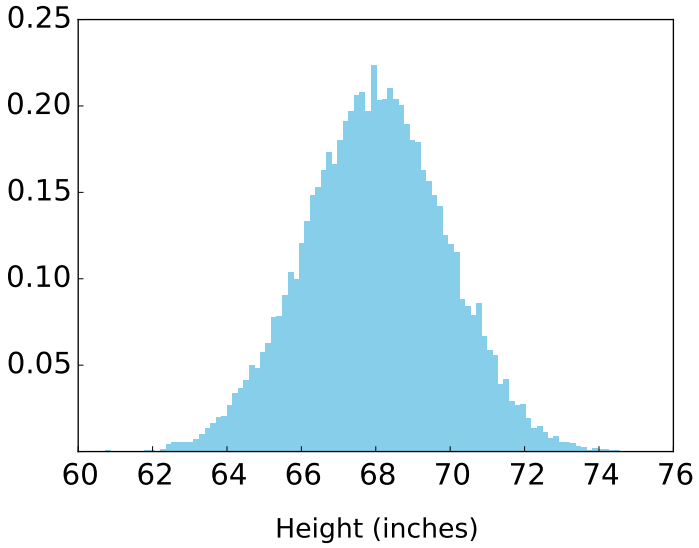


Figure 9.2: Histogram of the heights of a group of 25 000 people.

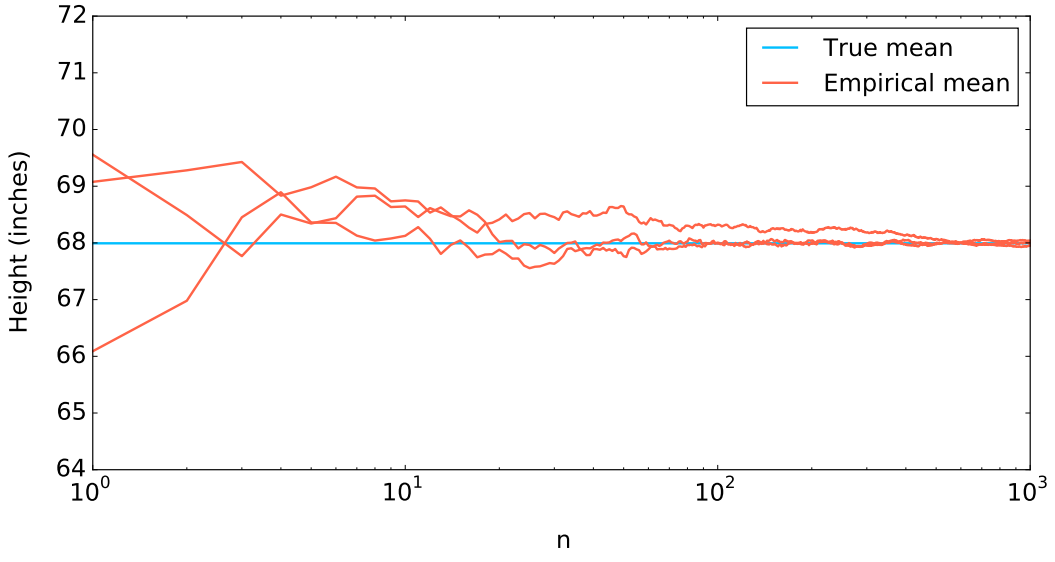


Figure 9.3: Different realizations of the sample mean when individuals from the population in Figure 9.2 are sampled with replacement.

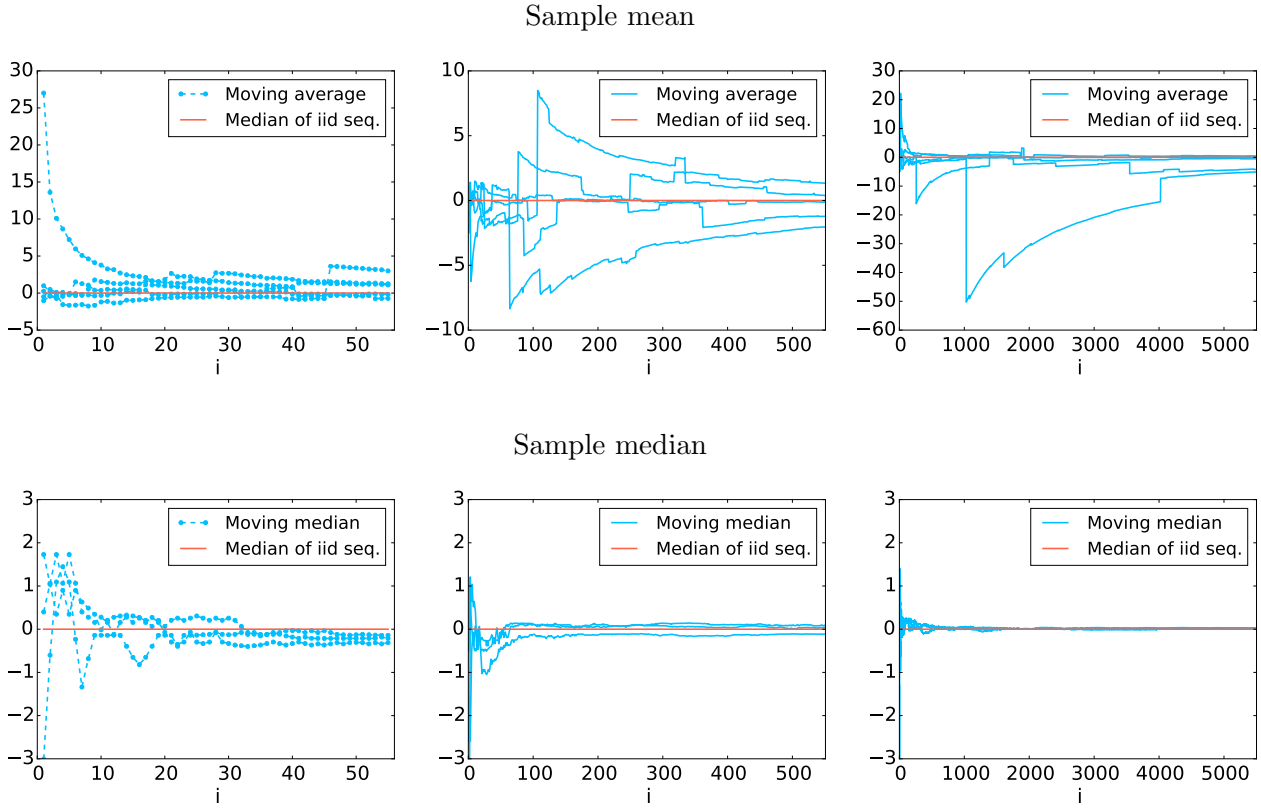


Figure 9.4: Realization of the moving average of an iid Cauchy sequence (top) compared to the moving median (bottom).

the sample mean can be severely affected by the presence of extreme values, as we discussed in Section 8.2. The sample median, in contrast, tends to be more robust in such situations, as discussed in Section 8.3. The following theorem establishes that the sample median is consistent under the iid assumption, even if the mean is not well defined or the variance is unbounded. The proof is in Section 9.7.2.

Theorem 9.3.4 (Sample median as an estimator of the median)**.** *The sample median is a consistent estimator of the median of an iid sequence of random variables.*

Figure 9.4 compares the moving average and the moving median of an iid sequence of Cauchy random variables for three different realizations. The moving average is unstable and does not converge no matter how many data are available, which is not surprising because the mean is not well defined. In contrast, the moving median does eventually converge to the true median as predicted by Theorem 9.3.4.

The sample variance and covariance are consistent estimators of the variance and covariance respectively, under certain assumptions on the higher moments of the underlying distributions. This provides an intuitive interpretation for principal component analysis (see Section 8.5.2) under the assumption that the data are realizations of an iid sequence of random vectors: the principal components approximate the eigenvectors of the true covariance matrix (see Section 4.3.3), and hence the directions of maximum variance of the multidimensional distribution. Figure 9.5

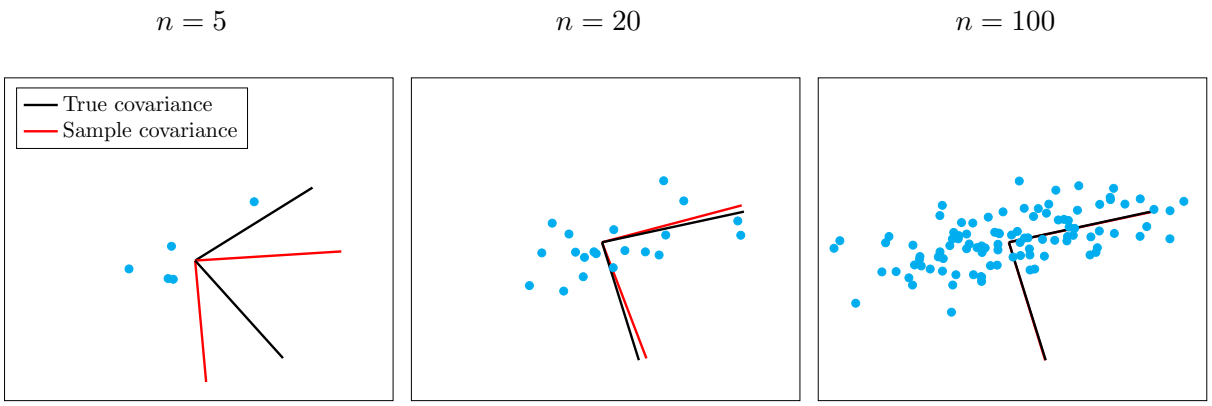


Figure 9.5: Principal directions of n samples from a bivariate Gaussian distribution (red) compared to the eigenvectors of the covariance matrix of the distribution (black).

illustrates this with a numerical example, where the principal components indeed converge to the eigenvectors as the number of data increases.

9.4 Confidence intervals

Consistency implies that an estimator will be perfect if we acquire infinite data, but this is of course impossible in practice. It is therefore important to quantify the accuracy of an estimator for a fixed number of data. Confidence intervals allow to do this from a frequentist point of view. A confidence interval can be interpreted as a *soft estimate* of the deterministic quantity of interest, which guarantees that the true value will belong to the interval with a certain probability.

Definition 9.4.1 (Confidence interval)**.** *A $1-\alpha$ confidence interval $\mathcal{I}$ for $\gamma \in \mathbb{R}$ satisfies*

$$\mathrm{P}\left(\gamma \in \mathcal{I}\right) \geq 1 - \alpha, \tag{9.16}$$

where $0 < \alpha < 1$.

Confidence intervals are usually of the form $[Y - c, Y + c]$ where Y is an estimator of the quantity of interest and c is a constant that depends on the number of data. The following theorem derives a confidence interval for the mean of an iid sequence. The confidence interval is centered at the sample mean.

Theorem 9.4.2 (Confidence interval for the mean of an iid sequence)**.** *Let $\widetilde{X}$ be an iid sequence with mean μ and variance $\sigma^2 \leq b^2$ for some $b > 0$. For any $0 < \alpha < 1$*

$$\mathcal{I}_n := \left[Y_n - \frac{b}{\sqrt{\alpha\, n}}, Y_n + \frac{b}{\sqrt{\alpha\, n}} \right], \qquad Y_n := \mathrm{av}\left(\widetilde{X}\left(1\right), \widetilde{X}\left(2\right), \ldots, \widetilde{X}\left(n\right) \right), \tag{9.17}$$

is a $1-\alpha$ confidence interval for μ.

Proof. Recall that the variance of Y_n equals $\text{Var}\left(\bar{X}_n\right) = \sigma^2/n$ (see equation (6.21) in the proof of Theorem 6.2.2). We have

$$P\left(\mu \in \left[Y_n - \frac{b}{\sqrt{\alpha\, n}}, Y_n + \frac{\sigma}{\sqrt{\alpha\, n}}\right]\right) = 1 - P\left(|Y_n - \mu| > \frac{b}{\sqrt{\alpha\, n}}\right) \tag{9.18}$$

$$\geq 1 - \frac{\alpha\, n \text{Var}\left(Y_n\right)}{b^2} \quad \text{by Chebyshev's inequality} \tag{9.19}$$

$$= 1 - \frac{\alpha\, \sigma^2}{b^2} \tag{9.20}$$

$$\geq 1 - \alpha. \tag{9.21}$$

□

The width of the interval provided in the theorem decreases with n for fixed α, which makes sense as incorporating more data reduces the variance of the estimator and hence our uncertainty about it.

Example 9.4.3 (Bears in Yosemite)**.** A scientist is trying to estimate the average weight of the black bears in Yosemite National Park. She manages to capture 300 bears. We assume that the bears are sampled uniformly at random with replacement (a bear can be weighed more than once). Under this assumptions, in Example 9.1.1 we show that the data can be modeled as iid samples and in Example 9.3.3 we show the sample mean is a consistent estimator of the mean of the whole population.

The average weight of the 300 captured bears is $Y := 200$ lbs. To derive a confidence interval from this information we need a bound on the variance. The maximum weight recorded for a black bear ever is 880 lbs. Let μ and σ^2 be the (unknown) mean and variance of the weights of the whole population. If X is the weight of a bear chosen uniformly at random from the whole population then X has mean μ and variance σ^2, so

$$\sigma^2 = \text{E}\left(X^2\right) - \text{E}^2\left(X\right) \tag{9.22}$$

$$\leq \text{E}\left(X^2\right) \tag{9.23}$$

$$\leq 880^2 \quad \text{because } X \leq 880. \tag{9.24}$$

As a result, 880 is an upper bound for the standard deviation. Applying Theorem 9.4.2,

$$\left[Y - \frac{b}{\sqrt{\alpha\, n}}, Y + \frac{b}{\sqrt{\alpha\, n}}\right] = [-27.2, 427.2] \tag{9.25}$$

is a 95% confidence interval for the average weight of the whole population. The interval is not very precise because n is not very large. △

As illustrated by this example, confidence intervals derived from Chebyshev's inequality tend to be very conservative. An alternative is to leverage the central limit theorem (CLT). The CLT characterizes the distribution of the sample mean asymptotically, so confidence intervals derived from it are not guaranteed to be precise. However, the CLT often provides a very accurate approximation to the distribution of the sample mean for finite n, as we show through some numerical examples in Chapter 6. In order to obtain confidence intervals for the mean of an iid sequence from the CLT as stated in Theorem 6.3.1 we would need to know the true variance of

the sequence, which is unrealistic in practice. However, the following result states that we can substitute the true variance with the sample variance. The proof is beyond the scope of these notes.

Theorem 9.4.4 (Central limit theorem with sample standard deviation). *Let $\widetilde{X}$ be an iid discrete random process with mean $\mu_{\widetilde{X}} := \mu$ such that its variance and fourth moment $\mathrm{E}(\widetilde{X}(i)^4)$ are bounded. The sequence*

$$\frac{\sqrt{n}\left(\mathrm{av}\left(\widetilde{X}(1),\ldots,\widetilde{X}(n)\right)-\mu\right)}{\mathrm{std}\left(\widetilde{X}(1),\ldots,\widetilde{X}(n)\right)} \tag{9.26}$$

converges in distribution to a standard Gaussian random variable.

Recall that the cdf of a standard Gaussian does not have a closed-form expression. To simplify notation we express the confidence interval in terms of the Q function.

Definition 9.4.5 (Q function). *$Q(x)$ is the probability that a standard Gaussian random variable is greater than x for positive x,*

$$Q(x) := \int_{u=x}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du, \quad x > 0. \tag{9.27}$$

By symmetry, if U is a standard Gaussian random variable and $y < 0$

$$\mathrm{P}(U < y) = Q(-y). \tag{9.28}$$

Corollary 9.4.6 (Approximate confidence interval for the mean). *Let $\widetilde{X}$ be an iid sequence that satisfies the conditions of Theorem 9.4.4. For any $0 < \alpha < 1$*

$$\mathcal{I}_n := \left[Y_n - \frac{S_n}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right), Y_n + \frac{S_n}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)\right], \tag{9.29}$$

$$Y_n := \mathrm{av}\left(\widetilde{X}(1),\widetilde{X}(2),\ldots,\widetilde{X}(n)\right), \tag{9.30}$$

$$S_n := \mathrm{std}\left(\widetilde{X}(1),\widetilde{X}(2),\ldots,\widetilde{X}(n)\right), \tag{9.31}$$

is an approximate $1-\alpha$ confidence interval for μ, i.e.

$$P(\mu \in \mathcal{I}_n) \approx 1-\alpha. \tag{9.32}$$

Proof. By the central limit theorem, when $n \to \infty$ $\bar{X}_n$ is distributed as a Gaussian random variable with mean μ and variance σ^2. As a result

$$P(\mu \in \mathcal{I}_n) = 1 - P\left(Y_n > \mu + \frac{S_n}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)\right) - P\left(Y_n < \mu - \frac{S_n}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)\right) \tag{9.33}$$

$$= 1 - P\left(\frac{\sqrt{n}(Y_n-\mu)}{S_n} > Q^{-1}\left(\frac{\alpha}{2}\right)\right) - P\left(\frac{\sqrt{n}(Y_n-\mu)}{S_n} < -Q^{-1}\left(\frac{\alpha}{2}\right)\right) \tag{9.34}$$

$$\approx 1 - 2Q\left(Q^{-1}\left(\frac{\alpha}{2}\right)\right) \quad \text{by Theorem 9.4.4} \tag{9.35}$$

$$= 1-\alpha. \tag{9.36}$$

□

It is important to stress that the result only provides an accurate confidence interval if n is large enough for the sample variance to converge to the true variance and for the CLT to take effect.

Example 9.4.7 (Bears in Yosemite (continued))**.** The sample standard deviation of the bears captured by the scientist equals 100 lbs. We apply Corollary 9.4.6 to derive an approximate confidence interval that is tighter than the one obtained applying Chebyshev's inequality. Given that $Q\left(1.95\right) \approx 0.025$,

$$\left[Y - \frac{\sigma}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right), Y + \frac{\sigma}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)\right] \approx [188.8, 211.3] \tag{9.37}$$

is an approximate 95% confidence interval for the mean weight of the population of bears.

△

Interpreting confidence intervals is somewhat tricky. After computing the confidence interval in Example 9.4.7 one is tempted to state:

The probability that the average weight is between 188.8 and 211.3 lbs is 0.95.

However we are modeling the average weight as a deterministic quantity, so there are no random quantities in this statement! The correct interpretation is that if we repeat the process of sampling the population and compute the confidence interval many times, then the true value will lie in the interval 95% of the time. This is illustrated in the following example and Figure 9.6.

Example 9.4.8 (Estimating the average height (continued))**.** Figure 9.6 shows several 95% confidence intervals for the average of the height population in Example 9.3.3. To compute each interval we select n individuals and then apply Corollary 9.4.6. The width of the intervals decreases as n grows, but because they are all 95% confidence intervals they all contain the true average with probability 0.95. Indeed this is the case for 113 out of 120 (94%) of the intervals that are plotted.

△

9.5 Nonparametric model estimation

In this section we consider the problem of estimating a distribution from multiple iid samples. This requires approximating the cdf, pmf or pdf of the distribution. If we assume that the distribution belongs to a predefined family, then the problem reduces to estimating the parameters that characterize that particular family, as we explain in detail in Section 9.6. Here we do not make such an assumption. Estimating a distribution directly is very challenging; clearly many (infinite!) different distributions could have generated the data. However with enough samples it is often possible to obtain models that produce an accurate approximation, as long as the iid assumption holds.

9.5.1 Empirical cdf

Under the assumption that a data set corresponds to iid samples from a certain distribution, a reasonable estimate for the cdf of the distribution at a given point x is the fraction of samples that are smaller than x. This results in a piecewise constant estimator known as the empirical cdf.

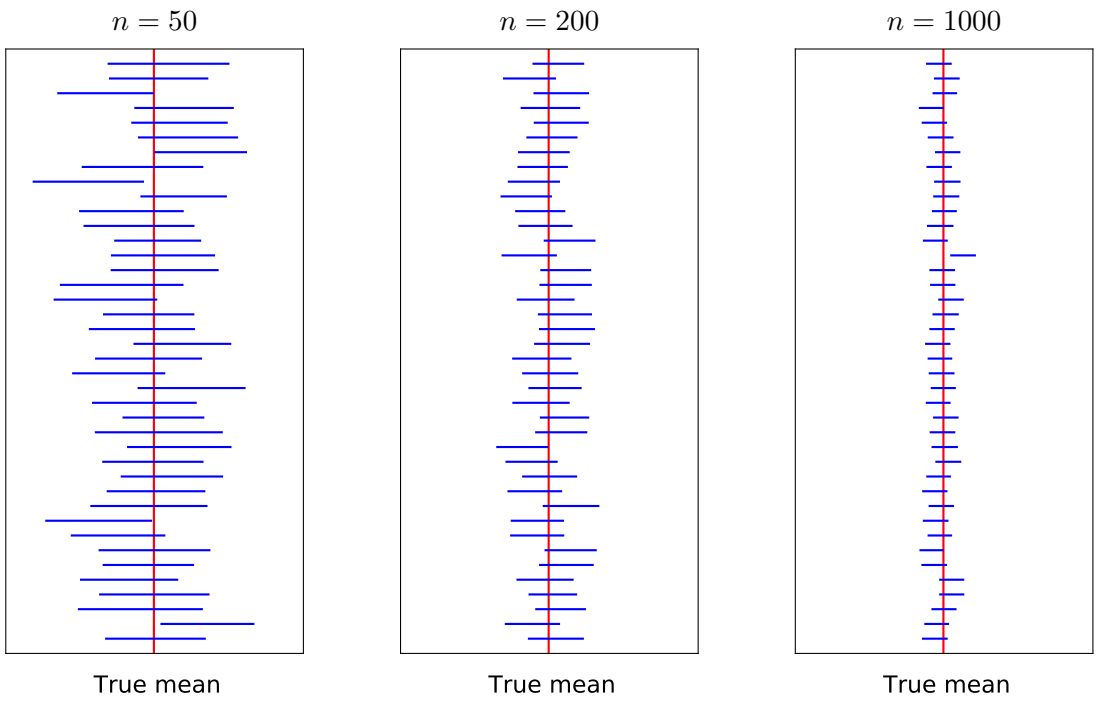


Figure 9.6: 95% confidence intervals for the average of the height population in Example 9.3.3.

Definition 9.5.1 (Empirical cdf)**.** *The empirical cdf corresponding to data* $x_1, \ldots, x_n$ *is*

$$\widehat{F}_n(x) := \frac{1}{n}\sum_{i=1}^{n} 1_{x_i \leq x}, \tag{9.38}$$

where $x \in \mathbb{R}$.

The empirical cdf is an unbiased and consistent estimator of the true cdf. This is established rigorously in Theorem 9.5.2 below and illustrated empirically in Figure 9.7. The cdf of the height data from 25,000 people is compared to three realizations of the empirical cdf computed from different numbers of iid samples. As the number of available samples grows, the approximation becomes very accurate.

Theorem 9.5.2. *Let* $\widetilde{X}$ *be an iid sequence with marginal cdf* F_X*. For any fixed* $x \in \mathbb{R}$ $\widehat{F}_n(x)$ *is an unbiased and consistent estimator of* $F_X(x)$*. In fact,* $\widehat{F}_n(x)$ *converges in mean square to* $F_X(x)$*.*

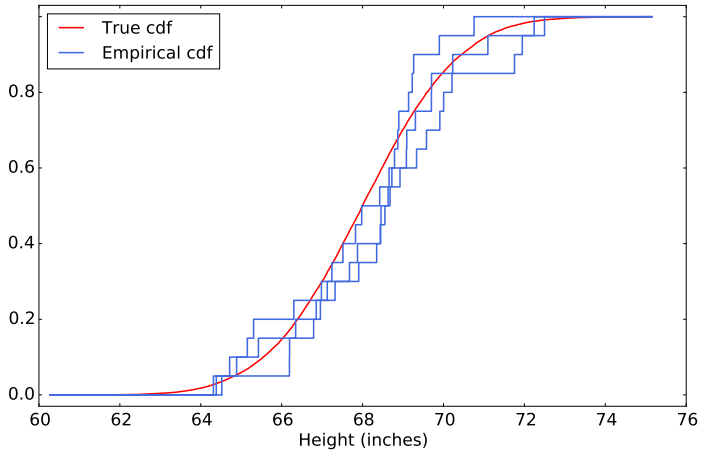
Proof. First, we verify

$$\mathrm{E}\left(\widehat{F}_n(x)\right) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} 1_{\widetilde{X}(i) \leq x}\right) \tag{9.39}$$

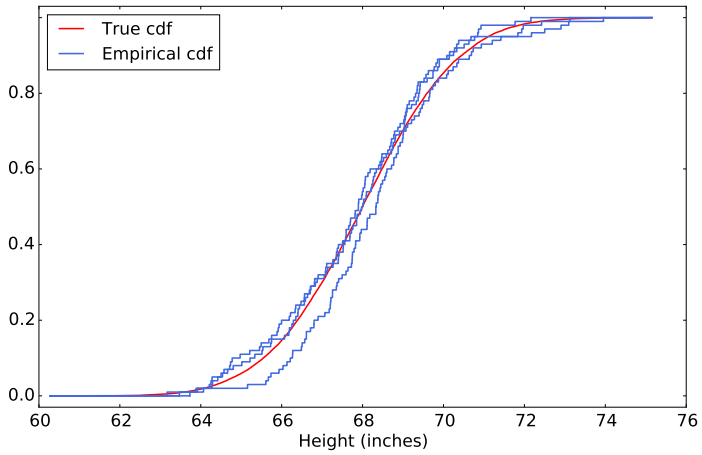
$$= \frac{1}{n}\sum_{i=1}^{n} \mathrm{P}\left(\widetilde{X}(i) \leq x\right) \quad \text{by linearity of expectation} \tag{9.40}$$

$$= F_X(x), \tag{9.41}$$

$n = 10$



$n = 100$



$n = 1000$

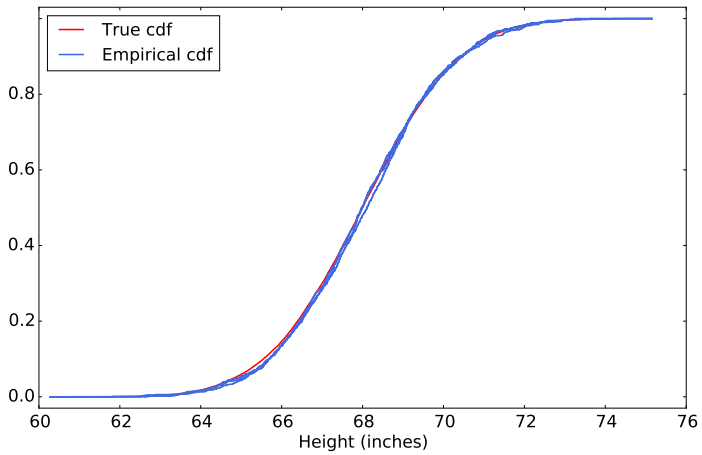


Figure 9.7: Cdf of the height data in Figure 2.13 along with three realizations of the empirical cdf computed with n iid samples for $n = 10, 100, 1000$.

so the estimator is unbiased. We now estimate its mean square

$$\mathrm{E}\left(\widehat{F}_n^2(x)\right) = \mathrm{E}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} 1_{\widetilde{X}(i)\leq x} 1_{\widetilde{X}(j)\leq x}\right) \tag{9.42}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{P}\left(\widetilde{X}(i)\leq x\right) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}\mathrm{P}\left(\widetilde{X}(i)\leq x, \widetilde{X}(j)\leq x\right) \tag{9.43}$$

$$= \frac{F_X(x)}{n} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} F_{\widetilde{X}(i)}(x) F_{\widetilde{X}(j)}(x) \quad \text{by independence} \tag{9.44}$$

$$= \frac{F_X(x)}{n} + \frac{n-1}{n}F_X^2(x). \tag{9.45}$$

The variance is consequently equal to

$$\mathrm{Var}\left(\widehat{F}_n(x)\right) = \mathrm{E}\left(\widehat{F}_n(x)^2\right) - \mathrm{E}^2\left(\widehat{F}_n(x)\right) \tag{9.46}$$

$$= \frac{F_X(x)(1-F_X(x))}{n}. \tag{9.47}$$

We conclude that

$$\lim_{n\to\infty}\mathrm{E}\left(\left(F_X(x)-\widehat{F}_n(x)\right)^2\right) = \lim_{n\to\infty}\mathrm{Var}\left(\widehat{F}_n(x)\right) = 0. \tag{9.48}$$

□

9.5.2 Density estimation

Estimating the pdf of a continuous quantity is much more challenging that estimating the cdf. If we have sufficient data, the fraction of samples that are smaller than a certain x provide a good estimate for the cdf at that point. However, no matter how much data we have, there is negligible probability that we will see any samples exactly at x: a pointwise empirical density estimator would equal zero almost everywhere (except at the available samples).

Our only hope to produce an accurate estimator is if the pdf that we aim to estimate is smooth. In that case, we can estimate its value at a point x from observed samples that are situated at neighboring locations. If there are many samples close to x then this suggests that the estimate at x should be large, whereas if all the samples are far away, then it should be small. **Kernel density estimation** achieves this by averaging the samples.

Definition 9.5.3 (Kernel density estimator)**.** *The kernel density estimate with bandwidth h of the distribution of $x_1, \ldots, x_n$ at $x \in \mathbb{R}$ is*

$$\widehat{f}_{h,n}(x) := \frac{1}{n\,h}\sum_{i=1}^{n} k\left(\frac{x-x_i}{h}\right), \tag{9.49}$$

where k is a kernel function centered at the origin that satisfies

$$k(x) \geq 0 \quad \textit{for all } x \in \mathbb{R}, \tag{9.50}$$

$$\int_{\mathbb{R}} k(x)\, dx = 1. \tag{9.51}$$

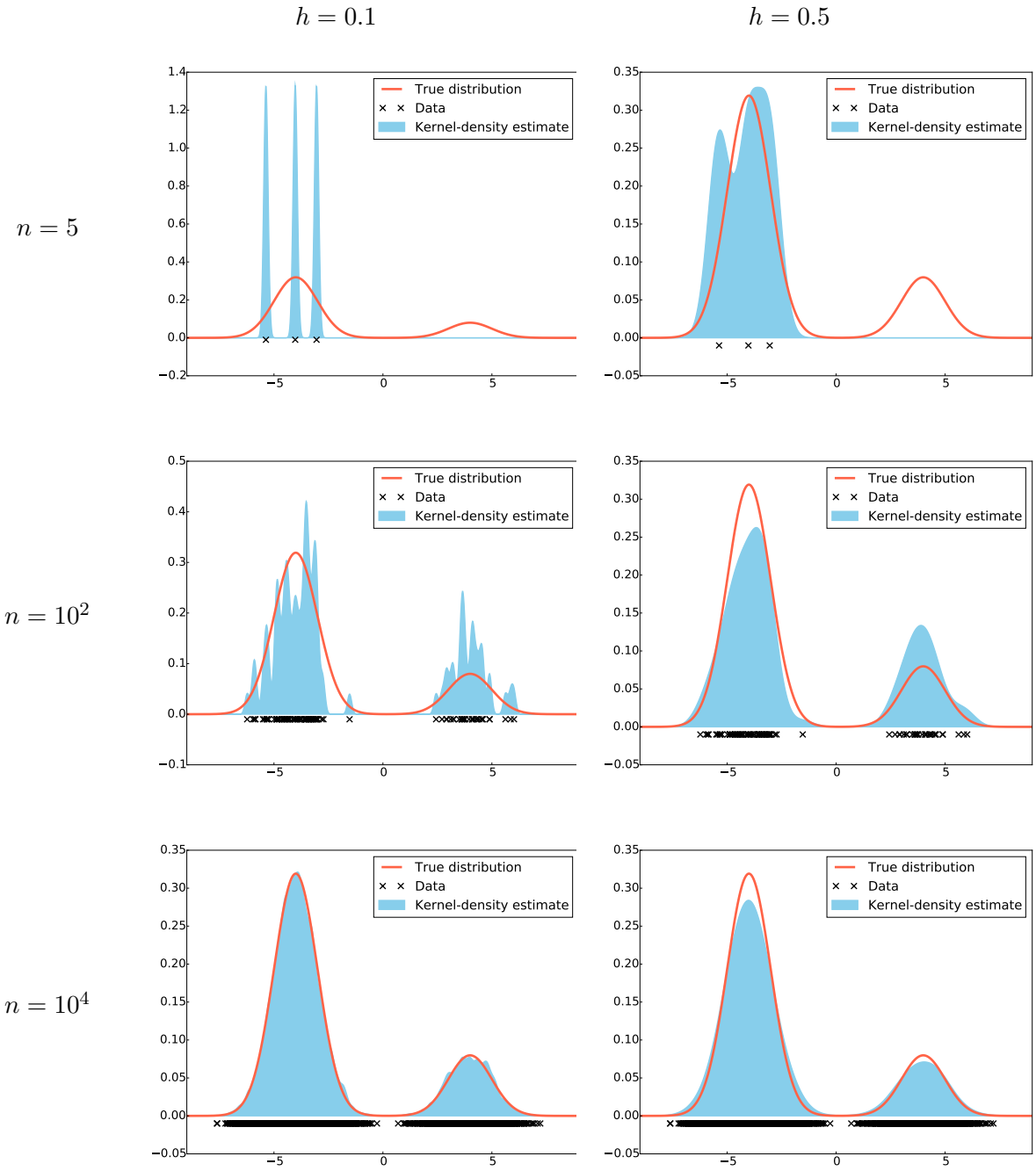


Figure 9.8: Kernel density estimation for the Gaussian mixture described in Example 9.6.5 for different number of iid samples and different values of the kernel bandwidth h.

The effect of the kernel is to weight each sample according to their distance to the point at which we are estimating the pdf x. Choosing a rectangular kernel yields an empirical density estimate that is piecewise constant and roughly looks like a histogram (the corresponding weights are constant or equal to zero). A popular alternative is the Gaussian kernel $k(x) = \exp\left(-x^2\right)/\sqrt{\pi}$, which produces a smooth density estimate. The kernel should decay so that $k\left(\left(x - x_i\right)/h\right)$ is large when the sample x_i is close to x and small when it is far. This decay is governed by the bandwidth h, which is chosen before hand based on our expectations about the smoothness of the pdf and on the amount of available data. If the bandwidth is very small, individual samples have a large influence on the density estimate. This allows to reproduce irregular shapes more easily, but also yields spurious fluctuations that are not present in the true curve, especially if we don't have a lot of samples. Increasing the bandwidth smooths out such fluctuations and yields more stable estimates when the number of data is small. However, it may also over-smooth the estimate. As a rule of thumb, we should decrease the bandwidth of the kernel as the number of data increases.

Figures 9.8 and 9.9 illustrate the effect of varying the bandwidth h at different sampling rates. In Figure 9.8 Gaussian kernel density estimation is applied to estimate the Gaussian mixture described in Example 9.6.5. Figure 9.9 shows an example where the same technique is used on real data: the aim is to estimate the density of the weight of a sea-snail population.[2] The whole population consists of 4,177 individuals. The kernel density estimate is computed from 200 iid samples for different values of the kernel bandwidth.

9.6 Parametric model estimation

In the previous section, we describe how to estimate a distribution by directly estimating the cdf or pdf generating the data. In this section, we discuss an alternative route based on the assumption that the type of distribution generating the data is known beforehand. If this is the case, the problem boils down to fitting the parameters characterizing the distribution to the data. Recall that from a frequentist viewpoint, the true distribution is fixed, so the corresponding parameters are modeled as deterministic quantities (in contrast, in a Bayesian framework they are modeled as random variables).

9.6.1 The method of moments

The method of moments adjusts the parameters of a distribution so that the moments of the distribution coincide with the sample moments of the data (i.e. its mean, mean square or variance, etc.). If the distribution only depends on one parameter, then we use the sample mean as a surrogate for the true mean and compute the corresponding value of the parameter. For an exponential with parameter λ and mean μ we have

$$\mu = \frac{1}{\lambda}. \tag{9.52}$$

Assuming that we have access to n iid samples $x_1, \ldots, x_n$ from the exponential distribution, the method-of-moments estimate of λ equals

$$\lambda_{\mathrm{MM}} := \frac{1}{\mathrm{av}\left(x_1, \ldots, x_n\right)}. \tag{9.53}$$

[2] The data are available at archive.ics.uci.edu/ml/datasets/Abalone

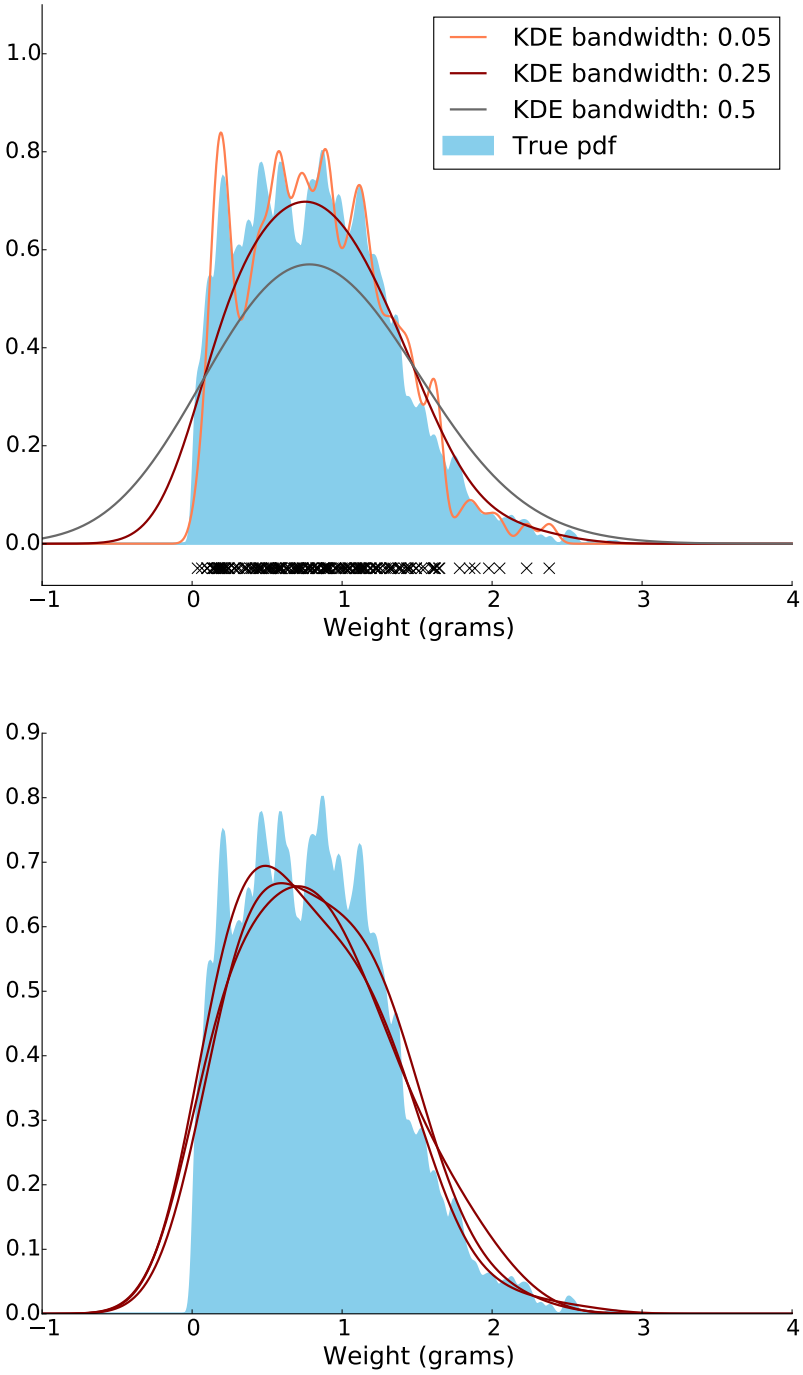


Figure 9.9: Kernel density estimate for the weight of a population of abalone, a species of sea snail. In the plot above the density is estimated from 200 iid samples using a Gaussian kernel with three different bandwidths. Black crosses representing the individual samples are shown underneath. In the plot below we see the result of repeating the procedure three times using a fixed bandwidth equal to 0.25.

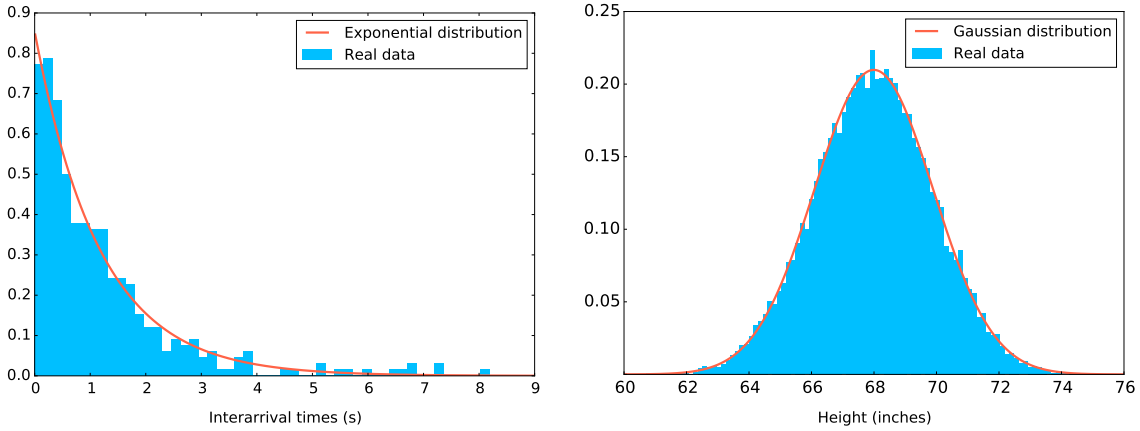


Figure 9.10: Exponential distribution fitted to data consisting of inter-arrival times of calls at a call center in Israel (left). Gaussian distribution fitted to height data (right).

The graph on the right of Figure 9.10 shows the result of fitting an exponential to the call-center data in Figure 2.11. Similarly, to fit a Gaussian using the method of moments we set the mean equal to its sample mean and the variance equal to the sample variance, as illustrated by the graph on the right of Figure 9.10 using the data from Figure 2.13.

9.6.2 Maximum likelihood

The most popular method for learning parametric models is maximum-likelihood fitting. The **likelihood** function is the joint pmf or pdf of the data, interpreted as a *function of the unknown parameters*. In more detail, let us denote the data by $x_1, \ldots, x_n$ and assume that they are realizations of a set of discrete random variables $X_1, \ldots, X_n$ which have a joint pmf that depends on a vector of parameters $\vec{\theta}$. To emphasize that the joint pmf depends on $\vec{\theta}$ we denote it by $p_{\vec{\theta}} := p_{X_1,\ldots,X_n}$. This pmf evaluated at the observed data

$$p_{\vec{\theta}}(x_1, \ldots, x_n) \tag{9.54}$$

is the likelihood function, when we interpret it as a function of $\vec{\theta}$. For continuous random variables, we use the joint pdf of the data instead.

Definition 9.6.1 (Likelihood function)**.** *Given a realization $x_1, \ldots, x_n$ of a set of discrete random variables $X_1, \ldots, X_n$ with joint pmf $p_{\vec{\theta}}$, where $\vec{\theta} \in \mathbb{R}^m$ is a vector of parameters, the likelihood function is*

$$\mathcal{L}_{x_1,\ldots,x_n}\left(\vec{\theta}\right) := p_{\vec{\theta}}(x_1, \ldots, x_n)\,. \tag{9.55}$$

If the random variables are continuous with pdf $f_{\vec{\theta}}$, where $\vec{\theta} \in \mathbb{R}^m$, the likelihood function is

$$\mathcal{L}_{x_1,\ldots,x_n}\left(\vec{\theta}\right) := f_{\vec{\theta}}(x_1, \ldots, x_n)\,. \tag{9.56}$$

*The **log-likelihood function** is equal to the logarithm of the likelihood function* $\log \mathcal{L}_{x_1,\ldots,x_n}\left(\vec{\theta}\right)$.

When the data are modeled as iid samples, the likelihood factors into a product of the marginal pmf or pdf, so the log likelihood can be decomposed into a sum.

In the case of discrete distributions, for a fixed $\vec{\theta}$ the likelihood is the probability that $X_1, \ldots, X_n$ equal the observed data. If we don't know $\vec{\theta}$, it makes sense to choose a value for $\vec{\theta}$ such that this probability is as high as possible, i.e. to maximize the likelihood. For continuous distributions we apply the same principle to the joint pdf of the data.

Definition 9.6.2 (Maximum-likelihood estimator)**.** *The* ***maximum likelihood*** *(ML) estimator for the vector of parameters $\vec{\theta} \in \mathbb{R}^m$ is*

$$\vec{\theta}_{\mathrm{ML}}\left(x_1, \ldots, x_n\right) := \arg\max_{\vec{\theta}} \mathcal{L}_{x_1,\ldots,x_n}\left(\vec{\theta}\right) \tag{9.57}$$

$$= \arg\max_{\vec{\theta}} \log \mathcal{L}_{x_1,\ldots,x_n}\left(\vec{\theta}\right). \tag{9.58}$$

The maximum of the likelihood function and that of the log-likelihood function are at the same location because the logarithm is monotone.

Under certain conditions, one can show that the maximum-likelihood estimator is consistent: it converges in probability to the true parameter as the number of data increases. One can also show that its distribution converges to that of a Gaussian random variable (or vector), just like the distribution of the sample mean. These results are beyond the scope of the course. Bear in mind, however, that they only hold if the data are indeed generated by the type of distribution that we are considering.

We now show how to derive the maximum-likelihood for a Bernoulli and a Gaussian distribution. The resulting estimators for the parameters are the same as the method-of-moments estimators (except for a slight difference in the estimate of the Gaussian variance parameter).

Example 9.6.3 (ML estimator of the parameter of a Bernoulli distribution)**.** We model a set of data $x_1, \ldots, x_n$ as iid samples from a Bernoulli distribution with parameter θ (in this case there is only one parameter). The likelihood function is equal to

$$\mathcal{L}_{x_1,\ldots,x_n}\left(\theta\right) = p_\theta\left(x_1, \ldots, x_n\right) \tag{9.59}$$

$$= \prod_{i=1} \left(1_{x_i=1}\theta + 1_{x_i=0}\left(1-\theta\right)\right) \tag{9.60}$$

$$= \theta^{n_1}\left(1-\theta\right)^{n_0} \tag{9.61}$$

and the log-likelihood function to

$$\log \mathcal{L}_{x_1,\ldots,x_n}\left(\theta\right) = n_1 \log\theta + n_0 \log\left(1-\theta\right), \tag{9.62}$$

where n_1 are the number of samples equal to one and n_0 the number of samples equal to zero. The ML estimator of the parameter θ is

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} \log \mathcal{L}_{x_1,\ldots,x_n}\left(\theta\right) \tag{9.63}$$

$$= \arg\max_{\theta} n_1 \log\theta + n_0 \log\left(1-\theta\right). \tag{9.64}$$

We compute the derivative and second derivative of the log-likelihood function,

$$\frac{\mathrm{d} \log \mathcal{L}_{x_1,\ldots,x_n}(\theta)}{\mathrm{d}\theta} = \frac{n_1}{\theta} - \frac{n_0}{1-\theta}, \tag{9.65}$$

$$\frac{\mathrm{d}^2 \log \mathcal{L}_{x_1,\ldots,x_n}(\theta)}{\mathrm{d}\theta^2} = -\frac{n_1}{\theta^2} - \frac{n_0}{(1-\theta)^2} < 0. \tag{9.66}$$

The function is concave, as the second derivative is negative. The maximum is consequently at the point where the first derivative equals zero, namely

$$\theta_{\mathrm{ML}} = \frac{n_1}{n_0 + n_1}. \tag{9.67}$$

The estimate is equal to the fraction of samples that are equal to one.

△

Example 9.6.4 (ML estimator of the parameters of a Gaussian distribution)**.** Let $x_1, x_2, \ldots$ be data that we wish to model as iid samples from a Gaussian distribution with mean μ and standard deviation σ. The likelihood function is equal to

$$\mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma) = f_{\mu,\sigma}(x_1,\ldots,x_n) \tag{9.68}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{9.69}$$

and the log-likelihood function to

$$\log \mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma) = -\frac{n \log(2\pi)}{2} - n \log\sigma - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}. \tag{9.70}$$

The ML estimator of the parameters μ and σ is

$$\{\mu_{\mathrm{ML}}, \sigma_{\mathrm{ML}}\} = \arg\max_{\{\mu,\sigma\}} \log \mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma) \tag{9.71}$$

$$= \arg\max_{\{\mu,\sigma\}} -n\log\sigma - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}. \tag{9.72}$$

We compute the partial derivatives of the log-likelihood function,

$$\frac{\partial \log \mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma)}{\partial\mu} = -\sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2}, \tag{9.73}$$

$$\frac{\partial \log \mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma)}{\partial\sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^3}. \tag{9.74}$$

The function we are trying to maximize is strictly concave in $\{\mu, \sigma\}$. To prove this, we would have to show that the Hessian of the function is positive definite. We omit the calculations that show that this is the case. Setting the partial derivatives to zero we obtain

$$\mu_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{9.75}$$

$$\sigma^2_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu_{\mathrm{ML}})^2. \tag{9.76}$$

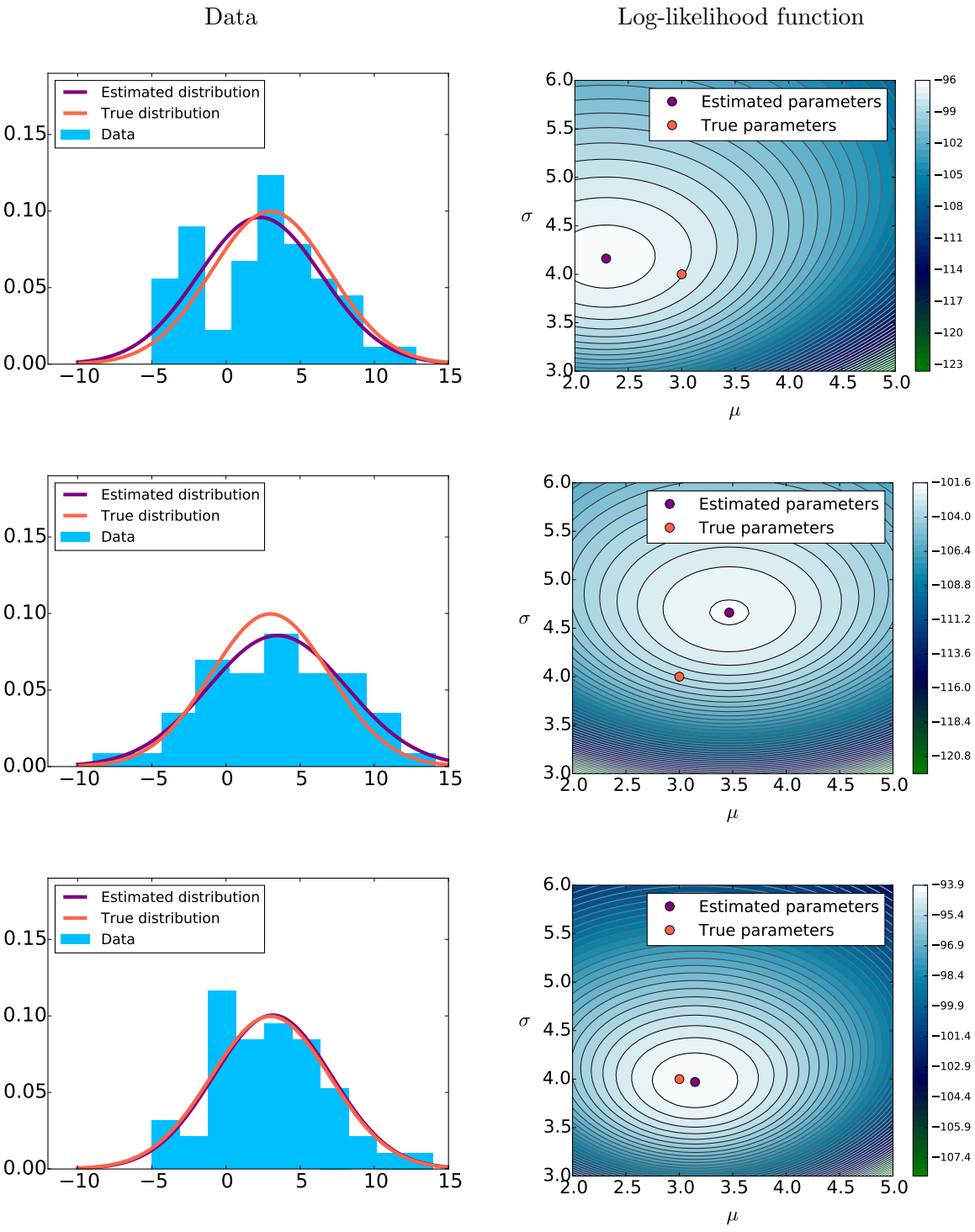


Figure 9.11: The left column shows histograms of 50 iid samples from a Gaussian distribution, together with the pdf of the original distribution, as well as the maximum-likelihood estimate. The right column shows the log-likelihood function corresponding to the data and the location of its maximum and of the point corresponding to the true parameters.

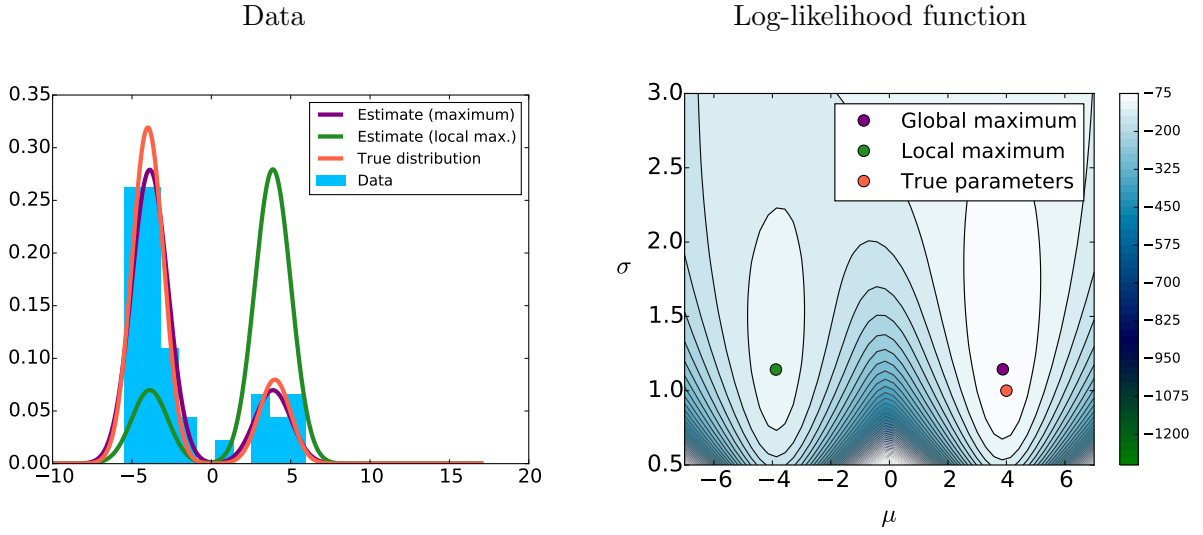


Figure 9.12: The left image shows a histogram of 40 iid samples from the Gaussian mixture defined in Example 9.6.5, together with the pdf of the original distribution. The right image shows the log-likelihood function corresponding to the data, which has a local maximum apart from the global maximum. The density estimates corresponding to the two maxima are shown on the left.

The estimator for the mean is just the sample mean. The estimator for the variance is a rescaled sample variance.

△

Figure 9.11 displays the log-likelihood function corresponding to 50 iid samples from a Gaussian distribution with $\mu := 3$ and $\sigma := 4$. It also shows the approximation to the true pdf obtained by maximum likelihood. In Examples 9.6.3 and 9.6.4 the log-likelihood function is strictly concave. This means that the function has a unique maximum that can be located by setting the gradient to zero. When this yields nonlinear equations that cannot be solved directly, we can leverage optimization methods such as gradient ascent that will converge to the maximum. However, the log-likelihood function is not always concave. As illustrated by the following example, in such cases it can have multiple local maxima, which may make it intractable to compute the maximum-likelihood estimator.

Example 9.6.5 (Log-likelihood function of a Gaussian mixture)**.** Let X be a Gaussian mixture defined as

$$X := \begin{cases} G_1 & \text{with probability } \frac{1}{5}, \\ G_2 & \text{with probability } \frac{4}{5}, \end{cases} \tag{9.77}$$

where G_1 is a Gaussian random variable with mean $-\mu$ and variance σ^2, whereas G_2 is also Gaussian with mean μ and variance σ^2. We have parameterized the mixture with just two parameters so that we can visualize the log-likelihood in two dimensions. Let $x_1, x_2, \ldots$ be data

modeled as iid samples from X. The likelihood function is equal to

$$\mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma) = f_{\mu,\sigma}(x_1,\ldots,x_n) \tag{9.78}$$

$$= \prod_{i=1}^{n} \frac{1}{5\sqrt{2\pi}\sigma} e^{-\frac{(x_i+\mu)^2}{2\sigma^2}} + \frac{4}{5\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{9.79}$$

and the log-likelihood function to

$$\log \mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma) = \sum_{i=1}^{n} \log\left(\frac{1}{5\sqrt{2\pi}\sigma} e^{-\frac{(x_i+\mu)^2}{2\sigma^2}} + \frac{4}{5\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right). \tag{9.80}$$

Figure 9.12 shows the log-likelihood function for 40 iid samples of the distribution when $\mu := 4$ and $\sigma := 1$. The function has a local maximum away from the global maximum. This means that if we use a local ascent method to find the ML estimator, we might not find the global maximum, but remain stuck at the local maximum instead. The estimate corresponding to the local maximum (shown on the left) has the same variance as the global maximum but μ is close to -4 instead of 4. Although the estimate doesn't fit the data very well, it is locally optimal, small shifts of μ and σ yield worse fits (in terms of the likelihood).

△

To finish this section, we describe a machine-learning algorithm for supervised learning based on parametric fitting using ML estimation.

Example 9.6.6 (Quadratic discriminant analysis)**.** Quadratic discriminant analysis is an algorithm for supervised learning. The input to the algorithm are two sets of training data, consisting of d-dimensional vectors $\vec{a}_1,\ldots,\vec{a}_n$ and $\vec{b}_1,\ldots,\vec{b}_n$ which belong to two different classes (the method can easily be extended to deal with more classes). The goal is to classify new instances based on the structure of the data.

To perform quadratic discriminant analysis we first fit a d-dimensional Gaussian distribution to the data of each class using the ML estimator for the mean and covariance matrix, which correspond to the sample mean and covariance matrix of the training data (up to a slight rescaling of the sample covariance). In more detail, $\vec{a}_1,\ldots,\vec{a}_n$ are used to estimate a mean $\vec{\mu}_a$ and covariance matrix Σ_a, whereas $\vec{b}_1,\ldots,\vec{b}_n$ are used to estimate $\vec{\mu}_b$ and Σ_b,

$$\{\vec{\mu}_a,\Sigma_a\} := \arg\max_{\vec{\mu},\Sigma} \mathcal{L}_{\vec{a}_1,\ldots,\vec{a}_n}(\vec{\mu},\Sigma), \tag{9.81}$$

$$\{\vec{\mu}_b,\Sigma_b\} := \arg\max_{\vec{\mu},\Sigma} \mathcal{L}_{\vec{b}_1,\ldots,\vec{b}_n}(\vec{\mu},\Sigma). \tag{9.82}$$

Then for each new example $\vec{x}$, the value of the density function at the example for both classes is evaluated. If

$$f_{\vec{\mu}_a,\Sigma_a}(\vec{x}) > f_{\vec{\mu}_b,\Sigma_b}(\vec{x}) \tag{9.83}$$

then $\vec{x}$ is declared to belong to the first class, otherwise it is declared to belong to the second class. Figure 9.13 shows the results of applying the method to data simulated using two Gaussian distributions.

△

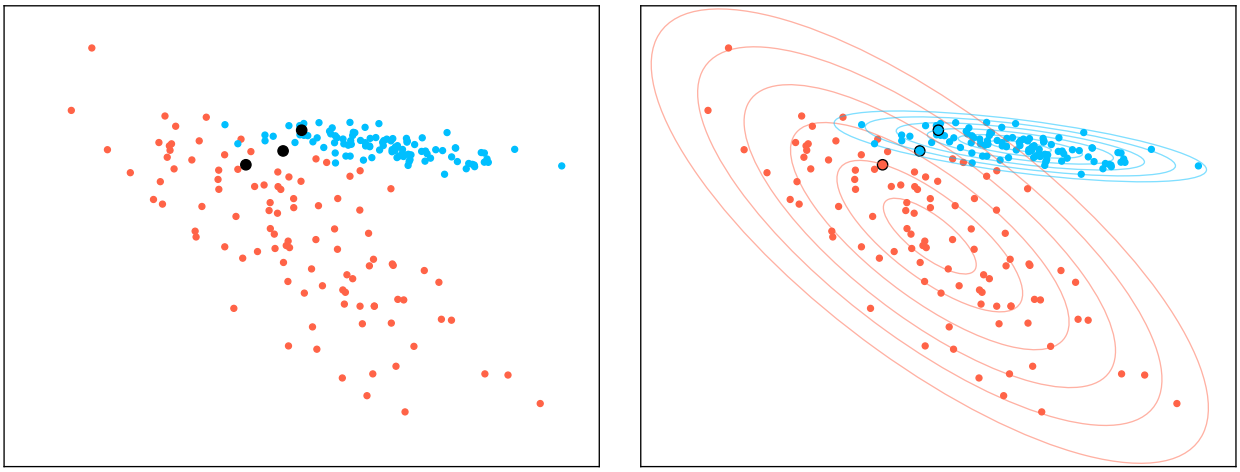


Figure 9.13: Quadratic-discriminant analysis applied to data from two different classes (left). The data corresponding to the two different classes are colored orange and blue. Three new examples are colored in black. Two bivariate Gaussians are fit to the data. Their contour lines are shown in the respective color of each class on the right. These distributions are used to classify the new examples, which are colored according to their estimated class.

9.7 Proofs

9.7.1 Proof of Lemma 9.2.5

We consider the sample variance of an iid sequence $\widetilde{X}$ with mean μ and variance σ^2,

$$\widetilde{Y}(n) := \frac{1}{n-1}\sum_{i=1}^{n}\left(\widetilde{X}(i) - \frac{1}{n}\sum_{j=1}^{1}\widetilde{X}(j)\right) \tag{9.84}$$

$$= \frac{1}{n-1}\left(\widetilde{X}(i) - \frac{1}{n}\sum_{j=1}^{n}\widetilde{X}(j)\right)^2 \tag{9.85}$$

$$= \frac{1}{n-1}\left(\widetilde{X}(i)^2 + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\widetilde{X}(j)\widetilde{X}(k) - \frac{2}{n}\sum_{j=1}^{n}\widetilde{X}(i)\widetilde{X}(j)\right) \tag{9.86}$$

To simplify notation, we denote the mean square $\mathrm{E}\left(\widetilde{X}(i)^2\right) = \mu^2 + \sigma^2$ by ξ. We have

$$\mathrm{E}\left(\widetilde{Y}(n)\right) = \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{E}\left(\widetilde{X}(i)^2\right) + \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{E}\left(\widetilde{X}(j)^2\right) + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\mathrm{E}\left(\widetilde{X}(j)\,\widetilde{X}(k)\right) \tag{9.87}$$

$$-\frac{2}{n}\mathrm{E}\left(\widetilde{X}(i)^2\right) - \frac{2}{n}\sum_{\substack{j=1\\j\neq i}}^{n}\mathrm{E}\left(\widetilde{X}(i)\,\widetilde{X}(j)\right) \tag{9.88}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\xi + \frac{n\,\xi}{n^2} + \frac{n(n-1)\,\mu^2}{n^2} - \frac{2\,\xi}{n} - \frac{2(n-1)\,\mu^2}{n} \tag{9.89}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\frac{n-1}{n}\left(\xi - \mu^2\right) \tag{9.90}$$

$$= \sigma^2. \tag{9.91}$$

9.7.2 Proof of Theorem 9.3.4

We denote the sample median by $\widetilde{Y}(n)$. Our aim is to show that for any $\epsilon > 0$

$$\lim_{n\to\infty}\mathrm{P}\left(\left|\widetilde{Y}(n) - \gamma\right| \geq \epsilon\right) = 0. \tag{9.92}$$

We will prove that

$$\lim_{n\to\infty}\mathrm{P}\left(\widetilde{Y}(n) \geq \gamma + \epsilon\right) = 0. \tag{9.93}$$

The same argument allows to establish

$$\lim_{n\to\infty}\mathrm{P}\left(\widetilde{Y}(n) \leq \gamma - \epsilon\right) = 0. \tag{9.94}$$

If we order the set $\left\{\widetilde{X}(1), \ldots, \widetilde{X}(n)\right\}$, then $\widetilde{Y}(n)$ equals the $(n+1)/2$th element if n is odd and the average of the $n/2$th and the $(n/2+1)$th element if n is even. The event $\widetilde{Y}(n) \geq \gamma + \epsilon$ therefore implies that at least $(n+1)/2$ of the elements are larger than $\gamma + \epsilon$.

For each individual $\widetilde{X}(i)$, the probability that $\widetilde{X}(i) > \gamma + \epsilon$ is

$$p := 1 - F_{\widetilde{X}(i)}(\gamma + \epsilon) = 1/2 - \epsilon' \tag{9.95}$$

where we assume that $\epsilon' > 0$. If this is not the case then the cdf of the iid sequence is *flat* at γ and the median is not well defined. The number of random variables in the set $\left\{\widetilde{X}(1), \ldots, \widetilde{X}(n)\right\}$ which are larger than $\gamma + \epsilon$ is distributed as a binomial random variable B_n with parameters n

and p. As a result, we have

$$\begin{aligned}
\mathrm{P}\left(\widetilde{Y}(n) \geq \gamma + \epsilon\right) &\leq \mathrm{P}\left(\frac{n+1}{2} \text{ or more samples are greater or equal to } \gamma + \epsilon\right) && (9.96)\\
&= \mathrm{P}\left(B_n \geq \frac{n+1}{2}\right) && (9.97)\\
&= \mathrm{P}\left(B_n - np \geq \frac{n+1}{2} - np\right) && (9.98)\\
&\leq \mathrm{P}\left(|B_n - np| \geq n\epsilon' + \frac{1}{2}\right) && (9.99)\\
&\leq \frac{\mathrm{Var}(B_n)}{\left(n\epsilon' + \frac{1}{2}\right)^2} \quad \text{by Chebyshev's inequality} && (9.100)\\
&= \frac{np(1-p)}{n^2\left(\epsilon' + \frac{1}{2n}\right)^2} && (9.101)\\
&= \frac{p(1-p)}{n\left(\epsilon' + \frac{1}{2n}\right)^2}, && (9.102)
\end{aligned}$$

which converges to zero as $n \to \infty$. This establishes (9.93).

Chapter 10

Bayesian Statistics

In the frequentist paradigm we model the data as realizations from a distribution that is fixed. In particular, if the model is parametric, the parameters are *deterministic* quantities. In contrast, in Bayesian parametric modeling the parameters are modeled as **random variables**. The goal is to have the flexibility to quantify our uncertainty about the underlying distribution beforehand, for example in order to integrate available prior information about the data.

10.1 Bayesian parametric models

In this section we describe how to fit a parametric model to a data set within a Bayesian framework. As in Section 9.6, we assume that the data are generated by sampling from known distributions with unknown parameters. The crucial difference is that we model the parameters as being random instead of deterministic. This requires selecting their prior distribution before fitting the data, which allows to quantify our uncertainty about the value of the parameters beforehand. A Bayesian parametric model is specified by:

1. The **prior** distribution is the distribution of $\vec{\Theta}$, which encodes our uncertainty about the model before seeing the data.
2. The **likelihood** is the conditional distribution of $\vec{X}$ given $\vec{\Theta}$, which specifies how the data depend on the parameters. In contrast to the frequentist framework, the likelihood is *not* interpreted as a deterministic function of the parameters.

Our goal when learning a Bayesian model is to compute the **posterior distribution** of the parameters Θ given $\vec{X}$. Evaluating this posterior distribution at the realization $\vec{x}$ allows to update our uncertainty about Θ using the data.

The following example fits a Bayesian model to iid samples from a Bernoulli random variable.

Example 10.1.1 (Bernoulli distribution)**.** Let $\vec{x}$ be a vector of data that we wish to model as iid samples from a Bernoulli distribution. Since we are taking a Bayesian approach we choose a prior distribution for the parameter of the Bernoulli. We will consider two different Bayesian estimators Θ_1 and Θ_2:

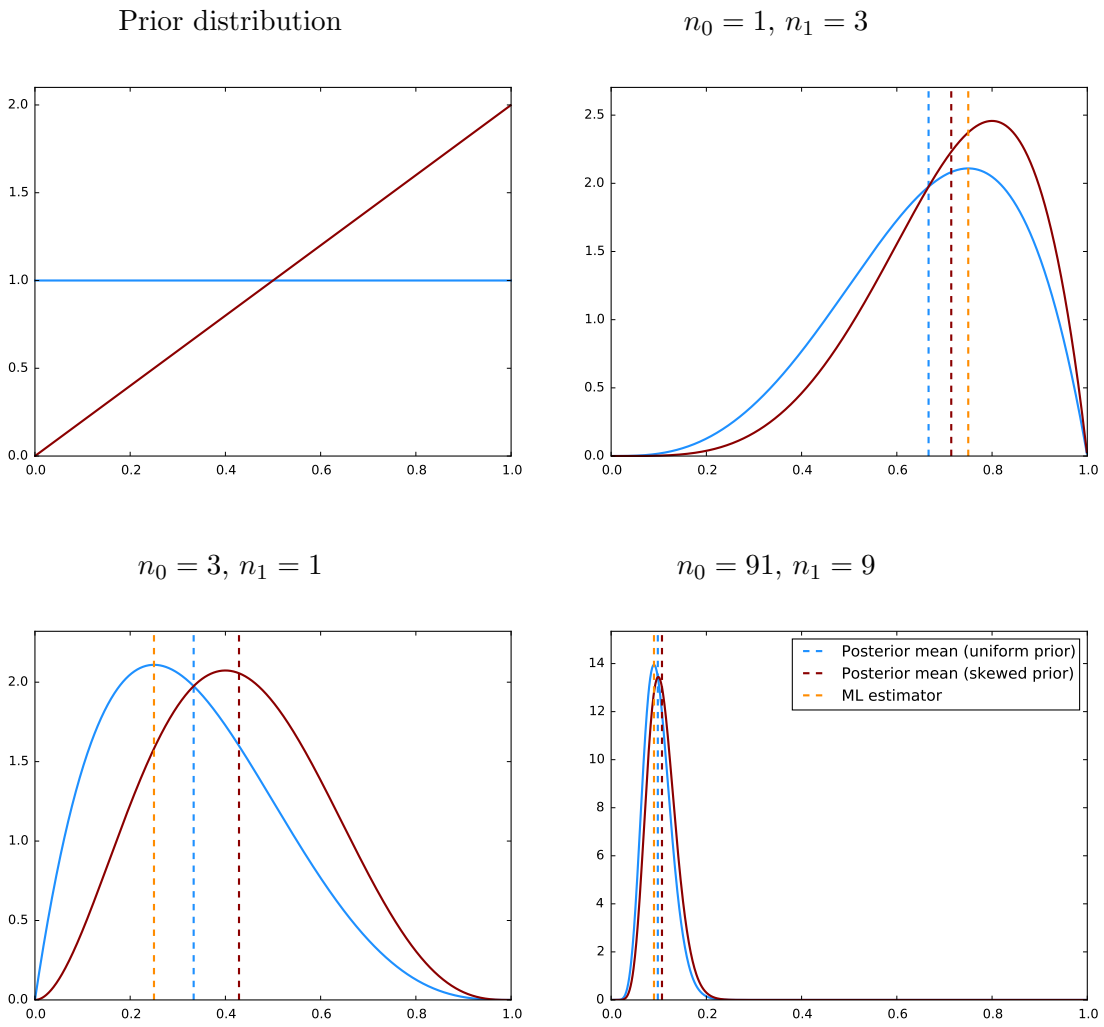


Figure 10.1: The prior distribution of Θ_1 (blue) and Θ_2 (dark red) in Example 10.1.1 are shown in the top-left graph. The rest of the graphs show the corresponding posterior distributions for different data sets.

1. Θ_1 represents a conservative estimator in terms of prior information. We assign a uniform pdf to the parameter. Any value in the unit interval has the same probability density:

$$f_{\Theta_1}(\theta) = \begin{cases} 1 & \text{for } 0 \leq \theta \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{10.1}$$

2. Θ_2 is an estimator that assumes that the parameter is closer to 1 than to 0. We could use it for instance to capture the suspicion that a coin is biased towards heads. We choose a skewed pdf that increases linearly from zero to one,

$$f_{\Theta_2}(\theta) = \begin{cases} 2\,\theta & \text{for } 0 \leq \theta \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{10.2}$$

By the iid assumption, the likelihood, which is just the conditional pmf of the data given the parameter of the Bernoulli, equals

$$p_{\vec{X}|\Theta}\left(\vec{x}|\theta\right) = \theta^{n_1}\left(1-\theta\right)^{n_0}, \tag{10.3}$$

where n_1 is the number of ones in the data and n_0 the number of zeros (see Example 9.6.3). The posterior pdfs of the two estimators are consequently equal to

$$\begin{aligned}
f_{\Theta_1|\vec{X}}\left(\theta|\vec{x}\right) &= \frac{f_{\Theta_1}\left(\theta\right) p_{\vec{X}|\Theta_1}\left(\vec{x}|\theta\right)}{p_{\vec{X}}\left(\vec{x}\right)} && (10.4)\\
&= \frac{f_{\Theta_1}\left(\theta\right) p_{\vec{X}|\Theta_1}\left(\vec{x}|\theta\right)}{\int_u f_{\Theta_1}\left(u\right) p_{\vec{X}|\Theta_1}\left(\vec{x}|u\right)\,\mathrm{d}u} && (10.5)\\
&= \frac{\theta^{n_1}\left(1-\theta\right)^{n_0}}{\int_u u^{n_1}\left(1-u\right)^{n_0}\,\mathrm{d}u} && (10.6)\\
&= \frac{\theta^{n_1}\left(1-\theta\right)^{n_0}}{\beta\left(n_1+1, n_0+1\right)}, && (10.7)\\
f_{\Theta_2|\vec{X}}\left(\theta|\vec{x}\right) &= \frac{f_{\Theta_2}\left(\theta\right) p_{\vec{X}|\Theta_2}\left(\vec{x}|\theta\right)}{p_{\vec{X}}\left(\vec{x}\right)} && (10.8)\\
&= \frac{\theta^{n_1+1}\left(1-\theta\right)^{n_0}}{\int_u u^{n_1+1}\left(1-u\right)^{n_0}\,\mathrm{d}u} && (10.9)\\
&= \frac{\theta^{n_1+1}\left(1-\theta\right)^{n_0}}{\beta\left(n_1+2, n_0+1\right)}, && (10.10)\\
& && (10.11)
\end{aligned}$$

where

$$\beta\left(a,b\right) := \int_u u^{a-1}\left(1-u\right)^{b-1}\,\mathrm{d}u \tag{10.12}$$

is a special function called the beta function or Euler integral of the first kind, which is tabulated.

Figure 10.1 shows the plot of the posterior distribution for different values of n_1 and n_0. It also shows the maximum-likelihood estimator of the parameter, which is just $n_1/\left(n_0+n_1\right)$ (see Example 9.6.3). For a small number of flips, the posterior pdf of Θ_2 is skewed to the right with respect to that of Θ_1, reflecting the prior belief that the parameter is closer to 1. However for a large number of flips both posterior densities are very close.

△

10.2 Conjugate prior

Both posterior distributions in Example 10.1.1 are beta distributions (see Definition 2.3.12), and so are the priors. The uniform prior of Θ_1 is beta with parameters $a=1$ and $b=1$, whereas the skewed prior of Θ_2 is beta distribution with parameters $a=2$ and $b=1$. Since the prior and the posterior belong to the same family, computing the posterior is equivalent to just updating the parameters. When the prior and posterior are guaranteed to belong to the same family of distributions for a particular likelihood, the distributions are called conjugate priors.

Definition 10.2.1 (Conjugate priors)**.** *A conjugate family of distributions for a certain likelihood satisfies the following property: if the prior belongs to the family, then the posterior also belongs to the family.*

Beta distributions are conjugate priors when the likelihood is binomial.

Theorem 10.2.2 (The beta distribution is conjugate to the binomial likelihood)**.** *If the prior distribution of Θ is a beta distributions with parameters a and b and the likelihood of the data X given Θ is binomial with parameters n and x, then the posterior distribution of Θ given X is a beta distribution with parameters $x+a$ and $n-x+b$.*

Proof.

$$f_{\Theta \,|\, X}\left(\theta \,|\, x\right) = \frac{f_\Theta\left(\theta\right) p_{X \,|\, \Theta}\left(x \,|\, \theta\right)}{p_X\left(x\right)} \tag{10.13}$$

$$= \frac{f_\Theta\left(\theta\right) p_{X \,|\, \Theta}\left(x \,|\, \theta\right)}{\int_u f_\Theta\left(u\right) p_{X \,|\, \Theta}\left(x \,|\, u\right) \,\mathrm{d}u} \tag{10.14}$$

$$= \frac{\theta^{a-1}\left(1-\theta\right)^{b-1}\binom{n}{x}\theta^x\left(1-\theta\right)^{n-x}}{\int_u u^{a-1}\left(1-u\right)^{b-1}\binom{n}{x}u^x\left(1-u\right)^{n-x}\,\mathrm{d}u} \tag{10.15}$$

$$= \frac{\theta^{x+a-1}\left(1-\theta\right)^{n-x+b-1}}{\int_u u^{x+a-1}\left(1-u\right)^{n-x+b-1}\,\mathrm{d}u} \tag{10.16}$$

$$= f_\beta\left(\theta; x+a, n-x+b\right). \tag{10.17}$$

□

Note that the posteriors obtained in Example 10.1.1 follow immediately from the theorem.

Example 10.2.3 (Poll in New Mexico)**.** In a poll in New Mexico for the 2016 US election, 429 participants, 227 people intend to vote for Clinton and 202 for Trump (the data are from a real poll[1], but for simplicity we are ignoring the other candidates and people that were undecided). Our aim is to use a Bayesian framework to predict the outcome of the election in New Mexico using these data.

We model the fraction of people that vote for Trump as a random variable Θ. We assume that the n people in the poll are chosen uniformly at random with replacement from the population, so given $\Theta = \theta$ the number of Trump voters is a binomial with parameters n and θ. We don't have any additional information about the possible value of Θ, so we assume it is uniform or equivalently a beta distribution with parameters $a := 1$ and $b := 1$.

By Theorem 10.2.2 the posterior distribution of Θ given the data that we observe is a beta distribution with parameters $a := 203$ and $b := 228$, depicted in Figure 10.2. The corresponding probability that $\Theta \geq 0.5$ is 11.4%, which is our estimate for the probability that Trump wins in New Mexico.

△

[1]The poll results are taken from
https://www.abqjournal.com/883092/clinton-still-ahead-in-new-mexico.html

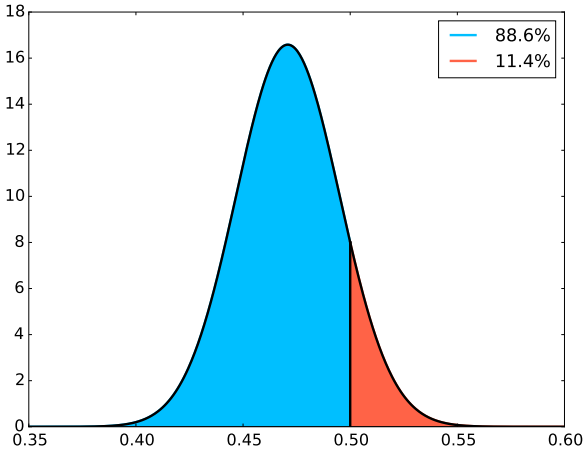


Figure 10.2: Posterior distribution of the fraction of Trump voters in New Mexico conditioned on the poll data in Example 10.2.3.

10.3 Bayesian estimators

The Bayesian approach to learning probabilistic models yields the whole posterior distribution of the parameters of interest. In this section we describe two alternatives for deriving a single estimate of the parameters from the posterior distribution.

10.3.1 Minimum mean-square-error estimation

The mean of the posterior distribution is the conditional expectation of the parameters given the data. Choosing the posterior mean as an estimator for the parameters $\vec{\Theta}$ has a strong theoretical justification: it is guaranteed to achieve the minimum mean square error (MSE) among *all possible estimators*. Of course, this only holds if all of the assumptions hold, i.e. the parameters are generated according to the prior and the data are then generated according to the likelihood, which may not be the case for real data.

Theorem 10.3.1 (The posterior mean minimizes the MSE)**.** *The posterior mean is the minimum mean-square-error (MMSE) estimate of the parameter $\vec{\Theta}$ given the data $\vec{X}$. To be more precise, let us define*

$$\theta_{\text{MMSE}}\left(\vec{x}\right) := \text{E}\big(\vec{\Theta}|\vec{X} = \vec{x}\big). \tag{10.18}$$

For any arbitrary estimator $\theta_{\text{other}}\left(\vec{x}\right)$,

$$\text{E}\left(\left(\theta_{\text{other}}(\vec{X}) - \vec{\Theta}\right)^2\right) \geq \text{E}\left(\left(\theta_{\text{MMSE}}(\vec{X}) - \vec{\Theta}\right)^2\right). \tag{10.19}$$

Proof. We begin by computing the MSE of the arbitrary estimator conditioned on $\vec{X} = \vec{x}$ in

terms of the conditional expectation of Θ given $\vec{X}$,

$$\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\vec{\Theta}\right)^2\left|\vec{X}=\vec{x}\right.\right) \tag{10.20}$$

$$=\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\theta_{\text{MMSE}}(\vec{X})+\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\left|\vec{X}=\vec{x}\right.\right) \tag{10.21}$$

$$=\left(\theta_{\text{other}}(\vec{x})-\theta_{\text{MMSE}}(\vec{x})\right)^2+\mathrm{E}\Big(\left(\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\left|\vec{X}=\vec{x}\right.\Big) \tag{10.22}$$

$$+2\left(\theta_{\text{other}}(\vec{x})-\theta_{\text{MMSE}}(\vec{x})\right)\mathrm{E}\left(\theta_{\text{MMSE}}(\vec{x})-\mathrm{E}\left(\vec{\Theta}\middle|\vec{X}=\vec{x}\right)\right)$$

$$=\left(\theta_{\text{other}}(\vec{x})-\theta_{\text{MMSE}}(\vec{x})\right)^2+\mathrm{E}\Big(\left(\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\left|\vec{X}=\vec{x}\right.\Big). \tag{10.23}$$

By iterated expectation,

$$\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\Theta\right)^2\right)=\mathrm{E}\left(\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\vec{\Theta}\right)^2\left|\vec{X}\right.\right)\right) \tag{10.24}$$

$$=\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\theta_{\text{MMSE}}(\vec{X})\right)^2\right)+\mathrm{E}\left(\mathrm{E}\Big(\left(\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\left|\vec{X}\right.\Big)\right)$$

$$=\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\theta_{\text{MMSE}}(\vec{X})\right)^2\right)+\mathrm{E}\left(\left(\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\right) \tag{10.25}$$

$$\geq\mathrm{E}\left(\left(\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\right), \tag{10.26}$$

since the expectation of a nonnegative quantity is nonnegative. □

Example 10.3.2 (Bernoulli distribution (continued))**.** In order to obtain point estimates for the parameter in Example 10.1.1 we compute the posterior means:

$$\mathrm{E}\left(\Theta_1|\vec{X}=\vec{x}\right)=\int_0^1\theta f_{\Theta_1|\vec{X}}\left(\theta|\vec{x}\right)\,\mathrm{d}\theta \tag{10.27}$$

$$=\frac{\int_0^1\theta^{n_1+1}\left(1-\theta\right)^{n_0}\,\mathrm{d}\theta}{\beta\left(n_1+1,n_0+1\right)} \tag{10.28}$$

$$=\frac{\beta\left(n_1+2,n_0+1\right)}{\beta\left(n_1+1,n_0+1\right)}, \tag{10.29}$$

$$\mathrm{E}\left(\Theta_2|\vec{X}=\vec{x}\right)=\int_0^1\theta f_{\Theta_2|\vec{X}}\left(\theta|\vec{x}\right)\,\mathrm{d}\theta \tag{10.30}$$

$$=\frac{\beta\left(n_1+3,n_0+1\right)}{\beta\left(n_1+2,n_0+1\right)}. \tag{10.31}$$

Figure 10.1 shows the posterior means for different values of n_0 and n_1.

△

10.3.2 Maximum-a-posteriori estimation

An alternative to the posterior mean is the posterior mode, which is the maximum of the pdf or the pmf of the posterior distribution.

Definition 10.3.3 (Maximum-a-posteriori estimator)**.** *The maximum-a-posteriori (MAP) estimator of a parameter $\vec{\Theta}$ given data $\vec{x}$ modeled as a realization of a random vector $\vec{X}$ is*

$$\theta_{\text{MAP}}\left(\vec{x}\right) := \arg\max_{\vec{\theta}} p_{\vec{\Theta}\,|\,\vec{X}}\left(\vec{\theta}\,|\,\vec{x}\right) \tag{10.32}$$

if $\vec{\Theta}$ is modeled as a discrete random variable and

$$\theta_{\text{MAP}}\left(\vec{x}\right) := \arg\max_{\vec{\theta}} f_{\vec{\Theta}\,|\,\vec{X}}\left(\vec{\theta}\,|\,\vec{x}\right) \tag{10.33}$$

if it is modeled as a continuous random variable.

In Figure 10.1 the ML estimator of Θ is the mode (maximum value) of the posterior distribution when the prior is uniform. This is not a coincidence, under a uniform prior the MAP and ML estimates are the same.

Lemma 10.3.4. *The maximum-likelihood estimator of a parameter Θ is the mode (maximum value) of the pdf of the posterior distribution given the data $\vec{X}$ if its prior distribution is uniform.*

Proof. We prove the result when the model for the data and the parameters is continuous, if any or both of them are discrete the proof is identical (in that case the ML estimator is the mode of the pmf of the posterior). If the prior distribution of the parameters is uniform, then $f_{\vec{\Theta}}\left(\vec{\theta}\right)$ is constant for any $\vec{\theta}$, which implies

$$\arg\max_{\vec{\theta}} f_{\vec{\Theta}\,|\,\vec{X}}\left(\vec{\theta}|\vec{x}\right) = \arg\max_{\vec{\theta}} \frac{f_{\vec{\Theta}}\left(\vec{\theta}\right) f_{\vec{X}|\vec{\Theta}}\left(\vec{x}|\vec{\theta}\right)}{\int_u f_{\vec{\Theta}}\left(u\right) f_{\vec{X}|\vec{\Theta}}\left(\vec{x}|u\right) \mathrm{d}u} \tag{10.34}$$

$$= \arg\max_{\vec{\theta}} f_{\vec{X}|\vec{\Theta}}\left(\vec{x}|\vec{\theta}\right) \qquad \text{(the rest of the terms do not depend on } \vec{\theta}\text{)}$$

$$= \arg\max_{\vec{\theta}} \mathcal{L}_{\vec{x}}\left(\vec{\theta}\right). \tag{10.35}$$

□

Note that uniform priors are only well defined in situations where the parameter is restricted to a bounded set.

We now describe a situation in which the MAP estimator is optimal. If the parameter Θ can only take a discrete set of values, then the MAP estimator minimizes the probability of making the wrong choice.

Theorem 10.3.5 (MAP estimator minimizes the probability of error)**.** *Let $\vec{\Theta}$ be a discrete random vector and let $\vec{X}$ be a random vector modeling the data. We define*

$$\theta_{\text{MAP}}\left(\vec{x}\right) := \arg\max_{\vec{\theta}} p_{\vec{\Theta}\,|\,\vec{X}}\left(\vec{\theta}\,|\,\vec{X} = \vec{x}\right). \tag{10.36}$$

For any arbitrary estimator $\theta_{\text{other}}\left(\vec{x}\right)$,

$$\mathrm{P}\left(\theta_{\text{other}}(\vec{X}) \neq \vec{\Theta}\right) \geq \mathrm{P}\left(\theta_{\text{MAP}}(\vec{X}) \neq \vec{\Theta}\right). \tag{10.37}$$

In words, the MAP estimator minimizes the probability of error.

Proof. We assume that $\vec{X}$ is a continuous random vector, but the same argument applies if it is discrete. We have

$$\mathrm{P}\left(\Theta = \theta_{\text{other}}(\vec{X})\right) = \int_{\vec{x}} f_{\vec{X}}(\vec{x})\, \mathrm{P}\left(\Theta = \theta_{\text{other}}(\vec{x}) \,|\, \vec{X} = \vec{x}\right) \mathrm{d}\vec{x} \tag{10.38}$$

$$= \int_{\vec{x}} f_{\vec{X}}(\vec{x})\, p_{\vec{\Theta}\,|\,\vec{X}}(\theta_{\text{other}}(\vec{x}) \,|\, \vec{x})\, \mathrm{d}\vec{x} \tag{10.39}$$

$$\leq \int_{\vec{x}} f_{\vec{X}}(\vec{x})\, p_{\vec{\Theta}\,|\,\vec{X}}(\theta_{\text{MAP}}(\vec{x}) \,|\, \vec{x})\, \mathrm{d}\vec{x} \tag{10.40}$$

$$= \mathrm{P}\left(\Theta = \theta_{\text{MAP}}(\vec{X})\right), \tag{10.41}$$

where (10.40) follows from the definition of the MAP estimator as the mode of the posterior. □

Example 10.3.6 (Sending bits)**.** We consider a very simple model for a communication channel in which we aim to send a signal Θ consisting of a single bit. Our prior knowledge indicates that the signal is equal to one with probability 1/4.

$$p_\Theta(1) = \frac{1}{4}, \qquad p_\Theta(0) = \frac{3}{4}. \tag{10.42}$$

Due to the presence of noise in the channel, we send the signal n times. At the receptor we observe

$$\vec{X}_i = \Theta + \vec{Z}_i, \quad 1 \leq i \leq n, \tag{10.43}$$

where $\vec{Z}$ contains n iid standard Gaussian random variables. Modeling perturbations as Gaussian is a popular choice in communications. It is justified by the central limit theorem, under the assumption that the noise is a combination of many small effects that are approximately independent.

We will now compute and compare the ML and MAP estimators of Θ given the observations.

The likelihood is equal to

$$\mathcal{L}_{\vec{x}}(\theta) = \prod_{i=1}^{n} f_{\vec{X}_i|\Theta}(\vec{x}_i \,|\, \theta) \tag{10.44}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\vec{x}_i - \theta)^2}{2}}. \tag{10.45}$$

It is easier to deal with the log-likelihood function,

$$\log \mathcal{L}_{\vec{x}}(\theta) = -\sum_{i=1}^{n} \frac{(\vec{x}_i - \theta)^2}{2} - \frac{n}{2} \log 2\pi. \tag{10.46}$$

Since Θ only takes two values, we can compare directly. We will choose $\theta_{\text{ML}}(\vec{x}) = 1$ if

$$\log \mathcal{L}_{\vec{x}}(1) = -\sum_{i=1}^{n} \frac{\vec{x}_i^2 - 2\vec{x}_i + 1}{2} - \frac{n}{2} \log 2\pi \tag{10.47}$$

$$\geq -\sum_{i=1}^{n} \frac{\vec{x}_i^2}{2} - \frac{n}{2} \log 2\pi \tag{10.48}$$

$$= \log \mathcal{L}_{\vec{x}}(0). \tag{10.49}$$

Equivalently,

$$\theta_{\mathrm{ML}}\left(\vec{x}\right) = \begin{cases} 1 & \text{if } \frac{1}{n}\sum_{i=1}^{n} \vec{x}_i > \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{10.50}$$

The rule makes a lot of sense: if the sample mean of the data is closer to 1 than to 0 then our estimate is equal to 1. By the law of total probability, the probability of error of this estimator is equal to

$$\begin{aligned} \mathrm{P}\left(\Theta \neq \theta_{\mathrm{ML}}(\vec{X})\right) &= \mathrm{P}\left(\Theta \neq \theta_{\mathrm{ML}}(\vec{X}) \middle| \Theta = 0\right) \mathrm{P}\left(\Theta = 0\right) + \mathrm{P}\left(\Theta \neq \theta_{\mathrm{ML}}(\vec{X}) \middle| \Theta = 1\right) \mathrm{P}\left(\Theta = 1\right) \\ &= \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n} \vec{X}_i > \frac{1}{2} \middle| \Theta = 0\right) \mathrm{P}\left(\Theta = 0\right) + \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n} \vec{X}_i < \frac{1}{2} \middle| \Theta = 1\right) \mathrm{P}\left(\Theta = 1\right) \\ &= Q\left(\sqrt{n}/2\right), \end{aligned} \tag{10.51}$$

where the last equality follows from the fact that if we condition on $\Theta = \theta$ the empirical mean is Gaussian with variance σ^2/n and mean θ (see the proof of Theorem 6.2.2).

To compute the MAP estimate we must find the maximum of the posterior pdf of Θ given the observed data. Equivalently, we find the maximum of its logarithm (this is equivalent because the logarithm is a monotone function),

$$\log p_{\Theta|\vec{X}}\left(\theta|\vec{x}\right) = \log \frac{\prod_{i=1}^{n} f_{\vec{X}_i|\Theta}\left(\vec{x}_i|\theta\right) p_{\Theta}\left(\theta\right)}{f_{\vec{X}}\left(\vec{x}\right)} \tag{10.52}$$

$$= \sum_{i=1}^{n} \log f_{\vec{X}_i|\Theta}\left(\vec{x}_i|\theta\right) p_{\Theta}\left(\theta\right) - \log f_{\vec{X}}\left(\vec{x}\right) \tag{10.53}$$

$$= -\sum_{i=1}^{n} \frac{\vec{x}_i^2 - 2\vec{x}_i\theta + \theta^2}{2} - \frac{n}{2}\log 2\pi + \log p_{\Theta}\left(\theta\right) - \log f_{\vec{X}}\left(\vec{x}\right). \tag{10.54}$$

We compare the value of this function for the two possible values of Θ: 0 and 1. We choose $\theta_{\mathrm{MAP}}\left(\vec{x}\right) = 1$ if

$$\log p_{\Theta|\vec{X}}\left(1|\vec{x}\right) + \log f_{\vec{X}}\left(\vec{x}\right) = -\sum_{i=1}^{n} \frac{\vec{x}_i^2 - 2\vec{x}_i + 1}{2} - \frac{n}{2}\log 2\pi - \log 4 \tag{10.55}$$

$$\geq -\sum_{i=1}^{n} \frac{\vec{x}_i^2}{2} - \frac{n}{2}\log 2\pi - \log 4 + \log 3 \tag{10.56}$$

$$= \log p_{\Theta|\vec{X}}\left(0|\vec{x}\right) + \log f_{\vec{X}}\left(\vec{x}\right). \tag{10.57}$$

Equivalently,

$$\theta_{\mathrm{MAP}}\left(\vec{x}\right) = \begin{cases} 1 & \text{if } \frac{1}{n}\sum_{i=1}^{n} \vec{x}_i > \frac{1}{2} + \frac{\log 3}{n}, \\ 0 & \text{otherwise.} \end{cases} \tag{10.58}$$

The MAP estimate shifts the threshold with respect to the ML estimate to take into account that Θ is more prone to equal zero. However, the correction term tends to zero as we gather more evidence, so if a lot of data is available the two estimators will be very similar.

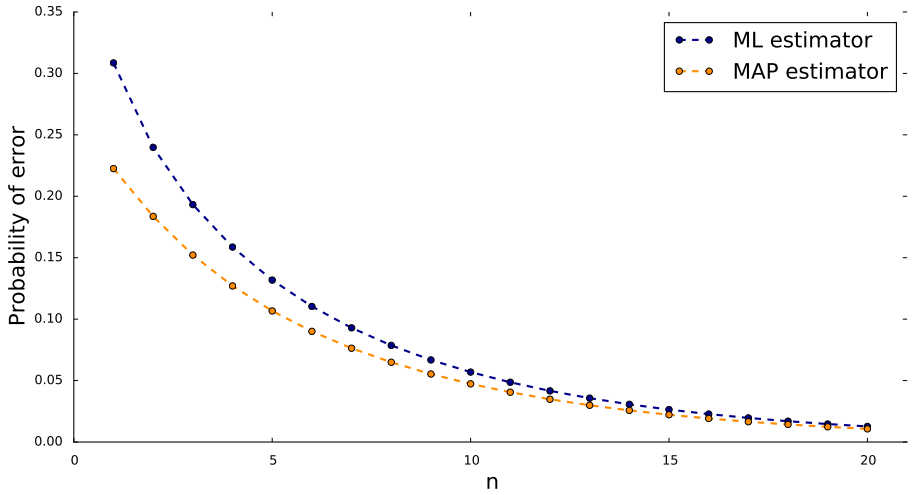


Figure 10.3: Probability of error of the ML and MAP estimators in Example 10.3.6 for different values of n.

The probability of error of the MAP estimator is equal to

$$\begin{aligned}
\mathrm{P}\left(\Theta \neq \theta_{\mathrm{MAP}}(\vec{X})\right) &= \mathrm{P}\left(\Theta \neq \theta_{\mathrm{MAP}}(\vec{X}) | \Theta = 0\right) \mathrm{P}\left(\Theta = 0\right) + \mathrm{P}\left(\Theta \neq \theta_{\mathrm{MAP}}(\vec{X}) | \Theta = 1\right) \mathrm{P}\left(\Theta = 1\right) \\
&= \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n} \vec{X}_i > \frac{1}{2} + \frac{\log 3}{n} \middle| \Theta = 0\right) \mathrm{P}\left(\Theta = 0\right) \qquad (10.59) \\
&\quad + \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n} \vec{X}_i < \frac{1}{2} + \frac{\log 3}{n} \middle| \Theta = 1\right) \mathrm{P}\left(\Theta = 1\right) \\
&= \frac{3}{4} Q\left(\sqrt{n}/2 + \frac{\log 3}{\sqrt{n}}\right) + \frac{1}{4} Q\left(\sqrt{n}/2 - \frac{\log 3}{\sqrt{n}}\right). \qquad (10.60)
\end{aligned}$$

We compare the probability of error of the ML and MAP estimators in Figure 10.3. MAP estimation results in better performance, but the difference becomes small as n increases.

△

Chapter 11

Hypothesis testing

In a medical study we observe that 10% of the women and 12.5% of the men suffer from heart disease. If there are 20 people in the study, we would probably be hesitant to declare that women are less prone to suffer from heart disease than men; it is very possible that the results occurred by chance. However, if there are 20,000 people in the study, then it seems more likely that we are observing a real phenomenon. Hypothesis testing makes this intuition precise; it is a framework that allows us to decide whether patterns that we observe in our data are likely to be the result of random fluctuations or not.

11.1 The hypothesis-testing framework

The aim of hypothesis testing is to evaluate a predefined conjecture. In the example above, this could be that heart disease is more prevalent in men than in women. The hypothesis that our conjecture is false is called the **null hypothesis**, denoted by H_0. In our example, the null hypothesis would be that heart disease is at least as prevalent in men as in women. If the null hypothesis holds, then whatever pattern we are detecting in our data that seems to support our conjecture is just a fluke. There just happen to be a lot of men with heart disease (or women without) in the study. In contrast, the hypothesis under which our conjecture is true is known as the **alternative hypothesis**, denoted by H_1. In this chapter we take a frequentist perspective: *the hypotheses either hold or not*, they are *not* modeled probabilistically.

A **test** is a procedure to determine whether we should *reject* the null hypothesis or not based on the data. Rejecting the null hypothesis means that we consider unlikely that it happened, which is evidence in favor of the alternative hypothesis. If we fail to reject the null hypothesis, this does *not* mean that we consider it likely, we just don't have enough information to discard it. Most tests produce a decision by thresholding a **test statistic**, which is a function that maps the data (i.e. a vector in $\mathbb{R}^n$) to a single number. The test rejects the null hypothesis if the test statistic belongs to a **rejection region** $\mathcal{R}$. For example, we could have

$$\mathcal{R} := \{t \mid t \geq \eta\}, \tag{11.1}$$

where t is the test statistic computed from the data and η is a predefined threshold. In this case, we would reject the null hypothesis only if t is larger than η.

As shown in Table 11.1, there are two possible errors that we can make. A **Type I error** is a *false positive*: our conjecture is false, but we reject the null hypothesis. A **Type II error** is

	Reject H_0?	
	No	Yes
H_0 is true	☺	Type I error
H_1 is true	Type II error	☺

Table 11.1: Type I and II errors.

a *false negative*: our conjecture holds, but we do not reject the null hypothesis. In hypothesis testing, our priority is to control Type I errors. When you read in a study that a result is **statistically significant** at a level of 0.05, this means that the probability of committing a Type I error is bounded by 5%.

Definition 11.1.1 (Significance level and size)**.** *The size of a test is the probability of making a Type I error. The significance level of a test is an upper bound on the size.*

Rejecting the null hypothesis does not give a quantitative sense of the extent to which the data are incompatible with the null hypothesis. The **p value** is a function of the data that plays this role.

Definition 11.1.2 (p value)**.** *The p value is the smallest significance level at which we would reject the null hypothesis for the data we observe.*

For a fixed significance level, it is desirable to select a test that minimizes the probability of making a Type II error. Equivalently, we would like to maximize the probability of rejecting the null hypothesis when it does not hold. This probability is known as the **power** of the test.

Definition 11.1.3 (Power)**.** *The power of a test is the probability of rejecting the null hypothesis if it does not hold.*

Note that in order to characterize the power of a test we need to know the distribution of the data under the alternative hypothesis, which is often unrealistic (recall that the alternative hypothesis is just the complement of the null hypothesis and consequently encompasses many different possibilities).

The standard procedure to apply hypothesis testing in the applied sciences is the following:

1. Choose a conjecture.
2. Determine the corresponding null hypothesis.
3. Choose a test.
4. Gather the data.
5. Compute the test statistic from the data.

6. Compute the p value and reject the null hypothesis if it is below a predefined limit (typically 1% or 5%).

Example 11.1.4 (Clutch)**.** We want to test the conjecture that a certain player in the NBA is *clutch*, i.e. that he scores more points at the end of close games than during the rest of the game. The null hypothesis is that there is no difference in his performance. The test statistic t that we choose is whether he makes more or less points per minute in the last quarter than in the rest of the game

$$t\left(\vec{x}\right) = \sum_{i=1}^{n} 1_{\vec{x}_i > 0}, \tag{11.2}$$

where $\vec{x}_i$ is the difference between the points per minute he scores in the 4th quarter and in the rest of the quarters of game i for $1 \leq i \leq n$.

The rejection region of the test is of the form

$$\mathcal{R} := \left\{ t\left(\vec{x}\right) \mid t\left(\vec{x}\right) \geq \eta \right\}, \tag{11.3}$$

for a fixed threshold η. Under the null hypothesis the probability of scoring more points per minute in the 4th quarter is 1/2 (for simplicity we ignore the possibility that he scores the same number of points), so we can model the test statistic under the null hypothesis as a binomial random variable with parameters n and 1/2. If η is an integer between 0 and n, then the probability that the test statistic is in the rejection region if the null hypothesis holds is

$$P\left(T_0 \geq \eta\right) = \frac{1}{2^n} \sum_{k=\eta}^{n} \binom{n}{k}. \tag{11.4}$$

So the size of the test is $\frac{1}{2^n} \sum_{k=\eta}^{n} \binom{n}{k}$. Table 11.2 shows this value for all possible values of η. If we want a significance level of 1% or 5% then we need to set the threshold at $\eta = 16$ or $\eta = 15$ respectively.

We gather the data from 20 games $\vec{x}$ and compute the value of the test statistic $t\left(\vec{x}\right)$ (note that we use a lowercase letter because it is a specific realization), which turns out to be 14 (he scores more points per minute in the fourth quarter in 14 of the games). This is not enough to reject the null hypothesis for our predefined level of 1% or 5%. Therefore the result is not statistically significant.

In any case, we compute the p value, which is the smallest level at which the result would have been significant. From the table it is equal to 0.058. Note that under a frequentist framework we *cannot* interpret this as the probability that the null hypothesis holds (i.e. that the player is not better in the fourth quarter) because the hypothesis is not random, it either holds or it doesn't. Our result is almost significant and although we do not have enough evidence to support our conjecture, it does seem plausible that the player performs better in the fourth quarter.

△

11.2 Parametric testing

In this section we discuss hypothesis testing under the assumption that our data are sampled from a known distribution with *unknown* parameters. We again take a frequentist perspective,

η	1	2	3	4	5	6	7	8	9	10
$P(T_0 \geq \eta)$	1.000	1.000	1.000	0.999	0.994	0.979	0.942	0.868	0.748	0.588
η	11	12	13	14	15	16	17	18	19	20
$P(T_0 \geq \eta)$	0.412	0.252	0.132	0.058	0.021	0.006	0.001	0.000	0.000	0.000

Table 11.2: Probability of committing a Type I error depending on the value of the threshold in Example 11.1.4. The values are rounded to three decimal points.

as is usually done in most studies in the applied sciences. The parameter is consequently deterministic and so are the hypotheses: the null hypothesis is true or not, there is no such thing as *the probability that the null hypothesis holds*.

To simplify the exposition, we assume that the probability distribution depends only on one parameter that we denote by θ. P_θ is the probability measure of our probability space if θ is the value of the parameter. $\vec{X}$ is a random vector distributed according to P_θ. The actual data that we observe, which we denote by $\vec{x}$ is assumed to be a realization from this random vector.

Assume that the null hypothesis is $\theta = \theta_0$. In that case, the size of a test with test statistic T and rejection region $\mathcal{R}$ is equal to

$$\alpha = \mathrm{P}_{\theta_0}\left(T(\vec{X}) \in \mathcal{R}\right). \tag{11.5}$$

For a rejection region of the form (11.1) we have

$$\alpha := \mathrm{P}_{\theta_0}\left(T(\vec{X}) \geq \eta\right). \tag{11.6}$$

If the realization of the test statistic is $T(x_1, \ldots, x_n)$ then the significance level at which we would reject H_0 would be

$$p = \mathrm{P}_{\theta_0}\left(T(\vec{X}) \geq T(\vec{x})\right), \tag{11.7}$$

which is the p value if we observe $\vec{x}$. The p value can consequently be interpreted as the probability of observing a result that is *more extreme* than what we observe in the data *if the null hypothesis holds*.

A hypothesis of the form $\theta = \theta_0$ is known as a **simple** hypothesis. If a hypothesis is of the form $\theta \in \mathcal{S}$ for a certain set $\mathcal{S}$ then the hypothesis is **composite**. For a composite null hypothesis $\theta \in \mathcal{H}_0$ we redefine the size and the p value in the following way,

$$\alpha = \sup_{\theta \in \mathcal{H}_0} \mathrm{P}_\theta\left(T(\vec{X}) \geq \eta\right), \tag{11.8}$$

$$p = \sup_{\theta \in \mathcal{H}_0} \mathrm{P}_\theta\left(T(\vec{X}) \geq T(\vec{x})\right). \tag{11.9}$$

In order to characterize the power of the test for a certain significance level, we compute the power function.

Definition 11.2.1 (Power function)**.** *Let P_θ be the probability measure parametrized by θ and let $\mathcal{R}$ the rejection region for a test based on the test statistic $T(\vec{x})$. The power function of the test is defined as*

$$\beta(\theta) := P_\theta\left(T(\vec{X}) \in \mathcal{R}\right) \tag{11.10}$$

Ideally we would like $\beta(\theta) \approx 0$ for $\theta \in \mathcal{H}_0$ and $\beta(\theta) \approx 1$ for $\theta \in \mathcal{H}_1$.

Example 11.2.2 (Coin flip)**.** We are interested in checking whether a coin is biased towards heads. The null hypothesis is that for each coin flip the probability of obtaining heads is $\theta \leq 1/2$. Consequently, the alternative hypothesis is $\theta > 1/2$. Let us consider a test statistic equal to the number of heads observed in a sequence of n iid flips,

$$T(\vec{x}) = \sum_{i=1}^{n} 1_{\vec{x}_i=1}, \tag{11.11}$$

where $\vec{x}_i$ is one if the ith coin flip is heads and zero otherwise. A natural rejection region is

$$T(\vec{x}) \geq \eta. \tag{11.12}$$

In particular, we consider two possible thresholds

1. $\eta = n$, i.e. we only reject the null hypothesis if *all* the coin flips are heads,
2. $\eta = 3n/5$, i.e. we reject the null hypothesis if at least three fifths of the coin flips are heads.

What test should we use if the number of coin flips is 5, 50 or 100? Do the tests have a 5% significance level? What is the power of the tests for these values of n?

To answer these questions, we compute the power function of the test for both options. If $\eta = n$,

$$\beta_1(\theta) = P_\theta\left(T(\vec{X}) \in \mathcal{R}\right) \tag{11.13}$$

$$= \theta^n. \tag{11.14}$$

If $\eta = 3n/5$,

$$\beta_2(\theta) = \sum_{k=3n/5}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k}. \tag{11.15}$$

Figure 11.1 shows the two power functions. If $\eta = n$, then the test has a significance level of 5% for the three values of n. However the power is very low, especially for large n. This makes sense: even if the coin is pretty biased the probability of n heads is extremely low. If $\eta = 3n/5$, then for $n = 5$ the test has a significance level way above 5%, since even if the coin is not biased the probability of observing 3 heads out of 5 flips is quite high. However for large n the test has much higher power than the first option. If the bias of the coin is above 0.7 we reject the null hypothesis with high probability.

△

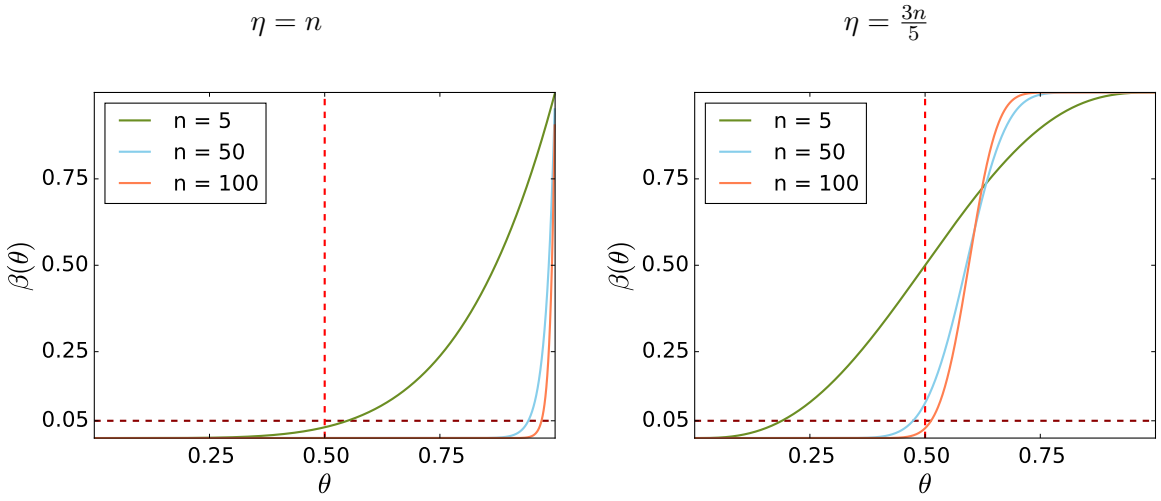


Figure 11.1: Power functions for the tests described in Example 11.2.2.

A systematic method for building tests under parametric assumptions is to threshold the ratio between the likelihood of the data under the null hypothesis and the likelihood of the data under the alternative hypothesis. If this ratio is high, the data are compatible with the null hypothesis, so it should not be rejected.

Definition 11.2.3 (Likelihood-ratio test)**.** *Let $\mathcal{L}_{\vec{x}}(\theta)$ denote the likelihood function corresponding to a data vector $\vec{x}$. $\mathcal{H}_0$ and $\mathcal{H}_1$ are the sets corresponding to the null and alternative hypotheses respectively. The likelihood ratio is*

$$\Lambda(\vec{x}) := \frac{\sup_{\theta \in \mathcal{H}_0} \mathcal{L}_{\vec{x}}(\theta)}{\sup_{\theta \in \mathcal{H}_1} \mathcal{L}_{\vec{x}}(\theta)}. \tag{11.16}$$

A likelihood-ratio test has a rejection region of the form $\{\Lambda(\vec{x}) \leq \eta\}$, for a constant threshold η.

Example 11.2.4 (Gaussian with known variance)**.** Imagine that you have some data that are well modeled as iid Gaussian with a known variance σ. The mean is unknown and we are interested in establishing that it is *not* equal to a certain value μ_0. What is the corresponding likelihood-ratio test and how should be set the threshold so that we have a significance level α?

First, from Example 9.6.4 the sample mean achieves the maximum of the likelihood function of a Gaussian

$$\text{av}(\vec{x}) := \arg\max_{\mu} \mathcal{L}_{\vec{x}}(\mu, \sigma) \tag{11.17}$$

for any value of σ. Using this result, we have

$$\Lambda(\vec{x}) = \frac{\sup_{\mu \in \mathcal{H}_0} \mathcal{L}_{\vec{x}}(\mu)}{\sup_{\mu \in \mathcal{H}_1} \mathcal{L}_{\vec{x}}(\mu)} \tag{11.18}$$

$$= \frac{\mathcal{L}_{\vec{x}}(\mu_0)}{\mathcal{L}_{\vec{x}}(\text{av}(\vec{x}))}. \tag{11.19}$$

Plugging in the expressions for the likelihood we obtain,

$$\Lambda\left(\vec{x}\right) = \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\left(\vec{x}_i - \text{av}\left(\vec{x}\right)\right)^2 - \left(\vec{x}_i - \mu_0\right)^2\right)\right\} \tag{11.20}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}\left(-2\,\text{av}\left(\vec{x}\right)\sum_{i=1}^{n}\vec{x}_i + n\,\text{av}\left(\vec{x}\right)^2 - 2\mu_0\sum_{i=1}^{n}\vec{x}_i + n\mu_0^2\right)\right\} \tag{11.21}$$

$$= \exp\left\{-\frac{n\left(\text{av}\left(\vec{x}\right) - \mu_0\right)^2}{2\sigma^2}\right\}. \tag{11.22}$$

Taking logarithms, the test is of the form

$$\frac{\left|\text{av}\left(\vec{x}\right) - \mu_0\right|}{\sigma} \geq \sqrt{\frac{-2\log\eta}{n}}. \tag{11.23}$$

The sample mean of n independent Gaussian random variables with mean μ_0 and variance σ^2 is Gaussian with mean μ_0 and variance σ^2/n, which implies

$$\alpha = \text{P}_{\mu_0}\left(\left|\frac{\text{av}\left(\vec{X}\right) - \mu_0}{\sigma/\sqrt{n}}\right| \geq \sqrt{-2\log\eta}\right) \tag{11.24}$$

$$= 2\,Q\left(\sqrt{-2\log\eta}\right). \tag{11.25}$$

If we fix a desired size α then the test becomes

$$\frac{\left|\text{av}\left(\vec{x}\right) - \mu_0\right|}{\sigma} \geq \frac{Q^{-1}\left(\alpha/2\right)}{\sqrt{n}}. \tag{11.26}$$

△

A motivating argument to employ the likelihood-ratio test is that if the null and alternative hypotheses are simple, then it is optimal in terms of power.

Lemma 11.2.5 (Neyman-Pearson Lemma)**.** *If both the null hypothesis and the alternative hypothesis are simple, i.e. the parameter θ can only have two values θ_0 and θ_1, then the likelihood-ratio test has the highest power among all tests with a fixed size.*

Proof. Recall that the power is the probability of rejecting the null hypothesis if it does not hold. If we denote the rejection region of the likelihood-ratio test by $\mathcal{R}_{LR}$ then its power is

$$P_{\theta_1}\left(\vec{X} \in \mathcal{R}_{LR}\right). \tag{11.27}$$

Assume that we have another test with rejection region $\mathcal{R}$. Its power is equal to

$$P_{\theta_1}\left(\vec{X} \in \mathcal{R}\right). \tag{11.28}$$

To prove that the power of the likelihood-ratio test is larger we only need to establish that

$$P_{\theta_1}\left(\vec{X} \in \mathcal{R}^c \cap \mathcal{R}_{LR}\right) \geq P_{\theta_1}\left(\vec{X} \in \mathcal{R}_{LR}^c \cap \mathcal{R}\right). \tag{11.29}$$

Let us assume that the data are continuous random variables (the argument for discrete random variables is practically the same) and that the pdf when the null and alternative hypotheses hold are f_{θ_0} and f_{θ_1} respectively. By the definition of the rejection region of the likelihood-ratio test, if $\Lambda(\vec{x}) \in \mathcal{R}_{LR}$

$$f_{\theta_1}(\vec{x}) \geq \frac{f_{\theta_0}(\vec{x})}{\eta}, \tag{11.30}$$

whereas if $\Lambda(\vec{x}) \in \mathcal{R}_{LR}^c$

$$f_{\theta_1}(\vec{x}) \leq \frac{f_{\theta_0}(\vec{x})}{\eta}. \tag{11.31}$$

If both tests have size α then

$$P_{\theta_0}\left(\vec{X} \in \mathcal{R}\right) = \alpha = P_{\theta_0}\left(\vec{X} \in \mathcal{R}_{LR}\right). \tag{11.32}$$

and consequently

$$\begin{aligned} P_{\theta_0}\left(\vec{X} \in \mathcal{R}^c \cap \mathcal{R}_{LR}\right) &= P_{\theta_0}\left(\vec{X} \in \mathcal{R}_{LR}\right) - P_{\theta_0}\left(\vec{X} \in \mathcal{R} \cap \mathcal{R}_{LR}\right) && (11.33)\\ &= P_{\theta_0}\left(\vec{X} \in \mathcal{R}\right) - P_{\theta_0}\left(\vec{X} \in \mathcal{R} \cap \mathcal{R}_{LR}\right) && (11.34)\\ &= P_{\theta_0}\left(\vec{X} \in \mathcal{R} \cap \mathcal{R}_{LR}^c\right). && (11.35) \end{aligned}$$

Now let us prove that (11.29) holds,

$$\begin{aligned} P_{\theta_1}\left(\vec{X} \in \mathcal{R}^c \cap \mathcal{R}_{LR}\right) &= \int_{\vec{x} \in \mathcal{R}^c \cap \mathcal{R}_{LR}} f_{\theta_1}(\vec{x})\, \mathrm{d}\vec{x} && (11.36)\\ &\geq \frac{1}{\eta} \int_{\vec{x} \in \mathcal{R}^c \cap \mathcal{R}_{LR}} f_{\theta_0}(\vec{x})\, \mathrm{d}\vec{x} \qquad \text{by (11.30)} && (11.37)\\ &= \frac{1}{\eta} P_{\theta_0}\left(\vec{X} \in \mathcal{R}^c \cap \mathcal{R}_{LR}\right) && (11.38)\\ &= \frac{1}{\eta} P_{\theta_0}\left(\vec{X} \in \mathcal{R} \cap \mathcal{R}_{LR}^c\right) \qquad \text{by (11.35)} && (11.39)\\ &= \frac{1}{\eta} \int_{\vec{x} \in \mathcal{R} \cap \mathcal{R}_{LR}^c} f_{\theta_0}(\vec{x})\, \mathrm{d}\vec{x} && (11.40)\\ &\geq \int_{\vec{x} \in \mathcal{R} \cap \mathcal{R}_{LR}^c} f_{\theta_1}(\vec{x})\, \mathrm{d}\vec{x} \qquad \text{by (11.31)} && (11.41)\\ &= P_{\theta_1}\left(\vec{X} \in \mathcal{R} \cap \mathcal{R}_{LR}^c\right). && (11.42) \end{aligned}$$

□

11.3 Nonparametric testing: The permutation test

In practical situations we may not able to design a parametric model that is adequate for our data. Nonparametric tests are hypothesis tests that do not assume that the data follow

any distribution with a predefined form. In this section we describe the permutation test, a nonparametric test that can be used to compare two data sets $\vec{x}_A$ and $\vec{x}_B$ in order to evaluate conjectures of the form $\vec{x}_A$ *is sampled from a distribution that has a higher mean than* $\vec{x}_B$ or $\vec{x}_B$ *is sampled from a distribution that has a higher variance than* $\vec{x}_A$. The null hypothesis is that the two data sets are actually sampled from the same distribution.

The test statistic in a permutation test is the difference between the values of a test statistic of interest t evaluated on the two data sets

$$t_{\text{diff}}\left(\vec{x}\right) := t\left(\vec{x}_A\right) - t\left(\vec{x}_B\right), \tag{11.43}$$

where $\vec{x}$ are all the data merged together. Our goal is to test whether $t\left(\vec{x}_A\right)$ is larger than $t\left(\vec{x}_B\right)$ at a certain significance level. The corresponding rejection region is of the form $\mathcal{R} := \{t \mid t \geq \eta\}$. The problem is how to fix the threshold so that the test has the desired significance level.

Imagine that we randomly permute the labels A and B in the merged data set $\vec{x}$. As a result, some of the data that were labeled as A will be labeled as B and vice versa. If we recompute $t_{\text{diff}}\left(\vec{x}\right)$ we will obviously obtain a different value. However, the *distribution* of the random variable $t_{\text{diff}}(\vec{X})$ under the hypothesis that the data are sampled from the same distribution has *not* changed. Indeed, the null hypothesis implies that the distribution of any function of $\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n$ that only depends on the class assigned to each variable is *invariant to permutations*. More formally, the random sequence is **exchangeable** with respect to such functions.

Consider the value of t_{diff} for all the possible permutations of the labels: $t_{\text{diff},1}, t_{\text{diff},2}, \ldots t_{\text{diff},n!}$. If the null hypothesis holds, then it would be surprising to find that $t_{\text{diff}}\left(\vec{x}\right)$ is larger than most of the $t_{\text{diff},i}$. In fact, under the null hypothesis, the random variable $t_{\text{diff}}(\vec{X})$ is uniformly distributed in the set $\{t_{\text{diff},1}, t_{\text{diff},2}, \ldots t_{\text{diff},n!}\}$, so that

$$\mathrm{P}\left(t_{\text{diff}}(\vec{X}) \geq \eta\right) = \frac{1}{n!}\sum_{i=1}^{n!} 1_{t_{\text{diff},i} \geq \eta}. \tag{11.44}$$

This is exactly to the size of the test. We can therefore compute the p value of the observed statistic $t_{\text{diff}}\left(\vec{x}\right)$ as

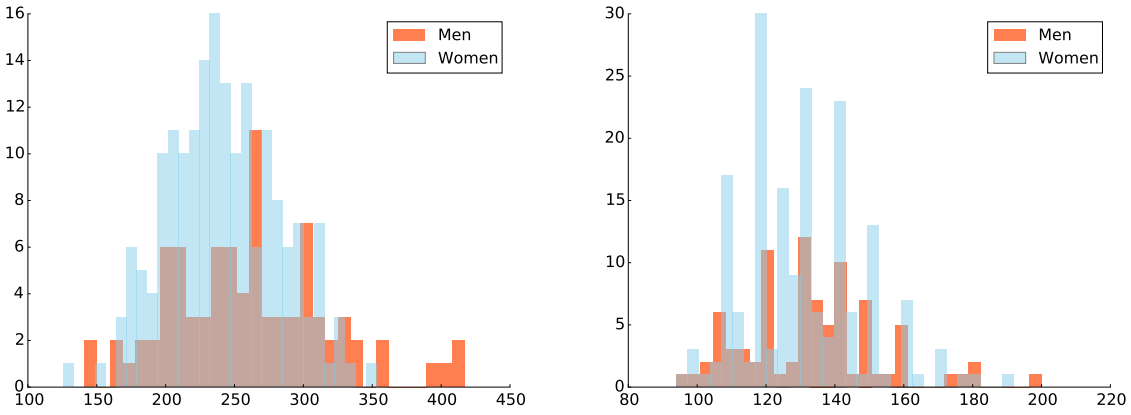
$$p = \mathrm{P}\left(t_{\text{diff}}(\vec{X}) \geq t_{\text{diff}}\left(\vec{x}\right)\right) \tag{11.45}$$

$$= \frac{1}{n!}\sum_{i=1}^{n!} 1_{t_{\text{diff},i} \geq t_{\text{diff}}(\vec{x})}. \tag{11.46}$$

In words, the p value is the fraction of permutations that yield a more extreme test statistic than the one we observe. Unfortunately, it is often challenging to compute (11.46) exactly. Even for moderately sized data sets the number of possible permutations is usually too large (for example, $40! > 8\,10^{47}$) for it to be computationally tractable. In such cases the p value can be approximated by sampling a large number of permutations and making a Monte Carlo approximation of (11.46) with its average.

Before looking at an example, let us review the steps to be followed when applying a permutation test.

1. Choose a conjecture as to how $\vec{x}_A$ and $\vec{x}_B$ are different.



Cholesterol Blood pressure

Figure 11.2: Histograms of the cholesterol and blood-pressure for men and women in Example 11.3.1.

2. Choose a test statistic t_{diff}.
3. Compute $t_{\text{diff}}\left(\vec{x}\right)$.
4. Permute the labels m times and compute the corresponding values of t_{diff}: $t_{\text{diff},1}$, $t_{\text{diff},2}$, $\ldots t_{\text{diff},m}$.
5. Compute the approximate p value

$$p = \mathrm{P}\left(t_{\text{diff}}(\vec{X}) \geq t_{\text{diff}}\left(\vec{x}\right)\right) \tag{11.47}$$

$$= \frac{1}{m}\sum_{i=1}^{m} 1_{t_{\text{diff},i} \geq t_{\text{diff}}(\vec{x})} \tag{11.48}$$

and reject the null hypothesis if it is below a predefined limit (typically 1% or 5%).

Example 11.3.1 (Cholesterol and blood pressure)**.** A scientist want to determine whether men have higher cholesterol and blood pressure. She gathers data from 86 men and 182 women. Figure 11.2 shows the histograms of the cholesterol and blood-pressure for men and women. From the histograms it seems that men have higher levels of cholesterol and blood pressure. The sample mean for cholesterol is 261.3 mg/dl amongst men and 242.0 mg/dl amongst women. The sample mean for blood pressure is 133.2 mmHg amongst men and 130.6 mmHg amongst women.

In order to quantify whether these differences are significant we compute the sample permutation distribution of the difference between the sample means using 10^6 permutations. To make sure that the results are stable, we repeat the procedure three times. The results are shown in Figure 11.3. For cholesterol, the p value is around 0.1%, so we have very strong evidence against the null hypothesis. In contrast, the p value for blood pressure is 13%, so the results are not very conclusive, we cannot reject the possibility that the difference is merely due to random fluctuations.

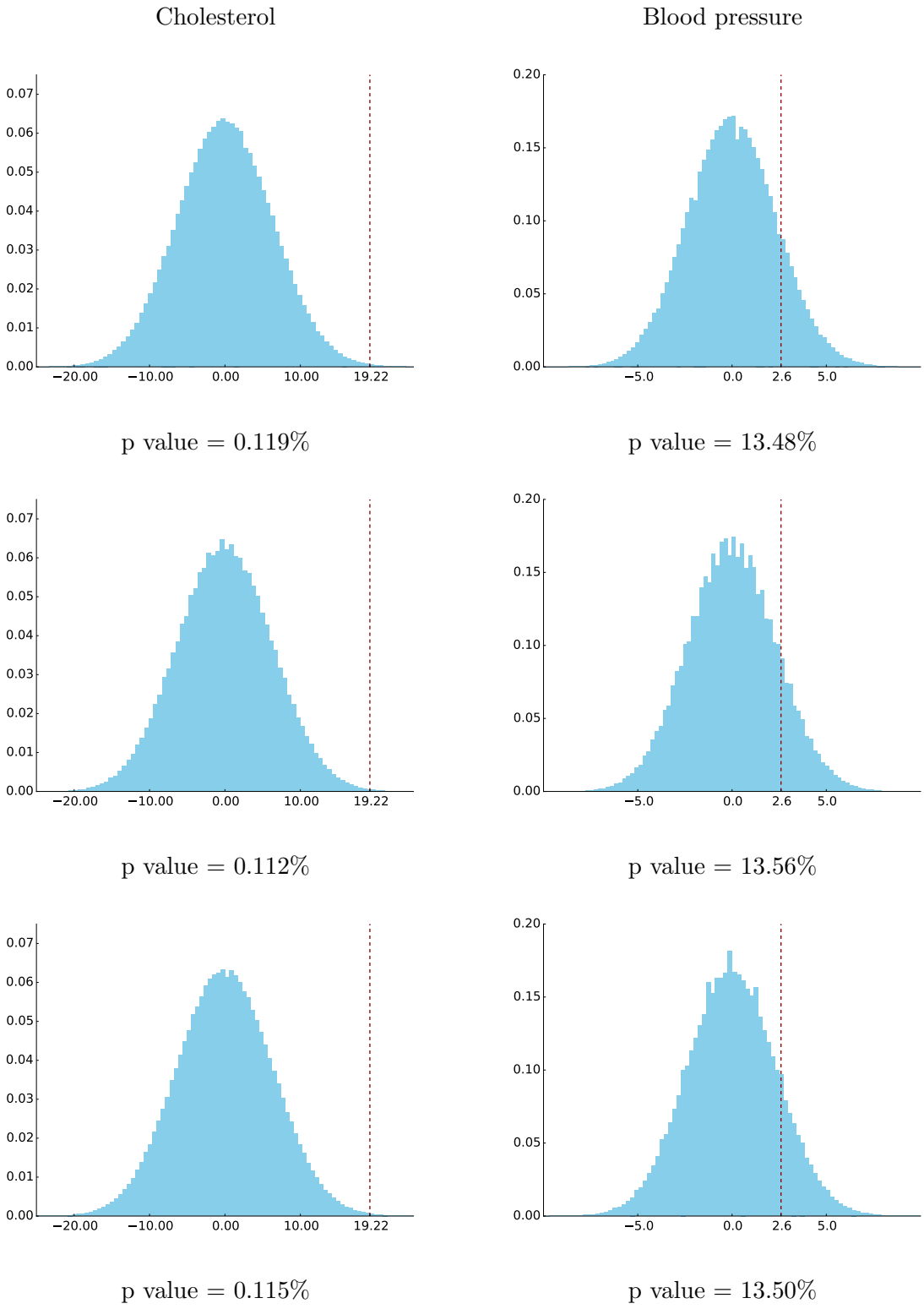


Figure 11.3: Approximate distribution under the null hypothesis of the difference between the sample means of cholesterol and blood pressure in men and women. The observed value for the test statistic is marked by a dashed line.

△

11.4 Multiple testing

In some applications, it is common to conduct many simultaneous hypothesis tests. For example, in computational genomics a researcher might be interested in testing whether any gene within a group of several thousand is relevant to a certain disease. If we apply a hypothesis test with size α in this setting, then the probability of obtaining a false positive for a particular gene is α. Now, assume that we test n genes and that the events *gene i is a false positive*, $1 \leq i \leq n$ are all mutually independent. The probability of obtaining at least one false positive is

$$\begin{aligned} \mathrm{P}\left(\text{at least one false positive}\right) &= 1 - \mathrm{P}\left(\text{no false positives}\right) && (11.49)\\ &= 1 - (1-\alpha)^n . && (11.50) \end{aligned}$$

For $\alpha = 0.01$ and $n = 500$ this probability is equal to 0.99! If we want to control the probability of making a Type I error we must take into account that we are carrying out multiple tests at the same time. A popular procedure to do this is **Bonferroni's method**.

Definition 11.4.1 (Bonferroni's method)**.** *Given n hypothesis tests, compute the corresponding p values $p_1, \ldots, p_n$. For a fixed significance level α reject the ith null hypothesis if*

$$p_i \leq \frac{\alpha}{n}. \tag{11.51}$$

The following lemma shows that the method guarantees that the desired significance level holds simultaneously for all the tests.

Lemma 11.4.2. *If we apply Bonferroni's method, the probability of making a Type I error is bounded by α.*

Proof. The result follows directly from the union bound, which controls the probability of a union of events with the sum of their individual probabilities.

Theorem 11.4.3 (Union bound)**.** *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $S_1, S_2, \ldots$ a collection of events in $\mathcal{F}$. Then*

$$\mathrm{P}\left(\cup_i S_i\right) \leq \sum_i \mathrm{P}\left(S_i\right). \tag{11.52}$$

Proof. Let us define the sets:

$$\tilde{S}_i = S_i \cap \cap_{j=1}^{i-1} S_j^c. \tag{11.53}$$

It is straightforward to show by induction that $\cup_{j=1}^n S_j = \cup_{j=1}^n \tilde{S}_j$ for any n, so $\cup_i S_i = \cup_i \tilde{S}_i$. The sets $\tilde{S}_1$, $\tilde{S}_2$, ... are disjoint by construction, so

$$\begin{aligned} \mathrm{P}\left(\cup_i S_i\right) = \mathrm{P}\left(\cup_i \tilde{S}_i\right) &= \sum_i \mathrm{P}\left(\tilde{S}_i\right) \quad \text{by Axiom 2 in Definition 1.1.4} && (11.54)\\ &\leq \sum_i \mathrm{P}\left(S_i\right) \quad \text{because } \tilde{S}_i \subseteq S_i. && (11.55) \end{aligned}$$

□

Applying the bound,

$$\begin{aligned} \mathrm{P}\left(\text{Type I error}\right) &= \mathrm{P}\left(\cup_{i=1}^{n} \text{Type I error for test } i\right) && (11.56)\\ &\leq \sum_{i=1}^{n} \mathrm{P}\left(\text{Type I error for test } i\right) \qquad \text{by the union bound} && (11.57)\\ &= n \cdot \frac{\alpha}{n} = \alpha. && (11.58) \end{aligned}$$

□

Example 11.4.4 (Clutch (continued))**.** If we apply the test in Example 11.1.4 to 10 players, the probability that one of them seems to be clutch just due to chance increases substantially. To control for this, by Bonferroni's method we must divide the p values of the individual tests by 10. As a result, to maintain a significance level of 0.05 we would require that each player score more points per minute during the last quarter in 17 of the 20 games instead of 15 (see Table 11.2) in order to reject the null hypothesis.

△

Chapter 12

Linear Regression

In statistics, regression is the problem of characterizing the relation between a certain quantity of interest y, called the **response** or the **dependent variable**, to several observed variables x_1, x_2, ..., x_p, known as **covariates**, **features** or **independent variables**. For example, the response could be price of a house and the covariates could correspond to the extension, the number of rooms, the year it was built, etc. A regression model would describe how house prices are affected by all of these factors.

More formally, the main assumption in regression models is that the predictor is generated according to a function h applied to the features and then perturbed by some unknown noise z, which is often additive,

$$y = h\left(\vec{x}\right) + z. \tag{12.1}$$

The aim is to learn h from n examples of responses and their corresponding features

$$\left(y^{(1)}, \vec{x}^{\,(1)}\right), \left(y^{(2)}, \vec{x}^{\,(2)}\right), \ldots, \left(y^{(n)}, \vec{x}^{\,(n)}\right). \tag{12.2}$$

In this chapter we focus on the case where h is a linear function.

12.1 Linear models

If the regression function h in a model of the form 12.1 is linear, then the response is modeled as a linear combination of the predictors:

$$y^{(i)} = \vec{x}^{\,(i)\,T}\vec{\beta}^{\,*} + z^{(i)}, \quad 1 \leq i \leq n, \tag{12.3}$$

where $z^{(i)}$ is an entry of the unknown noise vector. The function is parametrized by a vector of weights $\vec{\beta}^{\,*} \in \mathbb{R}^p$. All we need to fit the linear model to the data is to estimate these weights.

Expressing the linear system (12.3) in matrix form yields the following representation of the linear-regression model

$$\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdots \\ y^{(n)} \end{bmatrix} = \begin{bmatrix} \vec{x}_1^{\,(1)} & \vec{x}_2^{\,(1)} & \cdots & \vec{x}_p^{\,(1)} \\ \vec{x}_1^{\,(2)} & \vec{x}_2^{\,(2)} & \cdots & \vec{x}_p^{\,(2)} \\ \cdots & \cdots & \cdots & \cdots \\ \vec{x}_1^{\,(n)} & \vec{x}_2^{\,(n)} & \cdots & \vec{x}_p^{\,(n)} \end{bmatrix} \begin{bmatrix} \vec{\beta}_1^{\,*} \\ \vec{\beta}_2^{\,*} \\ \cdots \\ \vec{\beta}_p^{\,*} \end{bmatrix} + \begin{bmatrix} z^{(1)} \\ z^{(2)} \\ \cdots \\ z^{(n)} \end{bmatrix}. \tag{12.4}$$

Equivalently,

$$\vec{y} = \mathcal{X}\vec{\beta}^* + \vec{z}, \tag{12.5}$$

where $\mathcal{X}$ is a $n \times p$ matrix containing the features, $\vec{y}$ contains the response and $\vec{z} \in \mathbb{R}^n$ represents the noise.

Example 12.1.1 (Linear model for GDP)**.** We consider the problem of building a linear model to predict the gross domestic product (GDP) of a state in the US from its population and unemployment rate. We have available the following data:

	GDP (USD millions)	Population	Unemployment rate (%)
North Dakota	52 089	757 952	2.4
Alabama	204 861	4 863 300	3.8
Mississippi	107 680	2 988 726	5.2
Arkansas	120 689	2 988 248	3.5
Kansas	153 258	2 907 289	3.8
Georgia	525 360	10 310 371	4.5
Iowa	178 766	3 134 693	3.2
West Virginia	73 374	1 831 102	5.1
Kentucky	197 043	4 436 974	5.2
Tennessee	???	6 651 194	3.0

In this example, the GDP is the response, and the population and the unemployment rate are the features. Our goal is to fit a linear model to the data so that we can predict the GDP of Tennessee, using a linear model. We begin by centering and normalizing the data. The averages of the response and of the features are

$$\text{av}\,(\vec{y}) = 179\,236, \qquad \text{av}\,(X) = \begin{bmatrix} 3\,802\,073 & 4.1 \end{bmatrix}. \tag{12.6}$$

The empirical standard deviations are

$$\text{std}\,(\vec{y}) = 396\,701, \qquad \text{std}\,(X) = \begin{bmatrix} 7\,720\,656 & 2.80 \end{bmatrix}. \tag{12.7}$$

We subtract the average and divide by the standard deviations so that both the response and

the features are centered and on the same scale,

$$\vec{y} = \begin{bmatrix} -0.321 \\ 0.065 \\ -0.180 \\ -0.148 \\ -0.065 \\ 0.872 \\ -0.001 \\ -0.267 \\ 0.045 \end{bmatrix}, \qquad X = \begin{bmatrix} -0.394 & -0.600 \\ 0.137 & -0.099 \\ -0.105 & 0.401 \\ -0.105 & -0.207 \\ -0.116 & -0.099 \\ 0.843 & 0.151 \\ -0.086 & -0.314 \\ -0.255 & 0.366 \\ 0.082 & 0.401 \end{bmatrix}. \tag{12.8}$$

To obtain the estimate for the GDP of Tennessee we fit the model

$$\vec{y} \approx X\vec{\beta}, \tag{12.9}$$

rescale according to the standard deviations (12.7) and recenter using the averages (12.6). The final estimate is

$$\vec{y}^{\text{Ten}} = \text{av}\left(\vec{y}\right) + \text{std}\left(\vec{y}\right) \left\langle \vec{x}^{\text{Ten}}_{\text{norm}}, \vec{\beta} \right\rangle \tag{12.10}$$

where $\vec{x}^{\text{Ten}}_{\text{norm}}$ is centered using $\text{av}\left(X\right)$ and normalized using $\text{std}\left(X\right)$. △

12.2 Least-squares estimation

To calibrate the linear regression model, we need to estimate the weight vector so that it yields a good fit to the data. We can evaluate the fit for a specific choice of $\vec{\beta} \in \mathbb{R}^p$ using the sum of the squares of the error,

$$\sum_{i=1}^{n} \left(y^{(i)} - \vec{x}^{(i)\,T}\vec{\beta}\right)^2 = \left|\left|\vec{y} - \mathcal{X}\vec{\beta}\right|\right|_2^2. \tag{12.11}$$

The least-squares estimate $\vec{\beta}_{\text{LS}}$ is the vector of weights that minimizes this cost function,

$$\vec{\beta}_{\text{LS}} := \arg\min_{\vec{\beta}} \quad \left|\left|\vec{y} - \mathcal{X}\vec{\beta}\right|\right|_2. \tag{12.12}$$

The least-squares cost function is convenient from a computational view, since it is convex and can be minimized efficiently (in fact, as we will see in a moment it has a closed-form solution). In addition, it has intuitive geometric and probabilistic interpretations. Figure 12.1 shows the linear model learnt using least squares in a simple example where there is just one feature ($p = 1$) and 40 examples ($n = 40$).

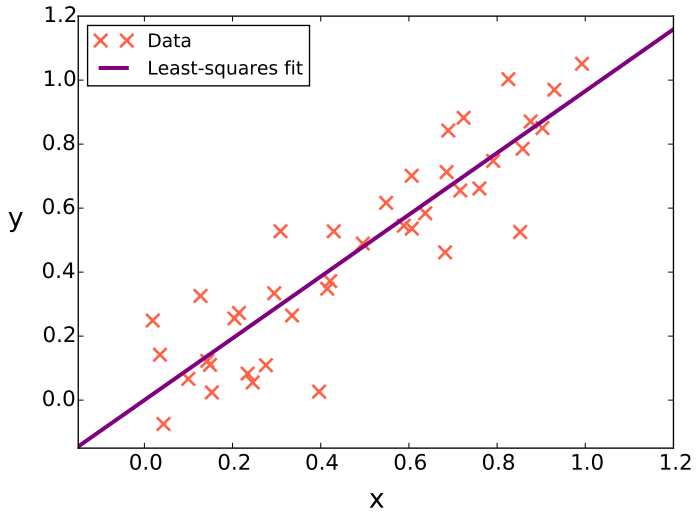


Figure 12.1: Linear model learnt via least-squares fitting for a simple example where there is just one feature ($p = 1$) and 40 examples ($n = 40$).

Example 12.2.1 (Linear model for GDP (continued))**.** The least-squares estimate for the regression coefficients in the linear GDP model is equal to

$$\vec{\beta}_{\text{LS}} = \begin{bmatrix} 1.019 \\ -0.111 \end{bmatrix}. \tag{12.13}$$

The GDP seems to be proportional to the population and inversely proportional to the unemployment rate. We now compare the fit provided by the linear model to the original data, as well as its prediction of the GDP of Tennessee:

	GDP	Estimate
North Dakota	52 089	46 241
Alabama	204 861	239 165
Mississippi	107 680	119 005
Arkansas	120 689	145 712
Kansas	153 258	136 756
Georgia	525 360	513 343
Iowa	178 766	158 097
West Virginia	73 374	59 969
Kentucky	197 043	194 829
Tennessee	328 770	345 352

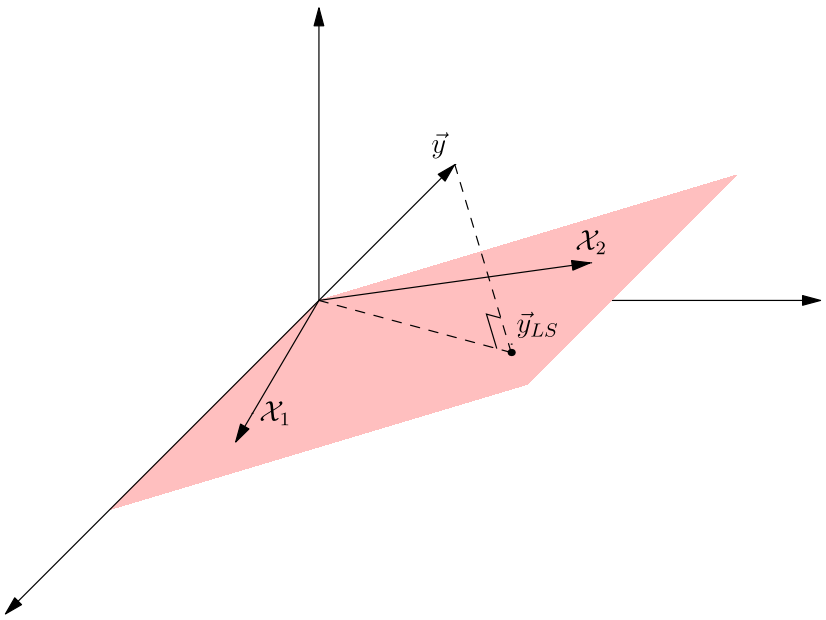


Figure 12.2: Illustration of Corollary 12.2.3. The least-squares solution is a projection of the data onto the subspace spanned by the columns of $\mathcal{X}$, denoted by $\mathcal{X}_1$ and $\mathcal{X}_2$.

△

12.2.1 Geometric interpretation

The following theorem, proved in Section 12.2.2, shows that the least-squares problem has a closed form solution.

Theorem 12.2.2 (Least-squares solution)**.** *For $p \geq n$, if $\mathcal{X}$ is full rank then the solution to the least-squares problem* (12.12) *is*

$$\vec{\beta}_{\text{LS}} := \left(\mathcal{X}^T \mathcal{X}\right)^{-1} \mathcal{X}^T \vec{y}. \tag{12.14}$$

A corollary to this result provides a geometric interpretation for the least-squares estimate of $\vec{y}$: it is obtained by projecting the response onto the column space of the matrix formed by the predictors.

Corollary 12.2.3. *For $p \geq n$, if $\mathcal{X}$ is full rank then $\mathcal{X}\vec{\beta}_{\text{LS}}$ is the projection of $\vec{y}$ onto the column space of $\mathcal{X}$.*

We provide a formal proof in Section 12.5.2 of the appendix, but the result is very intuitive. Any vector of the form $\mathcal{X}\vec{\beta}$ is in the span of the columns of $\mathcal{X}$. By definition, the least-squares estimate is the closest vector to $\vec{y}$ that can be represented in this way, so it is the projection of $\vec{y}$ onto the column space of $\mathcal{X}$. This is illustrated in Figure 12.2.

12.2.2 Probabilistic interpretation

If we model the noise in (12.5) as a realization from a random vector $\vec{Z}$ which has entries that are independent Gaussian random variables with mean zero and a certain variance σ^2, then we can

interpret the least-squares estimate as a maximum-likelihood estimate. Under that assumption, the data are a realization of the random vector

$$\vec{Y} := \mathcal{X}\vec{\beta} + \vec{Z}, \tag{12.15}$$

which is an iid Gaussian random vector with mean $\mathcal{X}\vec{\beta}$ and covariance matrix $\sigma^2 I$. The joint pdf of $\vec{Y}$ is equal to

$$f_{\vec{Y}}\left(\vec{a}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(\vec{a}_i - \left(\mathcal{X}\vec{\beta}\right)_i\right)^2\right) \tag{12.16}$$

$$= \frac{1}{\sqrt{(2\pi)^n}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\left|\left|\vec{a} - \mathcal{X}\vec{\beta}\right|\right|_2^2\right). \tag{12.17}$$

The likelihood is the probability density function of $\vec{Y}$ evaluated at the observed data $\vec{y}$ and interpreted as a function of the weight vector $\vec{\beta}$,

$$\mathcal{L}_{\vec{y}}\left(\vec{\beta}\right) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}\left|\left|\vec{y} - \mathcal{X}\vec{\beta}\right|\right|_2^2\right). \tag{12.18}$$

To find the ML estimate, we maximize the log likelihood. We conclude that it is given by the solution to the least-squares problem, since

$$\vec{\beta}_{\mathrm{ML}} = \arg\max_{\vec{\beta}} \mathcal{L}_{\vec{y}}\left(\vec{\beta}\right) \tag{12.19}$$

$$= \arg\max_{\vec{\beta}} \log \mathcal{L}_{\vec{y}}\left(\vec{\beta}\right) \tag{12.20}$$

$$= \arg\min_{\vec{\beta}} \left|\left|\vec{y} - \mathcal{X}\vec{\beta}\right|\right|_2^2 \tag{12.21}$$

$$= \vec{\beta}_{\mathrm{LS}}. \tag{12.22}$$

12.3 Overfitting

Imagine that a friend tells you:

I found a cool way to predict the temperature in New York: It's just a linear combination of the temperature in every other state. I fit the model on data from the last month and a half and it's perfect!

Your friend is not lying, but the problem is that she is using a number of data points to fit the linear model that is roughly the same as the number of parameters. If $n \leq p$ we can find a $\vec{\beta}$ such that $\vec{y} = \mathcal{X}\vec{\beta}$ exactly, even if $\vec{y}$ and $\mathcal{X}$ have nothing to do with each other! This is called overfitting and is usually caused by using a model that is too flexible with respect to the number of data that are available.

To evaluate whether a model suffers from overfitting we separate the data into a training set and a test set. The training set is used to fit the model and the test set is used to evaluate the error. A model that overfits the training set will have very low error when evaluated on the training examples, but will not generalize well to the test examples.

Figure 12.3 shows the result of evaluating the training error and the test error of a linear model with $p = 50$ parameters fitted from n training examples. The training and test data are generated by fixing a vector of weights $\vec{\beta}^*$ and then computing

$$\vec{y}_{\text{train}} = \mathcal{X}_{\text{train}} \, \vec{\beta}^* + \vec{z}_{\text{train}}, \tag{12.23}$$

$$\vec{y}_{\text{test}} = \mathcal{X}_{\text{test}} \, \vec{\beta}^*, \tag{12.24}$$

where the entries of $\mathcal{X}_{\text{train}}$, $\mathcal{X}_{\text{test}}$, $\vec{z}_{\text{train}}$ and $\vec{\beta}^*$ are sampled independently at random from a Gaussian distribution with zero mean and unit variance. The training and test errors are defined as

$$\text{error}_{\text{train}} = \frac{\left|\left| \mathcal{X}_{\text{train}} \, \vec{\beta}_{\text{LS}} - \vec{y}_{\text{train}} \right|\right|_2}{||\vec{y}_{\text{train}}||_2}, \tag{12.25}$$

$$\text{error}_{\text{test}} = \frac{\left|\left| \mathcal{X}_{\text{test}} \, \vec{\beta}_{\text{LS}} - \vec{y}_{\text{test}} \right|\right|_2}{||\vec{y}_{\text{test}}||_2}. \tag{12.26}$$

Note that even the true $\vec{\beta}^*$ does not achieve zero training error because of the presence of the noise, but the test error is actually zero if we manage to estimate $\vec{\beta}^*$ exactly.

The training error of the linear model grows with n. This makes sense as the model has to fit more data using the same number of parameters. When n is close to $p := 50$, the fitted model is much better than the true model at replicating the training data (the error of the true model is shown in green). This is a sign of overfitting: the model is adapting to the noise and not learning the true linear structure. Indeed, in that regime the test error is extremely high. At larger n, the training error rises to the level achieved by the true linear model and the test error decreases, indicating that we are learning the underlying model.

12.4 Global warming

In this section we describe an application of linear regression to climate data. In particular, we analyze temperature data taken in a weather station in Oxford over 150 years.[1] Our objective is not to perform prediction, but rather to determine whether temperatures have risen or decreased during the last 150 years in Oxford.

In order to separate the temperature into different components that account for seasonal effects we use a simple linear model with three predictors and an intercept

$$\vec{y}_t \approx \vec{\beta}_0 + \vec{\beta}_1 \cos\left(\frac{2\pi t}{12}\right) + \vec{\beta}_2 \sin\left(\frac{2\pi t}{12}\right) + \vec{\beta}_3 \, t \tag{12.27}$$

where $1 \leq t \leq n$ denotes the time in months (n equals 12 times 150). The corresponding matrix

[1]The data is available at http://www.metoffice.gov.uk/pub/data/weather/uk/climate/stationdata/oxforddata.txt.

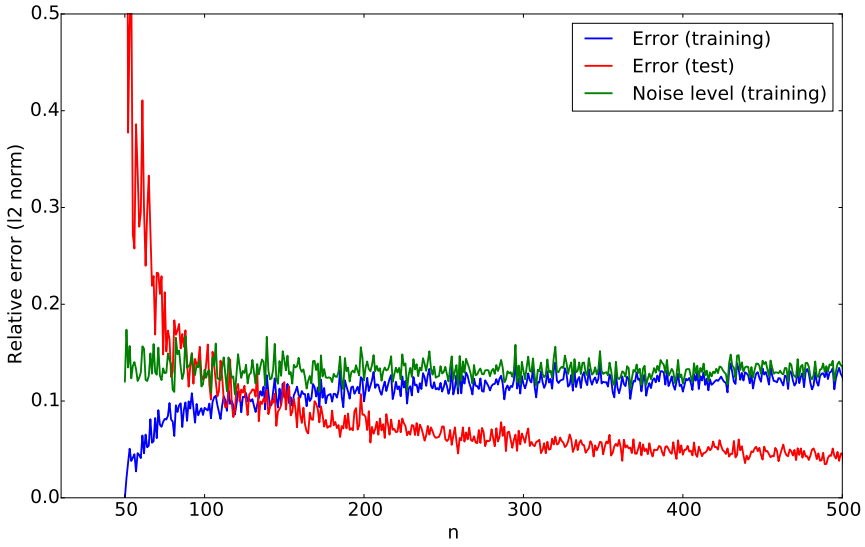


Figure 12.3: Relative ℓ_2-norm error in estimating the response achieved using least-squares regression for different values of n (the number of training data). The training error is plotted in blue, whereas the test error is plotted in red. The green line indicates the training error of the true model used to generate the data.

of predictors is

$$\mathcal{X} := \begin{bmatrix} 1 & \cos\left(\frac{2\pi t_1}{12}\right) & \sin\left(\frac{2\pi t_1}{12}\right) & t_1 \\ 1 & \cos\left(\frac{2\pi t_2}{12}\right) & \sin\left(\frac{2\pi t_2}{12}\right) & t_2 \\ \dots & \dots & \dots & \dots \\ 1 & \cos\left(\frac{2\pi t_n}{12}\right) & \sin\left(\frac{2\pi t_n}{12}\right) & t_n \end{bmatrix}. \tag{12.28}$$

The intercept $\vec{\beta}_0$ represents the mean temperature, $\vec{\beta}_1$ and $\vec{\beta}_2$ account for periodic yearly fluctuations and $\vec{\beta}_3$ is the overall trend. If $\vec{\beta}_3$ is positive then the model indicates that temperatures are increasing, if it is negative then it indicates that temperatures are decreasing.

The results of fitting the linear model are shown in Figures 12.4 and 12.5. The fitted model indicates that both the maximum and minimum temperatures have an increasing trend of about 0.8 degrees Celsius (around 1.4 degrees Fahrenheit).

12.5 Proofs

12.5.1 Proof of Proposition 12.2.2

Let $\mathcal{X} = U\Sigma V_T$ be the singular-value decomposition (SVD) of $\mathcal{X}$. Under the conditions of the theorem, $\left(\mathcal{X}^T\mathcal{X}\right)^{-1}\mathcal{X}^T y = V\Sigma U^T$. We begin by separating $\vec{y}$ into two components

$$y = UU^T y + \left(I - UU^T\right) y \tag{12.29}$$

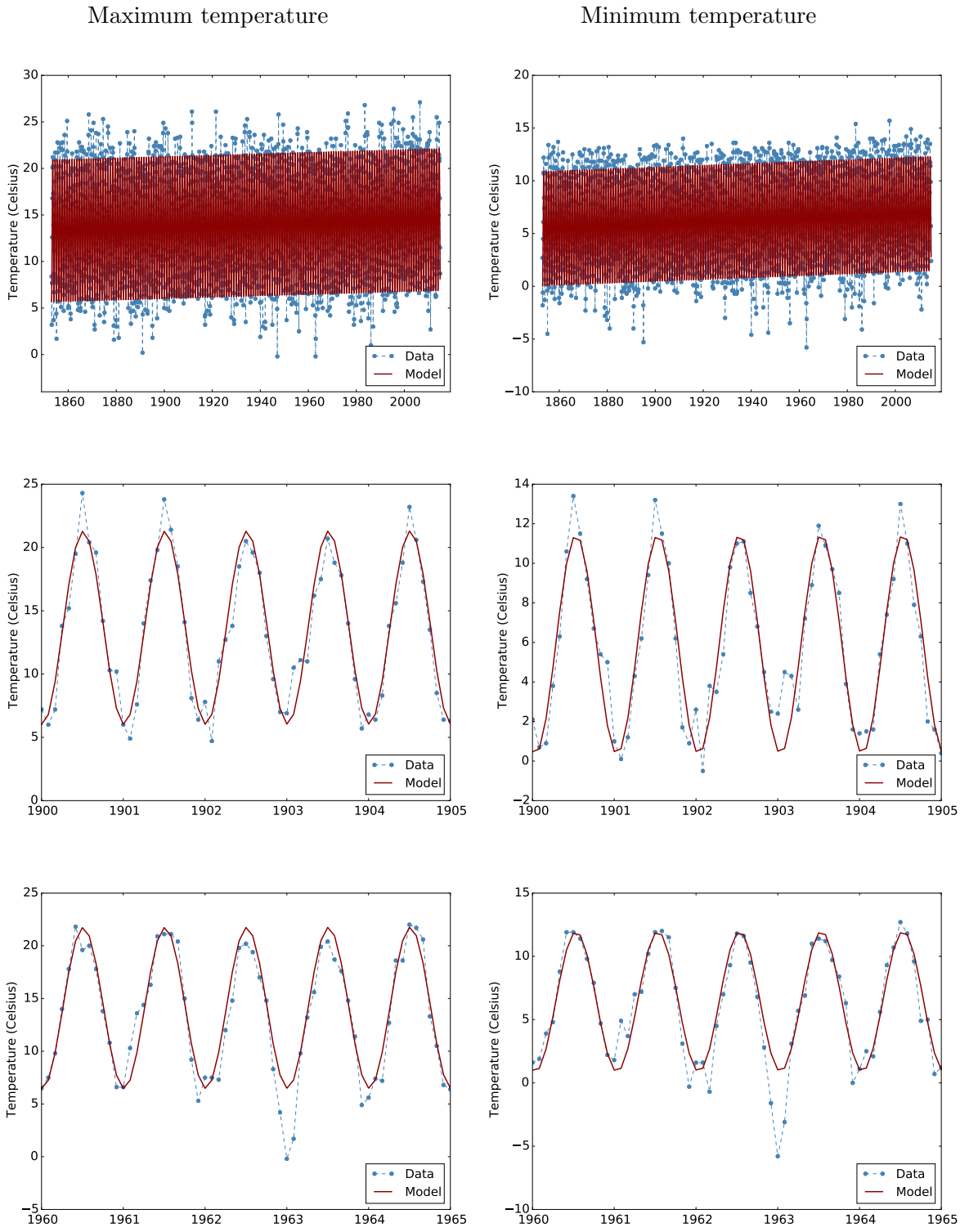


Figure 12.4: Temperature data together with the linear model described by (12.27) for both maximum and minimum temperatures.

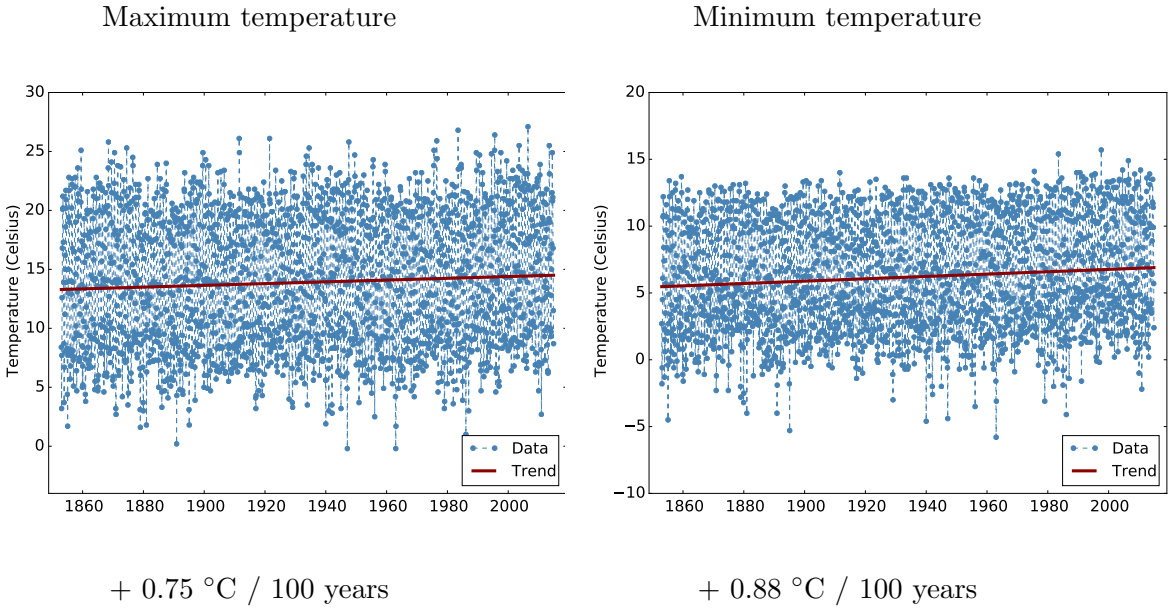


Figure 12.5: Temperature trend obtained by fitting the model described by (12.27) for both maximum and minimum temperatures.

where $UU^T y$ is the projection of $\vec{y}$ onto the column space of $\mathcal{X}$. Note that $\left(I - UU^T\right) y$ is orthogonal to the column space of $\mathcal{X}$ and consequently to both $UU^T y$ and $\mathcal{X}\vec{\beta}$ for any $\vec{\beta}$. By Pythagoras's Theorem

$$\left|\left|\vec{y} - \mathcal{X}\vec{\beta}\right|\right|_2^2 = \left|\left|\left(I - UU^T\right) y\right|\right|_2^2 + \left|\left|UU^T y - \mathcal{X}\vec{\beta}\right|\right|_2^2. \tag{12.30}$$

The minimum value of this cost function that can be achieved by optimizing over $\tilde{beta}$ is $||\vec{y}_{\mathcal{X}^\perp}||_2^2$. This can be achieved by solving the system of equations

$$UU^T y = \mathcal{X}\vec{\beta} = U\Sigma V_T \vec{\beta}. \tag{12.31}$$

Since $U^T U = I$ because $p \geq n$, multiplying both sides of the equality yields the equivalent system

$$U^T y = \Sigma V_T \vec{\beta}. \tag{12.32}$$

Since $\mathcal{X}$ is full rank, Σ and V are square and invertible (and by definition of the SVD $V^{-1} = V^T$), so

$$\vec{\beta}_{\text{LS}} = V\Sigma U^T y \tag{12.33}$$

is the unique solution to the system and consequently also of the least-squares problem.

12.5.2 Proof of Corollary 12.2.3

Let $\mathcal{X} = U\Sigma V^T$ be the singular-value decomposition of $\mathcal{X}$. Since $\mathcal{X}$ is full rank and $p \geq n$ we have $U^TU = I$, $V^TV = I$ and Σ is a square invertible matrix, which implies

$$\mathcal{X}\vec{\beta}_{\text{LS}} = \mathcal{X}\left(\mathcal{X}^T\mathcal{X}\right)^{-1}\mathcal{X}^T y \tag{12.34}$$

$$= U\Sigma V^T\left(V\Sigma U^T U\Sigma V^T\right)V\Sigma U^T y \tag{12.35}$$

$$= UU^T y. \tag{12.36}$$

Appendix A

Set theory

This chapter provides a review of basic concepts in set theory.

A.1 Basic definitions

A **set** is a collection of objects. The set containing every possible object that we consider in a certain situation is called the **universe** and is usually denoted by Ω. If an object x in Ω belongs to set S, we say that x is an **element** of S and write $x \in S$. If x is not an element of S then we write $x \notin S$. The **empty set**, usually denoted by $\emptyset$, is a set such that $x \notin$ for all $x \in \Omega$ (i.e. it has no elements). If all the elements in a set B also belong to a set A then B is a **subset** of A, which we denote by $B \subseteq A$. If in addition there is at least one element of A that does not belong to B then B is a proper subset of A, denoted by $B \subset A$.

The elements of a set can be arbitrary objects and in particular they can be sets themselves. This is the case for the power set of a set, defined in the next section.

A useful way of defining a set is through a statement concerning its elements. Let S be the set of elements such that a certain statement $s(x)$ holds, to define S we write

$$S := \{x \mid s(x)\}\,. \tag{A.1}$$

For example, $A := \{x \mid 1 < x < 3\}$ is the set of all elements greater than 1 and smaller than 3. Let us define some important sets and set operations using this notation.

A.2 Basic operations

Definition A.2.1 (Set operations)**.**

- *The **complement** S^c of a set S contains all elements that are not in S.*

$$S^c := \{x \mid x \notin S\}\,. \tag{A.2}$$

- *The **union** of two sets A and B contains the objects that belong to A or B.*

$$A \cup B := \{x \mid x \in A \text{ or } x \in B\}\,. \tag{A.3}$$

This can be generalized to a sequence of sets A_1, A_2, ...

$$\bigcup_n A_n := \{x \mid x \in A_n \text{ for some } n\}, \tag{A.4}$$

where the sequence may be infinite.

- *The **intersection** of two sets A and B contains the objects that belong to A and B.*

$$A \cap B := \{x \mid x \in A \text{ and } x \in B\}. \tag{A.5}$$

Again, this can be generalized to a sequence,

$$\bigcap_n A_n := \{x \mid x \in A_n \text{ for all } n\}. \tag{A.6}$$

- *The **difference** of two sets A and B contains the elements in A that are not in B.*

$$A/B := \{x \mid x \in A \text{ and } x \notin B\}. \tag{A.7}$$

- *The **power set** 2^S of a set S is the set of all possible subsets of S, including $\emptyset$ and S.*

$$2^S := \left\{S' \mid S' \subseteq S\right\}. \tag{A.8}$$

- *The **cartesian product** of two sets S_1 and S_2 is the set of all ordered pairs of elements in the sets*

$$S_1 \times S_2 := \{(x_1, x_2) \mid x_1 \in S_1, x_2 \in S_2\}. \tag{A.9}$$

An example is $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, the set of all possible pairs of real numbers.

Two sets are equal if they have the same elements, i.e. $A = B$ if and only if $A \subseteq B$ and $B \subseteq A$. It is easy to verify for instance that $(A^c)^c = A$, $S \cup \Omega = \Omega$, $S \cap \Omega = S$ or the following identities which are known as *De Morgan's laws*.

Theorem A.2.2 (De Morgan's laws)**.** *For any two sets A and B*

$$(A \cup B)^c = A^c \cap B^c, \tag{A.10}$$
$$(A \cap B)^c = A^c \cup B^c. \tag{A.11}$$

Proof. Let us prove the first identity; the proof of the second is almost identical.

First we prove that $(A \cup B)^c \subseteq A^c \cap B^c$. A standard way to prove the inclusion of a set in another set is to show that if an element belongs to the first set then it must also belong to the second. Any element x in $(A \cup B)^c$ (if the set is empty then the inclusion holds trivially, since $\emptyset \subseteq S$ for any set S) is in A^c; otherwise it would belong to A and consequently to $A \cup B$. Similarly, x also belongs to B^c. We conclude that x belongs to $A^c \cap B^c$, which proves the inclusion.

To complete the proof we establish $A^c \cap B^c \subseteq (A \cup B)^c$. If $x \in A^c \cap B^c$, then $x \notin A$ and $x \notin B$, so $x \notin A \cup B$ and consequently $x \in (A \cup B)^c$. □

Appendix B

Linear Algebra

This chapter provides a review of basic concepts in linear algebra.

B.1 Vector spaces

You are no doubt familiar with **vectors** in $\mathbb{R}^2$ or $\mathbb{R}^3$, i.e.

$$\vec{x} = \begin{bmatrix} 2.2 \\ 3 \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} -1 \\ 0 \\ 5 \end{bmatrix}. \tag{B.1}$$

From the point of view of algebra, vectors are much more general objects. They are elements of sets called **vector spaces** that satisfy the following definition.

Definition B.1.1 (Vector space)**.** *A vector space consists of a set $\mathcal{V}$ and two operations $+$ and $\cdot$ satisfying the following conditions.*

1. *For any pair of elements $\vec{x}, \vec{y} \in \mathcal{V}$ the **vector sum** $\vec{x} + \vec{y}$ belongs to $\mathcal{V}$.*
2. *For any $\vec{x} \in \mathcal{V}$ and any scalar $\alpha \in \mathbb{R}$ the **scalar multiple** $\alpha \cdot \vec{x} \in \mathcal{V}$.*
3. *There exists a **zero vector** or **origin** $\vec{0}$ such that $\vec{x} + \vec{0} = \vec{x}$ for any $\vec{x} \in \mathcal{V}$.*
4. *For any $\vec{x} \in \mathcal{V}$ there exists an additive inverse $\vec{y}$ such that $\vec{x} + \vec{y} = \vec{0}$, usually denoted as $-\vec{x}$.*
5. *The vector sum is commutative and associative, i.e. for all $\vec{x}, \vec{y} \in \mathcal{V}$*

$$\vec{x} + \vec{y} = \vec{y} + \vec{x}, \quad (\vec{x} + \vec{y}) + z = \vec{x} + (\vec{y} + z). \tag{B.2}$$

6. *Scalar multiplication is associative, for any $\alpha, \beta \in \mathbb{R}$ and $\vec{x} \in \mathcal{V}$*

$$\alpha \, (\beta \cdot \vec{x}) = (\alpha \, \beta) \cdot \vec{x}. \tag{B.3}$$

7. *Scalar and vector sums are both distributive, i.e. for all $\alpha, \beta \in \mathbb{R}$ and $\vec{x}, \vec{y} \in \mathcal{V}$*

$$(\alpha + \beta) \cdot \vec{x} = \alpha \cdot \vec{x} + \beta \cdot \vec{x}, \quad \alpha \cdot (\vec{x} + \vec{y}) = \alpha \cdot \vec{x} + \alpha \cdot \vec{y}. \tag{B.4}$$

*A **subspace** of a vector space $\mathcal{V}$ is a subset of $\mathcal{V}$ that is also itself a vector space.*

From now on, for ease of notation we will ignore the symbol for the scalar product $\cdot$, writing $\alpha \cdot \vec{x}$ as $\alpha\, \vec{x}$.

Remark B.1.2 (More general definition)**.** *We can define vector spaces over an arbitrary field, instead of $\mathbb{R}$, such as the complex numbers $\mathbb{C}$. We refer to any linear algebra text for more details.*

We can easily check that $\mathbb{R}^n$ is a valid vector space together with the usual vector addition and vector-scalar product. In this case the zero vector is the all-zero vector $\begin{bmatrix} 0 & 0 & 0 & \dots \end{bmatrix}^T$. When thinking about vector spaces it is a good idea to have $\mathbb{R}^2$ or $\mathbb{R}^3$ in mind to gain intuition, but it is also important to bear in mind that we can define vector sets over many other objects, such as infinite sequences, polynomials, functions and even random variables as in the following example.

The definition of vector space guarantees that any **linear combination** of vectors in a vector space $\mathcal{V}$, obtained by adding the vectors after multiplying by scalar coefficients, belongs to $\mathcal{V}$. Given a set of vectors, a natural question to ask is whether they can be expressed as linear combinations of each other, i.e. if they are **linearly dependent** or **independent**.

Definition B.1.3 (Linear dependence/independence)**.** *A set of m vectors $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_m$ is linearly dependent if there exist m scalar coefficients $\alpha_1, \alpha_2, \dots, \alpha_m$ which are **not** all equal to zero and such that*

$$\sum_{i=1}^{m} \alpha_i\, \vec{x}_i = \vec{0}. \tag{B.5}$$

*Otherwise, the vectors are **linearly independent**.*

Equivalently, at least one vector in a linearly dependent set can be expressed as the linear combination of the rest, whereas this is not the case for linearly independent sets.

Let us check the equivalence. Equation (B.5) holds with $\alpha_j \neq 0$ for some j if and only if

$$\vec{x}_j = \frac{1}{\alpha_j} \sum_{i \in \{1,\dots,m\}/\{j\}} \alpha_i\, \vec{x}_i. \tag{B.6}$$

We define the **span** of a set of vectors $\{\vec{x}_1, \dots, \vec{x}_m\}$ as the set of all possible linear combinations of the vectors:

$$\text{span}\,(\vec{x}_1, \dots, \vec{x}_m) := \left\{ \vec{y} \mid \vec{y} = \sum_{i=1}^{m} \alpha_i\, \vec{x}_i \quad \text{for some } \alpha_1, \alpha_2, \dots, \alpha_m \in \mathbb{R} \right\}. \tag{B.7}$$

This turns out to be a vector space.

Lemma B.1.4. *The span of any set of vectors $\vec{x}_1, \dots, \vec{x}_m$ belonging to a vector space $\mathcal{V}$ is a subspace of $\mathcal{V}$.*

Proof. The span is a subset of $\mathcal{V}$ due to Conditions 1 and 2 in Definition B.1.1. We now show that it is a vector space. Conditions 5, 6 and 7 in Definition B.1.1 hold because $\mathcal{V}$ is a vector space. We check Conditions 1, 2, 3 and 4 by proving that for two arbitrary elements of the span

$$\vec{y}_1 = \sum_{i=1}^{m} \alpha_i \, \vec{x}_i, \quad \vec{y}_2 = \sum_{i=1}^{m} \beta_i \, \vec{x}_i, \quad \alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_m \in \mathbb{R}, \tag{B.8}$$

$\gamma_1 \, \vec{y}_1 + \gamma_2 \, \vec{y}_2$ also belongs to the span. This holds because

$$\gamma_1 \, \vec{y}_1 + \gamma_2 \, \vec{y}_2 = \sum_{i=1}^{m} (\gamma_1 \, \alpha_i + \gamma_2 \, \beta_i) \, \vec{x}_i, \tag{B.9}$$

so $\gamma_1 \, \vec{y}_1 + \gamma_2 \, \vec{y}_2$ is in $\text{span}\,(\vec{x}_1, \ldots, \vec{x}_m)$. Now to prove Condition 1 we set $\gamma_1 = \gamma_2 = 1$, for Condition 2 $\gamma_2 = 0$, for Condition 3 $\gamma_1 = \gamma_2 = 0$ and for Condition 4 $\gamma_1 = -1, \gamma_2 = 0$. □

When working with a vector space, it is useful to consider the set of vectors with the smallest cardinality that spans the space. This is called a **basis** of the vector space.

Definition B.1.5 (Basis)**.** *A basis of a vector space $\mathcal{V}$ is a set of independent vectors $\{\vec{x}_1, \ldots, \vec{x}_m\}$ such that*

$$\mathcal{V} = \textit{span}\,(\vec{x}_1, \ldots, \vec{x}_m)\,. \tag{B.10}$$

An important property of all bases in a vector space is that they have the same cardinality.

Theorem B.1.6. *If a vector space $\mathcal{V}$ has a basis with finite cardinality then every basis of $\mathcal{V}$ contains the same number of vectors.*

This theorem, which is proved in Section B.8.1, allows us to define the **dimension** of a vector space.

Definition B.1.7 (Dimension)**.** *The dimension $\dim(\mathcal{V})$ of a vector space $\mathcal{V}$ is the cardinality of any of its bases, or equivalently the smallest number of linearly independent vectors that span $\mathcal{V}$.*

This definition coincides with the usual geometric notion of dimension in $\mathbb{R}^2$ and $\mathbb{R}^3$: a line has dimension 1, whereas a plane has dimension 2 (as long as they contain the origin). Note that there exist infinite-dimensional vector spaces, such as the continuous real-valued functions defined on $[0, 1]$.

The vector space that we use to model a certain problem is usually called the **ambient space** and its dimension the **ambient dimension**. In the case of $\mathbb{R}^n$ the ambient dimension is n.

Lemma B.1.8 (Dimension of $\mathbb{R}^n$)**.** *The dimension of $\mathbb{R}^n$ is n.*

Proof. Consider the set of vectors $\vec{e}_1, \ldots, \vec{e}_n \subseteq \mathbb{R}^n$ defined by

$$\vec{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \ldots \\ 0 \end{bmatrix}, \quad \vec{e}_2 = \begin{bmatrix} 0 \\ 1 \\ \ldots \\ 0 \end{bmatrix}, \quad \ldots, \quad \vec{e}_n = \begin{bmatrix} 0 \\ 0 \\ \ldots \\ 1 \end{bmatrix}. \tag{B.11}$$

One can easily check that this set is a basis. It is in fact the **standard basis** of $\mathbb{R}^n$. □

B.2 Inner product and norm

Up to now, the only operations we have considered are addition and multiplication by a scalar. In this section, we introduce a third operation, the **inner product** between two vectors.

Definition B.2.1 (Inner product)**.** *An inner product on a vector space $\mathcal{V}$ is an operation $\langle \cdot, \cdot \rangle$ that maps pairs of vectors to $\mathbb{R}$ and satisfies the following conditions.*

- *It is symmetric, for any $\vec{x}, \vec{y} \in \mathcal{V}$*

$$\langle \vec{x}, \vec{y} \rangle = \langle \vec{y}, \vec{x} \rangle . \tag{B.12}$$

- *It is linear, i.e. for any $\alpha \in \mathbb{R}$ and any $\vec{x}, \vec{y}, \vec{z} \in \mathcal{V}$*

$$\langle \alpha \, \vec{x}, \vec{y} \rangle = \alpha \, \langle \vec{y}, \vec{x} \rangle , \tag{B.13}$$

$$\langle \vec{x} + \vec{y}, \vec{z} \rangle = \langle \vec{x}, \vec{z} \rangle + \langle \vec{y}, \vec{z} \rangle . \tag{B.14}$$

- *It is positive semidefinite: $\langle \vec{x}, \vec{x} \rangle$ is nonnegative for all $\vec{x} \in \mathcal{V}$ and if $\langle \vec{x}, \vec{x} \rangle = 0$ then $\vec{x} = 0$.*

A vector space endowed with an inner product is called an **inner-product space**. An important instance of an inner product is the **dot product** between two vectors $\vec{x}, \vec{y} \in \mathbb{R}^n$ as

$$\vec{x} \cdot \vec{y} := \sum_i \vec{x}\,[i] \; \vec{y}\,[i] , \tag{B.15}$$

where $\vec{x}\,[i]$ is the ith entry of $\vec{x}$. In this section we use $\vec{x}_i$ to denote a vector, but in some other parts of the notes it may also denote an entry of a vector $\vec{x}$; this will be clear from the context. It is easy to check that the dot product is a valid inner product. $\mathbb{R}^n$ endowed with the dot product is usually called a Euclidean space of dimension n.

The **norm** of a vector is a generalization of the concept of *length*.

Definition B.2.2 (Norm)**.** *Let $\mathcal{V}$ be a vector space, a norm is a function $||\cdot||$ from $\mathcal{V}$ to $\mathbb{R}$ that satisfies the following conditions.*

- *It is homogeneous. For all $\alpha \in \mathbb{R}$ and $\vec{x} \in \mathcal{V}$*

$$||\alpha \, \vec{x}|| = |\alpha| \, ||\vec{x}|| . \tag{B.16}$$

- *It satisfies the* ***triangle inequality***

$$||\vec{x} + \vec{y}|| \leq ||\vec{x}|| + ||\vec{y}|| . \tag{B.17}$$

  In particular, it is nonnegative (set $\vec{y} = -\vec{x}$).

- *$||\vec{x}|| = 0$ implies that $\vec{x}$ is the zero vector $\vec{0}$.*

A vector space equipped with a norm is called a normed space. Distances in a normed space can be measured using the norm of the difference between vectors.

Definition B.2.3 (Distance)**.** *The distance between two vectors $\vec{x}$ and $\vec{y}$ in a normed space with norm $||\cdot||$ is*

$$d\left(\vec{x}, \vec{y}\right) := ||\vec{x} - \vec{y}||\,. \tag{B.18}$$

Inner-product spaces are normed spaces because we can define a valid norm using the inner product. The norm **induced** by an inner product is obtained by taking the square root of the inner product of the vector with itself,

$$||\vec{x}||_{\langle\cdot,\cdot\rangle} := \sqrt{\langle \vec{x}, \vec{x} \rangle}. \tag{B.19}$$

The norm induced by an inner product is clearly homogeneous by linearity and symmetry of the inner product. $||\vec{x}||_{\langle\cdot,\cdot\rangle} = 0$ implies $\vec{x} = 0$ because the inner product is positive semidefinite. We only need to establish that the triangle inequality holds to ensure that the inner-product is a valid norm. This follows from a classic inequality in linear algebra, which is proved in Section B.8.2.

Theorem B.2.4 (Cauchy-Schwarz inequality)**.** *For any two vectors $\vec{x}$ and $\vec{y}$ in an inner-product space*

$$|\langle \vec{x}, \vec{y} \rangle| \leq ||\vec{x}||_{\langle\cdot,\cdot\rangle}\, ||\vec{y}||_{\langle\cdot,\cdot\rangle}\,. \tag{B.20}$$

Assume $||\vec{x}||_{\langle\cdot,\cdot\rangle} \neq 0$,

$$\langle \vec{x}, \vec{y} \rangle = -\,||\vec{x}||_{\langle\cdot,\cdot\rangle}\, ||\vec{y}||_{\langle\cdot,\cdot\rangle} \iff \vec{y} = -\frac{||\vec{y}||_{\langle\cdot,\cdot\rangle}}{||\vec{x}||_{\langle\cdot,\cdot\rangle}} \vec{x}, \tag{B.21}$$

$$\langle \vec{x}, \vec{y} \rangle = ||\vec{x}||_{\langle\cdot,\cdot\rangle}\, ||\vec{y}||_{\langle\cdot,\cdot\rangle} \iff \vec{y} = \frac{||\vec{y}||_{\langle\cdot,\cdot\rangle}}{||\vec{x}||_{\langle\cdot,\cdot\rangle}} \vec{x}. \tag{B.22}$$

Corollary B.2.5. *The norm induced by an inner product satisfies the triangle inequality.*

Proof.

$$\begin{aligned} ||\vec{x} + \vec{y}||^2_{\langle\cdot,\cdot\rangle} &= ||\vec{x}||^2_{\langle\cdot,\cdot\rangle} + ||\vec{y}||^2_{\langle\cdot,\cdot\rangle} + 2\,\langle \vec{x}, \vec{y} \rangle && \text{(B.23)} \\ &\leq ||\vec{x}||^2_{\langle\cdot,\cdot\rangle} + ||\vec{y}||^2_{\langle\cdot,\cdot\rangle} + 2\,||\vec{x}||_{\langle\cdot,\cdot\rangle}\, ||\vec{y}||_{\langle\cdot,\cdot\rangle} \quad \text{by the Cauchy-Schwarz inequality} \\ &= \left(||\vec{x}||_{\langle\cdot,\cdot\rangle} + ||\vec{y}||_{\langle\cdot,\cdot\rangle} \right)^2. && \text{(B.24)} \end{aligned}$$

□

The Euclidean or ℓ_2 norm is the norm induced by the dot product in $\mathbb{R}^n$,

$$||\vec{x}||_2 := \sqrt{\vec{x} \cdot \vec{x}} = \sqrt{\sum_{i=1}^{n} \vec{x}[i]^2}. \tag{B.25}$$

In the case of $\mathbb{R}^2$ or $\mathbb{R}^3$ it is what we usually think of as the length of the vector.

B.3 Orthogonality

An important concept in linear algebra is orthogonality.

Definition B.3.1 (Orthogonality)**.** *Two vectors $\vec{x}$ and $\vec{y}$ are orthogonal if*

$$\langle \vec{x}, \vec{y} \rangle = 0. \tag{B.26}$$

A vector $\vec{x}$ is orthogonal to a set $\mathcal{S}$, if

$$\langle \vec{x}, \vec{s} \rangle = 0, \quad \text{for all } \vec{s} \in \mathcal{S}. \tag{B.27}$$

Two sets of $\mathcal{S}_1, \mathcal{S}_2$ are orthogonal if for any $\vec{x} \in \mathcal{S}_1, \vec{y} \in \mathcal{S}_2$

$$\langle \vec{x}, \vec{y} \rangle = 0. \tag{B.28}$$

The orthogonal complement of a subspace $\mathcal{S}$ is

$$\mathcal{S}^{\perp} := \{ \vec{x} \mid \langle \vec{x}, \vec{y} \rangle = 0 \quad \text{for all } \vec{y} \in \mathcal{S} \}. \tag{B.29}$$

Distances between orthogonal vectors measured in terms of the norm induced by the inner product are easy to compute.

Theorem B.3.2 (Pythagorean theorem)**.** *If $\vec{x}$ and $\vec{y}$ are orthogonal vectors*

$$||\vec{x} + \vec{y}||^2_{\langle \cdot, \cdot \rangle} = ||\vec{x}||^2_{\langle \cdot, \cdot \rangle} + ||\vec{y}||^2_{\langle \cdot, \cdot \rangle}. \tag{B.30}$$

Proof. By linearity of the inner product

$$||\vec{x} + \vec{y}||^2_{\langle \cdot, \cdot \rangle} = ||\vec{x}||^2_{\langle \cdot, \cdot \rangle} + ||\vec{y}||^2_{\langle \cdot, \cdot \rangle} + 2 \langle \vec{x}, \vec{y} \rangle \tag{B.31}$$

$$= ||\vec{x}||^2_{\langle \cdot, \cdot \rangle} + ||\vec{y}||^2_{\langle \cdot, \cdot \rangle}. \tag{B.32}$$

□

If we want to show that a vector is orthogonal to a certain subspace, it is enough to show that it is orthogonal to every vector in a basis of the subspace.

Lemma B.3.3. *Let $\vec{x}$ be a vector and $\mathcal{S}$ a subspace of dimension n. If for any basis $\vec{b}_1, \vec{b}_2, \ldots, \vec{b}_n$ of $\mathcal{S}$,*

$$\left\langle \vec{x}, \vec{b}_i \right\rangle = 0, \quad 1 \leq i \leq n, \tag{B.33}$$

then $\vec{x}$ is orthogonal to $\mathcal{S}$.

Proof. Any vector $v \in \mathcal{S}$ can be represented as $v = \sum_i \alpha^n_{i=1} \vec{b}_i$ for $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, from (B.33)

$$\langle \vec{x}, v \rangle = \left\langle \vec{x}, \sum_i \alpha^n_{i=1} \vec{b}_i \right\rangle = \sum_i \alpha^n_{i=1} \left\langle \vec{x}, \vec{b}_i \right\rangle = 0. \tag{B.34}$$

□

We now introduce orthonormal bases.

Definition B.3.4 (Orthonormal basis)**.** *A basis of mutually orthogonal vectors with norm equal to one is called an* ***orthonormal*** *basis.*

It is very easy to find the coefficients of a vector in an orthonormal basis: we just need to compute the dot products with the basis vectors.

Lemma B.3.5 (Coefficients in an orthonormal basis)**.** *If* $\{\vec{u}_1, \ldots, \vec{u}_n\}$ *is an orthonormal basis of a vector space* $\mathcal{V}$*, for any vector* $\vec{x} \in \mathcal{V}$

$$\vec{x} = \sum_{i=1}^{n} \langle \vec{u}_i, \vec{x} \rangle \, \vec{u}_i. \tag{B.35}$$

Proof. Since $\{\vec{u}_1, \ldots, \vec{u}_n\}$ is a basis,

$$\vec{x} = \sum_{i=1}^{m} \alpha_i \, \vec{u}_i \quad \text{for some } \alpha_1, \alpha_2, \ldots, \alpha_m \in \mathbb{R}. \tag{B.36}$$

Immediately,

$$\langle \vec{u}_i, \vec{x} \rangle = \left\langle \vec{u}_i, \sum_{i=1}^{m} \alpha_i \, \vec{u}_i \right\rangle = \sum_{i=1}^{m} \alpha_i \, \langle \vec{u}_i, \vec{u}_i \rangle = \alpha_i \tag{B.37}$$

because $\langle \vec{u}_i, \vec{u}_i \rangle = 1$ and $\langle \vec{u}_i, \vec{u}_j \rangle = 0$ for $i \neq j$. □

For *any* subspace of $\mathbb{R}^n$ we can obtain an orthonormal basis by applying the Gram-Schmidt method to a set of linearly independent vectors spanning the subspace.

Algorithm B.3.6 (Gram-Schmidt)**.** *Consider a set of linearly independent vectors* $\vec{x}_1, \ldots, \vec{x}_m$ *in* $\mathbb{R}^n$*. To obtain an orthonormal basis of the span of these vectors we:*

1. *Set* $\vec{u}_1 := \vec{x}_1 / \left|\left|\vec{x}_1\right|\right|_2$.
2. *For* $i = 1, \ldots, m$*, compute*

$$\vec{v}_i := \vec{x}_i - \sum_{j=1}^{i-1} \langle \vec{u}_j, \vec{x}_i \rangle \, \vec{u}_j. \tag{B.38}$$

and set $\vec{u}_i := \vec{v}_i / \left|\left|\vec{v}_i\right|\right|_2$.

It is not difficult to show that the resulting set of vectors $\vec{u}_1, \ldots, \vec{u}_m$ is an orthonormal basis for the span of $\vec{x}_1, \ldots, \vec{x}_m$. This implies in particular that we can always assume that a subspace has an orthonormal basis.

Theorem B.3.7. *Every finite-dimensional vector space has an orthonormal basis.*

Proof. To see that the Gram-Schmidt method produces an orthonormal basis for the span of the input vectors we can check that $\text{span}\,(\vec{x}_1, \ldots, \vec{x}_i) = \text{span}\,(\vec{u}_1, \ldots, \vec{u}_i)$ and that $\vec{u}_1, \ldots, \vec{u}_i$ is set of orthonormal vectors. □

B.4 Projections

The projection of a vector $\vec{x}$ onto a subspace $\mathcal{S}$ is the vector in $\mathcal{S}$ that is closest to $\vec{x}$. In order to define this rigorously, we start by introducing the concept of direct sum. If two subspaces are disjoint, i.e. their only common point is the origin, then a vector that can be written as a sum of a vector from each subspace is said to belong to their direct sum.

Definition B.4.1 (Direct sum)**.** *Let $\mathcal{V}$ be a vector space. For any subspaces $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{V}$ such that*

$$\mathcal{S}_1 \cap \mathcal{S}_2 = \{0\} \tag{B.39}$$

the direct sum is defined as

$$\mathcal{S}_1 \oplus \mathcal{S}_2 := \{\vec{x} \mid \vec{x} = \vec{s}_1 + \vec{s}_2 \quad \vec{s}_1 \in \mathcal{S}_1, \vec{s}_2 \in \mathcal{S}_2\}\,. \tag{B.40}$$

The representation of a vector in the direct sum of two subspaces is unique.

Lemma B.4.2. *Any vector $\vec{x} \in \mathcal{S}_1 \oplus \mathcal{S}_2$ has a **unique** representation*

$$\vec{x} = \vec{s}_1 + \vec{s}_2 \quad \vec{s}_1 \in \mathcal{S}_1, \vec{s}_2 \in \mathcal{S}_2. \tag{B.41}$$

Proof. If $\vec{x} \in \mathcal{S}_1 \oplus \mathcal{S}_2$ then by definition there exist $\vec{s}_1 \in \mathcal{S}_1, \vec{s}_2 \in \mathcal{S}_2$ such that $\vec{x} = \vec{s}_1 + \vec{s}_2$. Assume $\vec{x} = \vec{v}_1 + \vec{v}_2$, $\vec{v}_1 \in \mathcal{S}_1, \vec{v}_2 \in \mathcal{S}_2$, then $\vec{s}_1 - \vec{v}_1 = \vec{s}_2 - \vec{v}_2$. This implies that $\vec{s}_1 - \vec{v}_1$ and $\vec{s}_2 - \vec{v}_2$ are in $\mathcal{S}_1$ and also in $\mathcal{S}_2$. However, $\mathcal{S}_1 \cap \mathcal{S}_2 = \{0\}$, so we conclude $\vec{s}_1 = \vec{v}_1$ and $\vec{s}_2 = \vec{v}_2$. □

We can now define the projection of a vector $\vec{x}$ onto a subspace $\mathcal{S}$ by separating the vector into a component that belongs to $\mathcal{S}$ and another that belongs to its orthogonal complement.

Definition B.4.3 (Orthogonal projection)**.** *Let $\mathcal{V}$ be a vector space. The orthogonal projection of a vector $\vec{x} \in \mathcal{V}$ onto a subspace $\mathcal{S} \subseteq \mathcal{V}$ is a vector denoted by $\mathcal{P}_{\mathcal{S}}\,\vec{x}$ such that $\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{x} \in \mathcal{S}^{\perp}$.*

Theorem B.4.4 (Properties of orthogonal projections)**.** *Let $\mathcal{V}$ be a vector space. Every vector $\vec{x} \in \mathcal{V}$ has a **unique** orthogonal projection $\mathcal{P}_{\mathcal{S}}\,\vec{x}$ onto any subspace $\mathcal{S} \subseteq \mathcal{V}$ of finite dimension. In particular $\vec{x}$ can be expressed as*

$$\vec{x} = \mathcal{P}_{\mathcal{S}}\,\vec{x} + \mathcal{P}_{\mathcal{S}^{\perp}}\,\vec{x}. \tag{B.42}$$

For any vector $\vec{s} \in \mathcal{S}$

$$\langle \vec{x}, \vec{s} \rangle = \langle \mathcal{P}_{\mathcal{S}}\,\vec{x}, \vec{s} \rangle\,. \tag{B.43}$$

For any orthonormal basis $\vec{b}_1, \ldots, \vec{b}_m$ of $\mathcal{S}$,

$$\mathcal{P}_{\mathcal{S}}\,\vec{x} = \sum_{i=1}^{m} \left\langle \vec{x}, \vec{b}_i \right\rangle \vec{b}_i. \tag{B.44}$$

Proof. Let us denote the dimension of $\mathcal{S}$ by m. Since m is finite, the exists an orthonormal basis $\mathcal{S}$: $\vec{b}'_1, \ldots, \vec{b}'_m$. Consider the vector

$$\vec{p} := \sum_{i=1}^{m} \left\langle \vec{x}, \vec{b}'_i \right\rangle \vec{b}'_i. \tag{B.45}$$

It turns out that $\vec{x} - \vec{p}$ is orthogonal to every vector in the basis. For $1 \leq j \leq m$,

$$\left\langle \vec{x} - \vec{p}, \vec{b}'_j \right\rangle = \left\langle \vec{x} - \sum_{i=1}^{m} \left\langle \vec{x}, \vec{b}'_i \right\rangle \vec{b}'_i, \vec{b}'_j \right\rangle \tag{B.46}$$

$$= \left\langle \vec{x}, \vec{b}'_j \right\rangle - \sum_{i=1}^{m} \left\langle \vec{x}, \vec{b}'_i \right\rangle \left\langle \vec{b}'_i, \vec{b}'_j \right\rangle \tag{B.47}$$

$$= \left\langle \vec{x}, \vec{b}'_j \right\rangle - \left\langle \vec{x}, \vec{b}'_j \right\rangle = 0, \tag{B.48}$$

so $\vec{x} - \vec{p} \in \mathcal{S}^\perp$ and $\vec{p}$ is an orthogonal projection. Since $\mathcal{S} \cap \mathcal{S}^\perp = \{0\}$ [1] there cannot be two other vectors $\vec{x}_1 \in \mathcal{S}, \vec{x}_1 \in \mathcal{S}^\perp$ such that $\vec{x} = \vec{x}_1 + \vec{x}_2$ so the orthogonal projection is unique.

Notice that $\vec{o} := \vec{x} - \vec{p}$ is a vector in $\mathcal{S}^\perp$ such that $\vec{x} - \vec{o} = \vec{p}$ is in S and therefore in $\left(\mathcal{S}^\perp\right)^\perp$. This implies that $\vec{o}$ is the orthogonal projection of $\vec{x}$ onto $\mathcal{S}^\perp$ and establishes (B.42).

Equation (B.43) follows immediately from the orthogonality of any vector $\vec{s} \in \mathcal{S}$ and $\mathcal{P}_\mathcal{S}\, \vec{x}$.

Equation (B.44) follows from (B.43). □

Computing the norm of the projection of a vector onto a subspace is easy if we have access to an orthonormal basis (as long as the norm is induced by the inner product).

Lemma B.4.5 (Norm of the projection)**.** *The norm of the projection of an arbitrary vector $\vec{x} \in \mathcal{V}$ onto a subspace $\mathcal{S} \subseteq \mathcal{V}$ of dimension d can be written as*

$$||\mathcal{P}_\mathcal{S}\, \vec{x}||_{\langle \cdot, \cdot \rangle} = \sqrt{\sum_{i}^{d} \left\langle \vec{b}_i, \vec{x} \right\rangle^2} \tag{B.49}$$

for any orthonormal basis $\vec{b}_1, \ldots, \vec{b}_d$ of $\mathcal{S}$.

Proof. By (B.44)

$$||\mathcal{P}_\mathcal{S}\, \vec{x}||^2_{\langle \cdot, \cdot \rangle} = \langle \mathcal{P}_\mathcal{S}\, \vec{x}, \mathcal{P}_\mathcal{S}\, \vec{x} \rangle \tag{B.50}$$

$$= \left\langle \sum_{i}^{d} \left\langle \vec{b}_i, \vec{x} \right\rangle \vec{b}_i, \sum_{j}^{d} \left\langle \vec{b}_j, \vec{x} \right\rangle \vec{b}_j \right\rangle \tag{B.51}$$

$$= \sum_{i}^{d} \sum_{j}^{d} \left\langle \vec{b}_i, \vec{x} \right\rangle \left\langle \vec{b}_j, \vec{x} \right\rangle \left\langle \vec{b}_i, \vec{b}_j \right\rangle \tag{B.52}$$

$$= \sum_{i}^{d} \left\langle \vec{b}_i, \vec{x} \right\rangle^2. \tag{B.53}$$

□

[1] For any vector $\vec{v}$ that belongs to both $\mathcal{S}$ and $\mathcal{S}^\perp$ $\langle \vec{v}, \vec{v} \rangle = ||\vec{v}||_2^2 = 0$, which implies $\vec{v} = 0$.

Example B.4.6 (Projection onto a one-dimensional subspace)**.** To compute the projection of a vector $\vec{x}$ onto a one-dimensional subspace spanned by a vector $\vec{v}$, we use the fact that $\left\{\vec{v}/\left|\left|\vec{v}\right|\right|_{\langle\cdot,\cdot\rangle}\right\}$ is a basis for span $(\vec{v})$ (it is a set containing a unit vector that spans the subspace) and apply (B.44) to obtain

$$\mathcal{P}_{\text{span}(\vec{v})}\,\vec{x} = \frac{\langle\vec{v},\vec{x}\rangle}{\left|\left|\vec{v}\right|\right|^2_{\langle\cdot,\cdot\rangle}}\vec{v}. \tag{B.54}$$

△

Finally, we prove that the projection of a vector $\vec{x}$ onto a subspace $\mathcal{S}$ is indeed the vector in $\mathcal{S}$ that is closest to $\vec{x}$ in the distance induced by the inner-product norm.

Theorem B.4.7 (The orthogonal projection is closest)**.** *The orthogonal projection of a vector $\vec{x}$ onto a subspace $\mathcal{S}$ belonging to the same inner-product space is the closest vector to $\vec{x}$ that belongs to $\mathcal{S}$ in terms of the norm induced by the inner product. More formally, $\mathcal{P}_{\mathcal{S}}\,\vec{x}$ is the solution to the optimization problem*

$$\underset{\vec{u}}{\text{minimize}} \qquad \left|\left|\vec{x}-\vec{u}\right|\right|_{\langle\cdot,\cdot\rangle} \tag{B.55}$$

$$\text{subject to} \qquad \vec{u}\in\mathcal{S}. \tag{B.56}$$

Proof. Take any point $\vec{s}\in\mathcal{S}$ such that $\vec{s}\neq\mathcal{P}_{\mathcal{S}}\,\vec{x}$

$$\begin{aligned}
\left|\left|\vec{x}-\vec{s}\right|\right|^2_{\langle\cdot,\cdot\rangle} &= \left|\left|x-\mathcal{P}_{\mathcal{S}}\,\vec{x}+\mathcal{P}_{\mathcal{S}}\,\vec{x}-\vec{s}\right|\right|^2_{\langle\cdot,\cdot\rangle} && \text{(B.57)}\\
&= \left|\left|x-\mathcal{P}_{\mathcal{S}}\,\vec{x}\right|\right|^2_{\langle\cdot,\cdot\rangle}+\left|\left|\mathcal{P}_{\mathcal{S}}\,\vec{x}-\vec{s}\right|\right|^2_{\langle\cdot,\cdot\rangle} && \text{(B.58)}\\
&> \left|\left|x-\mathcal{P}_{\mathcal{S}}\,\vec{x}\right|\right|^2_{\langle\cdot,\cdot\rangle} \quad \text{because } \vec{s}\neq\mathcal{P}_{\mathcal{S}}\,\vec{x}, && \text{(B.59)}
\end{aligned}$$

where (B.58) follows from the Pythagorean theorem since because $\mathcal{P}_{\mathcal{S}^\perp}\,\vec{x} := x-\mathcal{P}_{\mathcal{S}}\,\vec{x}$ belongs to $S^\perp$ and $\mathcal{P}_{\mathcal{S}}\,\vec{x}-\vec{s}$ to S. □

B.5 Matrices

A **matrix** is a rectangular array of numbers. We denote the vector space of $m\times n$ matrices by $\mathbb{R}^{m\times n}$. We denote the ith row of a matrix A by $A_{i:}$, the jth column by $A_{:j}$ and the (i,j) entry by A_{ij}. The transpose of a matrix is obtained by switching its rows and columns.

Definition B.5.1 (Transpose)**.** *The **transpose** A^T of a matrix $A\in\mathbb{R}^{m\times n}$ is a matrix in $A\in\mathbb{R}^{m\times n}$*

$$\left(A^T\right)_{ij} = A_{ji}. \tag{B.60}$$

A **symmetric** matrix is a matrix that is equal to its transpose.

Matrices map vectors to other vectors through a linear operation called matrix-vector product.

Definition B.5.2 (Matrix-vector product)**.** *The product of a matrix $A \in \mathbb{R}^{m\times n}$ and a vector $\vec{x} \in \mathbb{R}^n$ is a vector $A\vec{x} \in \mathbb{R}^n$, such that*

$$(A\vec{x})_i = \sum_{j=1}^{n} A_{ij}\vec{x}\,[j] \tag{B.61}$$

$$= \langle A_{i:}, \vec{x}\rangle\,, \tag{B.62}$$

i.e. the ith entry of $A\vec{x}$ is the dot product between the ith row of A and $\vec{x}$.

Equivalently,

$$A\vec{x} = \sum_{j=1}^{n} A_{:j}\vec{x}\,[j]\,, \tag{B.63}$$

i.e. $A\vec{x}$ is a linear combination of the columns of A weighted by the entries in $\vec{x}$.

One can easily check that the transpose of the product of two matrices A and B is equal to the transposes multiplied in the inverse order,

$$(AB)^T = B^T A^T. \tag{B.64}$$

We can express the dot product between two vectors $\vec{x}$ and $\vec{y}$ as

$$\langle \vec{x}, \vec{y}\rangle = \vec{x}^T\,\vec{y} = \vec{y}^T\,\vec{x}. \tag{B.65}$$

The identity matrix is a matrix that maps any vector to itself.

Definition B.5.3 (Identity matrix)**.** *The identity matrix in $\mathbb{R}^{n\times n}$ is*

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{B.66}$$

Clearly, for any $\vec{x} \in \mathbb{R}^n$ $I\vec{x} = \vec{x}$.

Definition B.5.4 (Matrix multiplication)**.** *The product of two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$ is a matrix $AB \in \mathbb{R}^{m\times p}$, such that*

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{k_j} = \langle A_{i:}, B_{:,j}\rangle\,, \tag{B.67}$$

i.e. the (i, j) entry of AB is the dot product between the ith row of A and the jth column of B.

Equivalently, the jth column of AB is the result of multiplying A and the jth column of B

$$AB = \sum_{k=1}^{n} A_{ik}B_{k_j} = \langle A_{i:}, B_{:,j}\rangle\,, \tag{B.68}$$

and ith row of AB is the result of multiplying the ith row of A and B.

Square matrices may have an inverse. If they do, the inverse is a matrix that reverses the effect of the matrix of any vector.

Definition B.5.5 (Matrix inverse)**.** *The inverse of a square matrix $A \in \mathbb{R}^{n \times n}$ is a matrix $A^{-1} \in \mathbb{R}^{n \times n}$ such that*

$$AA^{-1} = A^{-1}A = \mathrm{I}. \tag{B.69}$$

Lemma B.5.6. *The inverse of a matrix is unique.*

Proof. Let us assume there is another matrix M such that $AM = \mathrm{I}$, then

$$M = A^{-1}AM \quad \text{by (B.69)} \tag{B.70}$$
$$= A^{-1}. \tag{B.71}$$

□

An important class of matrices are **orthogonal matrices**.

Definition B.5.7 (Orthogonal matrix)**.** *An orthogonal matrix is a square matrix such that its inverse is equal to its transpose,*

$$U^T U = UU^T = \mathrm{I} \tag{B.72}$$

By definition, the columns $U_{:1}, U_{:2}, \ldots, U_{:n}$ of any orthogonal matrix have unit norm and orthogonal to each other, so they form an orthonormal basis (it's somewhat confusing that orthogonal matrices are not called orthonormal matrices instead). We can interpret applying U^T to a vector $\vec{x}$ as computing the coefficients of its representation in the basis formed by the columns of U. Applying U to $U^T\vec{x}$ recovers $\vec{x}$ by scaling each basis vector with the corresponding coefficient:

$$\vec{x} = UU^T\vec{x} = \sum_{i=1}^{n} \langle U_{:i}, \vec{x} \rangle \, U_{:i}. \tag{B.73}$$

Applying an orthogonal matrix to a vector does not affect its norm, it just rotates the vector.

Lemma B.5.8 (Orthogonal matrices preserve the norm)**.** *For any orthogonal matrix $U \in \mathbb{R}^{n \times n}$ and any vector $\vec{x} \in \mathbb{R}^n$,*

$$||U\vec{x}||_2 = ||\vec{x}||_2 \,. \tag{B.74}$$

Proof. By the definition of an orthogonal matrix

$$||U\vec{x}||_2^2 = \vec{x}^T U^T U \vec{x} \tag{B.75}$$
$$= \vec{x}^T \vec{x} \tag{B.76}$$
$$= ||\vec{x}||_2^2 \,. \tag{B.77}$$

□

B.6 Eigendecomposition

An **eigenvector** $\vec{v}$ of a matrix A satisfies

$$A\vec{v} = \lambda\vec{v} \tag{B.78}$$

for a scalar λ which is the corresponding **eigenvalue**. Even if A is real, its eigenvectors and eigenvalues can be complex.

Lemma B.6.1 (Eigendecomposition)**.** *If a square matrix $A \in \mathbb{R}^{n\times n}$ has n linearly independent eigenvectors $\vec{v}_1, \ldots, \vec{v}_n$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ it can be expressed in terms of a matrix Q, whose columns are the eigenvectors, and a diagonal matrix containing the eigenvalues,*

$$A = \begin{bmatrix} \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \end{bmatrix}^{-1} \tag{B.79}$$

$$= Q\Lambda Q^{-1} \tag{B.80}$$

Proof.

$$AQ = \begin{bmatrix} A\vec{v}_1 & A\vec{v}_2 & \cdots & A\vec{v}_n \end{bmatrix} \tag{B.81}$$

$$= \begin{bmatrix} \lambda_1\vec{v}_1 & \lambda_2\vec{v}_2 & \cdots & \lambda_2\vec{v}_n \end{bmatrix} \tag{B.82}$$

$$= Q\Lambda. \tag{B.83}$$

If the columns of a square matrix are all linearly independent, then the matrix has an inverse, so multiplying the expression by Q^{-1} on both sides completes the proof. □

Lemma B.6.2. *Not all matrices have an eigendecomposition*

Proof. Consider for example the matrix

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}. \tag{B.84}$$

Assume λ has a nonzero eigenvalue corresponding to an eigenvector with entries $\vec{v}[1]$ and $\vec{v}[2]$, then

$$\begin{bmatrix} \vec{v}[2] \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \vec{v}[1] \\ \vec{v}[2] \end{bmatrix} = \begin{bmatrix} \lambda\vec{v}[2] \\ \lambda\vec{v}[2] \end{bmatrix}, \tag{B.85}$$

which implies that $\vec{v}[2] = 0$ and hence $\vec{v}[1] = 0$, since we have assumed that $\lambda \neq 0$. This implies that the matrix does not have nonzero eigenvalues associated to nonzero eigenvectors. □

An interesting use of the eigendecomposition is computing successive matrix products very fast. Assume that we want to compute

$$AA\cdots A\vec{x} = A^k\vec{x}, \tag{B.86}$$

i.e. we want to apply A to $\vec{x}$ k times. A^k *cannot* be computed by taking the power of its entries (try out a simple example to convince yourself). However, if A has an eigendecomposition,

$$\begin{aligned} A^k &= Q\Lambda Q^{-1}Q\Lambda Q^{-1}\cdots Q\Lambda Q^{-1} && \text{(B.87)}\\ &= Q\Lambda^k Q^{-1} && \text{(B.88)}\\ &= Q\begin{bmatrix} \lambda_1^k & 0 & \cdots & 0 \\ 0 & \lambda_2^k & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n^k \end{bmatrix} Q^{-1}, && \text{(B.89)} \end{aligned}$$

using the fact that for diagonal matrices applying the matrix repeatedly is equivalent to taking the power of the diagonal entries. This allows to compute the k matrix products using just 3 matrix products and taking the power of n numbers.

From high-school or undergraduate algebra you probably remember how to compute eigenvectors using determinants. In practice, this is usually not a viable option due to stability issues. A popular technique to compute eigenvectors is based on the following insight. Let $A \in \mathbb{R}^{n\times n}$ be a matrix with eigendecomposition $Q\Lambda Q^{-1}$ and let $\vec{x}$ be an arbitrary vector in $\mathbb{R}^n$. Since the columns of Q are linearly independent, they form a basis for $\mathbb{R}^n$, so we can represent $\vec{x}$ as

$$\vec{x} = \sum_{i=1}^{n} \alpha_i Q_{:i}, \quad \alpha_i \in \mathbb{R},\ 1 \leq i \leq n. \tag{B.90}$$

Now let us apply A to $\vec{x}$ k times,

$$\begin{aligned} A^k\vec{x} &= \sum_{i=1}^{n} \alpha_i A^k Q_{:i} && \text{(B.91)}\\ &= \sum_{i=1}^{n} \alpha_i \lambda_i^k Q_{:i}. && \text{(B.92)} \end{aligned}$$

If we assume that the eigenvectors are ordered according to their magnitudes and that the magnitude of one of them is larger than the rest, $|\lambda_1| > |\lambda_2| \geq \ldots$, and that $\alpha_1 \neq 0$ (which happens with high probability if we draw a random $\vec{x}$) then as k grows larger the term $\alpha_1\lambda_1^k Q_{:1}$ dominates. The term will blow up or tend to zero unless we normalize every time before applying A. Adding the normalization step to this procedure results in the **power method** or power iteration, an algorithm of great importance in numerical linear algebra.

Algorithm B.6.3 (Power method)**.**
Input*: A matrix A.*
Output*: An estimate of the eigenvector of A corresponding to the largest eigenvalue.*

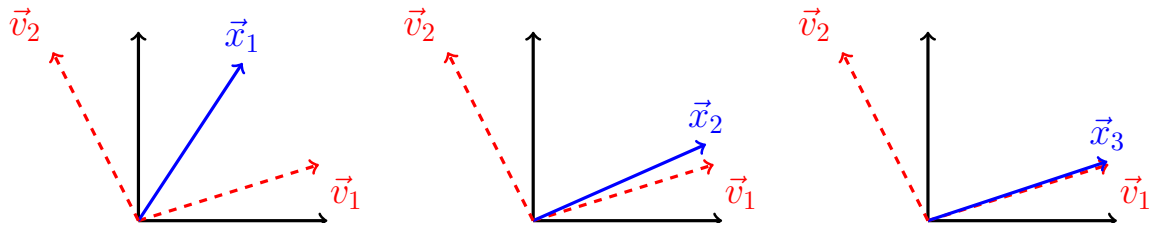


Figure B.1: Illustration of the first three iterations of the power method for a matrix with eigenvectors $\vec{v}_1$ and $\vec{v}_2$, whose corresponding eigenvalues are $\lambda_1 = 1.05$ and $\lambda_2 = 0.1661$.

Initialization*: Set* $\vec{x}_1 := \vec{x}/\left|\left|\vec{x}\right|\right|_2$*, where the entries of* $\vec{x}$ *are drawn at random.*
For $i = 1, \ldots, k$*, compute*

$$\vec{x}_i := \frac{A\vec{x}_{i-1}}{||A\vec{x}_{i-1}||_2}. \tag{B.93}$$

Figure B.1 illustrates the power method on a simple example, where the matrix is equal to

$$A = \begin{bmatrix} 0.930 & 0.388 \\ 0.237 & 0.286 \end{bmatrix}. \tag{B.94}$$

The convergence to the eigenvector corresponding to the eigenvalue with the largest magnitude is very fast.

B.7 Eigendecomposition of symmetric matrices

Real symmetric matrices always have an eigendecomposition. In addition, their eigenvalues are real and their eigenvectors are all orthogonal.

Theorem B.7.1 (Spectral theorem for real symmetric matrices)**.** *If* $A \in \mathbb{R}^{n \times n}$ *is symmetric, then it has an eigendecomposition of the form*

$$A = \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n \end{bmatrix}^T, \tag{B.95}$$

where the eigenvalues $\vec{u}_1$*,* $\vec{u}_2$*, . . . ,* $\vec{u}_n$ *are real and the eigenvectors* $\vec{u}_1$*,* $\vec{u}_2$*, . . . ,* $\vec{u}_n$ *are real and orthogonal.*

Proof. The proof that every real symmetric matrix has n eigenvectors is beyond the scope of these notes. Under the assumption that this is the case, we begin by proving that the eigenvalues are real. Consider an arbitrary eigenvalue λ_i and the corresponding normalized eigenvector $\vec{v}_i$, we have

$$\vec{v}_i^* A \vec{v}_i = \lambda \vec{v}_i^* \vec{v}_i = \lambda, \tag{B.96}$$

$$\vec{v}_i^* A \vec{v}_i = \left(A\vec{v}_i\right)^* \vec{v}_i = \left(\lambda \vec{v}_i\right)^* = \bar{\lambda} \vec{v}_i^* \vec{v}_i = \bar{\lambda}. \tag{B.97}$$

This implies that λ is real because $\lambda = \bar{\lambda}$, so we can restrict the eigenvectors to be real (since the eigenvalue is real, both the real and imaginary parts of the eigenvector are eigenvectors themselves and at least one of them must be nonzero). If several linearly independent eigenvectors have the same eigenvalue, an orthonormal basis of their span will also consist of eigenvectors of the matrix. All that is left to prove is that eigenvectors corresponding to different eigenvalues are orthogonal. Assume $\vec{v}_i$ and $\vec{v}_j$ are eigenvectors corresponding to different eigenvalues $\lambda_i \neq \lambda_j$, then

$$\vec{u}_i^T \vec{u}_j = \frac{1}{\lambda_i} \left(A\vec{u}_i\right)^T \vec{u}_j \tag{B.98}$$

$$= \frac{1}{\lambda_i} \vec{u}_i^T A^T \vec{u}_j \tag{B.99}$$

$$= \frac{1}{\lambda_i} \vec{u}_i^T A \vec{u}_j \tag{B.100}$$

$$= \frac{\lambda_j}{\lambda_i} \vec{u}_i^T \vec{u}_j. \tag{B.101}$$

This is only possible if $\vec{u}_i^T \vec{u}_j = 0$. □

The eigenvalues of a symmetric matrix determine the value of the **quadratic form**:

$$q\left(\vec{x}\right) := \vec{x}^T A \vec{x} = \sum_{i=1}^{n} \lambda_i \left(\vec{x}^T \vec{u}_i\right)^2 \tag{B.102}$$

If we order the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ then the first eigenvalue is the maximum value attained by the quadratic if its input has unit ℓ_2 norm, the second eigenvalue is the maximum value attained by the quadratic form if we restrict its argument to be normalized and orthogonal to the first eigenvector, and so on.

Theorem B.7.2. *For any symmetric matrix $A \in \mathbb{R}^n$ with normalized eigenvectors $\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_n$ with corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$*

$$\lambda_1 = \max_{||\vec{u}||_2 = 1} \vec{u}^T A \vec{u}, \tag{B.103}$$

$$\vec{u}_1 = \arg \max_{||\vec{u}||_2 = 1} \vec{u}^T A \vec{u}, \tag{B.104}$$

$$\lambda_k = \max_{||\vec{u}||_2 = 1, \vec{u} \perp \vec{u}_1, \ldots, \vec{u}_{k-1}} \vec{u}^T A \vec{u}, \tag{B.105}$$

$$\vec{u}_k = \arg \max_{||\vec{u}||_2 = 1, \vec{u} \perp \vec{u}_1, \ldots, \vec{u}_{k-1}} \vec{u}^T A \vec{u}. \tag{B.106}$$

Proof. The eigenvectors are an orthonormal basis (they are mutually orthogonal and we assume that they have been normalized), so we can represent any unit-norm vector $\vec{h}_k$ that is orthogonal to $\vec{u}_1, \ldots, \vec{u}_{k-1}$ as

$$\vec{h}_k = \sum_{i=k}^{m} \alpha_i \vec{u}_i \tag{B.107}$$

where

$$\left|\left|\vec{h}_k\right|\right|_2^2 = \sum_{i=k}^{m} \alpha_i^2 = 1, \tag{B.108}$$

by Lemma B.4.5. Note that $\vec{h}_1$ is just an arbitrary unit-norm vector.

Now we will show that the value of the quadratic form when the normalized input is restricted to be orthogonal to $\vec{u}_1, \ldots, \vec{u}_{k-1}$ cannot be larger than λ_k,

$$\vec{h}_k^T A \vec{h}_k = \sum_{i=1}^{n} \lambda_i \left(\sum_{j=k}^{m} \alpha_j \vec{u}_i^T \vec{u}_j \right)^2 \quad \text{by (B.102) and (B.107)} \tag{B.109}$$

$$= \sum_{i=1}^{n} \lambda_i \alpha_i^2 \quad \text{because } \vec{u}_1, \ldots, \vec{u}_m \text{ is an orthonormal basis} \tag{B.110}$$

$$\leq \lambda_k \sum_{i=k}^{m} \alpha_i^2 \quad \text{because } \lambda_k \geq \lambda_{k+1} \geq \ldots \geq \lambda_m \tag{B.111}$$

$$= \lambda_k, \quad \text{by (B.108)}. \tag{B.112}$$

This establishes (B.103) and (B.105). To prove (B.104) and (B.106) we just need to show that $\vec{u}_k$ achieves the maximum

$$\vec{u}_k^T A \vec{u}_k = \sum_{i=1}^{n} \lambda_i \left(\vec{u}_i^T \vec{u}_k \right)^2 \tag{B.113}$$

$$= \lambda_k. \tag{B.114}$$

□

B.8 Proofs

B.8.1 Proof of Theorem B.1.6

We prove the claim by contradiction. Assume that we have two bases $\{\vec{x}_1, \ldots, \vec{x}_m\}$ and $\{\vec{y}_1, \ldots, \vec{y}_n\}$ such that $m < n$ (or the second set has infinite cardinality). The proof follows from applying the following lemma m times (setting $r = 0, 1, \ldots, m-1$) to show that $\{\vec{y}_1, \ldots, \vec{y}_m\}$ spans $\mathcal{V}$ and hence $\{\vec{y}_1, \ldots, \vec{y}_n\}$ must be linearly dependent.

Lemma B.8.1. *Under the assumptions of the theorem, if $\{\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_r, \vec{x}_{r+1}, \ldots, \vec{x}_m\}$ spans $\mathcal{V}$ then $\{\vec{y}_1, \ldots, \vec{y}_{r+1}, \vec{x}_{r+2}, \ldots, \vec{x}_m\}$ also spans $\mathcal{V}$ (possibly after rearranging the indices $r+1, \ldots, m$) for $r = 0, 1, \ldots, m-1$.*

Proof. Since $\{\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_r, \vec{x}_{r+1}, \ldots, \vec{x}_m\}$ spans $\mathcal{V}$

$$\vec{y}_{r+1} = \sum_{i=1}^{r} \beta_i \vec{y}_i + \sum_{i=r+1}^{m} \gamma_i \vec{x}_i, \quad \beta_1, \ldots, \beta_r, \gamma_{r+1}, \ldots, \gamma_m \in \mathbb{R}, \tag{B.115}$$

where at least one of the γ_j is non zero, as $\{\vec{y}_1, \ldots, \vec{y}_n\}$ is linearly independent by assumption. Without loss of generality (here is where we might need to rearrange the indices) we assume that $\gamma_{r+1} \neq 0$, so that

$$\vec{x}_{r+1} = \frac{1}{\gamma_{r+1}} \left(\sum_{i=1}^{r} \beta_i \, \vec{y}_i - \sum_{i=r+2}^{m} \gamma_i \vec{x}_i \right) . \tag{B.116}$$

This implies that any vector in the span of $\{\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_r, \vec{x}_{r+1}, \ldots, \vec{x}_m\}$, i.e. in $\mathcal{V}$, can be represented as a linear combination of vectors in $\{\vec{y}_1, \ldots, \vec{y}_{r+1}, \vec{x}_{r+2}, \ldots, \vec{x}_m\}$, which completes the proof. □

B.8.2 Proof of Theorem B.2.4

If $||\vec{x}||_{\langle\cdot,\cdot\rangle} = 0$ then $\vec{x} = \vec{0}$ because the inner product is positive semidefinite, which implies $\langle \vec{x}, \vec{y} \rangle = 0$ and consequently that (B.20) holds with equality. The same is true if $||\vec{y}||_{\langle\cdot,\cdot\rangle} = 0$.

Now assume that $||\vec{x}||_{\langle\cdot,\cdot\rangle} \neq 0$ and $||\vec{y}||_{\langle\cdot,\cdot\rangle} \neq 0$. By semidefiniteness of the inner product,

$$0 \leq \left|\left| ||\vec{y}||_{\langle\cdot,\cdot\rangle} \, \vec{x} + ||\vec{x}||_{\langle\cdot,\cdot\rangle} \, \vec{y} \right|\right|^2 = 2 \, ||\vec{x}||^2_{\langle\cdot,\cdot\rangle} \, ||\vec{y}||^2_{\langle\cdot,\cdot\rangle} + 2 \, ||\vec{x}||_{\langle\cdot,\cdot\rangle} \, ||\vec{y}||_{\langle\cdot,\cdot\rangle} \, \langle \vec{x}, \vec{y} \rangle \, , \tag{B.117}$$

$$0 \leq \left|\left| ||\vec{y}||_{\langle\cdot,\cdot\rangle} \, \vec{x} - ||\vec{x}||_{\langle\cdot,\cdot\rangle} \, \vec{y} \right|\right|^2 = 2 \, ||\vec{x}||^2_{\langle\cdot,\cdot\rangle} \, ||\vec{y}||^2_{\langle\cdot,\cdot\rangle} - 2 \, ||\vec{x}||_{\langle\cdot,\cdot\rangle} \, ||\vec{y}||_{\langle\cdot,\cdot\rangle} \, \langle \vec{x}, \vec{y} \rangle \, . \tag{B.118}$$

These inequalities establish (B.20).

Let us prove (B.21) by proving both implications.

($\Rightarrow$) Assume $\langle \vec{x}, \vec{y} \rangle = - \, ||\vec{x}||_{\langle\cdot,\cdot\rangle} \, ||\vec{y}||_{\langle\cdot,\cdot\rangle}$. Then (B.117) equals zero, so $||\vec{y}||_{\langle\cdot,\cdot\rangle} \, \vec{x} = - \, ||\vec{x}||_{\langle\cdot,\cdot\rangle} \, \vec{y}$ because the inner product is positive semidefinite.

($\Leftarrow$) Assume $||\vec{y}||_{\langle\cdot,\cdot\rangle} \, \vec{x} = - \, ||\vec{x}||_{\langle\cdot,\cdot\rangle} \, \vec{y}$. Then one can easily check that (B.117) equals zero, which implies $\langle \vec{x}, \vec{y} \rangle = - \, ||\vec{x}||_{\langle\cdot,\cdot\rangle} \, ||\vec{y}||_{\langle\cdot,\cdot\rangle}$.

The proof of (B.22) is identical (using (B.118) instead of (B.117)).