Introduction to Data Science 1MS041

[1], Benny Avelin[1], and Raazesh Sainudiin[1]

[1]Department of Mathematics, Uppsala University, Uppsala, Sweden

September 20, 2022

[1]Partly based on CseBook by lamastex under CC BY NC SA 4.0

Preface

These set of lecture notes began as notes for the course Introduction to Data Science during the spring and summer of 2021. This course was designed as a mathematical introduction to datascience, covering most of the basics one would need to start their education in data science, in the sense of giving the reader a strong mathematical foundation on which to stand in the future. Our belief is that in order to develop new algorithms for data science / AI problems, one needs a strong mathematical intuition.

Another aim with these notes have been to bridge a gap between math, theoretical computer science and modern approaches concerning concentration of measure. Most of this material can be found in other texts, but scattered and with wildly differing levels of rigor.

Another novel point of these notes is that we focus quite a lot on separating the statistical model (assumptions about the data) from the estimation procedure (computer algorithms). This idea is hidden in most of modern data science and often goes against traditional parametric estimation, where the assumption is fairly often that the true underlying parameter one tries to estimate, is among those searched for. For instance, in linear regression, one assumes that the truth is linear and we are just trying to find that. In modern data-science where the goal is one of prediction (mostly) one instead does not assume that the truth is linear but one tries to approximate it with a linear function, and if the fit is good one is happy.

The concentration of measure phenomenon permeates these notes and we will use it to arrive at non-asymptotic estimates (finite sample bounds) in many practically useful cases, from performance metrics of classification to compression of data using dimensionality reduction. The goal has been to provide quantitative estimates for almost all problems in these notes, while some have been left out, they can be approached using the methods developed in these notes.

These notes suit students with little knowledge of probability theory, it is however easier to digest if you are familiar with the mathematical way of thinking, i.e. in that of abstraction.

Topics

- Axiomatic probability: Chapters 1 and 2. These chapters cover the mathematical basics needed for the rest of the notes. We have chosen a fairly rigorous way of presenting axiomatic probability which is very flexible and after you get to know it, very easy to use as there is very little ambiguity.

- Concentration of measure: Chapter 3. This is the main backbone of these notes, all chapters following rely on the results obtained here (except: Chapter 7). We have decided to only touch on the simplest concentration inequalities, i.e. Hoeffding's inequality and similar.

- Risk: Chapter 4. This chapter concerns the concept of Risk and how you can phrase common problems, like regression, pattern recognition and parameter estimation as risk minimization problems. All estimation problems that appear later in these notes will be a risk minimization problem, and specifically empirical risk minimization problems.

- Fundamentals of estimation: Chapter 5. Covers the traditional statistical terminology surrounding parameter estimation. Like consistency, bias or asymptotic properties.

- Random variable Generation: Chapter 6. Introduces the concept of pseudo-randomness, some ways to produce it on the computer, and how to use it to sample from arbitrary distributions.

- Finite Markov Chains: Chapter 7. This chapter introduces Markov chains as a means of modelling more than just i.i.d. samples. It is also where you will see a natural interpretation of the σ-algebra as history. Markov chains are essential in many sequential problems and is the simplest form of time-series.

- Pattern recognition and Regression: Chapters 8 and 9. This chapter covers pattern recognition and regression from the perspective of a-priori perfomance or a-posteriori testing. That is, what can we say about the performance of an algorithm without a test-set and what can we say once we have a test-set? These chapters relies on the fairly advanced topic of VC-dimension and growth functions. The a-posteriori testing is most important for a first course, as well as understanding the difference between guaranteeing performance beforehand or afterwards.

- High dimension and Dimensionality reduction: Chapters 10 and 11. These set of notes end with another look at concentration from the perspective of dimension and utilize this to perform dimensionality

reduction. We also cover singular value decomposition and its use in image compression/data.

Contents

List of Figures

Chapter 1

Probability Model

1.1 Experiments

Ideas about chance events and random behaviour arose out of thousands of years of game playing, long before any attempt was made to use mathematical reasoning about them. Board and dice games were well known in Egyptian times, and Augustus Caesar gambled with dice. Calculations of odds for gamblers were put on a proper theoretical basis by Fermat and Pascal in the early 17th century.

Definition 1.1. *An* **experiment** *is an activity or procedure that produces distinct, well-defined possibilities called* **outcomes**. *The set of all outcomes is called the* **sample space**, *and is denoted by* Ω.

The subsets of Ω *are called* **events**. *A single outcome,* ω, *when seen as a subset of* Ω, *as in* $\{\omega\}$, *is called a* **simple event**.

Given an outcome $\omega \in \Omega$ *we say that the event* $E \subset \Omega$ **occured** *if* $\omega \in E$.

Events, E_1, E_2 ... E_n, *that cannot occur at the same time are called* **mutually exclusive** *events, or* **pair-wise disjoint** *events. This means that* $E_i \cap E_j = \emptyset$ *where* $i \neq j$.

Example 1.2. *Some standard examples of experiments are the following:*

- $\Omega = \{$*Defective, Non-defective*$\}$ *if our experiment is to inspect a light bulb.*

  There are only two outcomes here, so $\Omega = \{\omega_1, \omega_2\}$ *where* $\omega_1 =$ *Defective and* $\omega_2 =$ *Non-defective.*

- $\Omega = \{$*Heads, Tails*$\}$ *if our experiment is to note the outcome of a coin toss.*

  This time, $\Omega = \{\omega_1, \omega_2\}$ *where* $\omega_1 =$ *Heads and* $\omega_2 =$ *Tails.*

- *If our experiment is to roll a die then there are six outcomes corresponding to the number that shows on the top. For this experiment,* $\Omega = \{1, 2, 3, 4, 5, 6\}$.

  Some examples of events are the set of odd numbered outcomes $A = \{1, 3, 5\}$*, and the set of even numbered outcomes* $B = \{2, 4, 6\}$.

  The simple events of Ω *are* $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}$*, and* $\{6\}$.

The outcome of a random experiment is uncertain until it is performed and observed. Note that sample spaces need to reflect the problem in hand.

Definition 1.3. *A* **trial** *is a single performance of an experiment and it results in an outcome.*

Example 1.4. *Some standard examples of a trial are:*

- *A roll of a die.*
- *A toss of a coin.*
- *A release of a chaotic double pendulum.*

An experimenter often performs more than one trial. Repeated trials of an experiment forms the basis of science and engineering as the experimenter learns about the phenomenon by repeatedly performing the same mother experiment with possibly different outcomes. This repetition of trials in fact provides the very motivation for the definition of probability.

Definition 1.5. *An* **n-product experiment** *is obtained by repeatedly performing n trials of some experiment. The experiment that is repeated is called the "mother" experiment.*

Example 1.6 (Toss a coin n times)**.** *Suppose our experiment entails tossing a coin n times and recording* `H` *for Heads and* `T` *for Tails. When* $n = 3$*, one possible outcome of this experiment is* `HHT`*, ie. a Head followed by another Head and then a Tail. Seven other outcomes are possible.*

The sample space for "toss a coin three times" experiment is:

$$\Omega = \{\mathtt{H}, \mathtt{T}\}^3 = \{\mathtt{HHH}, \mathtt{HHT}, \mathtt{HTH}, \mathtt{HTT}, \mathtt{THH}, \mathtt{THT}, \mathtt{TTH}, \mathtt{TTT}\} \ ,$$

with a particular sample point or outcome $\omega = \mathtt{HTH}$, and another distinct outcome $\omega' = \mathtt{HHH}$. An event, say A, that 'at least two Heads occur' is the following subset of Ω:

$$A = \{\mathtt{HHH}, \mathtt{HHT}, \mathtt{HTH}, \mathtt{THH}\} \ .$$

Another event, say B, that 'no Heads occur' is:

$$B = \{\texttt{TTT}\}$$

Note that the event B is also an outcome or sample point. Another interesting event is the empty set $\emptyset \subset \Omega$. The event that 'nothing in the sample space occurs' is $\emptyset$.

EXPERIMENT SUMMARY

Experiment	–	an activity producing distinct outcomes.
Ω	–	set of all outcomes of the experiment.
ω	–	an individual outcome in Ω, called a simple event.
$A \subseteq \Omega$	–	a subset A of Ω is an event.
Trial	–	one performance of an experiment resulting in 1 outcome.

1.2 Probability

The mathematical model for probability or the probability model is an axiomatic system that may be motivated by the intuitive idea of 'long-term relative frequency'. If the axioms and definitions are intuitively motivated, the probability model simply follows from the application of logic to these axioms and definitions. No attempt to define probability in the real world is made. However, the application of probability models to real-world problems through statistical experiments has a fruitful track record. In fact, you are here for exactly this reason.

Idea 1.7 (The long-term relative frequency (LTRF) idea)**.** *Suppose we are interested in the fairness of a coin, i.e. if landing Heads has the same "probability" as landing Tails. We can toss it n times and call $N(\texttt{H}, n)$ the fraction of times we observed Heads out of n tosses. Suppose that after conducting the tossing experiment* 1000 *times, we rarely observed Heads, e.g.* 9 *out of the* 1000 *tosses, then $N(\texttt{H}, 1000) = 9/1000 = 0.009$. Suppose we continued the number of tosses to a million and found that this number approached closer to* 0.1*, or, more generally, $N(\texttt{H}, n) \to 0.1$ as $n \to \infty$. We might, at least intuitively, think that the coin is unfair and has a lower "probability" of* 0.1 *of landing Heads. We might think that it is fair had we observed $N(\texttt{H}, n) \to 0.5$ as $n \to \infty$. Other crucial assumptions that we have made here are:*

1. **Something Happens***: Each time we toss a coin, we are certain to observe Heads* **or** *Tails, denoted by $\texttt{H} \cup \texttt{T}$. The probability that*

"something happens" is 1*. More formally:*

$$N(\mathtt{H} \cup \mathtt{T}, n) = \frac{n}{n} = 1.$$

This is an intuitively reasonable assumption that simply says that one of the possible outcomes is certain to occur, provided the coin is not so thick that it can land on or even roll along its circumference.

2. **Addition Rule***: Heads and Tails are mutually exclusive events in any given toss of a coin, i.e. they cannot occur simultaneously. The intersection of mutually exclusive events is the empty set and is denoted by* $\mathtt{H} \cap \mathtt{T} = \emptyset$*. The event* $\mathtt{H} \cup \mathtt{T}$*, namely that the event that "coin lands Heads* **or** *coin lands Tails" satisfies:*

$$N(\mathtt{H} \cup \mathtt{T}, n) = N(\mathtt{H}, n) + N(\mathtt{T}, n).$$

3. *The coin-tossing experiment is repeatedly performed in an* **independent** *manner, i.e. the outcome of any individual coin-toss does not affect that of another. This is an intuitively reasonable assumption since the coin has no memory and the coin is tossed identically each time.*

We will use the LTRF idea more generally to motivate a mathematical model of probability called probability model. Suppose A is an event associated with some experiment $\mathcal{E}$, so that A either does or does not occur when the experiment is performed. We want the probability that event A occurs in a specific performance of $\mathcal{E}$, denoted by $\mathbb{P}(A)$, to intuitively mean the following: if one were to perform a super-experiment $\mathcal{E}^\infty$ by independently repeating the experiment $\mathcal{E}$ and recording $N(A, n)$, the fraction of times A occurs in the first n performances of $\mathcal{E}$ within the super-experiment $\mathcal{E}^\infty$. Then the LTRF idea suggests:

$$N(A, n) := \frac{\text{Number of times } A \text{ occurs}}{n = \text{Number of performances of } \mathcal{E}} \to \mathbb{P}(A), \ as \quad n \to \infty \quad (1.1)$$

We first begin by defining certain collections of sets that will be the prototype for collections of events:

Definition 1.8 (Sigma algebras)**.** *Let* Ω *be a set: We say that a collection of subsets of* Ω*,* $\mathcal{F}$ *is a* **sigma-algebra/ sigma-field/** σ**-algebra** *if it satisfies the following properties:*

1. $\mathcal{F}$ *contains* Ω*, i.e.* $\Omega \in \mathcal{F}$*.*

2. *The collection* $\mathcal{F}$ *is closed under complementation*

$$A \in \mathcal{F} \implies A^C \in \mathcal{F}.$$

3. *The collection $\mathcal{F}$ is closed under countable unions*

$$A_1, A_2, \ldots \in \mathcal{F} \implies \bigcup_i A_i \in \mathcal{F}.$$

Remark 1.9. *For those not familiar with unions of a countable collection of sets, we define*

$$\bigcup_i A_i := \{\omega : \text{ there exists an } i \text{ such that } \omega \in A_i\}.$$

That is, $\omega \in \bigcup_i A_i$ if there is a set in the sequence $A_1, A_2, \ldots$ that contain ω.

Similarly we can define the countable intersection as

$$\bigcap_i A_i := \{\omega : \text{ for all } i \ \omega \in A_i\}.$$

That is, $\omega \in \bigcap_i A_i$ if it is in all $A_1, A_2, \ldots$.

Now, we are finally ready to define probability and events.

Definition 1.10 (Probability)**.** *Let $\mathcal{E}$ be an experiment with sample space Ω. Let $\mathcal{F}$ denote σ-algebra as in Definition 1.8. A* **probability measure** *is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying the following conditions:*

1. *The 'Something Happens' axiom holds, i.e. $\mathbb{P}(\Omega) = 1$.*
2. *The 'Addition Rule' axiom holds, i.e. for $A, B \in \mathcal{F}$:*

$$A \cap B = \emptyset \quad \implies \quad \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) \ .$$

We call elements of $\mathcal{F}$, events and we will call $(\Omega, \mathcal{F}, \mathbb{P})$ a **probability triple**.

1.2.1 Consequences of our Definition of Probability

It is important to realize that we accept the 'addition rule' as an axiom in our mathematical definition of probability (or our probability model) and we do **not** prove this rule. However, the facts which are stated (with proofs) below, are logical consequences of our definition of probability:

Lemma 1.11. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple, then*

1. *For any event $A \in \mathcal{F}$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.*

2. *For any two events $A, B \in \mathcal{F}$, we have the* **inclusion-exclusion principle***:*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

3. *From inclusion-exclusion principle we get* **Boole's inequality***: for any two events $A, B \in \mathcal{F}$*

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

An immediate consequence of 1 is: If $A = \Omega$ then $A^c = \Omega^c = \emptyset$ and $\mathbb{P}(\emptyset) = 1 - \mathbb{P}(\Omega) = 1 - 1 = 0$.

Proof. To prove 1 we proceed as follows

$$\overbrace{\mathbb{P}(A) + \mathbb{P}(A^c)}^{LHS} \underbrace{=}_{+ \text{ rule } \because A \cap A^c = \emptyset} \mathbb{P}(A \cup A^c) \underbrace{=}_{A \cup A^c = \Omega} \mathbb{P}(\Omega)$$

$$\underbrace{=}_{\because \ \mathbb{P}(\Omega)=1} \overbrace{1}^{RHS} \qquad \underbrace{\Longrightarrow}_{LHS-\mathbb{P}(A) \ \& \ RHS-\mathbb{P}(A)} \mathbb{P}(A^c)$$

$$= 1 - \mathbb{P}(A)$$

To prove 2 we note that since:

$$\begin{aligned} A = (A \setminus B) \cup (A \cap B) \quad &\text{and} \quad (A \setminus B) \cap (A \cap B) = \emptyset, \\ A \cup B = (A \setminus B) \cup B \quad &\text{and} \quad (A \setminus B) \cap B = \emptyset \end{aligned}$$

the addition rule implies that:

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B) \\ \mathbb{P}(A \cup B) &= \mathbb{P}(A \setminus B) + \mathbb{P}(B) \end{aligned}$$

Substituting the first equality above into the second, we get:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A \setminus B) + \mathbb{P}(B) = \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(B)$$

Finally we note that Booles inequality, 3, follows immediately from 2 since $-\mathbb{P}(A \cap B) \leq 0$. □

These basic properties can then be iterated to obtain similar statements when there is more than 2 events.

Lemma 1.12. *(Extended properties) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple, then*

1. *The inclusion-exclusion principle extends similarly to any n events $A_1, A_2, \ldots, A_n \in \mathcal{F}$ as follows:*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n-1} \sum_{i<\cdots<n} \mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right)$$

2. *Once again by the inclusion-exclusion principle, the Boole's inequality* **(Union bound)** *generalises to any n events $A_1, A_2, \ldots, A_n \in \mathcal{F}$ as follows:*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$$

3. *For a sequence of mutually disjoint events $A_1, A_2, A_3, \ldots, A_n \in \mathcal{F}$:*

$$A_i \cap A_j = \emptyset \quad \text{for any } i \neq j \quad \Longrightarrow \quad \mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots + \mathbb{P}(A_n).$$

Proof. For the proof of 1 and 2, see the counting argument in https://en.wikipedia.org/wiki/Inclusion%E2%80%93exclusion_principle if you are curious.

3 follows from 1 since all intersections are empty. □

We have formally defined the **probability model** specified by the **probability triple** $(\Omega, \mathcal{F}, \mathbb{P})$ that can be used to model an **experiment** $\mathcal{E}$.

Next, let us take a detour into how one might interpret it in the real world. The following is an adaptation from Williams D, *Weighing the Odds: A Course in Probability and Statistics*, Cambridge University Press, 2001, which henceforth is abbreviated as WD2001.

Probability Model	**Real-world Interpretation**
Sample space Ω	Set of all outcomes of an experiment
Sample point ω	Possible outcome of an experiment
(No counterpart)	Actual outcome $\omega^\star$ of an experiment
Event A, a (suitable) subset of Ω	The real-world event corresponding to A occurs if and only if $\omega^\star \in A$
$\mathbb{P}(A)$, a number between 0 and 1	Probability that A will occur for an experiment yet to be performed

Events in Probability Model	**Real-world Interpretation**
Sample space Ω	The certain even 'something happens'
The $\emptyset$ of Ω	The impossible event 'nothing happens'
The intersection $A \cap B$	'Both A and B occur'
$A_1 \cap A_2 \cap \cdots \cap A_n$	'All of the events $A_1, A_2, \ldots, A_n$ occur simultaneously'
The union $A \cup B$	'At least one of A and B occurs'
$A_1 \cup A_2 \cup \cdots \cup A_n$	'At least one of the events $A_1, A_2, \ldots, A_n$ occurs'
A^c, the complement of A	'A does not occur'
$A \setminus B$	'A occurs, but B does not occur'
$A \subset B$	'If A occurs, then B must occur'

1.2.2 More on Sigma Algebras

Generally one encounters four types of sigma algebras (you will understand the last two types after taking more advanced courses in mathematics, so it is fine to understand the ideas intuitively for now!) and they are:

- When the sample space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$ is a finite set with k outcomes and $\mathbb{P}(\omega_i)$, the probability for each outcome $\omega_i \in \Omega$ is known, then one typically takes the sigma-algebra $\mathcal{F}$ to be the set of all subsets of Ω called the **power set** and denoted by 2^Ω. The probability of each event $A \in 2^\Omega$ can be obtained by adding the probabilities of the outcomes in A, i.e., $\mathbb{P}(A) = \sum_{\omega_i \in A} \mathbb{P}(\omega_i)$. Clearly, 2^Ω is indeed a sigma-algebra and it contains $2^{\#\Omega}$ events in it.

- When the sample space $\Omega = \{\omega_1, \omega_2, \ldots\}$ is a countable set then one typically takes the sigma-algebra $\mathcal{F}$ to be the set of all subsets of Ω. Note that this is very similar to the case with finite Ω except now $\mathcal{F} = 2^\Omega$ could have uncountably many events in it.

- If $\Omega = \mathbb{R}^d$ for finite $d \in \{1, 2, 3, \ldots\}$ then the **Borel sigma-algebra** is the smallest sigma-algebra containing all **half-spaces**, i.e., sets of the form

$$\{x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d : x_1 \leq c_1, x_2 \leq c_2, \ldots, x_d \leq c_d\},$$

for any $c = (c_1, c_2, \ldots, c_d) \in \mathbb{R}^d$. When $d = 1$ the half-spaces are the half-lines $\{(-\infty, c] : c \in \mathbb{R}\}$ and when $d = 2$ the half-spaces are the south-west quadrants $\{(-\infty, c_1] \times (-\infty, c_2] : (c_1, c_2) \in \mathbb{R}^2\}$, etc. (Equivalently, the Borel sigma-algebra is the smallest sigma-algebra containing all open sets in $\mathbb{R}^d$).

- Given a finite set $\mathbb{S} = \{s_1, s_2, \ldots, s_k\}$, let Ω be the sequence space $\mathbb{S}^\infty := \mathbb{S} \times \mathbb{S} \times \mathbb{S} \times \cdots$, i.e., the set of sequences of infinite length that are made up of elements from $\mathbb{S}$. A set of the form

  $$A_1 \times A_2 \times \cdots \times A_n \times \mathbb{S} \times \mathbb{S} \times \cdots , \quad A_k \subset \mathbb{S} \text{ for all } k \in \{1, 2, \ldots, n\} ,$$

  is called a **cylinder set**. The set of events in $\mathbb{S}^\infty$ is the smallest sigma-algebra containing the cylinder sets.

1.3 Conditional Probability

Conditional probabilities arise when we have partial information about the result of an experiment which restricts the sample space to a range of outcomes. For example, if there has been a lot of recent seismic activity in Christchurch, then the probability that an already damaged building will collapse tomorrow is clearly higher than if there had been no recent seismic activity.

Conditional probabilities are often expressed in English by phrases such as:

"If A happens, what is the probability that B happens?"

or

"What is the probability that A happens if B happens?"

or

" What is the probability that A occurs given that B occurs?"

Definition 1.13 (Conditional Probability)**.** *Suppose we are given an experiment $\mathcal{E}$ with a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. Let $A, B \in \mathcal{F}$ (events), such that $\mathbb{P}(A) \neq 0$. Then, we define the* **conditional probability** *of B given A by,*

$$\mathbb{P}(B|A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} . \tag{1.2}$$

It turns out that the conditional probability is just a restriction of the events to A and it is as such, a probability measure.

Lemma 1.14. *Given a probability triple* $(\Omega, \mathcal{F}, \mathbb{P})$ *then for* $A \in \mathcal{F}$ *with* $\mathbb{P}(A) \neq 0$,

$$\mathbb{P}(\cdot|A) : \mathcal{F} \to [0,1]$$

is a probability measure as in Definition 1.10 over $(\Omega, \mathcal{F})$.

Proof. Exercise! □

It is now clear from Lemma 1.14 that $(\Omega, \mathcal{F}, \mathbb{P}(\cdot|A))$ is a probability triple and as such Lemmas 1.11 and 1.12 holds. Hence, there is no distinction in how we can work with conditional probabilities versus regular probabilities.

1.3.1 Bayes' Theorem

Next we look at one of the most elegant applications of the definition of conditional probability along with the addition rule for a partition of Ω called *Bayes' Theorem*. We will present a two event case first called *Bayes' Rule* and then present the more general case of the Theorem.

This is useful because many problems involve reversing the order of conditional probabilities. Suppose we want to investigate some phenomenon A and have an observation B that is evidence about A: for example, A may be breast cancer and B may be a positive mammogram. Then Bayes' Theorem tells us how we should update our probability of A, given the new evidence B.

Or, put more simply, Bayes' Rule is useful when you know $P(B|A)$ but want $P(A|B)$!

Proposition 1.15 (Bayes' Rule)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability triple, let* $A, B \in \mathcal{F}$ *with* $\mathbb{P}(A), \mathbb{P}(B) > 0$, *then*

$$\mathbb{P}(A|B) \;=\; \frac{\mathbb{P}(A)\,\mathbb{P}(B|A)}{\mathbb{P}(B)} \;. \tag{1.3}$$

Proof. From the definition of conditional probability we have

$$\begin{aligned} \mathbb{P}(A \mid B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\ \mathbb{P}(B \mid A) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}. \end{aligned} \tag{1.4}$$

From this we can see that

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)}{\mathbb{P}(B)} \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A)}{\mathbb{P}(B)} \mathbb{P}(B \mid A),$$

the first and last equality used (1.4) and in the second equality we just multiplied and divided by $\mathbb{P}(A)$. □

Before we see the more general form of Bayes' Rule, let us make a simple observation called the *total probaility theorem.*

Theorem 1.16 (Total probability)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability triple, suppose* $A_1 \cup A_2 \ldots \cup A_k \in \mathcal{F}$ *is a sequence of events with positive probability that partition the sample space, that is,* $A_1 \cup A_2 \cdots \cup A_k = \Omega$ *and* $A_i \cap A_j = \emptyset$ *for any* $i \neq j$*, then for some arbitrary event* $B \in \mathcal{F}$*.*

$$\mathbb{P}(B) \;=\; \sum_{h=1}^{k} \mathbb{P}(B \cap A_h) \;=\; \sum_{h=1}^{k} \mathbb{P}(B|A_h)\,\mathbb{P}(A_h) \tag{1.5}$$

Proof. Since $A_1, \ldots, A_k$ is a partition of Ω they are mutually exclusive (disjoint), hence

$$B \cap A_1, B \cap A_2, \ldots, B \cap A_k$$

are mutually exclusive. Thus the first equality follows from Lemma 1.12:3. The last equality follows from the definition of conditional probability. □

Reference to the Venn digram (you should draw below as done in lectures) will help you understand this idea for the four event case.

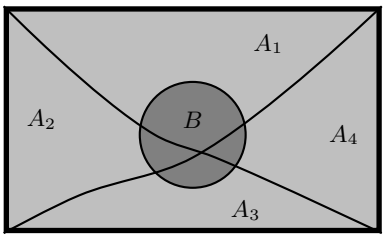


Figure 1.1: Reference to the Venn digram will help you understand this idea behind the proof of the total probability theorem in Theorem 1.16 for the four event case.

Theorem 1.17 (Bayes', 1763)**.** *Let everything be as in Theorem 1.16, and in addition assume that* $\mathbb{P}(B) > 0$*, then for any* $i = 1, \ldots, k$ *we have*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\,\mathbb{P}(A_i)}{\sum_{j=1}^{k} \mathbb{P}(B|A_j)\,\mathbb{P}(A_j)} \tag{1.6}$$

Proof. From Definition 1.13, Lemmas 1.12 and 1.14, and Theorem 1.16 we

have

$$\begin{aligned}\mathbb{P}(A_h|B) &= \frac{\mathbb{P}(A_h \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \cap A_h)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_h)\,\mathbb{P}(A_h)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B|A_h)\,\mathbb{P}(A_h)}{\mathbb{P}\left(\bigcup_{h=1}^{k}(B \cap A_h)\right)} = \frac{\mathbb{P}(B|A_h)\,\mathbb{P}(A_h)}{\sum_{h=1}^{k}\mathbb{P}\left(B \cap A_h\right)} \\ &= \frac{\mathbb{P}(B|A_h)\,\mathbb{P}(A_h)}{\sum_{h=1}^{k}\mathbb{P}(B|A_h)\,\mathbb{P}(A_h)}.\end{aligned}$$

□

It is customary to call $\mathbb{P}(A_h)$ the **prior probability of** A_h, i.e., before observing B or *a priori*, and $\mathbb{P}(A_h|B)$ the **posterior probability of** A_h, i.e., after observing B or *a posteriori*. Note that these names only make sense in the context of how you are modeling, in essence the above theorem does not differentiate between what is observed and not.

1.3.2 Independence and Dependence

In general, $P(A|B)$ and $P(A)$ are different, but sometimes the occurrence of B makes no difference, and gives no new information about the chances of A occurring. This is the idea behind independence. Events like "having blue eyes" and "having blond hair" are associated due to common genetic ancestry, but events like "my neighbour wins Lotto" and "I win Lotto" are not due to the Lotto machine being chaotically whirled around before ejection (as modelled by a well-stirred urn).

Definition 1.18 (Independence of two events)**.** *Given a probability triple* $(\Omega, \mathcal{F}, \mathbb{P})$*, any two events* $A, B \in \mathcal{F}$ *are said to be* **independent** *if and only if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)\ . \tag{1.7}$$

Another way of making sense of the above definition is through the following lemma.

Lemma 1.19. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability triple, let* $A, B \in \mathcal{F}$ *be independent, then if* $P(B) > 0$ *we have*

$$\mathbb{P}(A \mid B) = \mathbb{P}(A)$$

Proof. From Definition 1.13 we have

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\,\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A).$$

□

The above lemma says that information about the occurrence of B does not affect the occurrence of A. If $\mathbb{P}(A) > 0$ we can use Lemma 1.19 with A, B reversed to get

$$\mathbb{P}(B \mid A) = \mathbb{P}(B),$$

which says that information about the occurrence of A does not affect the occurrence of B. So in a way, independence means that they have no effect on eachother.

If we have more than two events we can extend the notion of pairwise independence to an independent sequence.

Definition 1.20 (Independence of a sequence of events)**.** *Given a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, we say that a finite or infinite sequence of events $A_1, A_2, \ldots \in \mathcal{F}$ are independent if whenever $i_1, i_2, \ldots, i_k$ are distinct elements from the set of indices $\mathbb{N}$, then*

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1})\,\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k})$$

It should be noted that for a sequence of more than two events $A_1, A_2, \ldots \in \mathcal{F}$, pairwise independence (see Definition 1.18) is a weaker requirement than sequentially independent (see Definition 1.20). To make this clear, see the next example:

Example 1.21 (Pairwise independent events that are not jointly independent)**.** *Let a ball be drawn from an well-stirred urn containing four balls labelled 1,2,3,4. Consider the events $A = \{1, 2\}$, $B = \{1, 3\}$ and $C = \{1, 4\}$. Then,*

$$\begin{aligned}
\mathbb{P}(A \cap B) &= \mathbb{P}(A)\,\mathbb{P}(B) = \frac{2}{4} \times \frac{2}{4} = \frac{1}{4}, \\
\mathbb{P}(A \cap C) &= \mathbb{P}(A)\,\mathbb{P}(C) = \frac{2}{4} \times \frac{2}{4} = \frac{1}{4}, \\
\mathbb{P}(B \cap C) &= \mathbb{P}(B)\,\mathbb{P}(C) = \frac{2}{4} \times \frac{2}{4} = \frac{1}{4},
\end{aligned}$$

but,

$$\frac{1}{4} = \mathbb{P}(\{1\}) = \mathbb{P}(A \cap B \cap C) \neq \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C) = \frac{2}{4} \times \frac{2}{4} \times \frac{2}{4} = \frac{1}{8} \ .$$

Therefore, inspite of being pairwise independent, the events A, B and C are not jointly independent.

1.4 Extension of probability*

Definition 1.10 is limited to only finite collections of sets, i.e. the additivity works for finitely many sets, which is certainly enough to obtain an understanding of probability. However, in the later stages we actually need the following extension of the definition of probability.

Definition 1.22 (Probability (Full))**.** *Let $\mathcal{E}$ be an experiment with sample space Ω. Let $\mathcal{F}$ denote σ-algebra as in Definition 1.8. A* **probability measure** *is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying the following conditions:*

1. *The 'Something Happens' axiom holds, i.e. $\mathbb{P}(\Omega) = 1$.*
2. *The 'Countably additive' axiom holds, i.e. let $\{E_i\}$ be a countable collection of events in $\mathcal{F}$ that are pairwise disjoint then:*

$$\mathbb{P}(\cup_i E_i) = \sum_i \mathbb{P}(E_i) \ .$$

Chapter 2

Random Variables

2.1 Basic Definitions

To take advantage of our measurements over the real numbers, in terms of its metric structure and arithmetic, we need to formally define this measurement process using the notion of a random variable.

Definition 2.1 (Random Variable)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be some probability triple. Then, a* **Random Variable (RV)***, say* X*, is a function from the sample space* Ω *to the set of real numbers* $\mathbb{R}$

$$X : \Omega \to \mathbb{R}$$

such that for every $x \in \mathbb{R}$*, the inverse image of the half-open real interval* $(-\infty, x]$ *is an element of the collection of events* $\mathcal{F}$*, i.e.:*

$$\text{for every } x \in \mathbb{R}, \qquad X^{[-1]}(\ (-\infty, x]\ ) := \{\omega : X(\omega) \leq x\} \in \mathcal{F}\ .$$

This definition can be summarised by the statement that a RV is an $\mathcal{F}$*-measurable map. We assign probability to the RV* X *as follows:*

$$\mathbb{P}(X \leq x) = \mathbb{P}(\ X^{[-1]}(\ (-\infty, x]\ )\ ) := \mathbb{P}(\ \{\omega : X(\omega) \leq x\}\ )\ . \tag{2.1}$$

Definition 2.2 (Distribution Function)**.** *The* **Distribution Function (DF)** *or* **Cumulative Distribution Function (CDF)** *of any RV* X*, over a probability triple* $(\Omega, \mathcal{F}, \mathbb{P})$*, denoted by* F *is:*

$$F(x) := \mathbb{P}(X \leq x) = \mathbb{P}(\ \{\omega : X(\omega) \leq x\}\ ), \qquad \text{for any} \quad x \in \mathbb{R}\ . \tag{2.2}$$

Thus, $F(x)$ *or simply* F *is a non-decreasing, right continuous,* $[0, 1]$*-valued function over* $\mathbb{R}$*. When a RV* X *has DF* F *we write* $X \sim F$*.*

Remark 2.3 (Notation)**.** *It is enough to understand the idea of random variables as explained above, and work with random variables using simplified notation like*

$$\mathbb{P}(2 \leq X \leq 3)$$

rather than

$$\mathbb{P}(\{\omega : 2 \leq X(\omega) \leq 3\})$$

but note that when learning or doing more advanced work this sample space notation is usually needed to clarify the true meaning of the simplified notation.

From the idea of a distribution function, we get something that resembles the fundamental theorem of calculus:

Proposition 2.4. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let X be a random variable with DF F. Then for $a < b$ we get*

$$\mathbb{P}(a < X \leq b) = F(b) - F(a). \tag{2.3}$$

Proof. For this proof we will be very formal for the sake of clarity, later on we will adopt the more relaxed notation. Define the sets

$$\begin{aligned} A &= \{\omega : \omega \in \Omega, X(\omega) \leq a\} \\ B &= \{\omega : \omega \in \Omega, X(\omega) \leq b\} \\ C &= \{\omega : \omega \in \Omega, a < X(\omega) \leq b\} \end{aligned}$$

Note that $A, B, C \in \mathcal{F}$, which follows from Definition 2.1. Furthermore, note that $B = A \cup C$ and that $A \cap C = \emptyset$. Now using Definition 2.2 and Lemma 1.11 and our above construction, we get

$$F(b) = \mathbb{P}(B) = \mathbb{P}(A \cup C) = \mathbb{P}(A) + \mathbb{P}(C) = F(a) + \mathbb{P}(C).$$

Rearranging the above,

$$F(b) - F(a) = \mathbb{P}(C) = \mathbb{P}(a < X \leq b).$$

which is (2.3). □

A special RV that often plays the role of 'building-block' in Probability and Statistics is the indicator function of an event A that tells us whether the event A has occurred or not. Recall that an event belongs to the collection of possible events $\mathcal{F}$ for our experiment.

Definition 2.5. *[Indicator Function] Given a probability triple* $(\Omega, \mathcal{F}, \mathbb{P})$*, the* **Indicator Function** *of an event* $A \in \{$ *which is denoted* $\mathbb{1}_A$ *is defined as follows:*

$$\mathbb{1}_A(\omega) := \begin{cases} 1 & \text{if} \quad \omega \in A \\ 0 & \text{if} \quad \omega \notin A \end{cases} \tag{2.4}$$

Lemma 2.6. *The indicator function* $\mathbb{1}_A$ *as in Definition 2.5 is a random variable.*

Proof. For $\mathbb{1}_A$ to be a RV, we need to verify that for any real number $x \in \mathbb{R}$, the inverse image $\mathbb{1}_A^{[-1]}(\ (-\infty, x]\ )$ is an event, ie :

$$\mathbb{1}_A^{[-1]}(\ (-\infty, x]\ ) := \{\omega : \mathbb{1}_A(\omega) \leq x\} \in \mathcal{F}\ .$$

Since the indicator function is either 0 or 1 we observe that we have to get the empty event if $x < 0$, furthermore we have to get the entire sample space if $x > 1$ (since it is always true). The last case is when $0 \leq x < 1$, which is only ok when $\mathbb{1}_A = 0$. Summarised below:

$$\mathbb{1}_A^{[-1]}(\ (-\infty, x]\ ) := \{\omega : \mathbb{1}_A(\omega) \leq x\} = \begin{cases} \emptyset & \text{if} \quad x < 0 \\ A^c & \text{if} \quad 0 \leq x < 1 \\ A \cup A^c = \Omega & \text{if} \quad 1 \leq x \end{cases}$$

Thus, $\mathbb{1}_A^{[-1]}(\ (-\infty, x]\ )$ is one of the following three sets that belong to $\mathcal{F}$; (1) $\emptyset$, (2) A^c and (3) Ω depending on the value taken by x relative to the interval $[0, 1]$. We have proved that $\mathbb{1}_A$ is indeed a RV. □

Model 2.7 (Indicator of an event as Bernoulli RV)**.** *Given a probability triple* $(\Omega, \mathcal{F}, \mathbb{P})$ *and an event* $A \in \mathcal{F}$*, the random variable* $\mathbb{1}_A$ *is called the* Bernoulli *RV for event* A *with a known probability* $\mathbb{P}(A)$*. We will adopt the notation* Bernoulli(θ) *for the RV by introducing a paramater* $\theta \in [0, 1]$ *for the typically unknown probability* $\mathbb{P}(A)$*.*

Lemma 2.8. *Given a probability triple* $(\Omega, \mathcal{F}, \mathbb{P})$ *and an event* $A \in \mathcal{F}$*, the following properties hold:*

1. $\mathbb{1}_A = 1 - \mathbb{1}_{A^c}$*, (complementation behaves like the probability)*
2. $\mathbb{1}_{A \cap B} = \mathbb{1}_A \mathbb{1}_B$ *(intersection becomes product)*
3. $\mathbb{1}_{A \cup B} = \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_A \mathbb{1}_B$ *(union becomes addition - intersection)*

As you can see, there is a lot of similarities between the properties of indicator function and the properties of the probability measure $\mathbb{P}$.

Proof. Exercise! □

We slightly abuse notation when A is a single element set by ignoring the curly braces.

2.2 Discrete Random Variables

When a RV takes at most countably many values from a discrete set $\mathbb{X}$, we call it a **discrete** RV. Recall that a set $\mathbb{X}$ is said to be discrete if we can enumerate its elements, i.e., find an enumerating or counting function $\mathbb{X} \ni x \mapsto i \in \mathbb{N}$ that associates each element $x \in \mathbb{X}$ to a natural number $i \in \mathbb{N}$. So, $\mathbb{X}$ is either finite with k elements in $\mathbb{X} = \{x_1, x_2, \ldots, x_k\}$ or countably infinite with the same cardinality as $\mathbb{N}$ with $\mathbb{X} = \{x_1, x_2, \ldots\}$.

Definition 2.9. *Let X be a $\mathbb{R}$-valued RV over a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. If X takes values in an enumerable set $\mathbb{X} \subset \mathbb{R}$ then we call X a $\mathbb{R}$-valued* **discrete** *random variable.*

The concept of distribution function (DF) does not differentiate between types of random variables, the next definition does:

Definition 2.10. *[probability mass function (PMF)] Let X be a $\mathbb{R}$-valued discrete RV over a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. We define the* **probability mass function** *(PMF) f of X to be the function $f : \mathbb{R} \to [0, 1]$ defined as follows:*

$$f(x) := \mathbb{P}(X = x) = \mathbb{P}(\ \{\omega : X(\omega) = x\}\ ) = \begin{cases} \theta_i & \text{if } x = x_i \in \mathbb{X}. \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

Theorem 2.11. *The relation between the DF F and PMF f for a discrete RV X is as follows:*

1. *For any $x \in \mathbb{R}$,*
$$F(x) = \sum_{x_i \le x} f(x_i) = \sum_{x_i \le x} \theta_i \ . \tag{2.6}$$
2. *For any $a, b \in \mathbb{R}$ with $a < b$,*
$$F(b) - F(a) = \sum_{a < x_i \le b} \theta_i \ . \tag{2.7}$$

   This is just the sum of all probabilities θ_i for which x_i satisfies $a < x_i \le b$.

3. *From the fact that* $\mathbb{P}(\Omega) = 1$*, we get that the sum of all the probabilties is* 1*:*

$$\sum_i \theta_i = 1 \ . \tag{2.8}$$

Proof. Let us prove the first equality. First, recall that the definition of a discrete random variable required that X takes values in an enumerable $\mathbb{X}$, i.e. $\mathbb{X} = \{x_1, \ldots\}$, then for each $x_i \in \mathbb{X}$ define the sets

$$A_i = \{\omega : \omega \in \Omega, i - 1 < X(\omega) \leq i\} = \{X = i\},$$

and note that $A_i \in \mathcal{F}$, $\cup_i A_i = \Omega$ and they are mutually exclusive. Now using Definitions 1.22 and 2.2 we get

$$F(x) = \mathbb{P}(X \leq x) = \mathbb{P}\left(\cup_{i:x_i \leq x} A_i\right) = \sum_{i:x_i \leq x} \mathbb{P}(A_i) = \sum_{x_i \leq x} f(x_i).$$

The last equality of (2.6) is just the definition of θ_i in Definition 2.10.

The proof of the other properties is an Exercise! □

Remark 2.12. *When X only has finitely many possibilities, say k with $\mathbb{X} = \{x_1, x_2, \ldots, x_k\}$, then we may think of the probability $\mathbb{P}$ specified by $(\theta_1, \theta_2, \ldots, \theta_k)$ as a point in the* ***unit*** $(k-1)$ ***simplex****:*

$$\Delta^{k-1} := \{(\theta_1, \theta_2, \ldots, \theta_k) \in \mathbb{R}^k : \sum_i \theta_i = 1 \text{ and } \theta_i \geq 0, \text{ for all } i\} \tag{2.9}$$

In particular when X has only two possible values with $\mathbb{X} = \{x_1, x_2\}$ then $\theta_2 = 1 - \theta_1$, so we can avoid subscripts and take $\theta := \theta_1$ and realize that the probability $\mathbb{P}$ is now specified by the point $(\theta, 1-\theta)$ in the ***unit*** 1 ***simplex****:*

$$\Delta^1 := \{(\theta, 1-\theta) \in \mathbb{R}^2 : 0 \leq \theta \leq 1\} \ . \tag{2.10}$$

See https://en.wikipedia.org/wiki/Simplex.

Out of the class of discrete random variables we will define specific kinds as they arise often in applications. We classify discrete random variables into three types for convenience as follows:

- Discrete uniform random variables with finitely many possibilities
- Discrete non-uniform random variables with finitely many possibilities
- Discrete non-uniform random variables with (countably) infinitely many possibilities

Model 2.13 (Discrete Uniform)**.** *We say that a discrete random variable X is uniformly distributed over k possible values in $\mathbb{X} = \{x_1, x_2, \ldots, x_k\}$ if its probability mass function is:*

$$f(x) = \begin{cases} \theta_i = \frac{1}{k} & \text{if } x = x_i, \quad \text{where } i = 1, 2, \ldots, k \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{2.11}$$

The distribution function for the discrete uniform random variable X is:

$$F(x) = \sum_{x_i \le x} f(x_i) = \sum_{x_i \le x} \theta_i = \begin{cases} 0 & \text{if } -\infty < x < x_1 \ , \\ \frac{1}{k} & \text{if } x_1 \le x < x_2 \ , \\ \frac{2}{k} & \text{if } x_2 \le x < x_3 \ , \\ \vdots & \\ \frac{k-1}{k} & \text{if } x_{k-1} \le x < x_k \ , \\ 1 & \text{if } x_k \le x < \infty \ . \end{cases} \tag{2.12}$$

The discrete uniform RV with values in $\mathbb{X} = \{1, 2, \ldots, k\}$ is called the equiprobable de Moivre(k) *RV.*

Example 2.14. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $A_1, \ldots, A_k \in \mathcal{F}$, then*

$$X = \sum_{i=1}^{k} \mathbb{1}_{A_i}$$

is a $\mathbb{R}$-valued discrete random variable, taking values in $\{0, \ldots, k\}$.

Can you write down the probability mass function and the distribution function for X when there is only two sets A_1, A_2?

2.3 Continuous Random Variables

If we have a random variable that does not take values in an enumerable set, then it is not discrete. However we often wish to allow the random variable to take any value in $\mathbb{R}$ or just every value in the interval $[0, 1]$. Let us define such a class of random variables.

Definition 2.15 (Continuous random variable)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let X be a $\mathbb{R}$-valued random variable with distribution function F. We say that X is a* **continuous** *RV if there exists a piecewise-continuous function $f : \mathbb{R} \to [0, \infty]$, called the* **probability density function (PDF)** *of X, such that*

$$F(x) \ = \ \mathbb{P}(X \le x) \ = \ \int_{-\infty}^{x} f(v) \, dv. \tag{2.13}$$

Remark 2.16. *see https://en.wikipedia.org/wiki/Piecewise.*

Remark 2.17. *There are actually random variables which are neither discrete or continuous, for instance, the product of a discrete and a continuous random variable!! We will deal with these later on.*

Theorem 2.18. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let X be a $\mathbb{R}$-valued continuous random variable, then the following holds:*

1. *For any $x \in \mathbb{R}$, the probability of observing a single value is zero:*

$$\mathbb{P}(X = x) = 0.$$

2. *The probability density function (density function) is the derivative of the distribution function. That is:*

$$f(x) = \frac{d}{dx}F(x) =: F'(x), \tag{2.14}$$

for every x at which $f(x)$ is continuous.

3. *For any $a, b \in \mathbb{R}$ with $a < b$,*

$$\begin{aligned}\mathbb{P}(a < X < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b) \quad &= \mathbb{P}(a \leq X < b) \\ = F(b) - F(a) \ = \int_a^b f(v)dv \ . \end{aligned} \tag{2.15}$$

4. *Finally:*

$$\int_{-\infty}^{\infty} f(x)\ dx = 1\ .$$

Proof. Fix x and let $\epsilon > 0$ be arbitrary, then from the properties of the probability measure, we get

$$F(x - \epsilon) \leq \mathbb{P}(X < x) \leq F(x)$$

However we know that integrals are continuous and as such $\lim_{\epsilon \to 0^+} F(x - \epsilon) = F(x)$, this proves that

$$\mathbb{P}(X < x) = \mathbb{P}(X \leq x). \tag{2.16}$$

Now to prove 1 we simply write

$$\mathbb{P}(X = x) = \mathbb{P}(x \leq X \leq x) = \mathbb{P}(X \leq x) - \mathbb{P}(X < x) = 0$$

where the last step follows from (2.16). Let $a \leq b$ and note that

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X < a)$$

Property 2 is the first fundamental theorem of calculus.
Property 3 follows from Proposition 2.4, (2.16), and Definition 2.15.
Finally propety 4 is an exercise!! □

The standard normal distribution is the most important continuous probability distribution. It was first described by De Moivre in 1733 and subsequently by C. F. Gauss (1777 - 1885). Many random variables have a normal distribution, or they are approximately normal, or can be transformed into normal random variables in a relatively simple fashion. Furthermore, the normal distribution is a useful approximation of more complicated distributions.

Model 2.19 (Normal$(0,1)$ or standard normal or Gaussian RV)**.** *A continuous random variable Z is called* ***standard normal*** *or* **standard Gaussian** *if its probability density function is*

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \ . \tag{2.17}$$

An exercise in calculus yields the first two derivatives of ϕ as follows:

$$\frac{d\phi}{dz} = -\frac{1}{\sqrt{2\pi}} z \exp\left(-\frac{z^2}{2}\right) = -z\phi(z),$$
$$\frac{d^2\phi}{dz^2} = \frac{1}{\sqrt{2\pi}}(z^2-1)\exp\left(-\frac{z^2}{2}\right) = (z^2-1)\phi(z) \ .$$

Thus, ϕ has a global maximum at 0, it is concave down if $z \in (-1,1)$ and concave up if $z \in (-\infty,-1) \cup (1,\infty)$. This shows that the graph of ϕ is shaped like a smooth symmetric bell centred at the origin over the real line.

2.3.1 Viewing a deterministic real variable as a random variable

Consider the class of discrete RVs with distributions that place all probability mass on a single real number. This is the probability model for the deterministic real variable, which is often thought of as an unknown constant $\theta \in \mathbb{R}$.

Model 2.20 (Point Mass(θ))**.** *Given a specific point $\theta \in \mathbb{R}$, we say an RV X has point mass at θ or is* Point Mass(θ) *distributed if the DF is:*

$$F(x;\theta) = \begin{cases} 0 & \text{if } x < \theta \\ 1 & \text{if } x \geq \theta \end{cases} \tag{2.18}$$

and the PMF is:

$$f(x;\theta) = \begin{cases} 0 & \text{if } x \neq \theta \\ 1 & \text{if } x = \theta \end{cases} \tag{2.19}$$

Thus, Point Mass(θ) RV X is deterministic in the sense that every realisation of X is exactly equal to $\theta \in \mathbb{R}$. We will see that this distribution plays a central limiting role in asymptotic statistics.

2.4 Transformations of random variables

Suppose we know the distribution of a random variable X. How do we find the distribution of a transformation of X, say $g(X)$?

Now, let us return to our question of determining the distribution of the transformation $g(X)$.

2.4.1 Transformations of discrete random variables

To answer this question we must first observe that the inverse image $g^{[-1]}$ satisfies the following properties:

- $g^{[-1]}(\mathbb{Y}) = \mathbb{X}$
- For any set A, $g^{[-1]}(A^c) = \left(g^{[-1]}(A)\right)^c$
- For any collection of sets $\{A_1, A_2, \ldots\}$,

$$g^{[-1]}\left(A_1 \cup A_2 \cup \cdots\right) = g^{[-1]}(A_1) \cup g^{[-1]}(A_2) \cup \cdots \ .$$

Let $\mathbb{X}$ be an enumerable subset of $\mathbb{R}$, then $\mathbb{Y} := g(\mathbb{X})$ is also enumerable and $(g(X))^{[-1]}((-\infty, x)) \in \mathcal{F}$. This is a subtle point, and I strongly encourage you to try to figure out why this is!!

Consequentially,

$$\mathbb{P}_g(A) = P\left(g(X) \in A\right) = P\left(X \in g^{[-1]}(A)\right) \tag{2.20}$$

satisfies the axioms of probability and gives the desired probability of the event A from the transformation $Y = g(X)$ in terms of the probability of the event given by the inverse image of A underpinned by the random variable X. It is crucial to understand this from the sample space Ω of the underlying experiment in the sense that (2.20) is just short-hand for its actual meaning:

$$\begin{aligned} P\left(\{\omega \in \Omega : g(X(\omega)) \in A\}\right) &= P\left(\left\{\omega \in \Omega : X(\omega) \in g^{[-1]}(A)\right\}\right) \\ &= P\left(X^{[-1]}(g^{[-1]}(A))\right). \end{aligned}$$

Because we have more than one random variable to consider, namely, X and its transformation $Y = g(X)$ we will subscript the probability density or mass function and the distribution function by the random variable itself. For example we denote the distribution function of X by $F_X(x)$ and that of Y by $F_Y(y)$.

For a discrete random variable X with probability mass function f_X we can obtain the probability mass function f_Y of $Y = g(X)$ using (2.20) as follows:

$$\begin{aligned} f_Y(y) &= \mathbb{P}(Y = y) = \mathbb{P}(Y \in \{y\}) \\ &= P\left(g(X) \in \{y\}\right) = P\left(X \in g^{[-1]}(\{y\})\right) \\ &= P\left(X \in g^{[-1]}(y)\right) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{x : g(x) = y\}} f_X(x) \ . \end{aligned}$$

This gives the formula:

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{x : g(x) = y\}} f_X(x) \ . \tag{2.21}$$

2.4.2 Transformations of continuous random variables

Suppose we know F_X and/or f_X of a continuous random variable X. Recall (2.20), in this formula it is essential that

$$\{\omega \in \Omega : X^{[-1]}(g^{[-1]}(A))\} \in \mathcal{F}.$$

That is, what we need to know is, when is $g(X)$ itself a random variable with respect to $(\Omega, \mathcal{F}, \mathbb{P})$?

Definition 2.21. *We say that a function $g : \mathbb{R} \to \mathbb{R}$ is Borel, if we denote Σ the Borel sigma algebra on $\mathbb{R}$ (see Section 1.2.2) and for every $A \in \Sigma$*

$$g^{[-1]}(A) \in \Sigma.$$

Remark 2.22. *The reason we need this seemingly abstract notion, is to guarantee that if $g : \mathbb{R} \to \mathbb{R}$ is Borel, then if X is any $\mathbb{R}$-valued RV, then $g(X)$ is an $\mathbb{R}$-valued RV.*

Recall that the definition of a $\mathbb{R}$-valued RV X in some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ was that

$$X^{[-1]}((-\infty, x]) \in \mathcal{F}, \quad \text{for all } x \in \mathbb{R}.$$

The Borel sigma-algebra Σ is the smallest sigma-algebra that contains the half open intervals $(-\infty, x]$ and thus if we take $A \in \Sigma$ then we also get $X^{[-1]}(A) \in \mathcal{F}$ by the properties of inverses. Thus it is immediate that $(g(X))^{[-1]}(A) \in \mathcal{F}$.

Remark 2.23. *Recall that in the case when X is a $\mathbb{R}$-valued discrete random variable, we don't need to assume anything about g. This was because of the fact that all enumerable sets of $\mathbb{R}$ is Borel, but we only mentioned it in passing.*

Remark 2.24. *A small subset of Borel functions are the continuous, piecewise continuous and monotone functions, which will be the ones we use mostly.*

Our objective now, is to obtain F_Y and/or f_Y of Y from F_X and/or f_X.

One-to-one transformations

The easiest case for transformations of continuous random variables is when g is **one-to-one and monotone** (monotone implies Borel).

- First, let us consider the case when g is **increasing** (monotone) on the range of the random variable X. In this case g^{-1} is also an increasing function and we can obtain the distribution function of $Y = g(X)$ in terms of the distribution function of X as

  $$F_Y(y) = P\left(Y \leq y\right) = P\left(g(X) \leq y\right) = P\left(X \leq g^{-1}(y)\right) = F_X(g^{-1}(y)) \ .$$

  Now, let us use the chainrule to compute the density of Y as follows:

  $$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X\left(g^{-1}(y)\right) = f_X\left(g^{-1}(y)\right)\frac{d}{dy}\left(g^{-1}(y)\right) \ .$$

- Second, let us consider the case when g is **decreasing** (monotone) on the range of the random variable X. In this case g^{-1} is also a decreasing function and we can obtain the distribution function of $Y = g(X)$ in terms of the distribution function of X as

  $$F_Y(y) = P\left(Y \leq y\right) = P\left(g(X) \leq y\right) = P\left(X \geq g^{-1}(y)\right) = 1 - F_X(g^{-1}(y)) \ ,$$

  and the density of Y as

  $$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}\left(1 - F_X\left(g^{-1}(y)\right)\right) = -f_X\left(g^{-1}(y)\right)\frac{d}{dy}\left(g^{-1}(y)\right) \ .$$

  For a decreasing g, its inverse function g^{-1} is also decreasing and consequently the density f_Y is indeed positive because $\frac{d}{dy}\left(g^{-1}(y)\right)$ is negative.

We can combine the above two cases and obtain the following

Proposition 2.25 (Change of variable formula)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let X be a $\mathbb{R}$-valued RV. If $g : \mathbb{R} \to \mathbb{R}$ is one-to-one and monotone (increasing or decreasing) on the range of X, i.e $X(\Omega) \subset \mathbb{R}$, then*

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right| \quad . \tag{2.22}$$

The next example yields the *l*ocation-scale family of normal random variables via a family of linear transformations of the standard normal random variable.

Example 2.26. *Let Z be the standard Gaussian or standard normal random variable with probability density function $\phi(z)$ given by Equation (2.17). For real numbers $\sigma > 0$ and μ consider the linear transformation of Z given by*

$$Y = g(Z) = \sigma Z + \mu \quad .$$

We are interested in the density of the tranformed random variable $Y = g(Z) = \sigma Z + \mu$. Once again, since g is a one-to-one monotone function let us follow the four steps and use the change of variable formula to obtain f_Y from $f_Z = \phi$ and g.

1. *$y = g(z) = \sigma z + \mu$ is a monotone increasing function over $-\infty < z < \infty$, the range of Z. So, we can apply the change of variable formula.*
2. *$z = g^{-1}(y) = (y - \mu)/\sigma$ is a monotone increasing function over the range of y given by, $-\infty < y < \infty$.*
3. *For $-\infty < y < \infty$,*
$$\left|\frac{d}{dy}g^{-1}(y)\right| = \left|\frac{d}{dy}\left(\frac{y-\mu}{\sigma}\right)\right| = \left|\frac{1}{\sigma}\right| = \frac{1}{\sigma} \quad .$$
4. *we can use (2.17) and (2.22) which gives*
$$f_Z(z) = \phi(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right) \quad ,$$
to find the density of Y as follows:
$$f_Y(y) = f_Z\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right| = \phi\left(\frac{y-\mu}{\sigma}\right)\frac{1}{\sigma} = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right] \quad ,$$
for $-\infty < y < \infty$.

Thus, we have obtained the expression for the probability density function of the linear transformation $\sigma Z + \mu$ of the standard normal random variable Z. This analysis leads to the following definition.

Model 2.27 (Normal(μ, σ^2) RV)**.** *Given a location parameter* $\mu \in (-\infty, +\infty)$ *and a scale parameter* $\sigma^2 > 0$*, the* Normal(μ, σ^2) *or Gaussian*(μ, σ^2) *random variable* X *has probability density function:*

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad (\sigma > 0) \ . \tag{2.23}$$

The normal distribution has the **distribution function**

$$F(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2\right] dv \ . \tag{2.24}$$

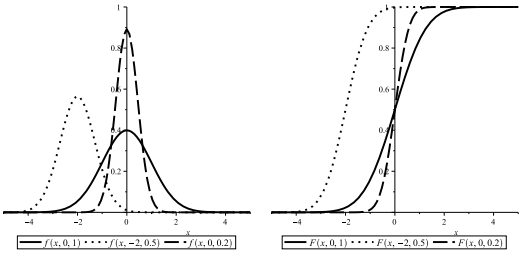


Figure 2.1: PDF and DF of a Normal(μ, σ^2) RV for different values of μ and σ^2

Direct method

If the transformation g in $Y = g(X)$ is not necessarily one-to-one then special care is needed to obtain the distribution function or density of Y. For a continuous random variable X with a known distribution function F_X we can obtain the distribution function F_Y of $Y = g(X)$ using (2.20) as follows:

$$\begin{aligned} F_Y(y) &= P\left(Y \leq y\right) = P\left(Y \in (-\infty, y]\right) \\ &= P\left(g(X) \in (-\infty, y]\right) = P\left(X \in g^{[-1]}((-\infty, y])\right) \\ &= P\left(X \in \{x : g(x) \in (-\infty, y]\}\right) \ . \end{aligned} \tag{2.25}$$

In words, the above equalities just mean that the probability that $Y \leq y$ is the probability that X takes a value x that satisfies $g(x) \leq y$. We can use this approach if it is reasonably easy to find the set $g^{[-1]}((-\infty, y]) = \{x : g(x) = (-\infty, y]\}$.

Example 2.28. *Let X be any random variable with distribution function F_X. Let $Y = g(X) = X^2$. Then we can find F_Y, the distribution function of Y from F_X as follows:*

- *Since $Y = X^2 \geq 0$, if $y < 0$ then $F_Y(y) = P\left(X \in \{x : x^2 < y\}\right) = \mathbb{P}(X \in \emptyset) = 0$.*
- *If $y \geq 0$ then*

$$\begin{aligned} F_Y(y) = P\left(Y \leq y\right) &= P\left(X^2 \leq y\right) \\ &= P\left(-\sqrt{y} \leq X \leq \sqrt{y}\right) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \ . \end{aligned}$$

By differentiation we get:

- *If $y < 0$ then $f_Y(y) = \frac{d}{dy}(F_Y(y)) = \frac{d}{dy}0 = 0$.*
- *If $y \geq 0$ then*

$$\begin{aligned} f_Y(y) = \frac{d}{dy}\left(F_Y(y)\right) &= \frac{d}{dy}\left(F_X(\sqrt{y}) - F_X(-\sqrt{y})\right) \\ &= \frac{d}{dy}\left(F_X(\sqrt{y})\right) - \frac{d}{dy}\left(F_X(-\sqrt{y})\right) \\ &= \frac{1}{2}y^{-\frac{1}{2}} f_X(\sqrt{y}) - \left(-\frac{1}{2}y^{-\frac{1}{2}} f_X(-\sqrt{y})\right) \\ &= \frac{1}{2\sqrt{y}}\left(f_X(\sqrt{y}) + f_X(-\sqrt{y})\right) \ . \end{aligned}$$

Therefore, the distribution function of $Y = X^2$ is:

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ F_X(\sqrt{y}) - F_X(-\sqrt{y}) & \text{if } y \geq 0 \ . \end{cases} \tag{2.26}$$

and the probability density function of $Y = X^2$ is:

$$f_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \frac{1}{2\sqrt{y}}\left(f_X(\sqrt{y}) + f_X(-\sqrt{y})\right) & \text{if } y \geq 0 \ . \end{cases} \tag{2.27}$$

Using the direct method's (2.25), we can obtain the distribution function of the Normal(μ, σ^2) random variable from that of the tabulated distribution function of the Normal$(0, 1)$.

Proposition 2.29 (One Table to Rule Them All Gaussians)**.** *The distribution function $F_X(x;\mu,\sigma^2)$ of the* Normal(μ,σ^2) *random variable X and the distribution function $F_Z(z)=\Phi(z)$ of the standard normal random variable Z are related by:*

$$F_X(x;\mu,\sigma^2) \;=\; F_Z\left(\frac{x-\mu}{\sigma}\right) \;=\; \Phi\left(\frac{x-\mu}{\sigma}\right) \;.$$

Proof. Let Z be a Normal$(0,1)$ random variable with distribution function $\Phi(z)=P(Z\leq z)$. We know that if $X=g(Z)=\sigma Z+\mu$ then X is the Normal(μ,σ^2) random variable. Therefore,

$$\begin{aligned} F_X(x;\mu,\sigma^2) &= P(X\leq x) = P\left(g(Z)\leq x\right) = P(\sigma Z+\mu\leq x) \\ &= P\left(Z\leq \frac{x-\mu}{\sigma}\right) \\ &= F_Z\left(\frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right) \;. \end{aligned}$$

□

Hence we often transform a general Normal(μ,σ^2) random variable, X, to a standardised Normal$(0,1)$ random variable, Z, by the substitution:

$$Z \;=\; \frac{X-\mu}{\sigma}\,.$$

2.5 Expectations and L^p spaces

Expectation is perhaps the most fundamental concept in probability theory. In fact, probability is itself an expectation as you will soon see!

Expectation is one of the fundamental concepts in probability. The expected value of a real-valued random variable gives the population mean, a measure of the centre of the distribution of the variable in some sense. Its variance measures its spread and so on.

Definition 2.30 (Expectation of a RV)**.** *The* **expectation**, *or* **expected value**, *or* **mean**, *or* **first moment**, *of a random variable X, with distribution function F and density f, is defined to be*

$$\mathbb{E}(X) := \int x\, dF(x) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int x f(x)\, dx & \text{if } X \text{ is continuous} , \end{cases} \tag{2.28}$$

provided the sum or integral is well-defined. We say the expectation exists if

$$\int |x|\ dF(x) < \infty . \tag{2.29}$$

Sometimes, we denote $\mathbb{E}(X)$ by $\mathbb{E}\,X$ for brevity. Thus, the expectation is a single-number summary of the RV X and may be thought of as the average.

Definition 2.31. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability triple, then for a random variable $X : \Omega \to \mathbb{R}$, we say that X is in $L^p(\mathbb{P})$ for some $1 \leq p < \infty$ if,*

$$\int |x|^p dF_X < \infty,$$

where F_X is the distribution function for X. If there is no fear of ambiguity we will simply write L^2 without referring to the measure $\mathbb{P}$.

So another way of saying that the expectation of the random variable X exists is the same as saying that $X \in L^1(\mathbb{P})$.

Definition 2.32 (Variance of a RV)**.** *Let X be a RV with mean or expectation $\mathbb{E}(X)$. The* **variance** *of X denoted by $\mathbb{V}(X)$ or simply $\mathbb{V}\,X$ is*

$$\mathbb{V}(X) := \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \int (x - \mathbb{E}(X))^2\, dF(x) ,$$

provided this expectation exists. The **standard deviation** *denoted by $\mathrm{sd}(X) := \sqrt{\mathbb{V}(X)}$. Thus variance is a measure of "spread" of a distribution.*

Another way of saying that the variance exists is to say that $X \in L^2(\mathbb{P})$.

Definition 2.33 (k-th moment of a RV)**.** *Let $k = 1, \ldots,$ we call*

$$\mathbb{E}\Big(X^k\Big) = \int x^k \, dF(x)$$

as the k-th moment of the RV X and say that the k-th moment exists when $X \in L^k$. We call the following expectation as the k-th central moment:

$$\mathbb{E}\Big((X - \mathbb{E}(X))^k\Big) \ .$$

2.6 Multivariate Random Variables

Often, in experiments we are measuring two or more aspects simultaneously. For example, we may be measuring the diameters and lengths of cylindrical shafts manufactured in a plant or heights, weights and blood-sugar levels of individuals in a clinical trial. Thus, the underlying outcome $\omega \in \Omega$ needs to be mapped to measurements as realizations of random vectors in the real plane $\mathbb{R}^2 = (-\infty, \infty) \times (-\infty, \infty)$ or the real space $\mathbb{R}^3 = (-\infty, \infty) \times (-\infty, \infty) \times (-\infty, \infty)$:

$$\omega \mapsto (X(\omega), Y(\omega)) : \Omega \to \mathbb{R}^2 \qquad \omega \mapsto (X(\omega), Y(\omega), Z(\omega)) : \Omega \to \mathbb{R}^3$$

More generally, we may be interested in heights, weights, blood-sugar levels, family medical history, known allergies, etc. of individuals in the clinical trial and thus need to make m measurements of the outcome in $\mathbb{R}^m$ using a "measurable mapping" from $\Omega \to \mathbb{R}^m$. To deal with such multivariate measurements we need the notion of **random vectors** ($\mathrm{R}\vec{\mathrm{V}}$s), i.e. ordered pairs of random variables (X, Y), ordered triples of random variables (X, Y, Z), or more generally ordered m-tuples of random variables $(X_1, X_2, \ldots, X_m)$.

We begin by defining what we mean

Definition 2.34 (Random Variable)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability triple. Then, a $\mathbb{R}^m$ valued* **Random Variable (RV)***, say X, is a function from the sample space Ω to the set vectors $\mathbb{R}^m$*

$$X : \Omega \to \mathbb{R}^m$$

such that the inverse image of the half-open product intervals $(-\infty, x_1] \times \cdots \times (-\infty, x_m]$ for $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$ is an element of the collection of events $\mathcal{F}$, i.e., for every $x \in \mathbb{R}^m$ we have

$$X^{[-1]}(\ (-\infty, x_1] \times \cdots \times (-\infty, x_m]\ ) := \{\omega : X(\omega) \leq x\} \in \mathcal{F},$$

where we interpret the inequality $X \leq x$ to hold for all components. This definition can be summarised by the statement that a RV is an $\mathcal{F}$-measurable map from Ω to $\mathbb{R}^m$.

This definition looks remarkably similar to Definition 2.1, in fact it is the same. We can simply write it as follows.

Definition 2.35 (Abstract definition of a RV)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $(\mathbb{X}, \Sigma_X)$ be a space together with a sigma algebra Σ_X. Then we call a function $X : \Omega \to \mathbb{X}$ an $\mathbb{X}$-valued RV if*

$$X^{[-1]}(A) \in \mathcal{F}, \quad \text{for all } A \in \Sigma_X.$$

If we in the above definition just set $\mathbb{X} = \mathbb{R}^m$ and Σ_X the Borel sigma algebra (or any subset that generates the Borel sigma algebra, like half spaces) as in Section 1.2.2, we obtain Definition 2.34.

Let us leave the abstract notion and go back to $\mathbb{R}^m$ valued RV's. Specifically, let us define the corresponding distribution function:

Definition 2.36 (JDF)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let X be a $\mathbb{R}^m$ valued RV. Then the* **joint distribution function (JDF)** *or* **joint cumulative distribution function (JCDF)***, $F_X(x) : \mathbb{R}^m \to [0, 1]$ is defined as*

$$\begin{aligned} F_X(x) &= \mathbb{P}(\cap_{i=1}^m (X_i \leq x_i)) = \mathbb{P}(X_1 \leq x_1, \ldots, X_m \leq x_m) \\ &= P\left(\{\omega : X_1(\omega) \leq x_1, \ldots, X_m \leq x_m\}\right), \end{aligned}$$

where $X = (X_1, \ldots, X_m)$ and each $X_i \in \mathbb{R}$, and $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$.

Why do we need this? Well, lets say we do two measurements, X and Y. What does $\mathbb{P}(X \leq x, Y \leq y)$ mean? Let $Z = (X, Y)$, then consider this as a $\mathbb{R}^2$ valued RV, as such we can write

$$F_Z(z) = F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$

The above extends to any finite number of random variables.

Lets now say that we have two random variables X, Y and we wish to compute something like

$$\mathbb{E}[X + Y], \quad \mathbb{E}[XY], \quad \mathbb{E}[X^r Y^s]$$

etc. then we need the joint distribution function to compute the above expecations.

From the above motivation, we would expect that each component of a $\mathbb{R}^m$ random variable is again a random variable:

Theorem 2.37. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let Z be a $\mathbb{R}^m$ valued RV, then for any $i = 1, \ldots, m$, Z_i is a $\mathbb{R}$-valued RV. Its distribution function*

$$\begin{aligned} F_{Z_i}(z_i) &= \mathbb{P}(Z_i \leq z_i) \\ &= \mathbb{P}(Z_1 \leq \infty, \ldots, Z_{i-1} \leq \infty, Z_i \leq z_i, Z_{i+1} \leq \infty, \ldots, Z_m \leq \infty) \end{aligned}$$

is called the **marginal distribution**.

Proof. Let us prove this in the case when $m = 2$, that is, let $Z = (X, Y)$. Let us show that X is a $\mathbb{R}$ valued RV. We do this by showing that

$$X^{[-1]}((-\infty, x)) \in \mathcal{F}$$

What do we mean by $X^{[-1]}$? We mean,

$$\begin{aligned} X^{[-1]}((-\infty, x)) &= \{\omega \in \Omega : X(\omega) \leq x\} = \{\omega \in \Omega : X(\omega) \leq x, Y(\omega) < \infty\} \\ &= Z^{[-1]}((-\infty, x) \times (-\infty, \infty)) \in \mathcal{F} \end{aligned}$$

where the final step follows from the definition of Z being a $\mathbb{R}^2$-valued RV. □

We have seen the notion of independence of two events in Definition 1.18 or of a sequence of events in Definition 1.20. Recall that independence amounts to having the probability of the joint occurrence of the events to be given by the product of the probabilities of each of the events.

We can use the definition of independence of two events to define the independence of two random variables using their distribution functions.

Definition 2.38 (Independence of Two RVs)**.** *Consider an $\mathbb{R}^2$-valued RV $X := (X_1, X_2)$. Then the $\mathbb{R}$-valued RVs X_1 and X_2 are said to be independent or independently distributed if and only if*

$$\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2) = \mathbb{P}(X_1 \leq x_1)\, \mathbb{P}(X_2 \leq x_2)$$

or equivalently,

$$F_{X_1,X_2}(x_1, x_2) = F_{X_1}(x_1) F_{X_2}(x_2) \ ,$$

for any pair of real numbers $(x_1, x_2) \in \mathbb{R}^2$.

The above definition can be extended to a sequence of random variables in the same way as in Definition 1.20.

2.6.1 Discrete random vectors

Let us specify the above fairly abstract concepts into something tangible, we start with discrete random vectors.

Definition 2.39. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let Z be a $\mathbb{R}^m$ valued RV, we say that Z is a discrete random variable if it takes values in an enumerable set $\mathbb{Z}$. The* **joint probability mass function** *f_Z of Z is the function $f_Z : \mathbb{R}^m \to [0, 1]$ defined as follows*

$$f_Z(z) := \mathbb{P}(Z = z) = \mathbb{P}(\{\omega : Z(\omega) = z\}) = \begin{cases} \theta_h & \text{if } z = z_h \in \mathbb{Z} \\ 0 & \text{otherwise} \end{cases}$$

The above definition is easier to grasp in the 2-d case. Let $Z = (X, Y) \in \mathbb{R}^2$ and let $\mathbb{X}$ and $\mathbb{Y}$ be two enumerable sets and consider their product $\mathbb{Z} = \{(x_i, y_j) : i = 1, \ldots, j = 1, \ldots\} \subset \mathbb{R}^2$ and we have $\theta_{i,j} = \mathbb{P}(X = x_i, Y = y_j) > 0$, then the JPMF can be written as

$$\begin{aligned} f_{X,Y}(x, y) &:= \mathbb{P}(X = x, Y = y) = \mathbb{P}(\{\omega : X(\omega) = x, Y(\omega) = y\}) \\ &= \begin{cases} p_{i,j} & \text{if } x = x_i, y = y_i, (x_i, y_i) \in \mathbb{Z} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

In this case we can quite easily compute also the **marginal distribution** and the **marginal probability mass function** as follows: Say we have X and Y as above, and lets compute the marginal distribution for X, according to the definition it is

$$\begin{aligned} F_X(x) &= F_{X,Y}(x, \infty) = \mathbb{P}(X \leq x, Y \leq \infty) \\ &= \mathbb{P}(\{\omega : X(\omega) \leq x\} \cap \{\omega : Y(\omega) \leq \infty\}) \end{aligned}$$

now since Y is also discrete, i.e. taking values $y_1, y_2, \ldots$ we can write the above as

$$\{\omega : Y(\omega) \leq \infty\} = \bigcup_j \{\omega : Y(\omega) = y_j\} := \bigcup_j A_j$$

And clearly all A_j are mutually exclusive we can thus write

$$\begin{aligned} F_X(x) &= \mathbb{P}\left(\{\omega : X(\omega) \leq x\} \cap \left(\bigcup_j A_j\right)\right) \\ &= \sum_j \mathbb{P}(\{\omega : X(\omega) \leq x\} \cap A_j) \\ &= \sum_{x_i \leq x} \sum_j p_{i,j} \end{aligned}$$

if we now define $p_i = \sum_j p_{i,j}$ then we simply have

$$F_X(x) = \sum_{x_i \leq x} p_i$$

which is the same as if we had defined X alone as a discrete random variable. From the above it is also clear that the marginal probability mass function is given by

$$f_X(x) := \mathbb{P}(X = x) = \mathbb{P}(\{\omega : X(\omega) = x\}) = \begin{cases} p_i & \text{if } x = x_i \in \mathbb{X} \\ 0 & \text{otherwise} \end{cases}$$

2.6.2 Continuous random vectors

Arguably the continuous random variables are easier:

Definition 2.40. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let Z be a $\mathbb{R}^m$ valued RV. We say that Z is a continuous random variable if there exists a piecewise-continuous function $f_Z : \mathbb{R}^m \to [0, \infty)$, called the* **joint probability density function** *of Z such that*

$$F_Z(z) = \mathbb{P}(Z \leq z) = \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_m} f_Z(v_1, \ldots, v_m) dv_1 \ldots dv_m.$$

The above definition is easier to grasp in the 2-d case. Consider $Z = (X, Y) \in \mathbb{R}^2$ then the joint density function satisfies

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v) du dv$$

In this case we can quite easily compute also the **marginal distribution** and the **marginal probability mass function** as follows: Say we have X and Y as above, and lets compute the marginal distribution for X, according to the definition it is

$$\begin{aligned} F_X(x) &= F_{X,Y}(x, \infty) = \mathbb{P}(X \leq x, Y \leq \infty) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(u, v) du dv \end{aligned}$$

Now let us define

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, v) dv$$

then

$$F_X(x) = \int_{-\infty}^{x} f_X(u) du.$$

This means that by Definition 2.15 X is a continuous random variable with density f_X.

For independent random variables we get

Theorem 2.41. *Consider two independent continuous $\mathbb{R}$ valued RVs, X, Y. Then*

$$f_{X,Y}(x, y) = f_X(x) f_Y(y).$$

Proof. Now using Definition 2.38 together with properties of interals we get

$$\begin{aligned} F_{X,Y}(x,y) = F_X(x)F_Y(y) &= \int_{-\infty}^{x} f_X(u)du \int_{-\infty}^{y} f_Y(v)dv \\ &= \int_{-\infty}^{x} f_X(u) \left(\int_{-\infty}^{y} f_Y(v)dv \right) du \\ &= \int_{-\infty}^{y} \int_{-\infty}^{x} f_X(u)f_Y(v)dudv. \end{aligned}$$

□

As we saw in the chapter about $\mathbb{R}$ valued RVs, we can look at functions of RVs. The first thing we need is an extended notion of Borel.

Definition 2.42. *We say that a function $g : \mathbb{R}^m \to \mathbb{R}^k$ is Borel, if we denote Σ_r the Borel sigma algebra on $\mathbb{R}^r$, $r = 1, \ldots$ (see Section 1.2.2) and for every $A \in \Sigma_k$*

$$g^{[-1]}(A) \in \Sigma_m.$$

Lemma 2.43. *Let X be an $\mathbb{R}^m$ valued RV and let $g : \mathbb{R}^m \to \mathbb{R}^k$ be Borel. Then $g(X)$ is a $\mathbb{R}^k$ valued RV.*

Exercise 2.44. *Prove the above lemma in the case when $k = 1$.*

2.6.3 Properties of expectations

Now that we know what a joint distribution function is, we can make sense of for instance

$$\mathbb{E}[X + Y]$$

Let us list the immediate properties of the expectation here

Theorem 2.45. *Properties of the expectation*

1. *If $X \in L^1(\mathbb{P})$ is an $\mathbb{R}$ valued RV and $\alpha \in \mathbb{R}$, then*

$$\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X]$$

2. *If $X, Y \in L^1(\mathbb{P})$ are $\mathbb{R}$ valued RV, then*

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

3. *If $X, Y \in L^2(\mathbb{P})$ are independent $\mathbb{R}$ valued RV, then*

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

4. *If $X, Y \in L^1(\mathbb{P})$ are $\mathbb{R}$-valued RVs then if $X \leq Y$ a.s., then*

$$\mathbb{E}[X] \leq \mathbb{E}[Y].$$

5. *Let X be an $\mathbb{R}$-valued RV then if $A \subset \mathbb{R}$ is Borel, then*

$$\mathbb{E}[\mathbb{1}_A(X)] = \mathbb{P}(X \in A)$$

Proof. We will only prove the above in the case of continuous random variables: For 1 we note that the linearity of integrals we have

$$\mathbb{E}[\alpha X] = \int_{-\infty}^{\infty} (\alpha x) dF(x) = \alpha \int_{-\infty}^{\infty} x dF(x) = \alpha \mathbb{E}[X].$$

Let us now consider 2, which requires the concept of marginals,

$$\begin{aligned}\mathbb{E}[X+Y] = &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y) dF_{X,Y}(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x dF_{X,Y}(x,y) \\ &+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y dF_{X,Y}(x,y) = I_X + I_Y\end{aligned}$$

Now

$$\begin{aligned}I_X &= \int_{-\infty}^{\infty} x dF_{X,Y}(x,y) = \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f_{X,Y}(x,y) dy \right) dx \\ &= \int_{-\infty}^{\infty} x f_X(x) dx = \mathbb{E}[X].\end{aligned}$$

To prove 3 note that since X and Y are independent, then from Theorem 2.41 we have $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ which gives us

$$\begin{aligned}\mathbb{E}[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy dF_{X,Y}(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_X(x) f_Y(y) dx dy \\ &= \int_{-\infty}^{\infty} x f_X(x) \left(\int_{-\infty}^{\infty} y f_Y(y) dy \right) dx \\ &= \left(\int_{-\infty}^{\infty} x f_X(x) dx \right) \left(\int_{-\infty}^{\infty} y f_Y(y) dy \right) \\ &= \mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

□

2.6.4 L^p is a normed vector space

Definition 2.46. *We say that a set $\mathbb{X}$ is a vector space if there is a notion of addition, and a notion of multiplication with scalars, such that*

1. *For $X, Y \in \mathbb{X}$, we have $X + Y \in \mathbb{X}$,*
2. *For a scalar $\alpha \in \mathbb{R}$ we have $\alpha X \in \mathbb{X}$ if $X \in \mathbb{X}$.*

Definition 2.47. *We say that a set $\mathbb{X}$ is a normed vector space if it is a vector space and the following holds true: there is a function $\|\cdot\| : \mathbb{X} \to \mathbb{R}$ (called a* **norm***) such that*

1. *$\|X\| \geq 0$ for any $X \in \mathbb{X}$,*
2. *$\|X\| = 0$ if and only if $X = 0$,*
3. *For every vector $X \in \mathbb{X}$ and every $\alpha \in \mathbb{R}$, one has*

$$\|\alpha X\| = |\alpha| \|X\|$$

4. *The* triangle inequality *holds, that is, for $X \in \mathbb{X}$ and $Y \in \mathbb{X}$ we have*

$$\|X + Y\| \leq \|X\| + \|Y\|.$$

Theorem 2.48. *Let $(\Omega, \mathcal{F}, P)$ be a probability triple, then the set of random variables $L^p(\mathbb{P})$ for $1 \leq p < \infty$ is a normed vector space, with norm*

$$\|X\|_{L^p(\mathbb{P})} = \left(\mathbb{E}\left[|X|^p\right]\right)^{\frac{1}{p}}$$

and with $X = Y$ in $L^p(\mathbb{P})$ if $X(\omega) = Y(\omega)$ for a.e $\omega \in \Omega$ with respect to $\mathbb{P}$. That is $\mathbb{P}(\{\omega \in \Omega, X(\omega) = Y(\Omega)) = 0$.

How would we prove such a theorem? Well, we basically need to verify all conditions in Definitions 2.46 and 2.47. Verifying Definition 2.46 and conditions 1,2,3 of Definition 2.47 is left to you to verify. We will however prove the triangle inequality using Hölders inequality:

Theorem 2.49. *Let $X \in L^p(\mathbb{P})$ and $Y \in L^q(\mathbb{P})$ with $\frac{1}{p} + \frac{1}{q} = 1$, for $1 < p < \infty$, then*

$$\mathbb{E}\left[XY\right] \leq \mathbb{E}\left[|X|^p\right]^{1/p} \mathbb{E}\left[|Y|^q\right]^{1/q}.$$

Lemma 2.50. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function, and consider numbers x_1, x_2 and parameter $t \in [0, 1]$, then*

$$\phi(tx_1 + (1-t)x_2) \leq t\phi(x_1) + (1-t)\phi(x_2).$$

Lemma 2.51. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X \in L^1(\mathbb{P})$ be a $\mathbb{R}$-valued RV. Then if $\phi(X) \in L^1(\mathbb{P})$, where $\phi : \mathbb{R} \to \mathbb{R}$ is a convex function, we have*

$$\phi(\mathbb{E}\left[X\right]) \leq \mathbb{E}\left[\phi(X)\right].$$

We will prove the triangle inequality in the case that $p = 2$ (we leave $p \neq 2$ as an exercise), that is in Theorem 2.49 we have $p = q = 2$, hence

$$\begin{aligned}\mathbb{E}\left[|X+Y|^2\right] = \mathbb{E}\left[|X+Y||X+Y|\right] &\leq \mathbb{E}\left[(|X|+|Y|)|X+Y|\right] \\ &\leq \left(\mathbb{E}\left[|X|^2\right]^{1/2} + \mathbb{E}\left[|Y|^2\right]^{1/2}\right)\mathbb{E}\left[|X+Y|^2\right]^{1/2}\end{aligned}$$

in the first inequality we used the triangle inequality for the absolute value and in the last inequality we used Theorem 2.49 with $p = q = 2$. Dividing both sides by $\mathbb{E}\left[|X+Y|^2\right]^{1/2}$ gives

$$\mathbb{E}\left[|X+Y|^2\right]^{1/2} \leq \mathbb{E}\left[|X|^2\right]^{1/2} + \mathbb{E}\left[|Y|^2\right]^{1/2}$$

or equivalently

$$\|X+Y\|_{L^2(\mathbb{P})} \leq \|X\|_{L^2(\mathbb{P})} + \|Y\|_{L^2(\mathbb{P})},$$

which proves the triangle inequality. For the case of $p \neq 2$ one does

$$\mathbb{E}\left[|X+Y|^p\right] = \mathbb{E}\left[|X+Y||X+Y|^{p-1}\right] \leq \mathbb{E}\left[(|X|+|Y|)|X+Y|^{p-1}\right]$$

and then apply Hölders inequality with $q = \frac{p-1}{p}$.

Theorem 2.52. *The following are consequences of Hölders inequality*

1. *If $X \in L^p(\mathbb{P})$ and $Y \in L^q(\mathbb{P})$ then $XY \in L^1(\mathbb{P})$ if $\frac{1}{p} + \frac{1}{q} = 1$.*
2. *If $X \in L^r(\mathbb{P})$ then $X \in L^s(\mathbb{P})$ for $1 \leq s \leq r < \infty$.*
3. *If $X \in L^p(\mathbb{P})$ and $Y \in L^q(\mathbb{P})$ then $XY \in L^r(\mathbb{P})$ if $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$.*

Proof. Statement 1 follows immediately from Theorem 2.49.

To prove 2. First apply Hölders inequality with p to be chosen

$$\mathbb{E}\left[|X|^s\right] = \mathbb{E}\left[|X|^s \times 1\right] \leq \mathbb{E}\left[|X|^{sp}\right]^{1/p}\mathbb{E}\left[1^q\right]^{1/q}.$$

Now choose $p = r/s \geq 1$ which gives

$$\mathbb{E}\left[|X|^s\right] \leq \mathbb{E}\left[|X|^r\right]^{\frac{s}{r}}$$

or equivalently

$$\|X\|_{L^s(\mathbb{P})} \leq \|X\|_{L^r(\mathbb{P})}.$$

To prove 3. Apply Hölders inequality with the pair (s, h) and get

$$\mathbb{E}\left[|XY|^r\right] = \mathbb{E}\left[|X|^r|Y|^r\right] \leq \mathbb{E}\left[|X|^{rs}\right]^{1/s}\mathbb{E}\left[|Y|^{rh}\right]^{1/h}$$

now choose $h = \frac{p}{p-r}$ and $s = \frac{s-1}{s} = \frac{p}{r}$ which gives

$$\mathbb{E}\left[|XY|^r\right] \leq \mathbb{E}\left[|Y|^{\frac{pr}{p-r}}\right]^{(p-r)/p} \mathbb{E}\left[|X|^p\right]^{r/p}$$

the last term is finite but what about the expectation of Y, well note that

$$\frac{pr}{p-r} = q.$$

□

The same ideas as in Theorem 2.52 can be extended to a product of k random variables. Here the Hölder exponents become $\sum_{i=1}^{k} \frac{1}{p_i} = 1/r$. Try this out for yourself!

Functions of random variables

What we saw above with moments, i.e. powers of random variables is a special case of functions of random variables. Often we are interested in functions of random variables, like correlation etc. Let us define what we mean with the expectation of a random variable.

Definition 2.53 (Expectation of a function of a RV)**.** *The **Expectation** of a function $g(X)$ of a random variable X is defined as:*

$$\mathbb{E}(g(X)) := \int g(x)dF(x) = \begin{cases} \sum_x g(x)f(x) & \text{if } X \text{ is a discrete RV} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{if } X \text{ is a continuous RV} \end{cases}$$

provided $\mathbb{E}(g(X))$ exists, i.e., $\int |g(x)|dF(x) < \infty$.

Remark 2.54. *Notational convenience: Note in the above how we wrote*

$$\mathbb{E}\left[g(X)\right] = \int g(x)dF(x)$$

even if X was discrete. This can be made rigorous by the introduction of point measures (dirac measures or atomic measure), we will however skip that part in this course and instead work with the understanding that the integral is a sum in the discrete case.

The above can be taken as a defintion, but why does it make sense? It is actually something we can prove, and its called the law of the unconcious statistician. We will only prove this in the case of one-to-one monotone transformations. Lets do it for increasing functions f. If we think of the

transformation part we note that using the direct method that if X is a random variable and f is a function then for $Y = f(X)$ the distribution is

$$F_Y(y) = \mathbb{P}\left[X \in \{x : f(x) \in (-\infty, y]\}\right]$$

and $f_Y = \frac{d}{dy}F_Y(y)$, so by a change of variables we get $y = f(x)$

$$\mathbb{E}\left[Y\right] = \int y f_Y(y)dy = \int f(x) f_Y(f(x))\frac{df}{dx}dx$$

now, $\frac{d}{dx}F_Y(f(x)) = f_Y(f(x))\frac{df}{dx}$, but since f is monotone and increasing we get $F_Y(f(x)) = P(f(X) \leq f(x)) = \mathbb{P}\left[X \leq x\right] = F_X(x)$. This gives that

$$\mathbb{E}\left[Y\right] = \int y f_Y(y)dy = \int f(x) f_Y(f(x))\frac{df}{dx}dx = \int f(x) f_X(x)dx$$

Now all of this mean that $f(X)$ is a random variable and one can question in what $L^p(\mathbb{P})$ it lies, i.e. what value of p? Well, the existence of the expectation is equivalent to saying that $f(X) \in L^1(\mathbb{P})$. But actually, since if $X \in \mathbb{R}$ is a real valued random variable with density f_X where $f_X : \mathbb{R} \to \mathbb{R}$ then we can actually view everything from the perspective of integrability in $\mathbb{R}$. That is, we can define

$$L^p(F) = \{f \mid f : \mathbb{R} \to \mathbb{R}, \int |f(x)|^p dF(x) < \infty\}$$

this space of functions from $\mathbb{R}$ to $\mathbb{R}$ inherits all properties of $L^p(\mathbb{P})$ and as such it is a normed vector space.

2.6.5 Conditional Random Variables

Often we will have a condition where one of the two random variables that make up a random vector (X_1, X_2) already occurs and takes a value. And we might want to compute the probability of the occurrence of the other random variable given this conditional information. For this all we need to do is extend the idea of conditional probabilities to $\mathbb{R}^2$-valued random variables.

Definition 2.55. *Let (X, Y) be a $\mathbb{R}^2$ valued random variable, and let $A \subset \mathbb{R}$ be a Borel set such that $\mathbb{P}(Y \in A) > 0$ then define the conditional distribution function of X given that $Y \in A$ as*

$$F_{X|Y}(x \mid A) := \frac{\mathbb{P}(X \leq x, Y \in A)}{\mathbb{P}(Y \in A)}.$$

Lemma 2.56. *Let (X, Y) be a $\mathbb{R}^2$ valued random variable, then*

$$F_{X|Y}(x \mid (-\infty, y))F_Y(y) = \frac{F_{X,Y}(x, y)}{F_Y(y)}F_Y(y) = F_{X,Y}(x, y).$$

If Y is a discrete random variable and $f_Y(y) > 0$ the above definition is well defined for $A = \{y\}$ and we can write

$$F_{X|Y}(x \mid y) := \frac{\mathbb{P}(X \leq x, Y = y)}{\mathbb{P}(Y = y)}.$$

If however Y is continuous we know from Theorem 2.18 that $\mathbb{P}(Y = y) = 0$ and as such Definition 2.55 does not apply and we need another definition. Fix a point $y \in \mathbb{R}$ such that $f_Y(y) > 0$ and let $\epsilon > 0$, then define $U_\epsilon = [y - \epsilon, y + \epsilon]$, hence $\mathbb{P}(Y \in U_\epsilon) > 0$ since f_Y is piecewise continuous. We can now compute

$$\begin{aligned} F_{X|Y}(x|U_\epsilon) = \frac{\mathbb{P}(X \leq x, Y \in U_\epsilon)}{\mathbb{P}(Y \in U_\epsilon)} &= \frac{\int_{-\infty}^{x} \left(\int_{y-\epsilon}^{y+\epsilon} f_{X,Y}(u,v)dv \right) du}{\int_{y-\epsilon}^{y+\epsilon} f_Y(v)dv} \\ &= \frac{\int_{-\infty}^{x} \left(\frac{1}{2\epsilon} \int_{y-\epsilon}^{y+\epsilon} f_{X,Y}(u,v)dv \right) du}{\frac{1}{2\epsilon} \int_{y-\epsilon}^{y+\epsilon} f_Y(v)dv} \end{aligned}$$

from this we can "define"

$$F_{X|Y}(x \mid y) := \lim_{\epsilon \to 0} F_{X|Y}(x|U_\epsilon) = \int_{-\infty}^{x} \frac{f_{X,Y}(u,y)}{f_Y(y)} du.$$

Why "define", well basically we need to make sure what happens if we are at points of discontinuity of $f_{X,Y}(u,v)$ and $f_Y(y)$. This does not turn into any problems as there are only distinct points of discontinuity for $f_Y(y)$ and $f_{X,Y}(x,y)$ which all have probability zero and dont contribute to the integral.

Definition 2.57 (Conditional PDF or PMF)**.** *Let (X,Y) be a $\mathbb{R}^2$ valued RV. Then the* **conditional probability mass / density function** *is defined as*

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

if $f_Y(y) > 0$

Lemma 2.58. *Let (X,Y) be a $\mathbb{R}^2$ valued RV. Then*

$$f_{X|Y}(x \mid y) f_Y(y) = f_{X,Y}(x,y)$$

where the left hand side is interpreted as 0 if $f_Y(y) = 0$.

Exercise 2.59. *Consider two independent fair coin tosses (2 sided), i.e. $X, Y \sim$ Bernoulli$(1/2)$, and let 1 be heads and 0 is tails. Let $Z = X + Y$, what is the PMF of Z given X. Once, you have this, what is the joint PMF of (Z, X)?*

These definitions allows us to construct conditional expectations, like

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) dx.$$

But it also allows for the following very useful property

Theorem 2.60 (The tower property)**.** *Let (X, Y) be a $\mathbb{R}^2$ valued RV. Then*

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X].$$

In the above we introduced a new notation, namely $\mathbb{E}[X \mid Y]$, what is this? Denote $g(y) = \mathbb{E}[X \mid Y = y]$, then define

$$\mathbb{E}[X \mid Y] := g(Y).$$

Proof. We will also prove this only for continuous RVs, the discrete is an exercise!!

Lets begin by writing down the LHS

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[g(Y)] &= \int_{-\infty}^{\infty} g(y) f_Y(y) dy = \int_{-\infty}^{\infty} \mathbb{E}[X \mid Y = y] f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) dx \right) f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) f_Y(y) dx dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y) dx dy = \mathbb{E}[X].
\end{aligned}$$

In the last stages we used Lemma 2.58 and the definition of marginal density. □

2.6.6 Mixed random variables

We have dealt with both discrete and continuous random variables, but what about mixed random variables? For instance, if X is continuous and Y is discrete, what is (X, Y)? Actually it is neither, but we have all the tools to deal with it, we just have to mix the two concepts.

2.7 Examples Of Modeling

Now that we have developed the language of

1. Probability
2. Random variables
3. Dependence and independence
4. Conditional distribution etc.

We can take some common problems encountered in data science and model it.

2.7.1 Email spam filtering

Lets say that you wish to construct an email filter that takes as input an email and predicts if it is spam or not.

- The experiment, the next incoming email.
- $\Omega = \{\text{All strings of length 1000}\}$, that is, an outcome is one email (We limit to the first 1000 characters, but that is arbitrary).
- Ω is finite, so the σ-algebra is just all subsets of Ω, i.e. the power set.
- The probability measure $\mathbb{P}$ is unknown, but it is there, not all emails are equally probable. This is crucially the case, because we want to estimate probabilities based on the data.
- To each email, there is a function that tells us if it is a spam or not. We could take this as X and $X : \Omega \to \{0, 1\}$, 0 would be not spam and 1 would be spam.

This is our first setup of the problem. We are recieving an email, this is the experiment. It is either spam or not spam, and this is the value of the RV. X. However our initial problem was to predict if the email was a spam or not, i.e. we wish to use some information in the email and construct a decision function, which takes the email and outputs if it is a 1 or a 0. We pretty much would like to know how the unknown X works using a set of observations.

We can use many things, but one of the simplest would be. Let $Z : \Omega \to \{0, 1\}$ be a function that is 1 if the email contains the word "donate" and hope that Z is a good predictor of X. But how do we phrase that?

$$\mathbb{P}(X = 1 \mid Z = 1) > \mathbb{P}(X = 1)$$

if the above inequality holds, we know that if "donate" is in the email, then it is more likely to be spam. It is thus a valuable predictor. If $\mathbb{P}(X = 1 \mid Z = 1) = 1$ then it is a perfect predictor, since if you know $Z = 1$ then you know that $X = 1$.

If however

$$\mathbb{P}(X = 1 \mid Z = 1) = \mathbb{P}(X = 1)$$

then X and Z are independent, i.e. knowing Z gives nothing about X.

You have just seen the simplest case of a word based prediction model, in practice however you would use several words and in that case its called a **bag of words** model. One particularly successful model is **Naive Bayes** which is very close to what we did above, but with more words.

2.7.2 Number of website requests during a day

Lets say that you are monitoring the number of website requests, and let us assume that each request requires some processing work to be commissioned. That is, what you want to know is, how much resources should I give?

In this particular problem, the goal would be to predict the number of requests on a given day in advance so that there is time to add resources.

- The experiment, recording the website requests during a day. For each website request you log a bunch of information (where it came from, what day, what the request was, etc.)
- $\Omega = \{\text{all sequences of all valid website requests}\}$. Since we are recording during a day, we can have any number of requests. Often, however this set is enumerable. The choice here is fairly arbitrary of what Ω should be, we could just as well define Ω to be an abstract set which contains all possible outcomes of the experiment. With this we mean, all possible information that can be recorded is recorded in ω.
- Here we can as σ-algebra, $\mathcal{F}$ choose all possible subsets of Ω, i.e. the power set. If we would go the route of taking Ω to be the abstract set, then $\mathcal{F}$ is unknown, but it is here.
- Again, the probability measure $\mathbb{P}$ is here, but unknown. In this case, very very very hard to estimate, so we will not even try to.
- $X : \Omega \to \{0, 1, 2, \ldots\}$, a function that takes a sequence of valid website requests and outputs the number of requests. With the simpler Ω we know that X is measurable, however for the abstract Ω we have to assume it is. (This is crucial, we assume that what we observe is a random variable!!! This is a modeling assumption.)

Our goal here is to find a good way to predict X beforehand. Specifically, since it can take on many values we want to estimate the distribution function

$$F_X(x) = \mathbb{P}(X \leq x).$$

Lets say that if $X > 10$ we need to commission extra resources, that is we would then like to know

$$\mathbb{P}(X > 10) = 1 - F_X(10).$$

Let's say that some information about the outcome can be known, like the day of the week. Call this random variable $Z : \Omega \to \{1, 2, 3, 4, 5, 6, 7\}$, where the number is the day of week. It is likely that X and Z are not independent and that the conditional distribution of X given Z, which is

$$F_{X|Z}(x \mid z)$$

can instead be estimated. It is fairly reasonable to assume that the amount of traffic is different for different days, (compare wednesday to saturday for a work related website). Here we can take decisions based on

$$\mathbb{P}(X > 10 \mid Z = z) = 1 - F_{X|Z}(x \mid z).$$

In the case of a more abstract Ω we could envision other random variables Y which contains some information, say its 1 if today is a holiday. We could then try to estimate

$$\mathbb{P}(X > 10 \mid Z = z, Y = y) = 1 - F_{X|Z,Y}(x \mid z, y).$$

2.7.3 Summary

You have now seen to common estimation examples where we defined what could be known and left the unknown. We assumed that the unknown quantities existed (this is a modeling assumption).

Here is a modeling procedure which is always defined

- Lets say you are doing some experiment, say, looking at an image.
- Let Ω be all possible outcomes where you do not specify what an outcome is, it can be seen as unknown.
- Assume that there are unknown $\mathcal{F}$ and $\mathbb{P}$, that is we assume that underlying our problem there is a $(\Omega, \mathcal{F}, \mathbb{P})$, i.e. a probability triple, which is unknown.

- We assume that some information about the image, say if it contains a cat or not (1 or 0) is knowable, call that value X. We further assume that X is a random variable with respect to the probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.

Now, since everything is unknown but assumed to be there, and all we care about is whether or not there is a cat in the image. We could try to estimate

$$F_X(x) = \mathbb{P}(X \leq x)$$

and since X only takes values 0 and 1, it would be enough to estimate

$$\mathbb{P}(X = 1).$$

If we do independent repeats of our experiment and record if its a cat or not, then it is reasonable to expect that the average would be a good estimate for $\mathbb{P}(X = 1)$. Point being, we did not need to know $(\Omega, \mathcal{F}, \mathbb{P})$ in order to estimate $\mathbb{P}(X = 1)$.

This idea is underlying many models, we assume that our repeated experiments are independent and that the value we are looking at is a random variable. This means we can "ignore" what $(\Omega, \mathcal{F}, \mathbb{P})$ actually is, but know that we assume its well defined and there.

Chapter 3

Concentration and Limits

3.1 Concentration inequalities

In probability theory, concentration inequalities provide bounds on how a random variable deviates from some value (typically, its expected value).

Theorem 3.1 (**Markov's inequality**). *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability triple and let* $X \in L^1(\mathbb{P})$ *be a non-negative* $\mathbb{R}$*-valued RV. Then,*

$$\mathbb{P}(X \geq \epsilon) \leq \frac{\mathbb{E}(X)}{\epsilon}, \qquad \text{for any} \quad \epsilon > 0 \ . \tag{3.1}$$

Proof. Let $A_\epsilon = [\epsilon, \infty)$ then by Lemma 2.8

$$\mathbb{1}_{A_\epsilon}(x) + \mathbb{1}_{A_\epsilon^c}(x) = 1$$

as such we can write

$$X = X\mathbb{1}_{A_\epsilon}(X) + X\mathbb{1}_{A_\epsilon^c}(X) \geq X\mathbb{1}_{A_\epsilon}(X) \geq \epsilon\mathbb{1}_{A_\epsilon}(X).$$

Now the inequalities are preserved when taking the expectation of both sides (Theorem 2.45), and we get

$$\mathbb{E}[X] \geq \epsilon\,\mathbb{E}[\mathbb{1}_{A_\epsilon}(X)] = \epsilon\,\mathbb{P}(X \in A_\epsilon) = \epsilon\,\mathbb{P}(X \geq \epsilon)$$

□

Let us look at some immediate consequences of Markov's inequality.

Proposition 3.2 (**Chebychev's inequality**). *For* **any** *RV X and any $\epsilon > 0$,*

$$\mathbb{P}(|X| > \epsilon) \leq \frac{\mathbb{E}(|X|)}{\epsilon}$$
$$\mathbb{P}(|X| > \epsilon) = \mathbb{P}(X^2 \geq \epsilon^2) \leq \frac{\mathbb{E}(X^2)}{\epsilon^2}$$
$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \epsilon) = \mathbb{P}((X - \mathbb{E}(X))^2 \geq \epsilon^2) \leq \frac{\mathbb{E}(X - \mathbb{E}(X))^2}{\epsilon^2} = \frac{\mathbb{V}(X)}{\epsilon^2}$$

In the above we interpret the expectations as ∞ if they dont exist. However if we want finite expressions then we would need $X \in L^1(\mathbb{P})$ for the first inequality and $X \in L^2(\mathbb{P})$ for the second and third inequality.

Proof. All three forms of Chebychev's inequality are mere corollaries (careful reapplications) of Markov's inequality. □

Definition 3.3. *We say that the sequence $X_1, X_2, \ldots$ of $\mathbb{R}$-valued RVs is an independent and identically distributed (i.i.d.) sequence of $\mathbb{R}$-valued random variables with distribution F, if for any $n \in \mathbb{N}$ we have that $X_1, \ldots, X_n \sim F$ and that $Z = (X_1, \ldots, X_n)$ is an $\mathbb{R}^n$-valued RV with distribution function*

$$F_Z(x_1, \ldots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n) = \prod_{i=1}^{n} F(x_i).$$

We usually denote this with $X_1, \ldots, X_n \overset{\text{IID}}{\sim} F$.

Remark 3.4. *This can quite easily be extended to sequences i.i.d. random vectors with the trivial modifications.*

One of the most fundamental aspects of statistics is the concept of "concentration of measure". As we saw in Chapter 1 one can motivate the concept of probability as a long-term relative frequency. That is if we toss a fair coin we expect that $N(\mathtt{H}, n) \to \frac{1}{2}$ as $n \to \infty$, however for any finite number of tosses there is a probability that this can deviate from $\frac{1}{2}$ (we could for instance observe the unlikely event that we get all heads). We do however expect that the probability of a large deviation to become smaller as we observe more tosses, this is the phenomenon of concentration of measure. We will begin with a "helper lemma" and then move on to prove a fundamental concentration inequality called Hoeffdings inequality.

Lemma 3.5 (**Hoeffdings lemma**). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and suppose that X is a $\mathbb{R}$-valued RV such that $\mathbb{P}(X \in [a, b]) = 1$ for $a < b$.*

Then, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right)$$

Proof. First we need to make sure that the left hand side is OK. First we need to make sure that $X \in L^1(\mathbb{P})$ and then we need to make sure that $e^{\lambda X} \in L^1(\mathbb{P})$. However the boundedness condition $a \leq X \leq b$ immediately implies this, since

$$a = \mathbb{E}[a] \leq \mathbb{E}[X] \leq \mathbb{E}[b] = b$$

and the same thing holds for the exponential, i.e.

$$e^{\lambda a} = \mathbb{E}[e^{\lambda a}] \leq \mathbb{E}[e^{\lambda X}] \leq \mathbb{E}[e^{\lambda a}] \leq e^{\lambda a}.$$

Now since $e^{\lambda x}$ is convex we have

$$e^{\lambda x} \leq \frac{b-x}{b-a}e^{\lambda a} + \frac{x-a}{b-a}e^{\lambda b}$$

for all $a \leq x \leq b$. Hence if we let $Y = X - \mathbb{E}[X]$ we get

$$\mathbb{E}[e^{\lambda Y}] \leq \frac{b-\mathbb{E}[Y]}{b-a}e^{\lambda a} + \frac{\mathbb{E}[Y]-a}{b-a}e^{\lambda b} = \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b}.$$

Now let $h = \lambda(b-a)$, $p = \frac{-a}{b-a}$ and $L(h) = -hp + \ln(1-p+pe^h)$, then

$$\frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b} = e^{L(h)}. \tag{3.2}$$

Let us bound L from above, we will do that using basic calculus, first note that

$$L(0) = \ln(1) = 0$$

and

$$L'(h) = -p + \frac{pe^h}{1-p+pe^h} \implies L'(0) = 0.$$

Let us now consider the second derivative,

$$L''(h) = \frac{pe^h}{1-p+pe^h} - \frac{(pe^h)^2}{(1-p+pe^h)^2}$$

this is of the form $y - y^2$ which cannot be larger than $1/4$, as such we get using Taylors theorem, that

$$L(h) \leq \frac{h^2}{8} = \frac{\lambda^2(b-a)^2}{8}$$

from (3.2) we now complete the lemma. □

Theorem 3.6 (Hoeffdings inequality (simple case))**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, \ldots, X_n \overset{\text{IID}}{\sim} F$ be $\mathbb{R}$-valued RVs such that $\mathbb{P}(X_i \in [a,b]) = 1$, then for any $\epsilon > 0$ we get for $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$,*

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \geq \epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

Proof. Let $S_n = \sum_{i=1}^n X_i$. Let $s, t > 0$ be positive numbers to be chosen, then using Theorem 3.1 we get

$$\begin{aligned}
\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) &= \mathbb{P}(e^{s(S_n - \mathbb{E}[S_n])} \geq e^{st}) && (3.3)\\
&\leq e^{-st}\,\mathbb{E}(e^{s(S_n - \mathbb{E}[S_n])}) && (3.4)\\
&= e^{-st}\prod_{i=1}^n \mathbb{E}(e^{s(X_i - \mathbb{E}[X_i])}) && (3.5)
\end{aligned}$$

where in the last step we used the independence of $X_1, \ldots$ together with Theorem 2.45. Now using Lemma 3.5 with $\lambda = s$ for each term in the product, we get

$$e^{-st}\prod_{i=1}^n \mathbb{E}(e^{s(X_i - \mathbb{E}[X_i])}) \leq e^{-st} e^{s^2(b-a)^2 n/8} \tag{3.6}$$

Notice now, that the value s was arbitrarily chosen and we can choose it to make the right hand side as small as possible. That is we want to minimize

$$h(s) = s^2\frac{n(b-a)^2}{8} - st. \tag{3.7}$$

This function is minimized at $s^* = \frac{4t}{n(b-a)^2}$, plugging that in we get

$$h(s^*) = s^2\frac{n(b-a)^2}{8} - st = -\frac{2t^2}{n(b-a)^2}. \tag{3.8}$$

Assembling (3.3) and (3.6)–(3.8) we get

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq e^{-\frac{2t^2}{n(b-a)^2}}.$$

Replacing $t = n\epsilon$ we get

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \geq \epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}},$$

which proves the theorem. □

Corollary 3.7. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, \ldots, X_n \overset{\text{IID}}{\sim} F$ $\mathbb{R}$-valued RVs such that $\mathbb{P}(X_i \in [a,b]) = 1$, then for any $\epsilon > 0$ we get for $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$,*

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \leq -\epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}},$$

furthermore

$$\mathbb{P}(\left|\overline{X}_n - \mathbb{E}[\overline{X}_n]\right| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}},$$

Proof. Exercise!! Hint: In Theorem 3.6 we did not assume anything about the sign of X_i. □

Let us look at an application of this, namely that of constructing confidence regions:

Lemma 3.8. *[Estimating p in Bernoulli] Let $X_1, \ldots, X_n \overset{\text{IID}}{\sim} \text{Bernoulli}(p)$. Then for $\alpha \in (0,1)$ we have for $\delta = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{2}\ln\left(\frac{2}{\alpha}\right)}$*

$$\mathbb{P}(\overline{X}_n - \delta \leq p \leq \overline{X}_n + \delta) \geq 1 - \alpha.$$

Remark 3.9. *In the above, if we fix $\alpha = 0.05$ we get $\delta \approx \frac{1.36}{\sqrt{n}}$.*

Proof. We wish to apply Corollary 3.7. Note that $a = 0, b = 1$ in the Bernoulli case, hence for a fix α and fix n we need to solve

$$\alpha = 2e^{-2n\epsilon^2}$$

with respect to ϵ. We get

$$\epsilon = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{2}\ln\left(\frac{2}{\alpha}\right)}$$

as such we get from Corollary 3.7 that

$$\begin{aligned}\mathbb{P}\left(\overline{X}_n - \frac{1}{\sqrt{n}}\sqrt{\frac{1}{2}\ln\left(\frac{2}{\alpha}\right)} \leq p \leq \overline{X}_n + \frac{1}{\sqrt{n}}\sqrt{\frac{1}{2}\ln\left(\frac{2}{\alpha}\right)}\right) \\ = 1 - \mathbb{P}\left(\left|\overline{X}_n - p\right| > \frac{1}{\sqrt{n}}\sqrt{\frac{1}{2}\ln\left(\frac{2}{\alpha}\right)}\right) \geq 1 - \alpha.\end{aligned}$$

□

Remark 3.10. *Computing multiple intervals for random variables which are not necessarily independent is quite easy using the union bound. Actually, we did it above when going from the one-sided to the two sided inequality, see Corollary 3.7.*

Suppose we have m sequences of random variables $Z_1 = (X_{11}, X_{12}, \dots, X_{1n})$, $\dots, Z_m = (X_{m1}, \dots, X_{mn})$. Where for each i the sequence Z_i is i.i.d. but Z_i and Z_j are not necessarily independent (they could all be the same for instance). Assume that each one of them satisfies

$$\mathbb{P}(|\frac{1}{n}\sum_{j=1}^{n} X_{ij} - \mathbb{E}[X_{1j}]| \geq \epsilon) \leq C_i$$

for every i and for some number C_i. Then

$$\mathbb{P}(|\frac{1}{n}\sum_{j=1}^{n} X_{ij} - \mathbb{E}[X_{1j}]| \geq \epsilon \text{ for some } i) \leq \sum_{i=1}^{m} C_i$$

the complement of this is

$$\mathbb{P}(|\frac{1}{n}\sum_{j=1}^{n} X_{ij} - \mathbb{E}[X_{1j}]| < \epsilon \text{ for all } i) \geq 1 - \sum_{i=1}^{m} C_i.$$

So for instance if all of where Bernoulli(p_i) *then from the above and Lemma 3.8 we get that*

$$\mathbb{P}(\frac{1}{n}\sum_{j=1}^{n} X_{ij} - \delta \leq p_i \leq \frac{1}{n}\sum_{j=1}^{n} X_{ij} + \delta \text{ for all } i) \geq 1 - \alpha.$$

where $\delta = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{2}\ln\left(\frac{2m}{\alpha}\right)}$. This means that m appears in the logarithm, and the increase of δ w.r.t. m is fairly slow. We will use this fact later when we produce multiple intervals for different metrics. This is equivalent to **Bonferroni correction** *in multiple testing.*

So the Hoeffding inequality is actually quite useful (extremely), but the restriction that the random variables are bounded is a heavy restriction. However, if we look at the proof of Theorem 3.6 we note that everything follows from Lemma 3.5, so if the estimate in Lemma 3.5 holds, then so should Theorem 3.6. With this in mind, let us define

Definition 3.11. *A $\mathbb{R}$ valued random variable X is said to be* **sub-Gaussian** *with parameter λ if*

$$\mathbb{E}[e^{s(X-\mathbb{E}[X])}] \leq e^{\frac{s^2\lambda^2}{2}}, \quad \text{for all } s.$$

and a local version

Definition 3.12. *A $\mathbb{R}$ valued random variable X is said to be* **sub-exponential** *with parameter λ if*

$$\mathbb{E}\left[e^{s(X-\mathbb{E}[X])}\right] \leq e^{\frac{s^2\lambda^2}{2}}, \qquad \text{for all } |s| \leq \frac{1}{\lambda}.$$

Both of these can be used in the proof of Theorem 3.6, however in the case of sub-exponential we have to take care of the restriction on s which actually yields a weaker bound.

Theorem 3.13. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, \ldots, X_n \overset{\text{IID}}{\sim} F$ be $\mathbb{R}$-valued sub-Gaussian RVs with parameter σ then for any $\epsilon > 0$ we get for $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$,*

$$\mathbb{P}(\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right] \geq \epsilon) \leq e^{-\frac{n\epsilon^2}{2\sigma^2}}.$$

For the sub-exponential case we get a weaker bound for the tails. The reason for this is the fact that the bound on $\mathbb{E}^{sX}$ only holds for small s, the resulting estimate thus differentiates between small and big ϵ. We can see in the estimate below that for large ϵ the tail is exponential, i.e. $e^{-\epsilon}$, this in one of the reasons for the name sub-exponential.

Theorem 3.14. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, \ldots, X_n \overset{\text{IID}}{\sim} F$ be $\mathbb{R}$-valued sub-exponential RVs with parameter λ then for any $\epsilon > 0$ we get for $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$,*

$$\mathbb{P}(\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right] \geq \epsilon) \leq e^{-\frac{\epsilon^2 n}{2\lambda^2}} \wedge e^{-\frac{(\epsilon+1)n}{2\lambda}}.$$

Proof. Proceeding as in the proof of (3.3) we get

$$\mathbb{P}(S_n - \mathbb{E}\left[S_n\right] \geq t) = e^{-st}\prod_{i=1}^n \mathbb{E}(e^{s(X_i - \mathbb{E}[X_i])}) \tag{3.9}$$

Consider now $s \leq \frac{1}{\lambda}$ and apply Definition 3.12 for each term in the product, we get

$$e^{-st}\prod_{i=1}^n \mathbb{E}(e^{s(X_i - \mathbb{E}[X_i])}) \leq e^{-st}e^{\frac{s^2\lambda^2 n}{2}} \tag{3.10}$$

If we proceed as in Theorem 3.6 we consider

$$h(s) = \frac{s^2\lambda^2 n}{2} - st$$

and note again that the function is minimized at $s^* = \frac{t}{n\lambda^2}$. If now $s^* \leq \frac{1}{\lambda}$ we get

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq e^{-\frac{t^2}{2n\lambda^2}} \tag{3.11}$$

However if $s^* > 1/\lambda$, that is if $t > n\lambda$ we get can only take

$$h(1/\lambda) = \frac{n}{2} - \frac{1}{\lambda}t < -\frac{n}{2} - \frac{t}{2\lambda}$$

that is,

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq e^{-\frac{n}{2} - \frac{t}{2\lambda}}. \tag{3.12}$$

Assembling (3.11) and (3.12) we get

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq e^{-\frac{t^2}{2n\lambda^2}} \wedge e^{-\frac{n}{2} - \frac{t}{2\lambda}}, \tag{3.13}$$

where $a \wedge b = \min(a, b)$. Now again replacing $t = n\epsilon$ we get

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \geq \epsilon) \leq e^{-\frac{\epsilon^2 n}{2\lambda^2}} \wedge e^{-\frac{(\epsilon+1)n}{2\lambda}}.$$

□

Lemma 3.15. *The following properties hold*

1. *Let X be a sub-Gaussian RV with parameter λ, then αX is sub-Gaussian with parameter $|\alpha|\lambda$.*
2. *Let X be a sub-exponential RV with parameter λ, then αX is sub-exponential with parameter $|\alpha|\lambda$.*
3. *A sub-Gaussian RV X with parameter λ is sub-Exponential with parameter λ.*
4. *A bounded RV X, i.e. $\mathbb{P}(X \in [a, b]) = 1$, then X is sub-Gaussian with parameter $(b-a)/2$. Specifically a Bernoulli RV is sub-Gaussian with parameter $1/2$.*
5. *If X is sub-Gaussian with parameter λ then $Z = X^2$ is sub-exponential with parameter $8\lambda^2$.*
6. *if X, Y are independent and sub-Gaussian with parameter σ_1, σ_2, then $X + Y$ is sub-Gaussian with parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$.*

Proof. Let $Z = X^2 - \mathbb{E}[X^2]$, use the power series representation of the exponential (we have not really gone through the theory for this one, but this is OK by the dominated convergence theorem which is outside the scope of this course)

$$\mathbb{E}[e^{sZ}] = 1 + \sum_{k=2}^{\infty} \frac{s^k \mathbb{E}[Z^k]}{k!}$$

First use the elementary fact that $(a+b)^k \leq 2^{k-1}(a^k + b^k)$ (for $k > 1$ since the power function is convex), we get

$$\mathbb{E}[Z^k] = \mathbb{E}[(X^2 - \mathbb{E}[X^2])^k] \leq 2^{k-1}(\mathbb{E}[X^{2k}] + (\mathbb{E}[X^2])^k)$$

Now by Hölders inequality Theorem 2.49

$$\mathbb{E}[X^2]^k \leq \mathbb{E}[X^{2k}]$$

Thus we get

$$\mathbb{E}[e^{sZ}] \leq 1 + \sum_{k=2}^{\infty} \frac{s^k 2^k \mathbb{E}[X^{2k}]}{k!}$$

Now, note that Theorem 3.13 gives us bounds for the moments of X above, i.e. using the fact that (for the first inequality see Exercise 5.18) we get

$$\begin{aligned}\mathbb{E}[X^k] &= \int_0^\infty \mathbb{P}(|X|^k > t)dt \leq 2\int_0^\infty e^{-\frac{t^{2/k}}{2\lambda^2}} dt = (2\lambda^2)^{k/2} k \int_0^\infty e^{-u} u^{k/2-1} du \\ &= (2\lambda^2)^{k/2} k\Gamma(k/2)\end{aligned}$$

Going back to our problem we get (using that $k\Gamma(k) = k!$)

$$\begin{aligned}\mathbb{E}[e^{sZ}] &\leq 1 + \sum_{k=2}^{\infty} \frac{s^k 2^k (2\lambda^2)^k 2k!}{k!} \\ &\leq 1 + 2\sum_{k=2}^{\infty} (4s\lambda^2)^k \\ &\leq 1 + 2(4s\lambda^2)^2 \sum_{k=0}^{\infty} (4s\lambda^2)^k \\ &\leq 1 + 64s^2\lambda^4 \leq e^{32s^2\lambda^4}\end{aligned}$$

the last sum is a geometric sum and is less than 2 if $8s\lambda^2 < 1$, i.e. $s < \frac{1}{4\lambda^2}$. Thus we see that X^2 is sub-exponential with parameter $8\lambda^2$. □

Distribution	sub-exponential	sub-Gaussian
Gaussian	Yes	Yes
Bernoulli	Yes	Yes
Uniform	Yes	Yes
Bounded	Yes	Yes
Exponential	Yes	No
χ^2	Yes	No
Weibull ($k \geq 1$)	Yes	No
Laplace	Yes	No
Pareto	No	No
Lognormal	No	No

Figure 3.1: Examples of distributions that are sub-exponential and sub-Gaussian

The question is now, what distributions are sub-Gaussian and which are sub-exponential? See Fig. 3.1. We will be using these concentration inequalities in the course to prove that the algorithms we are interested in is actually doing what we want with high probability.

Exercise 3.16. *For the Poisson distribution, we have*

$$\mathbb{E}\left[e^{sX}\right] = e^{\lambda(e^s - 1)}$$

is this sub-Gaussian, sub-exponential or neither?

3.1.1 Random variables that are not exponentially integrable*

Both the sub-Gaussian and sub-exponential rely on the fact that $\mathbb{E}\left[e^{sX}\right] < \infty$, if we rewrite this for a continuous RV we get

$$\int_{-\infty}^{\infty} e^{sx} f_X(x) dx < \infty$$

hence we need either that f_X has finite support or we need that it decays exponentially at infinity. This is why the sub-exponential for instance has the restriction on the size of s, as in that case f_X behaves like $e^{-\frac{1}{\lambda}|x|}$, that is

$$\int_{-\infty}^{\infty} e^{sx} e^{-\frac{1}{\lambda}|x|} dx < \infty, \quad \text{if and only if } s < \frac{1}{\lambda}.$$

One could hope that there is still some concentration of measure in this case. Our first observation is that the exponential integrability implies that all moments exists, i.e. $X \in L^p(\mathbb{P})$ for every $1 \leq p < \infty$. However, what if we only have $X \in L^p(\mathbb{P})$ for some $1 \leq p < \infty$, what can we say then?

Theorem 3.17. *Lets say that* $X_1, \ldots, X_n \overset{\text{IID}}{\sim} F$ *be* $\mathbb{R}$ *valued RVs. Suppose that* $X_i \in L^{2s}(\mathbb{P})$ *and*

$$|\,\mathbb{E}\,[(X_i - \mathbb{E}\,[X_i])^r]| \leq \sigma^2 r!, \text{ for } r = 2, 3, \ldots, 2s$$

for a positive integer $s > 1$. *Then if* $\epsilon \in [0, \sqrt{2}n\sigma^2]$ *and* $s \leq n\sigma^2$ *we have*

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}\,[\overline{X}_n]| \geq \epsilon) \leq \left(\frac{4s\sigma^2}{\epsilon^2 n}\right)^{s/2}.$$

If further, $s \geq \epsilon^2/(2n\sigma^2)$ *then we also have*

$$\mathbb{P}(|\overline{X_n}| \geq \epsilon) \leq 3e^{-\frac{\epsilon^2 n}{12\sigma^2}}.$$

Proof. The proof uses Theorem 3.1 as Theorem 3.6, however now we cannot use the exponential, instead we use powers. The power of $\mathbb{E}\,[|\sum_i X_i|^k]$ has to computed, this can be done by carefully checking the combinatorics of the terms and using the independence assumption. See the FDS book Theorem 12.5. □

3.2 Convergence of Random Variables

This important topic is concerned with the limiting behavior of sequences of RVs. We want to understand what it means for a sequence of random variables $\{X_n\}_{n=1}^{\infty} := X_1, X_2, \ldots$ to converge to another random variable X, when all RVs are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

$$\{X_i\}_{i=1}^{n} := X_1, X_2, X_3, \ldots X_{n-1}, X_n \qquad \text{as } n \to \infty\ .$$

From a statistical or decision-making viewpoint, $n \to \infty$ is associated with the amount of data or information $\to \infty$. More abstractly, we are interested in what happens to the limiting RV $X := \lim_{n\to\infty} X_n$ when given the DFs $F_n(x)$ for each X_n.

We need different notions of convergence to characterize such a behavior: two simplest behaviors are that the sequence eventually takes a constant value θ, i.e. X_n approaches $X \sim$ Point Mass(θ) RV, or that values in the sequence continue to change but can be described by an unchanging probability distribution, i.e., X_n approaches $X \sim F(x)$. See https://en.wikipedia.org/wiki/Convergence_of_random_variables.

Definition 3.18 (**Convergence in Distribution** (or Weakly, or in Law))**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability triple and let* $X_1, X_2, \ldots,$ *be a sequence of RVs*

and let X be another RV. Let F_n denote the DF of X_n and F denote the DF of X. The we say that X_n converges to X in distribution, and write:

$$X_n \rightsquigarrow X$$

if for any real number t at which F is continuous,

$$\lim_{n\to\infty} F_n(t) = F(t).$$

The above limit, by (2.2) in our Definition 2.2 of a DF, can be equivalently expressed as follows:

$$\lim_{n\to\infty} \mathbb{P}(\ \{\omega : X_n(\omega) \le t\}\ ) = \mathbb{P}(\ \{\omega : X(\omega) \le t\}\ ).$$

Convergence in distribution does not in general imply that the sequence of corresponsing probability density functions will also converge. Consider for example RV X_n with density $\mathbb{1}_{(0,1)}(x)(1 - \cos(2\pi n x))$. These RVs converge in distribution to $X \sim \text{Uniform}(0, 1)$, but their densities (PDFs) do not converge at all as evident in Fig. 3.2.

The other way around is however true:

Lemma 3.19. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, X_2, \ldots,$ be a sequence of RVs and let X be another RV. Let f_n denote the PDF of X_n and f denote the PDF/PMF of X. If*

$$f_n(x) \to f(x), \quad \forall x \in \mathbb{R},$$

then

$$X_n \rightsquigarrow X.$$

From Lemma 3.19 we see that for a discrete sequence of RVs X_n to converge in distribution to another discrete RV X taking values in $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$, it is sufficient to show that $\lim_{n\to\infty} \mathbb{P}(X_n = x) = \mathbb{P}(X = x)$ for each $x \in \mathbb{Z}_+$. We will use this fact to prove why we can approximate Binomial RVs by a Poisson under some limiting conditions.

Theorem 3.20. *Let $X_n \sim \text{Binomial}(n, \lambda/n)$ for $n = 1, \ldots$ and let $Y \sim \text{Poisson}(\lambda)$, then*

$$X_n \rightsquigarrow Y.$$

Proof. Let $X_n \sim \text{Binomial}(n, \theta = \lambda/n)$ and $Y \sim \text{Poisson}(\lambda)$ for a fixed λ. We need to show that

$$\lim_{n\to\infty} \mathbb{P}(X_n = x) = \mathbb{P}(Y = x) = e^{-\lambda}\lambda^x/x!$$

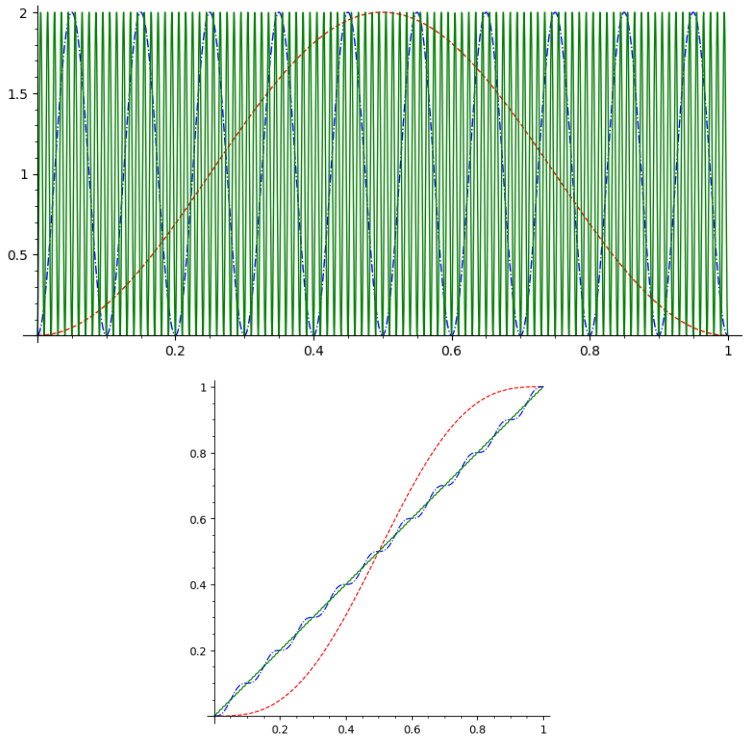


Figure 3.2: PDF $f_{X_n}(x) := \mathbb{1}_{(0,1)}(x)(1 - \cos(2\pi n x))$ of the RV X_n [the left sub-figure] and its DF $F_n(x) := \int_{-\infty}^{x} \mathbb{1}_{(0,1)}(v)(1 - \cos(2\pi n v))dv$ [the right sub-figure], for $n = 1$ [red '- -'], $n = 10$ [blue '-.'], and $n = 100$ [green '-'], respectively. One can see clear convergence of the DFs F_n to $\mathbb{1}_{(0,1)}(x)x$, the DF of the Uniform$(0, 1)$ RV, while the corresponding PDFs $f_n(x)$ keep oscillating wildly with n across $[0, 2]$ about $\mathbb{1}_{(0,1)}(x)$, the PDF of the Uniform$(0, 1)$ RV X. Thus giving a counter-example to the claim that convergence in DFs does not imply convergence in PDFs.

for any $x \in \{0, 1, 2, 3, \ldots, n\}$.

$$\mathbb{P}(X_n = x) \quad = \quad \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

First note that

$$\binom{n}{x} \left(\frac{\lambda}{n}\right)^x = \frac{n!}{(n-x)!n^x} \frac{\lambda^x}{x!}$$

By Stirlings formula this now converges to $\frac{\lambda^x}{x!}$. The last term

$$\left(1 - \frac{\lambda}{n}\right)^{n-x} = \left(1 - \frac{\lambda}{n}\right)^{n} \left(1 - \frac{\lambda}{n}\right)^{-x}$$

Is the product of two terms, where the first tends to $e^{-\lambda}$ and the second tends 1. We thus get

$$\lim_{n\to\infty} \mathbb{P}(X_n = x) = \mathbb{P}(Y = x)$$

which according to Lemma 3.19 gives $X_n \rightsquigarrow Y$. □

The second notion of convergence of RVs is convergence in probability.

Definition 3.21 (**Convergence in Probability**). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, X_2, \ldots,$ be a sequence of RVs and let X be another RV. We say that X_n converges to X in probability, and write:*

$$X_n \xrightarrow{\mathbb{P}} X$$

if for every real number $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

Once again, the above limit, by (2.1) in our Definition 2.1 of a RV, can be equivalently expressed as follows:

$$\lim_{n\to\infty} \mathbb{P}(\ \{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}\ ) = 0.$$

Definition 3.22 (**Convergence Almost Surely** (or with Probability 1)). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, X_2, \ldots,$ be a sequence of RVs and let X be another RV. We say that X_n converges to X almost surely (or with probability 1/strongly) if*

$$\mathbb{P}\left(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\right) = 1,$$

denoted as

$$X_n \overset{a.s.}{\to} X$$

This means that the values of X_n approach the value of X, in the sense that events for which X_n does not converge to X have probability 0,

$$\mathbb{P}\left(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\}\right) = 0,$$

Another notion which is quite useful is the mean-square convergence (or just L^2 convergence) which is a special case of

Definition 3.23 (**Convergence in L^p**). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, X_2, \ldots, \in L^p(\mathbb{P})$ be a sequence of RVs and let $X \in L^p(\mathbb{P})$ be another RV. We say that X_n converges to X in $L^p(\mathbb{P})$ if*

$$\|X_n - X\|_{L^p(\mathbb{P})} \to 0.$$

Recall the definition of the $L^p(\mathbb{P})$ norm, Section 2.6.4, hence the above is equivalent to

$$\mathbb{E}\left[|X_n - X|^p\right] \to 0.$$

Other notions of convergence are termed sure convergence or pointwise convergence, such as convergence in mean. But the above types of convergence are elementary.

3.2.1 Properties of Convergence of RVs**

We will merely state some properties (without proofs that are hyper-linked for the curious student as they are advanced for this course) and relations between the three notions of convergence with some examples to better appreciate the subtleties among them. Just remember that subtle implication relations exist between the notions.

Theorem 3.24. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, X_2, \ldots,$ be a sequence of RVs and let X be another RV. The following are equivalent: Let $f : \mathbb{R} \to \mathbb{R}$ be a function*

- $\mathbb{E}\left[f(X_n)\right] \to \mathbb{E}\left[f(X)\right]$ *for all bounded, countinuous f.*
- $\mathbb{E}\left[f(X_n)\right] \to \mathbb{E}\left[f(X)\right]$ *for all bounded, Lipschitz countinuous f.*
- $X_n \rightsquigarrow X$.
- *For any "good enough" set $A \subset \mathbb{R}$, $\mathbb{P}(X_n \in A) \to \mathbb{P}(X \in A)$.*

Remark 3.25. *For those interested, good enough in the above means that $\mathbb{P}(X \in \partial A) = 0$. This is equivalent to the convergence happening only on the points of continuity of the distribution function, as in Definition 3.18.*

- Convergence almost surely implies convergence in probability[1]

$$\boxed{X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{\mathbb{P}} X \ .}$$

- By the Borel-Cantelli Lemma [2], convergence in probability does not imply almost sure convergence in the discrete case [3]

- Convergence in probability implies convergence in distribution [4]

$$\boxed{X_n \xrightarrow{\mathbb{P}} X \implies X_n \rightsquigarrow X \ .}$$

- Convergence in distribution to a constant θ implies convergence in probability to θ: [5]

$$X_n \rightsquigarrow \text{Point Mass}(\theta) \implies X_n \xrightarrow{\mathbb{P}} \text{Point Mass}(\theta) \ .$$

- Convergence in L^p for $1 \leq q \leq p < \infty$ implies convergence in L^q. (Follows from Theorem 2.49)

- Convergence in L^p for $1 \leq p < \infty$ implies convergence in probability. Follows from Theorem 3.1.

- In general, convergence in distribution does not imply convergence in probability.

3.3 Law of Large Numbers

Theorem 3.26. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, X_2, \ldots, \in L^2(\mathbb{P})$ be a sequence of i.i.d. RVs with $\mathbb{E}[X_i] = \mu$. Then*

$$\overline{X}_n \xrightarrow{\mathbb{P}} \mu.$$

Proof. We need to prove that for a fixed $\epsilon > 0$ that

$$\lim_{n \to \infty} \mathbb{P}(|X_n - \mu| > \epsilon) \to 0.$$

[1] https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables#Convergence_almost_surely_implies_convergence_in_probability

[2] https://en.wikipedia.org/wiki/Borel%E2%80%93Cantelli_lemma

[3] https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables#Convergence_in_probability_does_not_imply_almost_sure_convergence_in_the_discrete_case

[4] https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables#Convergence_in_probability_implies_convergence_in_distribution

[5] https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables#Convergence_in_distribution_to_a_constant_implies_convergence_in_probability

But note that from Theorem 3.1 we have

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\mathbb{E}\left[|\overline{X}_n - \mu|^2\right]}{\epsilon^2} = \frac{1}{n}\frac{\mathbb{E}\left[|X_1 - \mu|^2\right]}{\epsilon^2}. \tag{3.14}$$

The last step used

$$\begin{aligned}
\mathbb{E}\left[|\overline{X}_n - \mu|^2\right] &= \mathbb{E}\left[\frac{1}{n^2}|\sum_i X_i - n\mu|^2\right] \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i,j}(X_i - \mu)(X_j - \mu)\right] \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i}(X_i - \mu)^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[|X_1 - \mu|^2\right],
\end{aligned}$$

where in the second to last step in the above we used the independence assumption as $\mathbb{E}\left[(X_i - \mu)(X_j - \mu)\right] = \mathbb{E}\left[(X_i - \mu)\right]\mathbb{E}\left[(X_j - \mu)\right] = 0$ if $i \neq j$. (This is the famous "variance of the sum is the sum of the variance" for independent random variables).

Now (3.14) completes the proof. □

3.4 Central Limit Theorem

What if we scale the sum of X_i's by $\sqrt{n}$ instead of n?

Theorem 3.27. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, X_2, \ldots, \in L^2(\mathbb{P})$ be a sequence of i.i.d. RVs with $\mathbb{E}\left[X_i\right] = \mu$ and $\mathbb{V}\left[X_i\right] = \sigma^2$. Then if we denote*

$$Z_n := \frac{\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right]}{\sqrt{\mathbb{V}\left[\overline{X}_n\right]}}$$

we get

$$Z_n \rightsquigarrow Z$$

where $Z \sim N(0, 1)$.

Rewriting this in $L^p(\mathbb{P})$ space notation we get for $Y_n = X_n - \mathbb{E}\left[X_n\right]$ that

$$Z_n := \frac{\overline{Y}_n}{\|\overline{Y_n}\|_{L^2(\mathbb{P})}}$$

that is, we normalized Y_n to have L^2 norm 1, $\|Z_n\|_{L^2(\mathbb{P})} = 1$. The central limit theorem tells us that there is a $Z \in L^2(\mathbb{P})$ that is the distributional limit of Z_n. (Warning we cannot expect stronger convergence, this is due to non-compactness of the unit ball in L^2).

Proof. We will skip the proof of the CLT, as it is not particularly useful for us. □

Chapter 4

Risk

Definition 4.1. *A* **statistical model** *is an indexed family of distributions (or densities or regression functions)* $\mathcal{F} = \{f(x;\theta), \theta \in \Theta\}$.

- A **parametric model** is a model where the indexing parameter θ is a vector in k-dimensional Euclidean space. That is, θ is finite dimensional.
- A **non-parametric model** is a model where Θ is infinite dimensional.

Example 4.2. $\mathcal{N} = \{N(\mu, \sigma), \mu \in \boldsymbol{R}, \sigma > 0\}$. *This is a parametric model.*

Example 4.3. $\mathcal{E} = \{F : F \text{ is a CDF}\}$. *This is a non-parametric model.*

A statistical model is a model of the data generation, that is, it is what we assume the truth is. As you can see above, a parametric model is more restrictive, this usually means that drawing conclusions (estimation) from data is "easier" (higher precision with less data).

Let us quote Merriam-Webster:

"Main Entry: **in·fer·ence**
Pronunciation: 'in-f(&-)r&n(t)s, -f&rn(t)s
Function: *noun*
Date: 1594

1. the act or process of inferring: as
   (a) the act of passing from one proposition, statement, or judgment considered as true to another whose truth is believed to follow from that of the former
   (b) the act of passing from statistical sample data to generalizations (as of the value of population parameters) usually with calculated degrees of certainty."

Inference lies at the heart of statistics and learning. The question is: what do we want to know?

Under a statistical model $\mathcal{F}$, there is a hidden $f^* \in \mathcal{F}$ that generates the data, we would like to infer something about f^* using observations.

Here are some examples of inference problems:

1. Density estimation, or consequences of the density, like estimating the probability of an event.
2. Estimating the distribution function. Can be used to answer questions about probabilities of simpler events, but can also be functionals of the distribution.
3. Functional dependence, usually regression, or pattern recognition.

4.1 The supervised learning problem

As we will be working with machine learning and data science let us describe the learning problem as seen from the field of computer science and let us interpret each piece using our probabilistic terminology. We will begin with learning a functional dependency, the model contains three elements

1. The generator of the data G
2. The supervisor S
3. The learning machine LM.

The generator G is a source of situations, we will make the simplest assumptions, that G generates vectors X_i i.i.d. according to some unknown but fixed distribution $F(x)$. These vectors X_i are inputs to the *supervisor* that outputs a value Y_i, we know the supervisor has an unknown function transforming X_i into Y_i. At this point we could also consider that $Y_i|X_i$ has some noise in it (perhaps the supervisor is measuring something about X_i but that measurement has a random error in it). The learning machine observes a realization of the n pairs

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

we denote this realized set as $(x_1, y_1), \ldots, (x_n, y_n)$ (the training set). In this course we make the assumption that the supervisor generates Y_i from X_i according to an unknown conditional distribution $F(y|x)$, that is the conditional distribution of $Y_i|X_i$. Recall that this includes the case of a functional dependecy $y = f(x)$. The learning machine thus observes pairs $(X_i, Y_i) \sim F_{X,Y}(x, y)$ where $f_{X,Y}(x, y) = f_{Y|X}(y|x)f_X(x)$ is the joint distribution.

> The goal is to approximate this functional dependency in some way using the observations!

Whenever we want to produce an approximation it often makes sense to come up with a measure of quality. Denote $z = (x, y)$ and consider a function $g(z)$ that is of a type we are interested in and define a loss functional $L(z, g)$ that measures the quality of g at the point z. Now consider the expected loss, usually denoted as *Risk*

$$R(g) = \int L(z, g) dF(z) = \mathbb{E}\left[L(Z, g)\right]$$

where $Z = (X, Y) \sim F(x, y)$.

> Goal of the learning machine: Define a class of functions g to search from and minimize the risk inside this class.

We will now work through how this is formulated mathematically in some special cases that will be general enough for us in this course. The purpose, for now, is to get a feeling for the concepts. Later we will move on to, how to actually minimize the risk using empirical data (Empirical Risk Minimization) and some guarantees we can make under certain assumptions.

4.1.1 Mathematical description of the learning problem "find f"

Let us begin describing the learning problem for a simple case of the supervisor is using a real valued continuous function $y = f(x)$, for $x \in [0, 1]$, i.e. $f \in C([0, 1], \mathbb{R})$, that is $F(y|x) = \mathbb{1}_{y \geq f(x)}(y)$, i.e. a point mass centered at $f(x)$.

Example 4.4. *Let us motivate the above setup with a typical example. In many image analysis problems you need a way to determine the scale of the objects in view. This is usually done using a fiducial mark, that is a reference shape of known size. This is often a solid circle of known size.*

What we need in order to determine the scale of the image is to detect the fiducial mark and measure how many pixels it corresponds to in the image. Since it is a circle, we only need to figure out the radius of it in the image (in terms of the number of pixels).

- *The data generator G is the process that produces the images, i.e. the experiment. The data that is generated is X which is the image.*
- *The supervisor is anyone or anything that knows the answer of the radius of the circle. That is, if you give the supervisor the image X, the output will be the correct radius of the circle Y. Our assumption is that $Y = f(X)$.*

- *The goal of the learning machine is to figure out how to get from the image X to the radius Y simply by observing examples of pairs (X, Y).*

To describe the learning problem in the above setting we need to set-up the following things

1. Statistical model: $\mathcal{F} = \{F(x, y) = \mathbb{1}_{y \geq f(x)}(y)F(x), f \in C([0, 1], \mathbb{R})\}$
2. Model space: $\mathcal{M} = \{g_\lambda(x) : \lambda \in \Lambda\}$, a parametrized space of functions in which we are searching for f, or an approximation thereof. (The kind of functions the learning machine can represent).
3. A loss function $L : [0, 1] \times \mathbb{R} \to \mathbb{R}$.

In this setting the learning problem becomes the risk minimization problem below

$$g_* = \operatorname*{argmin}_{g \in \mathcal{M}} R(g)$$

However, since the model space is parametrized we can actually rewrite the minimization problem to be over Λ instead, as follows

$$\lambda_* = \operatorname*{argmin}_{\lambda \in \Lambda} R(g_\lambda).$$

This also allows us to write the loss function as a function of z and the parameter λ as $L(z, g_\lambda(z)) = Q(z, \lambda)$ and our risk can be written as

$$R(\lambda) = \int Q(z, \lambda) dF(z) = \int Q((x, f(x)), \lambda) dF(x)$$

where the first integral is a double integral over $z = (x, y)$ and the second integral is a single integral over x.

As you have seen above the loss function is not specified and can be chosen quite freely. Let us next consider the problem of regression.

4.1.2 Finding the regression function $r(x) = \mathbb{E}[Y|X]$

Let us now assume that the supervisor is generating Y from $F(y|x)$ given the value of X. In this setup, perhaps we would like to estimate the full conditional distribution $F(y|x)$, but this is a hard problem. Instead one could try to estimate some of its properties, for instance, we could try to estimate a functional of $F(y|x)$. The concept of regression is that we are interested in the following functional

$$r(x) = \int y dF(y|x) = \mathbb{E}[Y \mid X = x].$$

Here, r is called the regression function. Let us assume that $Y \in L^2(\mathbb{P})$ and $r \in L^2(dF_X)$, that is

$$\mathbb{E}[Y^2] < \infty, \quad \mathbb{E}[r^2(X)] < \infty.$$

Example 4.5. *Finding the regression function is an extension of the "finding the function", therefore we could use the same example with the fiducial point, but now we could assume that the supervisor doesn't know the exact size of the circle. But instead is performing a measurement which has some error connected to it.*

Example 4.6. *We have already seen an example of a regression problem. Namely Section 2.7.2. Here the goal was to estimate*

$$F_{X|Z}(x \mid z)$$

if we change this to the notation above, the Z is the data coming from our data-generator and the supervisor gives us X sampled from $F_{X|Z}(x \mid z)$.

In this case the statistical model is

$$\mathcal{F} = \Big\{ F(x,y) = F_{Y|X}(y \mid x)F(x); \\ r(x) = \int y dF(y \mid x), \mathbb{E}\left[r(X)^2\right] < \infty, \mathbb{E}\left[Y^2\right] < \infty \Big\}$$

Consider now an model space $\mathcal{M} = \{g_\lambda(x) : \lambda \in \Lambda, g_\lambda \in L^2(dF_X)\}$ of some functions g_λ parametrized by λ.

For a function $g_\lambda \in \mathcal{M}$ we consider the following risk

$$R(\lambda) = \int (y - g_\lambda(x))^2 dF(x,y)$$

WARNING: we have not specified if $r \in \mathcal{M}$!!

Assume there exists a $\lambda^* \in \Lambda$ such that for $g^* = g_{\lambda^*}$

$$R(g^*) = \inf_{g \in \mathcal{M}} R(g)$$

then write the risk as

$$\begin{aligned} R(\lambda) &= \mathbb{E}\left[(Y - g_\lambda(X))^2\right] = \mathbb{E}\left[(Y - g_\lambda(X) + r(X) - r(X))^2\right] \\ &= \mathbb{E}\left[(Y - r)^2\right] + \mathbb{E}\left[(r(X) - g_\lambda(X))^2\right] + 2\,\mathbb{E}\left[(Y - r)(r - g_\lambda)\right] \\ &= I + II + III. \end{aligned}$$

The assumption that Y, $r(X)$ and $g_\lambda(X)$ all have finite second moment is what allows us to do the computation above, and know that all terms involved are finite. Let us now consider III and note that by the tower property and the definition of r that

$$\begin{aligned} III &= 2\,\mathbb{E}\left[(Y - r)(r - g_\lambda)\right] = 2\,\mathbb{E}\left[\mathbb{E}\left[(Y - r)(r - g_\lambda)|X\right]\right] \\ &= 2\,\mathbb{E}\left[(\mathbb{E}\left[Y|X\right] - r)(r - g_\lambda)\right] = 0 \end{aligned}$$

From this we now see that

$$\operatorname*{argmin}_{\lambda\in\Lambda} R(\lambda) = \operatorname*{argmin}_{\lambda\in\Lambda} \mathbb{E}\left[(r(X) - g_\lambda(X))^2\right]$$

which means that the minimizer g^* will be the function in $\mathcal{M}$ that is closest to r in the mean square sense (or in L^2 if using function space notation).

NOTE: if $r \in \mathcal{M}$ then $g^* = r$ a.e. with respect to dF_X.

4.1.3 The pattern recognition problem (classification)

In the pattern recognition model we assume that the supervisors conditional distribution $F(y|x)$ is discrete, and can take k different values, $y = 0, \ldots, k-1$. Consider a model space $\mathcal{M} = \{g_\lambda(x) : g_\lambda(x) \in \{0, \ldots, k-1\}\}$, that is, functions g_λ that takes values in $\{0, \ldots, k-1\}$. It is common to call the functions in the pattern recognition problem, g_λ a **decision function** or **decision rule**. With this at hand, we define the $0-1$ loss function for $z = (x, y)$

$$L(z, u) = \begin{cases} 0 & \text{if } y = u \\ 1 & \text{if } y \neq u \end{cases}$$

that is, the loss is 1 if u is the wrong value and 0 if it is correct. The pattern recognition problem is the problem of minimizing the functional

$$R(\lambda) = \int L(y, g_\lambda(x)) dF(x, y) = \mathbb{E}\left[L(Y, g_\lambda(X))\right]$$

where $(X, Y) \sim F(x, y)$.

Exercise 4.7. *What is a reasonable statistical model for the Pattern Recognition problem?*

The classification problem is in modern times very often associated with the prototypical example of classification of images of dogs and cats. In that example, the images is X and the class is given by Y.

The risk above has a natural interpretation, given the "decision rule" g_λ, the risk $R(\lambda)$ is the probability of an incorrect classification by the rule g_λ,

$$\mathbb{E}\left[L(Y, g_\lambda(X))\right] = \mathbb{P}(\{Y \neq g_\lambda(X)\}).$$

Bayes rule

What is the optimal decision rule? Recall that in the regression setting we had the regression function as the minimizer, but what is it in the pattern recognition problem? Consider the case when $k = 2$ and denote

$$r(x) = \mathbb{E}\left[Y \mid X = x\right] = \mathbb{P}(Y = 1 \mid X = x)$$

Definition 4.8. *The* **Bayes classification rule** h^* *is*

$$h^*(x) = \begin{cases} 1 & \text{if } r(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Let us prove that the Bayes classification rule is the rule that optimizes the risk in the pattern recognition problem.

Theorem 4.9. *For any decision function $g(x)$ taking values in $\{0, 1\}$, we have*

$$R(h^*) \leq R(g).$$

Proof. Note that we can write

$$R(g) = \mathbb{E}\left[L(Y, g(X))\right] = \mathbb{E}\left[\mathbb{E}\left[L(Y, g(X)) \mid X\right]\right]$$

we will work only with the inner part, i.e. now

$$\begin{aligned} \mathbb{E}\left[L(Y, g(X)) \mid X = x\right] &= 1 - \mathbb{E}\left[\mathbb{1}_{\{y=g(x)\}} \mid X = x\right] \\ &= 1 - \mathbb{E}\left[\mathbb{1}_{\{1=g(x)\}}\mathbb{1}_{\{y=1\}} + \mathbb{1}_{\{0=g(x)\}}\mathbb{1}_{\{y=0\}} \mid X = x\right] \\ &= 1 - \mathbb{1}_{\{1=g(x)\}}\,\mathbb{E}\left[\mathbb{1}_{\{y=1\}} \mid X = x\right] - \mathbb{1}_{\{0=g(x)\}}\,\mathbb{E}\left[\mathbb{1}_{\{y=0\}} \mid X = x\right] \\ &= 1 - \mathbb{1}_{\{1=g(x)\}} r(x) - \mathbb{1}_{\{0=g(x)\}}(1 - r(x)) \end{aligned}$$

Now

$$\begin{aligned} &\mathbb{E}[L(Y, g(X)) \mid X = x] - \mathbb{E}\left[L(Y, h^*(X)) \mid X = x\right] = \\ &\quad = -\mathbb{1}_{\{1=g(x)\}} r(x) - \mathbb{1}_{\{0=g(x)\}}(1 - r(x)) + \mathbb{1}_{\{1=h^*(x)\}} r(x) + \mathbb{1}_{\{0=h^*(x)\}}(1 - r(x))) \\ &\quad = r(x)(\mathbb{1}_{\{1=h^*(x)\}} - \mathbb{1}_{\{1=g(x)\}}) + (1 - r(x))(\mathbb{1}_{\{0=h^*(x)\}} - \mathbb{1}_{\{0=g(x)\}}) \\ &\quad = r(x)(\mathbb{1}_{\{1=h^*(x)\}} - \mathbb{1}_{\{1=g(x)\}}) - (1 - r(x))(\mathbb{1}_{\{1=h^*(x)\}} - \mathbb{1}_{\{1=g(x)\}}) \\ &\quad = (2r(x) - 1)(\mathbb{1}_{\{1=h^*(x)\}} - \mathbb{1}_{\{1=g(x)\}}) \geq 0. \end{aligned}$$

This immediately implies the statement of the theorem. □

From the above proof we see that if we are minimizing the risk inside a class $\mathcal{M}$ that does not include h^*, we can write g^* as a minimizer in $\mathcal{M}$ and get

$$R(g^*) = R(h^*) + \mathbb{E}\left[(2r(X) - 1)(I_{\{1=h^*(X)\}} - I_{\{1=g^*(X)\}})\right]$$

Remark 4.10. *The above expression is interesting.*

- *$h^*(x) = 1$ which means $r(x) > 1/2$ then the cost of misclassifying is $(2r-1)$, which means that the cost is higher for higher values of r.*
- *$h^*(x) = 0$ which means that $r(x) \leq 1/2$ then the cost of misclassifying is higher for r close to 0.*
- *In the case $r = 1/2$ there is always zero cost and it does not matter if we misclassify.*

4.2 Maximum Likelihood Estimation

We will now derive the Maximum Likelihood as a special case of risk minimization.

Assume that we have a parametric model $\mathcal{E} = \{p_\alpha(z), \alpha \in \mathbb{R}^n\}$, for some given family of densities p_α. For example we can take

$$p_\alpha(z) = \frac{1}{\sqrt{2\pi\alpha_2}} e^{-\frac{|z-\alpha_1|^2}{\alpha_2}},$$

where $\alpha = (\alpha_1, \alpha_2)$ which is the Gaussian family. Assume that our underlying model is given by a hidden parameter α^*, then consider the loss function $L(z, \alpha) = -\ln p_\alpha(z)$ then the risk becomes

$$R(\alpha) = -\int \ln(p_\alpha(z)) p_{\alpha^*}(z) dx$$

If we let Z be a random variable with law p_{α^*} then we can write the above as

$$R(\alpha) = \mathbb{E}\left[-\ln(p_\alpha(Z))\right]$$

Given a sequence of i.i.d. random variables $Z_1, \ldots, Z_n$ sampled from p_{α^*} the empirical Risk just becomes

$$\hat{R}(\alpha) = -\frac{1}{n}\sum_{i=1}^{n} \ln(p_\alpha(Z_i)).$$

This is nothing but the negative log likelihood of the observations $Z_1, \ldots, Z_n$ under the model p_α. Thus to minimize the risk with respect to α is the same as maximizing the log likelihood.

So, is the choice of loss L any good? can we say that the minimum is attained at α^*?

How do we prove that? Well, we prove it using Jensens inequality (Lemma 2.51). Consider the function $\psi(u) = \ln(u)$ (concave, Jensen is reversed) and $\Phi(x) = \frac{p_\alpha(x)}{p_{\alpha^*}(x)}$, using Jensens inequality we get

$$R(\alpha^*) - R(\alpha) = \int \psi(\Phi(x)) p_{\alpha^*} dx \leq \psi\left(\int \Phi(x) p_{\alpha^*} dx\right) = \ln 1 = 0$$

as such we have that $R(\alpha^*) \leq R(\alpha)$ and hence α^* is the global minimum of the risk. Is there any other α that also minimizes the risk? We have

$$\int \psi(\Phi(x))p_{\alpha^*}dx = 0$$

this implies that $\psi(\Phi(x)) = 0$ a.e. with respect to p_{α^*}, so if our family is well behaved (identifiable) then the minimum is unique.

4.2.1 Maximum Likelihood and regression

Suppose that we have a pair (X, Y) of random variable. Denote a proposal joint density of (X, Y) as $f_{X,Y}$. Consider a sequence of i.i.d. samples (X_i, Y_i), $i = 1, \ldots, n$ with the same law as (X, Y), then the negative log-likelihood (which is just the empirical risk under loss ln, see Section 4.2) is given by

$$-\sum_{i=1}^{n} \ln(f_{X,Y}(X_i, Y_i))$$

if we condition on X we get

$$\begin{aligned} -\sum_{i=1}^{n} \ln(f_{X,Y}(X_i, Y_i)) &= -\sum_{i=1}^{n} \ln(f_{Y|X}(Y_i \mid X_i)f_X(X_i)) \\ &= -\sum_{i=1}^{n} \ln(f_{Y|X}(Y_i \mid X_i)) - \sum_{i=1}^{n} \ln(f_X(X_i)) \end{aligned}$$

Now, consider a parametrized family of proposal joint distributions where the marginal density f_X does not depend on any parameter, only the conditional distribution $f_{Y|X}$ then if we want to minimize the negative log-likelihood over this particular proposal family, only the first summand can change, as such, it is enough to minimize over this. If we flip the sign we get that we would like to maximize the conditional likelihood. This is the main idea behind linear regression and logistic regression, both of which are ubiquitous in the field of data science. To see how this looks like in the context of linear regression, see [W, Chapter 13].

Example 1: Linear regression

In this case we make the assumption that $f_{a,b,\sigma} := f_{a,b,\sigma;Y|X}$ is the density of $N(aX + b, \sigma^2)$ where the parameters of interest are a, b, σ. We assume that f_X is some fixed proposal density, actually it will not matter (see above).

$$-\sum_{i=1}^{n} \ln(f_{Y|X}(Y_i \mid X_i)) = -\sum_{i=1}^{n} \ln(\frac{1}{\sigma}e^{-\frac{1}{2\sigma^2}(Y_i-(aX_i+b))^2}) - Cn$$

The first term on the right can be rewritten as

$$\sum_{i=1}^{n} \ln(\sigma) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - (aX_i + b))^2$$

The main realization is that minimizing the likelihood gives the same parameters as minimizing the conditional likelihood which gives the same parameters as minimizing the sum of squares. I.e. linear regression in this case is equivalent to mean square regression, as in Section 4.1.2.

Example 2: Logistic regression

Here we assume on the contrary to linear regression, that the proposal density $f_{\beta_0,\beta_1;Y|X}$ is the density of a Bernoulli$(G(\beta_0+\beta_1 X))$ where the function G is defined as

$$G(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x},$$

and is called the logistic function. Here we assume that $Y \in \{0, 1\}$

If we call $p(X) = G(\beta_0 + \beta_1 X)$ then

$$\begin{aligned} -\sum_{i=1}^{n} \ln(f_{Y|X}(Y_i \mid X_i)) &= -\sum_{i=1}^{n} \ln(p(X_i)^{Y_i}(1-p(X_i))^{1-Y_i}) \\ &= -\sum_{i=1}^{n} (Y_i \ln(p(X_i)) + (1-Y_i)\ln(1-p(X_i))) \end{aligned}$$

Now

$$\begin{aligned} \ln(p(X_i)) &= \ln(1/(1+e^{-(\beta_0+\beta_1 X_i)})) = -\ln(1+e^{-(\beta_0+\beta_1 X_i)}) \\ \ln(1-p(X_i)) &= \ln(1-1/(1+e^{-(\beta_0+\beta_1 X_i)})) = -\ln(1+e^{\beta_0+\beta_1 X_i}). \end{aligned}$$

When $Y_i = 0$ we get

$$-\ln(p(X_i)^{Y_i}(1-p(X_i))^{1-Y_i}) = -\ln(1-p(X_i)) = \ln(1+e^{\beta_0+\beta_1 X_i})$$

and when $Y_i = 1$ we get

$$-\ln(p(X_i)^{Y_i}(1-p(X_i))^{1-Y_i}) = -\ln(p(X_i)) = \ln(1+e^{-(\beta_0+\beta_1 X_i)}).$$

Thus the only thing that changes is the sign of the exponent, so if we write $Z_i = 2Y_i - 1$ then $Z_i = 1$ if $Y_i = 1$ and $Z_i = -1$ if $Y_i = 0$ and we can write

$$-\sum_{i=1}^{n} \ln(p(X_i)^{Y_i}(1-p(X_i))^{1-Y_i}) = \sum_{i=1}^{n} \ln(1+e^{-Z_i(\beta_0+\beta_1 X_i)}).$$

Now, you might wonder, why the specific form of $G(x)$ other than the fact that it outputs numbers between 0 and 1? To see why this formula is used, consider the log-odds ratio given X, i.e.

$$\ln\left(\frac{\mathbb{P}(Y=1 \mid X)}{\mathbb{P}(Y=0 \mid X)}\right) = \ln\left(\frac{p(X)}{1-p(X)}\right) = \ln\left(\frac{G(\beta_0+\beta_1 X)}{1-G(\beta_0+\beta_1 X)}\right)$$
$$= \ln(e^{\beta_0+\beta_1 X}) = \beta_0+\beta_1 X$$

Thus for the logistic regression the log odds ratio is linear.

Chapter 5

Fundamentals of Estimation

5.1 Introduction

Now that we have been introduced to two notions of convergence for RV sequences, we can begin to appreciate the basic limit theorems used in statistical inference. The problem of estimation is of fundamental importance in statistical inference and learning. We will formalise the general estimation problem here. There are two basic types of estimation. In point estimation we are interested in estimating a particular point of interest that is supposed to belong to a set of points. In (confidence) set estimation, we are interested in estimating a set with a particular form that has a specified probability of "trapping" the particular point of interest from a set of points. Here, a point should be interpreted as an element of a collection of elements from some space.

5.2 Point Estimation

Point estimation is any statistical methodology that provides one with a "**single best guess**" of some specific quantity of interest. Traditionally, we denote this **quantity of interest as** θ^*. This quantity of interest, which is usually unknown, can be:

- an **integral** $\vartheta^* := \int_A h(x)\,dx \in \boldsymbol{\Theta}$. If ϑ^* is finite, then $\boldsymbol{\Theta} = \mathbb{R}$. The risk is a prime example, or
- a **parameter** θ^* which is an element of the **parameter space** $\boldsymbol{\Theta}$, denoted $\theta^* \in \boldsymbol{\Theta}$,
- a **distribution function (DF)** $F^* \in \mathbb{F} :=$ the set of all DFs
- a **density function (pdf)** $f \in \{$"not too wiggly Sobolev functions"$\}$, or

- a **regression function** $g^* \in \mathbb{G}$, where $\mathbb{G}$ is a class of regression functions in a regression experiment, or
- a **classifier** $g^* \in \mathbb{G}$.

Definition 5.1 (Data)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple, assume that $X = (X_1, \ldots, X_n)$ is a sequence of $\mathbb{R}^m$ valued random variables taking values in the* **data space** $\mathbb{X}$*:*

$$X(\omega) : \Omega \to \mathbb{X} \ .$$

Note that $\mathbb{X} \subset (\mathbb{R}^m)^{\otimes n}$. The realisation of the RV X when an experiment is performed is the observation or data $x \in \mathbb{X}$. That is, when the experiment is performed once and it yields a specific $\omega \in \Omega$, the data $X(\omega) = x \in \mathbb{X}$ is the corresponding realisation of the RV X.

Definition 5.2 (Statistic)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple, assume that $X : \Omega \to \mathbb{X}$ is a random variable (sequence of $\mathbb{R}^m$ valued) taking values in the data space $\mathbb{X}$, then a* **statistic** *T is any Borel (see Definition 2.42) function on the data space:*

$$T(x) : \mathbb{X} \to \mathbb{T} \ .$$

Remark 5.3. *Thus, given a statistic T, we can associate with it a RV $T(X)$ that takes values in the space $\mathbb{T}$. Sometimes we use $\mathbb{X}_n$, $T_n(X)$ and $\mathbb{T}_n$ to emphasise that X is a sequence of n random variables, i.e. $\mathbb{X}_n \subset (\mathbb{R}^m)^{\otimes n}$*

Definition 5.4. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let*

$$\mathcal{E} = \{F(x; \lambda) : \mathbb{X} \to [0,1] : \lambda \in \mathbf{\Lambda}, \ F \text{ is a DF}\}$$

be a statistical model of distribution functions. Let a parameter map be given $\theta : \mathbf{\Lambda} \to \mathbf{\Theta}$. Consider the sequence $X = (X_1, \ldots, X_n) \overset{\text{IID}}{\sim} F(\cdot; \lambda^*) \in \mathcal{E}$ be $\mathbb{R}^m$-valued RVs. A **point estimator** *of $\theta^* := \theta(\lambda^*) \in \mathbf{\Theta}$ is a statistic, i.e.*

$$\hat{\Theta} : \mathbb{X} \to \mathbf{\Theta},$$

sometimes we denote it as $\hat{\Theta}_n$ to highlight that it depends on n values.

The bias *of an estimator $\widehat{\Theta}_n$ of $\theta^* \in \mathbf{\Theta}$ is:*

$$\mathsf{bias}(\widehat{\Theta}_n(X)) := \mathbb{E}(\widehat{\Theta}_n(X)) - \theta^* = \int \widehat{\Theta}_n(x)\, dF(x; \lambda^*) - \theta(\lambda^*) \ . \tag{5.1}$$

Some comments are in order to connect these concepts to the risk minimization problems in supervised learning, as see in Chapter 4. Let us be

given a statistical model $\mathcal{E}$ of distribution functions $F_{X,Y}$, and let us consider the regression example, i.e. we wish to estimate

$$r(x) = \int y dF_{Y|X}(y \mid x)$$

this means that for each fix x the parameter map is $\theta_x(F_{X,Y}) = r(x)$, and if given $X_1, \ldots, X_n$ we come up with a proposal function from $\mathcal{M}$, i.e. $g_n(X_1, \ldots, X_n) \in \mathcal{M}$ then $g_n(X_1, \ldots, X_n; x)$ becomes a point estimator of $r(x)$. Sometimes however, in regression experiments you assume that the model space and the statistical model are the same and parametric (finite dimensional). In this case the parameter becomes easy to define and much of this simplifies.

As we shall see later, we are often not so concerned with the statistical properties of the specific estimator of the regression function but we are interested in some functional of it. Usually we are interested in the Risk, and in this case it is as simple as an expectation, see Example 5.6.

5.2.1 Some Properties of Point Estimators

Given that an estimator is merely a function from the data space to the parameter space, we need a way to define what a good estimator is. Recall that a point estimator $\widehat{\Theta}_n$, being a statistic, has a corresponding RV $\widehat{\Theta}_n(X)$ which has a probability distribution over its range $\boldsymbol{\Theta}$. This distribution over $\boldsymbol{\Theta}$ is called the **sampling distribution** of $\widehat{\Theta}_n(X)$.

Definition 5.5 (Bias of a Point Estimator)**.** *We say that the estimator* $\widehat{\Theta}_n$ *is* **unbiased** *if*

$$\mathsf{bias}(\widehat{\Theta}_n(X)) = 0,$$

for every n*. If*

$$\lim_{n\to\infty} \mathsf{bias}_n(\widehat{\Theta}_n) = 0,$$

we say that the estimator is **asymptotically unbiased**.

Since the expectation of the sampling distribution of the point estimator $\widehat{\Theta}_n$ depends on the unknown λ^*, we emphasise the λ^*-dependence by $\mathbb{E}_{\lambda^*}(\widehat{\Theta}_n(X))$.

Example 5.6. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability triple and let*

$$\mathcal{E} = \{F(x;\lambda) : \mathbb{X} \to [0,1] : \lambda \in \boldsymbol{\Lambda},\ F \text{ is a DF}\}$$

Let the parameter map $\theta(\lambda) := \int x dF(\lambda)$ *be the expectation. Let* $X_1, \ldots, X_n \overset{\text{IID}}{\sim} F(\cdot, \lambda^*)$*, that is, with some unknown parameter* λ^* *and hence some unknown mean* $\theta^* = \theta(\lambda^*)$*. Consider the* **sample mean** *estimator*

$$\widehat{\Theta}_n(X) := \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then

$$\mathsf{bias}(\widehat{\Theta}_n(X)) = \mathbb{E}_{\lambda^*}\left[\overline{X}_n\right] - \theta(\lambda_*) = 0,$$

hence the sample mean estimator is unbiased in our statistical model $\mathcal{E}$ *with respect to the parameter map* θ*.*

Remark 5.7. *In the above example our statistical model is parametrized by* Λ *which is infinite dimensional, we would thus say that this is a nonparametric model. Another example of estimation would be that we assume that the statistical model is that of normal distributions with mean* μ *and variance* σ^2*. In this case we could take* $\mathbf{\Lambda}$ *to be two dimensional and the parameter map be the identity map from* $\mathbb{R}^2 \to \mathbb{R}^2$*.*

Remark 5.8. *The bias of an estimator is a term that we use to theoretically study the properties of the estimator. In a real setting,* θ^* *is unknown so we could never compute the bias, but perhaps we can get bounds for it. In certain cases we can prove that an estimator is asymptotically unbiased without knowing* θ^**.*

For instance if $X_1, \ldots, X_n$ *as in our example above, if we furthermore assumed in our statistical model that* $X_i \in L^2(\mathbb{P})$ *then the law of large numbers Theorem 3.26 implies the asymptotic unbiasedness.*

Definition 5.9 (Standard Error of a Point Estimator)**.** *The standard deviation of the point estimator* $\widehat{\Theta}_n(X)$ *of* $\theta^* \in \mathbf{\Theta}$ *is called the* **standard error***:*

$$\mathsf{se}(\widehat{\Theta}_n(X)) := \sqrt{\mathbb{V}_{\lambda^*}(\widehat{\Theta}_n)} := \sqrt{\int \left(\widehat{\Theta}_n(x) - \mathbb{E}_{\lambda^*}(\widehat{\Theta}_n)\right)^2 dF(x;\lambda^*)}\ . \quad (5.2)$$

Since the variance of the sampling distribution of the point estimator $\widehat{\Theta}_n$ depends on the fixed and possibly unknown λ^*, as emphasised by $\mathbb{V}_{\lambda^*}$ in (5.2), the $\mathsf{se}(\widehat{\Theta}_n(X))$ is also a possibly unknown quantity and may itself be estimated from the data.

Example 5.10 (Standard Error of our Estimator of θ^*)**.** *Consider the sample mean estimator* $\widehat{\Theta}_n := \overline{X}_n$ *of* θ^**, from* $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta^*)$*. Observe that the statistic:*

$$T_n((X_1, X_2, \ldots, X_n)) := n\,\widehat{\Theta}_n((X_1, X_2, \ldots, X_n)) = \sum_{i=1}^{n} X_i$$

is the Binomial(n, θ^*) *RV. The standard error* se_n *of this estimator is:*

$$\begin{aligned}\mathsf{se}(\widehat{\Theta}_n) &= \sqrt{\mathbb{V}_{\lambda^*}\left(\sum_{i=1}^{n} \frac{X_i}{n}\right)} = \sqrt{\left(\sum_{i=1}^{n} \frac{1}{n^2}\,\mathbb{V}_{\lambda^*}(X_i)\right)} \\ &= \sqrt{\frac{n}{n^2}\,\mathbb{V}_{\lambda^*}(X_i)} = \sqrt{\theta^*(1-\theta^*)/n}\ .\end{aligned}$$

Another reasonable property of an estimator is that it converge to the "true" parameter θ^* – here "true" means the supposedly fixed and possibly unknown θ^*, as we gather more and more IID data from a θ^*-specified DF $F(x; \theta^*)$. This property is stated precisely next.

Definition 5.11 (Asymptotic Consistency of a Point Estimator)**.** *A point estimator* $\widehat{\Theta}_n$ *of* $\theta^* \in \Theta$ *is said to be* **asymptotically consistent** *if:*

$$\widehat{\Theta}_n \xrightarrow{P} \theta^*.$$

Definition 5.12 (Mean Squared Error (MSE) of a Point Estimator)**.** *Often, the quality of a point estimator* $\widehat{\Theta}_n$ *of* $\theta^* \in \Theta$ *is assessed by the* **mean squared error** *or* MSE *defined by:*

$$\mathsf{MSE}_n(\widehat{\Theta}_n(X)) := \mathbb{E}_{\lambda^*}\left((\widehat{\Theta}_n(X) - \theta^*)^2\right)\ .$$

The following proposition shows a simple relationship between the mean square error, bias and variance of an estimator $\widehat{\Theta}_n$ of θ^*.

Proposition 5.13 (The $\sqrt{\mathsf{MSE}_n} : \mathsf{se}_n : \mathsf{bias}_n$–Sided Right Triangle of an Estimator)**.** *Let* $\widehat{\Theta}_n$ *be an estimator of* $\theta^* \in \Theta$*. Then:*

$$\mathsf{MSE}_n(\widehat{\Theta}_n) = (\mathsf{se}_n(\widehat{\Theta}_n))^2 + (\mathsf{bias}_n(\widehat{\Theta}_n))^2\ . \tag{5.3}$$

Proof. Consider the mean squared error

$$\begin{aligned}\mathbb{E}_{\lambda^*}\left[(\widehat{\Theta}_n(X) - \theta^*)^2\right] &= \mathbb{E}_{\lambda^*}\left[(\widehat{\Theta}_n(X) - \theta^* - \mathbb{E}\left[\widehat{\Theta}_n(X)\right] + \mathbb{E}\left[\widehat{\Theta}_n(X)\right])^2\right] \\ &= \mathbb{E}_{\lambda^*}\left[(\widehat{\Theta}_n(X) - \mathbb{E}\left[\widehat{\Theta}_n(X)\right])^2\right] \\ &\quad + \mathbb{E}_{\lambda^*}\left[(\theta^* - \mathbb{E}\left[\widehat{\Theta}_n(X)\right])^2\right] \\ &\quad + 2\,\mathbb{E}_{\lambda^*}\left[(\widehat{\Theta}_n(X) - \mathbb{E}\left[\widehat{\Theta}_n(X)\right])(\mathbb{E}\left[\widehat{\Theta}_n(X)\right] - \theta^*)\right]\end{aligned}$$

the first expectation on the RHS is just the standard error, the second is the bias and the last expectation is 0 because

$$\mathbb{E}_{\lambda^*}\left[(\widehat{\Theta}_n(X) - \mathbb{E}\left[\widehat{\Theta}_n(X)\right](\mathbb{E}\left[\widehat{\Theta}_n(X)\right] - \theta^*)\right] \\ = \mathsf{bias}(\widehat{\Theta}_n(X))\,\mathbb{E}\left[\widehat{\Theta}_n(X) - \mathbb{E}\left[\widehat{\Theta}_n(X)\right]\right] = 0.$$

□

Proposition 5.14 (Asymptotic consistency of a point estimator)**.** *Let $\widehat{\Theta}_n$ be an estimator of $\theta^* \in \Theta$. Then, if $\mathsf{bias}_n(\widehat{\Theta}_n) \to 0$ and $\mathsf{se}_n(\widehat{\Theta}_n) \to 0$ as $n \to \infty$, the estimator $\widehat{\Theta}_n$ is asymptotically consistent:*

$$\widehat{\Theta}_n \xrightarrow{P} \theta^* \ .$$

Proof. If $\mathsf{bias}(\widehat{\Theta}_n) \to 0$ and $\mathsf{se}(\widehat{\Theta}_n) \to 0$, then by (5.3), $\mathsf{MSE}(\widehat{\Theta}_n) \to 0$, i.e.

$$\mathbb{E}_{\lambda^*}\left[(\widehat{\Theta}_n - \theta^*)^2\right] \to 0.$$

That is, $\widehat{\Theta}_n(X) \to \theta^*$ in $L^2(\mathbb{P})$ which implies convergence in probablity, see Section 3.2.1. □

We want our estimator to be unbiased with small standard errors as the sample size n gets large.

Example 5.15 (Asymptotic consistency of our Estimator of θ^*)**.** *Consider the sample mean estimator $\widehat{\Theta}_n(X) := \overline{X}_n$ of θ^*, from $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim}$ Bernoulli(θ^*). Since $\mathsf{bias}_n(\widehat{\Theta}_n) = 0$ for any n and $\mathsf{se}_n = \sqrt{\theta^*(1-\theta^*)/n} \to 0$, as $n \to \infty$, by Proposition 5.14, $\widehat{\Theta}_n \xrightarrow{P} \theta^*$. That is $\widehat{\Theta}_n$ is an* **asymptotically consistent estimator** *of θ^*.*

We saw in Section 3.1 that the concentration inequalities gives us quite some control, let us see an application to the mean square error of the sample mean estimator.

Lemma 5.16. *Let Y be a RV satisfying the estimate for fixed $c_0 \geq 1$ and for all $\epsilon > 0$*

$$\mathbb{P}(|Y| \geq \epsilon) < 2e^{-c_0\epsilon^2}. \tag{5.4}$$

Then

$$\mathbb{E}\left[|Y|^2\right] \leq \frac{5}{c_0}$$

Before we proceed with the proof let us take an example

Example 5.17. *Let us revisit the problem of estimating the mean of an L^2 RV. Let $X_1, \ldots, X_n$ be i.i.d. RVs in $L^2(\mathbb{P})$ that are also sub-Gaussian with parameter σ, then using Theorem 3.6 and Lemma 5.16 we get*

$$\mathsf{MSE}(\overline{X}_n) = \mathbb{E}\left[|\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right]|^2\right] \leq \frac{10\sigma^2}{n}$$

So for sub-Gaussian RVs we have almost the same standard error as if the random variables where Gaussian. Optimizing the proof of Lemma 5.16 we can eek out a smaller constant.

Proof. Let $\delta > 0$, we will choose it later

$$\mathbb{E}\left[|Y|^2\right] \leq \mathbb{E}\left[\sum_{k=2}^{\infty} \delta^2 k^2 \mathbb{1}_{\delta(k-1)\leq Y<\delta k}\right] + \mathbb{E}\left[Y^2 \mathbb{1}_{Y\leq\delta}\right] = I + II$$

We first estimate II by noting that

$$\mathbb{E}\left[Y^2 \mathbb{1}_{Y\leq\delta}\right] \leq \delta^2$$

To estimate I note that according to (5.4) we get

$$\begin{aligned} I = \mathbb{E}\left[\sum_{k=2}^{\infty} \delta^2 k^2 \mathbb{1}_{\delta(k-1)\leq Y<\delta k}\right] &\leq \mathbb{E}\left[\sum_{k=2}^{\infty} \delta^2 k^2 \mathbb{1}_{\delta(k-1)\leq Y}\right] \\ &\leq 2\sum_{k=2}^{\infty} \delta^2 k^2 e^{-c_0\delta^2(k-1)^2} \end{aligned}$$

Now choose $\delta^2 = 1/c_0$, from this we get

$$\sum_{k=2}^{\infty} \delta^2 k^2 e^{-c_0\delta^2(k-1)^2} \leq \frac{1}{c_0}\sum_{k=2}^{\infty} k^2 e^{-(k-1)^2} \leq \frac{2}{c_0}.$$

Putting it all together we get

$$\mathbb{E}\left[|Y|^2\right] \leq \frac{1}{c_0} + \frac{4}{c_0} \leq \frac{5}{c_0}.$$

□

Exercise 5.18. *If you use the equality*

$$\mathbb{E}\left[X\right] = \int_0^{\infty} \mathbb{P}(X > t)dt$$

(valid for non-negative RV's) can you improve upon the constant in Lemma 5.16?

Exercise 5.19. *Advanced!! What happens in Lemma 5.16 if we replace sub-Gaussian with sub-exponential?*

5.3 Non-parametric DF Estimation

So far, we have been interested in some estimation problems where the parameter map has a finite dimensional range. For instance, in the mean estimation problem of $\mathbb{R}$ valued RVs, the space Θ is of dimension 1. Similarly, if we are estimating the mean and variance the space Θ is of dimension 2.

Next we consider a non-parametric experiment in which n IID samples are drawn according to some fixed and possibly unknown DF F^* from the space of **All Distribution Functions**: That is, our statistical model is

$$\mathcal{E} := \{\text{All DFs}\} := \{\, F(x; F) : F \text{ is a DF} \,\}$$

and we assume that there is an $F^* \in \mathcal{E}$, which is the DF for an i.i.d. sequence $X_1, \ldots, X_n$. Here the parameter space is $\mathcal{E}$ itself, which is infinite dimensional and the parameter map θ is the identity map, so $\Theta = \mathcal{M}$.

Consider now a Model space which is

$$\mathcal{M}_0 := \{F(x) = \int_{-\infty}^{x} p(x; p)dx, \quad p \text{ is a PDF}\}$$

that is the space of all DFs from all continuous random variables. Let $F^* \in \mathcal{M}_0$ be a DF, let $X = (X_1, \ldots, X_n)$ be i.i.d. from DF F^*, and for simplicity of presentation assume that X is continuous and $f^* = (F^*)'$. For any distribution function $G \in \mathcal{M}_0$ with density g, we define the relative entropy loss functional as

$$L(x, G) = \ln\left(\frac{f^*(x)}{g(x)}\right)$$

and the relative entropy risk becomes

$$R(G) = \int \ln\left(\frac{f^*(x)}{g(x)}\right) f^*(x)dx.$$

The relative entropy risk is just the relative entropy between F and G, it can also be identified with the Kullbach-Leibler divergence between F and G. We would like to minimize $R(G)$ over all distribution functions G, we know from Section 4.2 that F is the minimizer. However we only have access to the empirical risk, namely

$$\widehat{R}_n(p; X) = \frac{1}{n}\sum_i \ln\left(\frac{f^*(X_i)}{g(X_i)}\right)$$

Exercise 5.20. *Show that the relative entropy risk is the same risk as we saw in Section 4.2, it only differs by a constant.*

If the law of large numbers is applicable we know that

$$\lim_{n\to\infty} \frac{1}{n} \sum_i \ln\left(\frac{f^*(X_i)}{g(X_i)}\right) = R(G).$$

The infimum of the empirical risk over $\mathcal{M}_0$ is given by a CDF $F_n \notin \mathcal{M}_0$ of a discrete RV for which PMF that puts weight $1/n$ on each X_i, i.e. the PMF

$$p_n(x; X) = \frac{1}{n} \sum_i \mathbb{1}_{x=X_i}.$$

Exercise 5.21. *Prove that the minimizer is necessarily a discrete measure.*

Exercise 5.22. *Prove that among all the discrete distributions with support on the X_i the uniform one is minimizing the risk.*

> We know that in good cases the LLN implies that the empirical risk for a fixed DF converges to the risk.
>
> Question: What happens to the minimum of the empirical risk, does it converge to the minimum of the risk as $n \to \infty$?

Definition 5.23. *Let $X = (X_1, \ldots, X_n)$ be an i.i.d. sequence of RVs with DF F. We denote*

$$\widehat{F}_n(x; X) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \le x}$$

we call $\widehat{F}_n(x; X)$ the empirical distribution function. Often we will supress the X, and just write $\widehat{F}_n(x)$, but we must never forget X.

For each fixed $x \in \mathbb{R}$, $\widehat{F}_n(x; X)$ is a statistic and is thus a RV. We can say that $\widehat{F}_n(x; X) : \mathbb{R} \times \mathbb{X} \to [0, 1]$ is a random function. It has the following properties:

Lemma 5.24. *Let $X = (X_1, \ldots, X_n)$ be an i.i.d. sequence of RVs with DF F. Let $\widehat{F}_n(x; X)$ be the empirical distribution function, then*

1.

$$\mathbb{E}\left[\widehat{F}_n(x; X)\right] = F(x)$$

2.

$$\mathbb{V}\left[\widehat{F}_n(x; X)\right] = \frac{F(x)(1 - F(x))}{n}$$

Proof. To compute the expectation, note that

$$\mathbb{E}\left[\widehat{F}_n(x;X)\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i\leq x}\right] = F(x).$$

To compute the variance we simply compute

$$\begin{aligned}\mathbb{E}\left[\widehat{F}_n(x;X)^2\right] &= \mathbb{E}\left[\frac{1}{n^2}\sum_{i,j=1}^{n}\mathbb{1}_{X_i\leq x}\mathbb{1}_{X_j\leq x}\right] \\ &= \frac{1}{n^2}\sum_{i\neq j}F(x)^2 + \frac{1}{n^2}\sum_{i=j}F(x) = F(x)^2 - \frac{n}{n^2}F(x)^2 + \frac{n}{n^2}F(x)\end{aligned}$$

So,

$$\mathbb{V}\left[\widehat{F}_n(x;X)^2\right] = \frac{1}{n}F(x)(1-F(x)).$$

□

Remark 5.25. *We see from the above that the empirical distribution function is unbiased and asymptotically consistent. Note: We dont even need to know anything about the integrability of X as the empirical distribution function is always a bounded RV, i.e. takes values between* $[0,1]$.

If we use Hoeffdings inequality we get the following concentration:

Lemma 5.26. *Let $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} F$. Then, for any $\epsilon > 0$ and x;*

$$\mathbb{P}(|\widehat{F}_n(x;X) - F(x)| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

This is perhaps weaker than we would like, actually we can prove that the same estimate holds but over all x at the same time. The next proposition is often referred to as the **fundamental theorem of statistics** and is at the heart of non-parametric inference, empirical processes, and computationally intensive bootstrap techniques.

Theorem 5.27 (The Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality)**.** *Let $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} F$. Then, for any $\epsilon > 0$:*

$$P\left(\sup_x |\widehat{F}_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}. \tag{5.5}$$

You have now seen the first example of the empirical risk minimization framework. This lies at the heart of machine learning, we will see this later in the pattern recognition problem, where, for certain not overly complex model spaces $\mathcal{M}_0$ we can guarantee estimates like the DKW above. This is coined the uniform convergence of empirical means (ECEMP).

5.4 Plug-in Estimators of Statistical Functionals: Direct estimation

A **statistical functional** is simply any function of the DF F. For example, the median $T(F) = F^{[-1]}(1/2)$ is a statistical functional. Thus, $T(F) : \{\text{All DFs }\} \to \mathbb{T}$, being a map or function from the space of DFs to its range $\mathbb{T}$, is a functional. The idea behind the plug-in estimator for a statistical functional is simple: just plug-in the point estimate $\widehat{F}_n$ instead of the unknown DF F^* to estimate the statistical functional of interest.

Definition 5.28 (Plug-in estimator)**.** *Suppose, $X_1, \ldots, X_n \overset{IID}{\sim} F^*$. The plug-in estimator of a statistical functional of interest, namely, $T(F^*)$, is defined by:*

$$\widehat{T}_n := \widehat{T}_n(X_1, \ldots, X_n) = T(\widehat{F}_n) \ .$$

Definition 5.29 (Linear functional)**.** *If $T(F) = \int r(x)dF(x)$ for some function $r(x) : \mathbb{X} \to \mathbb{R}$, then T is called a* **linear functional**. *Thus, T is linear in its arguments:*

$$T(aF + a'F') = aT(F) + a'T(F') \ .$$

Proposition 5.30 (Plug-in Estimator of a linear functional)**.** *The plug-in estimator for a linear functional $T = \int r(x)dF(x)$ is:*

$$T(\widehat{F}_n) = \int r(x)d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(X_i) \ .$$

Furthermore, if $r(X) \in L^2(\mathbb{P})$ then the estimator $T(\widehat{F}_n)$ is unbiased and asymptotically consistent.

Proof. That $T(\widehat{F}_n)$ is asymptotically consistent follows from Theorem 3.26. □

Remark 5.31. *This means that any plug in estimator of a linear functional is actually the sum of independent RVs. If $r(X)$ is nice enough, say sub-Gaussian or sub-exponential, then we can utilize the concentration inequalities in Section 3.1.*

Remark 5.32. *However, there are non-linear functionals that one is often interested in. For instance the median $T(F) = F^{-1}(\frac{1}{2})$.*

Definition 5.33. *The influence function is defined as*

$$L_F(y) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon \mathbb{1}_{x \geq y}) - T(F)}{\epsilon}$$

If T is "nice enough" we can basically use the above definition of derivative to cook up a first order approximation which would look like

$$T(\widehat{F}_n) - T(F) \approx \int L_F(y) d\widehat{F}_n = \frac{1}{n}\sum L_F(X_i).$$

The point here being that the linear term is of leading order for large n, i.e. the quadratic term is quadratic in $1/n$.

If $L_F(X_i) \in L^2(\mathbb{P})$ then according to the central limit theorem Theorem 3.27 we have that $\frac{1}{\sqrt{n}}\sum L_F(X_i)$ is asymptotically normal. Keep this in mind when we later go in to **the bootstrap**.

Lemma 5.34. *Let F be a DF and $a \in (0,1)$ a quantile, then if F is differentiable at $m = F^{[-1]}(x)$ with positive derivative (actually F is invertible at this point). Then the influence function for the quantile a is*

$$L_F(y) = \frac{a - \mathbb{1}_{m \geq y}}{\frac{dF}{dx}(m)}$$

where $m = F^{[-1]}(a)$. We thus see that L_F is bounded if the density at the median is non-zero, and as such $L_F(X)$ is sub-Gaussian and we get good concentration for the first order term.

Remark 5.35. *Note that even though L_F is bounded, we cannot know this bound as it depends on the density at the quantile. It is fairly easy to construct a density which is zero at the quantile a. Think of the median in a symmetric bimodal distribution for which the density is 0 at the middle between the two modes.*

This is actually more problematic than it seems! If we define the median as

$$F^{[-1]}(1/2)$$

then in some cases this is a set and F is not invertible at $1/2$. Thus demanding consistency does not really make much sense. For this, there is a notion of weak consistency

Proof. Let y be given and consider

$$q = T((1-\epsilon)F + \epsilon \mathbb{1}_{x \geq y})$$

Let $m = F^{[-1]}(a)$ then if $y > m$ we get

$$(1-\epsilon)F(q) = a$$
$$q = F^{[-1]}(\frac{a}{1-t}).$$

In the case that $y \leq m$ we get

$$(1-\epsilon)F(q) + \epsilon = a$$
$$q = F^{[-1]}(\frac{a-t}{1-t})$$

We now know that

$$\left.\frac{dq}{dt}\right|_{t=0} = \frac{a - \mathbb{1}_{m \geq y}}{\frac{dF}{dx}(m)}$$

□

Let us dig deeper into the quantiles, let us define the quantile function

$$F^{-1}(p) = \inf\{x : F(x) \geq p\}$$

then F^{-1} is a left-continuous function with range equal to the support of F and is hence unbounded. Note the subtle difference between the set-valued formal inverse $F^{[-1]}$ and the quantile function F^{-1}. Let us record some interesting properties of F^{-1}.

Lemma 5.36. *For every $0 < p < 1$ and $x \in \mathbb{R}$,*

1. *$F^{-1}(p) \leq x$ if and only if $p \leq F(x)$*
2. *$F(F^{-1}(p)) \geq p$ with equality iff p is in the range of F.*
3. *$F^{-1}(F(x)) \leq x$, where equality fails iff x is in the interior or at the right end of a flat piece of F.*
4. *$F^{-1}(F(F^{-1})) = F^{-1}$, and $F(F^{-1}(F)) = F$.*

Exercise 5.37. *Prove this lemma!*

Recall that a uniform distribution function on the interval $[0,1]$ is given by $F_{unif}(x) = \min(\max(x,0),1)$. Let $U \sim F_{unif}$ then (1) above implies that

$$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = \min(\max(F(x),0),1) = F(x)$$

in other words $F^{-1}(U) \sim F$. Let us collect that as a theorem

Theorem 5.38 (Inversion sampling)**.** *If $U \sim$ Uniform($[0,1]$) and F is a DF, then $F^{-1}(U) \sim F$.*

Some specific examples of statistical functionals we have already seen include:

1. The **mean** of RV $X \sim F$ is a function of the DF F:
$$T(F) = \mathbb{E}(X) = \int x \, dF(x) \ .$$

2. The **variance** of RV $X \sim F$ is a function of the DF F:
$$T(F) = \mathbb{E}(X - \mathbb{E}(X))^2 = \int (x - \mathbb{E}(X))^2 \, dF(x) \ .$$

3. The **value of DF at a given** $x \in \mathbb{R}$ of RV $X \sim F$ is also a function of DF F:
$$T(F) = F(x) \ .$$

4. The $q^{\textbf{th}}$ **quantile** of RV $X \sim F$:
$$T(F) = F^{[-1]}(q) \ \text{ where } q \in [0, 1] \ .$$

5. The **first quartile** or the $0.25^{\textbf{th}}$ **quantile** of the RV $X \sim F$:
$$T(F) = F^{[-1]}(0.25) \ .$$

6. The **median** or the **second quartile** or the $0.50^{\textbf{th}}$ **quantile** of the RV $X \sim F$:
$$T(F) = F^{[-1]}(0.50) \ .$$

7. The **third quartile** or the $0.75^{\textbf{th}}$ **quantile** of the RV $X \sim F$:
$$T(F) = F^{[-1]}(0.75) \ .$$

Chapter 6

Random Variable Generation

Definition 6.1 (Informal)**.** *A* **uniform pseudorandom number generator** *(UPRNG) is an algorithm which starting from an initial value u_0 and a transformation D, produces a sequence $u_i = D(u_{i-1})$ in $[0,1]$ for $i = 1, \ldots$. For all n, $u_1, \ldots, u_n$ approximate the behavior of an i.i.d. sequence of* Uniform$([0,1])$ *random numbers.*

We could provide a mathematical definition of a UPRNG.

Definition 6.2 (Formal)**.** *Let $u_0 \in [0,1]$ and let $D : [0,1] \to [0,1]$, define the dynamical system*

$$u_i = D(u_{i-1}), \quad i = 1, 2, \ldots$$

For a set $A \subset [0,1]$, define $N_n(A)$ as the number of $u_i \in A$ for $i = 0, 1, 2, \ldots, n-1$. We call D a UPRNG if and only if for every $u_0 \in [0,1]$ and every Borel set $A \subset [0,1]$

$$\frac{N_n(A)}{n} \to \int_A dx.$$

In words, no matter the starting point u_0 the long term relative frequency of the event $u_i \in A$ approaches the probability of that event for a uniform random RV as $n \to \infty$.

But before we get to the UPRNG let us define a pseudorandom sequence

Definition 6.3 (pseudorandom)**.** *Consider the finite set $\mathcal{M} = \{0, 1, \ldots, M-1\}$ and consider the sequence $u_0, u_1, \ldots \in \mathcal{M}$. For every $a \in \mathcal{M}$, define $N_n(a)$ as the number of $u_i = a$ for $i = 0, 1, 2, \ldots, n-1$. We call the sequence $u_0, u_1, \ldots$* **pseudorandom** *on $\mathcal{M}$ if and only if for every $a \in \mathcal{M}$*

$$\frac{N_n(a)}{n} \to \frac{1}{M}.$$

6.1 Congruential Generators

Definition 6.4. *Let u_0 be fixed and let D be a map, define the dynamical system*

$$u_i = D(u_{i-1}), \quad i = 1, \ldots$$

We call T_0 the period of D started at u_0 the smallest positive integer such that

$$u_{i+T_0} = u_i, \text{ for some } i.$$

The smallest period T for all admissible starting points u_0 is called the period for D.

Exercise 6.5. *If we start at a fixed point u_0, and let T_0 be the period of D w.r.t u_0, then if*

$$u_{i+T_0} = u_i$$

holds for some i, then it holds for all i.

Definition 6.6. *A* **congruential generator** *with parameters (a, b, M) on $\{0, 1, \ldots, M-1\}$ is defined by the function*

$$D(x) = (ax + b) \mod M.$$

Example 6.7. *The congruential generator $(3, 0, 16)$ on $\{0, 1, \ldots, 16\}$, has many different periods. For instance if $u_0 = 0$ then the period for 0 is 0. If we instead start at 1 then the period is 4. If we start at 2 the period is 2. etc.*

Is it possible for a congruential generator to generate something pseudo-random?

Lemma 6.8. *Consider a congruential generator D on $\mathcal{M} = \{0, 1, \ldots, M-1\}$ with period M, then for any starting point $u_0 \in \mathcal{M}$, the sequence $u_i = D(u_{i-1})$ is pseudorandom on $\mathcal{M}$.*

Exercise 6.9. *Prove the above lemma.*

The problem with a congruential generator on $\mathcal{M}$ is that the period is as long as the number of unique values, this will be problematic if M is small. What we can do is to use the congruential generator for a larger set and restrict it to a smaller to get a better generator.

Lemma 6.10. *Consider a congruential generator D on $\mathcal{M} = \{0, 1, \ldots, M-1\}$ with period M, then for any starting point $u_0 \in \mathcal{M}$, define $u_i = D(u_{i-1})$ then the sequence $v_i = u_i \mod K$ for $1 \leq K \leq M$ is pseudorandom on $\{0, 1, \ldots, K-1\}$ if M is a multiple of K.*

Exercise 6.11. *Prove the above lemma.*

The following number theoretical theorem tells us exactly when we can expect period M.

Theorem 6.12 (Hull–Dobell Theorem)**.** *The congruential generator (a, b, M) has period M iff*

- $\gcd(b, M) = 1$,
- *p divides $a - 1$ for every prime p that divides M*
- *4 divides $a - 1$ if 4 divides M.*

See [HD, K].

Let us check the moments of the pseudorandom numbers generated

Lemma 6.13. *Let $u_0, u_1, \ldots$ be a psuedo random sequence over $\mathcal{M} = \{0, 1, \ldots, M-1\}$. Then the empirical mean and variance has limits as follows*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} u_i = \frac{M-1}{2}$$

and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} u_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} u_i \right)^2 = \frac{M^2 - 1}{12}.$$

Proof. From Definition 6.3

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} u_i = \lim_{n\to\infty} \sum_{i=0}^{M-1} i N_n(i) = \sum_{i=0}^{M-1} \frac{i}{M} = \frac{M-1}{2}$$

The empirical variance follows similarly. □

The conclusion is that the long term empirical moments converge to the discrete uniform over $\mathcal{M}$.

We saw earlier that rescaling the result by taking the modulus gives us a generator over a smaller set. Our initial problem of generating number from the uniform distribution can be partially solved by a generator of large period.

Corollary 6.14. *Let $u_0, u_1, \ldots$ be a psuedo random sequence over $\mathcal{M} = \{0, 1, \ldots, M-1\}$. Then $v_i = u_i/M$ has the empirical mean and variance limits as follows*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} v_i = \frac{1}{2} - \frac{1}{2M}$$

and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} v_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} v_i \right)^2 = \frac{1}{12} - \frac{1}{12M^2}.$$

From the above we see that if we have a generator with a large period M, which could be over 64-bit integers, i.e. we could have period 2^{64}, then the resulting u_i/M will have mean $\frac{1}{2} - 2^{-65}$ and variance $\frac{1}{12} - \frac{1}{3}2^{-130}$. Will such a generator be an UPRNG? Actually no, but we can achieve the following

Lemma 6.15. *Let $v_0, v_1, \ldots$ be a pseudorandom sequence in $\mathcal{M} = \{0, 1, \ldots, M-1\}$, define $u_i = v_i/M$. For any interval $A = (a, b) \subset [0, 1]$, define $N_n(A)$ as the number of $u_i \in A$ for $i = 0, 1, 2, \ldots, n-1$. We have*

$$\left| \lim_{n\to\infty} \frac{N_n(A)}{n} - \int_A dx \right| \leq \frac{1}{M}.$$

For reasons a little bit beyond this course, it turns out that we cannot represent a UPRNG on a finite machine, we can however generate almost an UPRNG as seen in the lemma above.

Now all of what we have done so-far is to check whether the sequence has the right limiting distribution and the right moments. But, this doesn't tell us anything about the randomness, for instance the sequence $0, 1, 0, 1, 0, 1, \ldots$ is a pseudorandom sequence over $\{0, 1\}$, it is however very structrured and thus not random. There is a series of tests one usually does to verify if a pseudorandom sequence is good, we will see this a bit more in the computer exercises.

6.2 Sampling

The previous section was just to give you a flavor of what random in a computer represents, namely it is not random but a deterministic sequence that "looks" random. We will leave the pseudorandom part for now and instead attack the problem of sampling from a generic distribution given a random sample from Uniform($[0, 1]$).

> Recall from Theorem 5.38 that given a distribution function F and $X \sim \text{Uniform}([0, 1])$, then $F^{-1}(X) \sim F$.

However, sometimes finding the quantile function F^{-1} can be analytically impossible or if done numerically, very expensive. There are other ways to sample that are more costly than inversion sampling (given the inverse) but sometimes cheaper than computing the inverse.

Algorithm 1 Accept-Reject Sampler

1: *input:*

(1) a target density $f(x)$,

(2) a sampling density $g(x)$ that satisfies $f(x) \leq Mg(x)$.

2: *output:* a sequence of samples $x_0, \ldots$ with distribution f
3: Sample initial state $X^{(0)}$ from g.
4: **repeat**
5: At iteration t,
6: Generate x from g and compute the ratio $r(x) = \frac{f(x)}{Mg(x)}$
7: Draw $U \sim \text{Uniform}([0,1])$ and set $X^{t+1} = x$, if $U \leq r(x)$, otherwise **goto** 6?
8: **until** desired number of samples are obtained.

As you can see in Algorithm 1, the updated variable is a conditional random variable on the event $U \leq r(X)$, the distribution of this conditional random variable is the distribution we are after, namely F. This is contained in the lemma below:

Lemma 6.16. *Let $X \sim G$ and let $U \sim \text{Uniform}([0,1])$, then if we define the RV $I = \mathbb{1}_{U \leq r(X)}$, the random variable $Y = X \mid (I = 1)$ satisfies $Y \sim F$.*

Proof. By the properties of conditional densities we have the equality (Bayes)

$$f_{X|I}(x \mid I = 1) = \frac{f_{I|X}(I = 1 \mid X = x) f_X(x)}{f_I(1)}.$$

Let us compute the constituents of the RHS of the above. Now since I is discrete we know that

$$f_{I|X}(I = 1 \mid X = x) = \mathbb{P}(I = 1 \mid X = x) = \mathbb{P}(U \leq r(x)) = r(x)$$

and from the law of total probability (Theorem 1.16)

$$f_I(1) = \mathbb{P}(I = 1) = \int \mathbb{P}(I = 1 \mid X = x) g(x) dx = \frac{1}{M} \int p(x) dx = \frac{1}{M}.$$

Finally we achieve

$$f_{X|I}(x \mid I = 1) = \frac{r(x)g(x)}{1/M} = \frac{f(x)g(x)M}{Mg(x)} = f(x)$$

□

In some cases there is not really simple density g to the density f that you wish to sample from, so the accept-reject does not work. An example of this is the Normal distribution, which basically requires us to have a density g which is Gaussian like if we want the algorithm to not reject way too much. So how do we do it? We have a few options

1. Since the distribution function for the Gaussian does not have a closed form, the inverse is hard to compute and requires a lot of computation. There does however exist approximations of the inverse that we can use.

$$\Phi^{-1}(\alpha) \approx t - \frac{a_0 + a_1 t}{1 + b_1 t + b_2 t}.$$

   You will find the constants in the notebooks.

2. Use a transformation method, i.e. find some kind of function $h(x)$ and a simpler distribution F such that $h(X)$ is Gaussian or close to it.

Theorem 6.17 (Box-Muller)**.** *Suppose that* $U_1, U_2 \overset{\text{IID}}{\sim} \text{Uniform}([0,1])$*, then*

$$Z_0 = \sqrt{-2\ln(U_1)}\cos(2\pi U_2)$$
$$Z_1 = \sqrt{-2\ln(U_1)}\sin(2\pi U_2)$$

are independent random variables, and $Z_0, Z_1 \sim \text{Normal}(0,1)$*.*

Proof. Consider bivariate normal RV. Z, then the distribution of $Y = |Z|^2$ is χ^2 distributed with 2 degrees of freedom. Furthermore $W = Z/|Z|$, is uniformly distributed on the unit circle. We know that Z, W are independent (see exercise below). Thus to generate a bivariate normal it is enough to generate from a χ^2 distribution with 2 degrees of freedom and a point from the uniform distribution on the circle. The χ^2 with 2 degrees of freedom is just the exponential distribution with parameter $1/2$, as such we can generate it using the inversion sampling method (Theorem 5.38). The rest of the proof is left as an exercise. □

Exercise 6.18. *First show that* W, Y *in the proof above are independent. Then show that* W *generated using* $(\cos(2\pi U_2), \sin(2\pi U_2))$ *is uniform on the unit circle. Finally to show that* Z_0, Z_1 *are independent, since they are Gaussian it suffices to show that their covariance is zero.*

6.3 Practice exercises

Exercise 6.19.

- *Implement your own congruential generator* (a, b, M).
- *Use the congruential generator to generate pseudo random numbers from* Uniform$([0, 1])$. *Test out a combination* (a, b, M) *that seems to work well when tested with for instance, Kolmogorov Smirnoff. This is easy to do, let* $\hat{F}_n$ *denote the empirical distribution function according to the* n *samples drawn, compare this to the distribution function for the uniform distribution* F *and consider*

$$\sup_{x\in[0,1]} |\hat{F}_n(x) - F(x)|.$$

Derive a statistical test based on Theorem 5.27 and test whether your sampler passes.

Exercise 6.20.

- *Now that you can sample from the uniform distribution, generate samples from* $N(10, 5)$ *using the Box-Muller method, Theorem 6.17.*
- *Repeat the testing from Exercise 6.19 but now compare to* F *being the normal distribution.*
- *What did you actually assume for the test above? Is it accurate? Does it satisfy the conditions of Theorem 5.27?*

Exercise 6.21. *Consider the continuous distribution with density*

$$p(x) = \frac{1}{2}\cos(x), \quad -\frac{\pi}{2} < x < \frac{\pi}{2}.$$

- *Plot the distribution function* $F(x)$.
- *Find the inverse distribution function* F^{-1}.
- *Implement an inversion sampler to sample from* F.
- *Implement an Accept-Reject Sampler, Algorithm 1 with sampling density* Uniform$([-\pi/2, \pi/2])$. *On average, how many samples get rejected?*

Chapter 7

Finite Markov Chains

Markov chains that we will be studying in this chapter is a stochastic process, which we have yet to define:

Definition 7.1. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability triple. A* $\mathbb{R}$*-valued* **stochastic process** *is a parametrized set of RVs. That is, we denote the collection*

$$(X_\alpha)_{\alpha \in \mathbb{A}}$$

for a parameter set $\mathbb{A}$*, a* $\mathbb{R}$*-valued stochastic process.*

If the index set $\mathbb{A} = \mathbb{N}$ *we call it a* **discrete (or discrete-time) stochatic process**.

That is, previously we have used the concept of an i.i.d. sequence of random variables, that is, $X_1, X_2, \ldots \overset{\text{IID}}{\sim} F$. This sequence is a simple case of a discrete stochastic process. To study this sequence as a process is quite uninteresting as all X_i are independent, we will in this chapter introduce some dependency an analyze the resulting structure. These discrete stochastic processes are termed finite Markov chains. We will cover their properties and simulation methods.

7.1 Introduction

A finite Markov chain is a stochastic process that moves among elements in a finite set $\mathbb{X}$ as follows: when at $x \in \mathbb{X}$ the next position is chosen at random according to a fixed probability distribution $P(\cdot|x)$. We define such a process more formally below.

7.1.1 Advanced intro*

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $\mathcal{G}$ be a sigma algebra on Ω. Define for a $\mathbb{R}$ valued R.V. X the conditional expectation

$$\mathbb{E}[X \mid \mathcal{G}]$$

which is any $\mathcal{G}$ measurable function (is it unique?) $\Omega \to \mathbb{R}$ which satisfies for any $G \in \mathcal{G}$

$$\int_G \mathbb{E}[X \mid \mathcal{G}] dP = \int_G X dP$$

This can be thought of the best possible guess of X given the knowledge contained in $\mathcal{G}$. The conditional probability can be defined as

$$\mathbb{P}(X \in A \mid \mathcal{G}) := \mathbb{E}[\mathbb{1}_A(X) \mid \mathcal{G}]$$

which constitutes the single best guess for the if the event $X \in A$ happened given the information contained in $\mathcal{G}$.

Remark 7.2. *If we think of a random variable $X \in L^2(\mathbb{P})$, then for any σ-algebra $\mathcal{G}$ we see that*

$$\mathbb{E}[(X - \mathbb{E}[X \mid \mathcal{G}])\mathbb{1}_G] = 0$$

for all $G \in \mathcal{G}$, which says that the $L^2(\mathbb{P})$ random variable $X - \mathbb{E}[X \mid \mathcal{G}]$ is orthogonal to all indicators $\mathbb{1}_G$, $G \in \mathcal{G}$. In this case the conditional expecation is unique and can be thought of as a projection of X onto $\mathcal{G}$.

Remark 7.3. *Properties:*

- $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] = \mathbb{E}[X]$*, so the tower property still holds.*
- *If X is $\mathcal{G}$ measurable (i.e. we know X), then $\mathbb{E}[X \mid \mathcal{G}] = X$, i.e. we are allowed to guess X itself since we know it. There is no better guess.*
- *If X is independent of $\mathcal{G}$, i.e. the information in $\mathcal{G}$ is irrelevant for X, then $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$, i.e. we gained nothing.*

Consider now a stochastic process $X_1, \ldots, X_n$ with index set $\mathbb{N}$. Define $\mathcal{F}_n$ as the smallest σ-algebra on Ω such $(X_1, \ldots, X_n)$ is an $\mathbb{R}^n$ valued RV.

We can evaluate this on a particular realization of $(X_1, \ldots, X_n)$ as follows

$$\mathbb{P}(X_t \in A \mid \mathcal{F}_n) = \mathbb{P}(X_t \in A \mid x_1, \ldots, x_n)((X_1, \ldots, X_n)).$$

From the above it is clear that $\mathbb{P}(X_t \in A \mid \mathcal{F}_n)$ is a random variable that depends on $(X_1, \ldots, X_n)$. Note that $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdot$ is an increasing family of σ-algebras, all of which are subsets of $\mathcal{F}$, such a sequence is denoted a filtration.

Remark 7.4. *This can be seen as over-complicating it. Why do we even need this level of formalism? Since the statespace $\mathbb{X}$ is finite we could equally well say*

$$\mathbb{P}(A \mid X_1 = x, X_2 = x_2, \ldots, X_n = x_n)$$

which we already know what it means. The reason is

1. *It simplifies notation.*
2. *If you take a course in continuous stochastic processes then you would have to use this notation.*
3. *If the statespace is not enumerable, then you would also have to use this formalism.*
4. *The object $\mathcal{F}_n$ has a natural interpretation as the "history", or specifically when mapped by $(X_1, \ldots, X_n)$ as the set of trajectories of the stochastic process up to and including time n.*

Definition 7.5 (Finite Markov Chain)**.** *A stochastic process,*

$$\{X_n : n \in \mathbb{N}\}$$

is a **Markov chain** *with* **state space** $\mathbb{X}$*, if for any $t \in \mathbb{N}$ the following holds*

$$\mathbb{P}(X_{t+1} = x|\mathcal{F}_t) = \mathbb{P}(X_{t+1} = x|X_t).$$

We say that the Markov chain is **homogeneous** *if*

$$\mathbb{P}(X_{t+1} = x|X_t) = \mathbb{P}(X_{s+1} = x|X_s)$$

for all $t, s \in \mathbb{N}$.

From the above and with the intuition that $\mathbb{P}(X_{t+1} = x \mid \mathcal{F}_t)$ constitutes our best guess for the event $X_{t+1} = x$ being true given the history of the process up to time t, we can interpret the Markov chain condition as saying that the only information from the history that we need is the value at time t.

7.1.2 Non advanced introduction

Definition 7.6 (Finite Markov Chain)**.** *A stochastic process,*

$$\{X_n : n \in \mathbb{N}\}$$

is a **Markov chain** *with* **state space** $\mathbb{X}$*, if for any* $t \in \mathbb{N}$ *the following holds*

$$\mathbb{P}(X_{t+1} = x | X_0, X_1, \ldots, X_t) = \mathbb{P}(X_{t+1} = x | X_t).$$

We say that the Markov chain is **homogeneous** *if*

$$\mathbb{P}(X_{t+1} = x | X_t) = \mathbb{P}(X_{s+1} = x | X_s)$$

for all $t, s \in \mathbb{N}$.

Note that if the Markov chain is homogeneous then it is enough to know

$$\mathbb{P}(X_1 = x_1 | X_0 = x_0) : \mathbb{X} \times \mathbb{X} \to [0, 1]$$

i.e. we can identify it with a $|\mathbb{X}| \times |\mathbb{X}|$ matrix

$$P_{xy} = \mathbb{P}(X_1 = y | X_0 = x).$$

Usually we identify $x \in \mathbb{X}$ with the enumeration of elements in $\mathbb{X}$ and thus we can write for $N = |\mathbb{X}|$ an $N \times N$ matrix P_{ij}. This matrix is denoted the **transition matrix**.

Lemma 7.7. *Let* $\{X_n, n \in \mathbb{N}\}$ *be a homogeneous Markov chain. Let the statespace* $\mathbb{X} = \{s_1, \ldots, s_N\}$ *be enumerated and let* μ_0 *be the PMF of* X_0*. Then the PMF* μ_n *for* X_n *is*

$$\mu_n = \mu_0 P^n$$

Proof. Let us start with applying the law of total probability

$$\begin{aligned}\mathbb{P}(X_n = x_n) &= \sum_{x_{n-1}} \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}) \, \mathbb{P}(X_{n-1} = x_{n-1}) \\ &= \sum_{x_{n-1}} P_{x_{n-1} x_n} \, \mathbb{P}(X_{n-1} = x_{n-1})\end{aligned}$$

Since n was arbitrary we can apply it again until we reach X_0, namely

$$\mathbb{P}(X_n = x_n) = \sum_{x_0, \ldots, x_{n-1}} P_{x_{n-1} x_n} \ldots P_{x_0 x_1} \, \mathbb{P}(X_0 = x_0)$$

The above is just a sequence of matrix multiplications, we can write

$$\mu_n = \mu_0 P^n.$$

□

Since we will be interested in Markov chains on $\mathbb{X} = \{s_1, s_2, \ldots, s_k\}$ with the same transition matrix P but different initial distributions, we introduce $\mathbb{P}_\mu$ and $\mathbb{E}_\mu$ for probabilities and expectations given that the initial distribution is μ, respectively. When the initial distribution is concentrated at a single initial state x given by:

$$\mathbb{1}_{\{x\}}(y) := \begin{cases} 1 & \text{if } y = x \\ 0 & \text{if } y \neq x \end{cases}$$

we represent it by e_x, the $1 \times k$ ortho-normal basis row vector with a 1 in the x-th entry and a 0 elsewhere. We simply write $\mathbb{P}_x$ for $\mathbb{P}_{\mathbb{1}_{\{x\}}}$ or $\mathbb{P}_{e_x}$ and $\mathbb{E}_x$ for $\mathbb{E}_{\mathbb{1}_{\{x\}}}$ or $\mathbb{E}_{e_x}$. Thus, Lemma 7.7 along with our new notations means that:

$$\mathbb{P}_x(X_t = y) = (e_x P^t)(y) = P^t(x, y) \ .$$

In words, the probability of going to y from x in t steps is given by the (x, y)-th entry of P^t, the **t-step transition matrix**. We refer to the x-th row and the x-th column of P by $P(x, \cdot)$ and $P(\cdot, x)$, respectively.

Let a function $f(x) : \mathbb{X} \to \mathbb{R}$ be given, then we can define

$$(P^t f)(x) := \sum_y P^t(x, y) f(y) = \sum_y f(y)\, \mathbb{P}_x(X_t = y) = \mathbb{E}_x(f(X_t)) \ . \quad (7.1)$$

This is the expected value of f under the distribution of states in t steps given that we start at state x.

Identified in the above way, we see that $P^t : (\mathbb{X} \to \mathbb{R}) \to (\mathbb{X} \to \mathbb{R})$ i.e. it maps a $\mathbb{R}$ valued function to an $\mathbb{R}$ valued function. Let us look at some properties of P^t

Lemma 7.8. *Let $f : \mathbb{X} \to \mathbb{R}$, then the mapping P^t defined in (7.1) satisfies:*

- *Let $g : \mathbb{X} \to \mathbb{R}$, then $P^t(f + g) = P^t f + P^t g$, i.e. it is a linear functional.*
- *Let $t > s > 0$ be positive integers, then $P^t f = P^{t-s}(P^s f)$.*

Proof. Let us denote X_t, Y_t two homogeneous and independent Markov processes with the same transition Matrix. First, define $g(x) = (P^s f)(x) =$

$\mathbb{E}_x[f(X_s)]$, then consider

$$\begin{aligned}(P^{t-s}(P^s f))(x) &= \mathbb{E}_x[g(Y_{t-s})] = \sum_y g(y)\,\mathbb{P}(Y_{t-s} = y \mid Y_0 = x) \\ &= \sum_y \sum_z f(z)\,\mathbb{P}(X_s = z \mid X_0 = y)\,\mathbb{P}(Y_{t-s} = y \mid Y_0 = x) \\ &= \sum_y \sum_z f(z)\,\mathbb{P}(X_t = z \mid X_{t-s} = y)\,\mathbb{P}(X_{t-s} = y \mid X_0 = x) \\ &= \sum_z f(z)\,\mathbb{P}(X_t = z \mid X_0 = x) \\ &= (P^t f)(x)\end{aligned}$$

□

This is the prime example of a so called **Semigroup**,

Definition 7.9. *A semigroup is a set S together with a binary operator $\odot$, i.e. a function $\odot : S \times S \to S$ that satisfies the associative property*

$$(a \odot b) \odot c = a \odot (b \odot c)$$

for all $a, b, c \in S$.

Let $S = \{P^t, t > 0\}$ and define the operator $\odot$ as

$$P^t \odot P^s = P^{t+s}$$

then we see from the above lemma Lemma 7.8 that $\{P_t, t > 0\}$ forms a semigroup, specifically a one parameter semigroup.

Remark 7.10. *The semigroup property is retained when moving over to continuous time Markov processes. If the semigroup is what is called strongly continuous there is also a time dependent (parabolic) partial differential equation which $P^t f$ solves.*

Until now our Markov chains have been **homogeneous** in time according to Definition 7.5, i.e., the transition matrix P does not change with time. We define inhomogeneous Markov chains as Markov chains that are not homogeneous. Such Markov chains are more realistic as models in some situations and more flexible as algorithms in the sequel.

Lemma 7.11. *For a finite inhomogeneous Markov chain $(X_t)_{t\in\mathbb{Z}_+}$ with state space $\mathbb{X} = \{s_1, s_2, \ldots, s_k\}$, initial distribution*

$$\mu_0 := (\mu_0(s_1), \mu_0(s_2), \ldots, \mu_0(s_k)),$$

where $\mu_0(s_i) = \mathbb{P}(X_0 = s_i)$*, and transition matrices*

$$(P_1, P_2, \ldots), \quad P_t := (P_t(s_i, s_j))_{(s_i,s_j)\in\mathbb{X}\times\mathbb{X}}, \ t \in \{1, 2, \ldots\}$$

we have for any $t \in \mathbb{Z}_+$ *that the distribution at time* t *given by:*

$$\mu_t := (\mu_t(s_1), \mu_t(s_2), \ldots, \mu_t(s_k)),$$

where $\mu_t(s_i) = \mathbb{P}(X_t = s_i)$*, satisfies:*

$$\mu_t = \mu_0 P_1 P_2 \cdots P_t \ . \tag{7.2}$$

Proof. Left as Exercise 7.12. □

Exercise 7.12. *Prove Lemma 7.11 in a similar way to Lemma 7.7.*

7.2 Random Mapping Representation and Simulation

In order to simulate $(x_0, x_1, \ldots, x_n)$, a sequential realisation or sequence of states visited by a Markov chain, we need a random mapping representation of a Markov chain.

Definition 7.13 (Random mapping representation (RMR))**.** *A* **random mapping representation** *(RMR) of a transition matrix* $P := (P(x, y))_{(x,y)\in\mathbb{X}^2}$ *is a function*

$$\rho(x, w) : \mathbb{X} \times \mathbb{W} \to \mathbb{X} \ , \tag{7.3}$$

along with a $\mathbb{W}$*-valued random variable* W*, satisfying*

$$\mathbb{P}(\{\rho(x, W) = y\}) = P(x, y), \quad \text{for each } (x, y) \in \mathbb{X}^2 \ . \tag{7.4}$$

Theorem 7.14. *Every Markov chain on* $\mathbb{X}$ *has a random mapping representation.*

Proof. Let X_t be a Markov chain with transition matrix P_t at t, for simplicity assume that $\mathbb{X} \subset \mathbb{N}$. Let t be an arbitrary time. Let $Z_t \sim \text{Uniform}([0, 1])$. For any $i, j \in \mathbb{X}$, set

$$F_{i,j} = \sum_{m=1}^{j} P_t(i, m).$$

Define

$$f_t(i, z) := j \quad \text{when} \quad F_{i,j-1} < z \leq F_{i,j}.$$

We have

$$\begin{aligned}\mathbb{P}(f_t(i,Z)=j) &= \mathbb{P}(F_{i,j-1} < Z \leq F_{i,j}) = F_{i,j} - F_{i,j-1} \\ &= \sum_{m=1}^{j} P_t(i,m) - \sum_{m=1}^{j-1} P_t(i,m) = P_t(i,j).\end{aligned}$$

We see that (f_t, Z_t) is a RMR for X_t at t. □

Theorem 7.15. *Let $W_1, \ldots, \overset{\text{IID}}{\sim} F$ such that (ρ_t, W_t) is a RMR for a transition matrix P_t, for all $t \in \mathbb{N}$. Then if $X_0 \sim \mu_0$,*

$$X_t := \rho_t(X_{t-1}, W_t), t \in \mathbb{N},$$

is a Markov chain with initial distribution μ_0 and transition matrix P_t at time t.

Exercise 7.16. *Do the proof of Theorem 7.15 by using the necessary Definitions.*

Exercise 7.17. *Show that the RMR for a Markov chain is not necessarily unique.*

7.3 Irreducibility and Aperiodicity

The utility of our mathematical constructions with Markov chains depends on a delicate balance between generality and specificity. We introduce two specific conditions called irreducibility and aperiodicity that make Markov chains more useful to model real-word phenomena.

Definition 7.18. *Let X_t be a homogeneous Markov chain on state space $\mathbb{X} = \{s_1, \ldots, s_N\}$. We say that $s_i \to s_j$ (* **communicates***) if there exists a $t \in \mathbb{N}$ such that*

$$\mathbb{P}(X_t = s_j \mid X_0 = s_i) > 0.$$

We say that s_i, s_j **intercommunicates** *if $s_i \to s_j$ and $s_j \to s_i$, we write this as $s_i \leftrightarrow s_j$.*

Definition 7.19 (Irreducible)**.** *A homogeneous Markov chain $(X_t)_{t\in\mathbb{Z}_+}$ on state space $\mathbb{X} := \{s_1, s_2, \ldots, s_k\}$ is said to be* **irreducible** *if $s_i \leftrightarrow s_j$ for each $(s_i, s_j) \in \mathbb{X}^2$. Otherwise the chain is said to be* **reducible***.*

Definition 7.20 (Return times and period)**.** *Let* $\mathbb{T}(x) := \{t \in \mathbb{N} : P^t(x,x) > 0\}$ *be the set of* **possible return times** *to the starting state* x*. The* **period** *of state* x *is defined to be* $\gcd(\mathbb{T}(x))$*, the greatest common divisor of* $\mathbb{T}(x)$*. When the period of a state* x *is* 1*, i.e.,* $\gcd(\mathbb{T}(x)) = 1$*, then* x *is said to be an* **aperiodic state***.*

Proposition 7.21. *If the Markov chain* $(X_t)_{t\in\mathbb{Z}_+}$ *with transition matrix* P *on state space* $\mathbb{X}$ *is irreducible then* $\gcd(\mathbb{T}(x)) = \gcd(\mathbb{T}(y))$ *for any* $(x,y) \in \mathbb{X}^2$*.*

Proof. Fix any pair of states $(x,y) \in \mathbb{X}^2$. Since, P is irreducible, $x \leftrightarrow y$ and therefore there exists natural numbers $\eta(x,y)$ and $\eta(y,x)$ such that $P^{\eta(x,y)}(x,y) > 0$ and $P^{\eta(y,x)}(y,x) > 0$. Let $\eta' = \eta(x,y) + \eta(y,x)$ and observe that $\eta' \in \mathbb{T}(x) \cap \mathbb{T}(y)$, $\mathbb{T}(x) \subset \mathbb{T}(y) - \eta' := \{t - \eta' : t \in \mathbb{T}(y)\}$ and $\gcd(\mathbb{T}(y))$ divides all elements in $\mathbb{T}(x)$. Thus, $\gcd(\mathbb{T}(y)) \leq \gcd(\mathbb{T}(x))$. By a similar argument we can also conclude that $\gcd(\mathbb{T}(x)) \leq \gcd(\mathbb{T}(y))$. Therefore $\gcd(\mathbb{T}(x)) = \gcd(\mathbb{T}(y))$. □

Definition 7.22 (Aperiodic)**.** *A Markov chain* $(X_t)_{t\in\mathbb{Z}_+}$ *with transition matrix* P *on state space* $\mathbb{X}$ *is said to be aperiodic if all of its states are aperiodic, i.e.,* $\gcd(\mathbb{T}(x)) = 1$ *for every* $x \in \mathbb{X}$*. If a chain is not aperiodic, we call it* **periodic***.*

7.4 Stationarity

We are interested in statements about a Markov chain that has been running for a long time. For any nontrivial Markov chain $(X_0, X_1, \ldots)$ the value of X_t will keep fluctuating in the state space $\mathbb{X}$ as $t \to \infty$ and we cannot hope for convergence to a fixed point state $x^* \in \mathbb{X}$ or to a k-cycle of states $\{x_1, x_2, \ldots, x_k\} \subset \mathbb{X}$. However, we can look one level up into the space of probability distributions over $\mathbb{X}$ that give the probability of the Markov chain visiting each state $x \in \mathbb{X}$ at time t, and hope that the distribution of X_t over $\mathbb{X}$ settles down as $t \to \infty$. The Markov chain convergence theorem indeed sattes that the distribution of X_t over $\mathbb{X}$ settles down as $t \to \infty$, provided the Markov chain is irreducible and aperiodic.

Definition 7.23 (Stationary distribution)**.** *Let* $(X_t)_{t\in\mathbb{Z}_+}$ *be a Markov chain with state space* $\mathbb{X} = \{s_1, s_2, \ldots, s_k\}$ *and transition matrix* $P = (P(x,y))_{(x,y)\in\mathbb{X}^2}$*. A row vector*

$$\pi = (\pi(s_1), \pi(s_2), \ldots, \pi(s_k)) \in \mathbb{R}^{1\times k}$$

is said to be a **stationary distribution** *for the Markov chain, if it satisfies the conditions of being:*

1. a probability distribution*:* $\pi(x) \geq 0$ *for each* $x \in \mathbb{X}$ *and* $\sum_{x\in\mathbb{X}} \pi(x) = 1$*, and*

2. a fixed point*:* $\pi P = \pi$*, i.e.,* $\sum_{x\in\mathbb{X}} \pi(x)P(x, y) = \pi(y)$ *for each* $y \in \mathbb{X}$*.*

Proposition 7.24 (Existence of Stationary distribution)**.** *For any irreducible and aperiodic Markov chain there exists at least one stationary distribution.*

Proof. See the Perron Frobenius theorem, Wikipedia. □

7.5 Reversibility

We introduce another specific property called reversibility. This property will assist in conjuring Markov chains with a desired stationary distibution.

Definition 7.25 (Reversible)**.** *A probability distribution* π *on* $\mathbb{X} = \{s_1, s_2, \ldots, s_k\}$ *is said to be a* **reversible distribution** *for a Markov chain* $(X_t)_{t\in\mathbb{Z}}$ *on* $\mathbb{X}$ *with transition matrix* P *if for every pair of states* $(x, y) \in \mathbb{X}^2$*:*

$$\pi(x)P(x, y) = \pi(y)P(y, x) \ . \tag{7.5}$$

A Markov chain that has a reversible distribution is said to be a reversible Markov chain.

In words, $\pi(x)P(x, y) = \pi(y)P(y, x)$ says that if you start the chain at the reversible distribution π, i.e., $\mu_0 = \pi$, then the probability of going from x to y is the same as that of going from y to x.

Proposition 7.26 (A reversible π is a stationary π)**.** *Let* $(X_t)_{t\in\mathbb{Z}_+}$ *be a Markov chain on* $\mathbb{X} = \{s_1, s_2, \ldots, s_k\}$ *with transition matrix* P*. If* π *is a reversible distribution for* $(X_t)_{t\in\mathbb{Z}_+}$ *then* π *is a stationary distribution for* $(X_t)_{t\in\mathbb{Z}_+}$*.*

Exercise 7.27. *Prove Proposition 7.26.*

7.5.1 Random Walks on Graphs

Random walks on graphs is one of the most useful applications of Markov chains. In this section, we will see some basic definitions from graph theory and define simple random walks on graphs as finite Markov chains to shed light on the random surfer model of Google.

Definition 7.28 (Definitions in Graph Theory)**.** *Here we take a brief tour of the most basic defintions in graph theory. A* **Graph** $\mathbb{G} := (\mathbb{V}, \mathbb{E})$ *consists of a* **vertex set** $\mathbb{V} := \{v_1, v_2, \ldots, v_k\}$ *together with an* **edge set** $\mathbb{E} := \{e_1, e_2, \ldots, e_l\}$*. Each* **edge** *in* $\mathbb{E}$ *connects two of the* **vertices** *in* $\mathbb{V}$*. A* **directed edge** e_h *connecting vertex* v_i *to* v_j *is denoted by the ordered pair* (v_i, v_j)*. An* **undirected edge** *simply connects two vertices without regard to order and is denoted by* $\{v_i, v_j\}$ *to represent both of the directed edges* (v_i, v_j) *and* (v_j, v_i)*. Thus,* $\mathbb{E} \subset \mathbb{V}^2$ *and a graph* $\mathbb{G}$ *with directed eges in* $\mathbb{E}$ *is said to be an* **directed graph** *and a graph* $\mathbb{G}$ *with undirected edges is said to be an* **undirected graph***. Two vertices are* **neighbours** *if they share an edge, i.e., if they are connected by an edge. The* **neighbourhood** *of a vertex* v_i *denoted by* $\mathrm{nbhd}(v_i) := \{v_j : (v_i, v_j) \in \mathbb{E}\}$ *is the set of neighbouring vertices of* v_i*. The number of neighbours of a vertex* v_i *in an undirected graph is called its* **degree** *and is denoted by* $\deg(v_i)$*. Note that* $\deg(v_i) = \#\,\mathrm{nbhd}(v_i)$*. If there is a sequences of edges or a path from every vertex to every other vertex then the undirected graph is said to be connected. In a graph we only allow one edge per pair of vertices but in a* **multigraph** *we allow more than one edge per pair of vertices. An edge can be* **weighted** *by being associated with a real number called its weight. More generally, vertices and edges can be augmented with various properties, including addresses, names, etc., and weights, relation types, etc. Graphs whose vertices and edges are further augmented by various properties are called* **property graphs***, an extremely useful and versatile representation of data from different domains. We can represent a directed graph by its* **adjacency matrix** *given by:*

$$A := (A(v_i, v_j))_{(v_i, v_j) \in \mathbb{V} \times \mathbb{V}}, \quad A(v_i, v_j) = \begin{cases} 1 & \text{if } (v_i, v_j) \in \mathbb{E} \\ 0 & \text{otherwise .} \end{cases}$$

Thus the adjacency matrix of an undirected graph is symmetric. In a directed graph, each vertex v_i *has* **in-edges** *that come into it and* **out-edges** *that go out of it. The number of in-edges and out-edges of* v_i *is denoted by* $\mathrm{ideg}(v_i)$ *and* $\mathrm{odeg}(v_i)$ *respectively. Note that a transition diagram of a Markov chain is a weighted directed graph and is represented by the transition probability matrix.*

Model 7.29 (Random Walk on a Connected Undirected Graph)**.** *A random walk on a connected undirected graph* $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ *is a Markov chain with state space* $\mathbb{V} := \{v_1, v_2, \ldots, v_k\}$ *and the following transition rules: if the chain is at vertex* v_i *at time* t *then it moves uniformly at random to one of the neighbours of* v_i *at time* $t + 1$*. If* $\deg(v_i)$ *is the degree of* v_i *then the*

transition probabilities of this Markov chain is

$$P(v_i, v_j) = \begin{cases} \frac{1}{\deg(v_i)} & \text{if } (v_i, v_j) \in \mathbb{E} \\ 0 & \text{otherwise,} \end{cases}$$

Proposition 7.30. *The random walk on a connected undirected graph* $\mathbb{G} = (\mathbb{V}, \mathbb{E})$*, with vertex set* $\mathbb{V} := \{v_1, v_2, \ldots, v_k\}$ *and degree sum* $d = \sum_{i=1}^{k} \deg(v_i)$ *is a reversible Markov chain with the reversible distribution* π *given by:*

$$\pi = \left(\frac{\deg(v_1)}{d}, \frac{\deg(v_2)}{d}, \ldots, \frac{\deg(v_k)}{d} \right) .$$

Exercise 7.31. *Prove Proposition 7.30 by directly showing that* π *is reversible.*

Example 7.32 (Google's random surfer on the word wide web)**.** *Consider the huge graph with vertices as webpages and hyper-links as undirected edges. Then Model 7.29 gives a random walk on this graph. However if a page has no links to other pages, it becomes a sink and therefore terminates the random walk. Let us modify this random walk into a* **random surf** *to avoid getting stuck. If the random surfer arrives at a sink page, she picks another page at random and continues surfing at random again. Google's PageRank formula uses a random surfer model who gets bored after several clicks and switches to a random page. The PageRank value of a page reflects the chance that the random surfer will land on that page by clicking on a link. The stationary distribution of the random surfer on the world wide web is a very successful model for ranking pages and has grown into a Trillion dollar company.*

Chapter 8

Pattern recognition

Let us introduce the pattern recognition problem use computer science notation (its good for you to see that too, as you will probably be reading a bunch of cs papers in the future).

Suppose we have n training data points $T_n := ((X_i, Y_i))_{i=1}^n$ and are interested in a classification rule $h(X)$ that uses T_n to *predict*, i.e., assign labels to previously unseen data X.

Thus, we want our classification rule h, which is typically an algorithm, to perform well on previously unseen data by learning from the training data. This is known as *generalization*.

The space $\mathcal{X}$ where X_i belongs to is called the *instance space* or *feature space* and the space $\mathcal{Y}$ where Y_i belongs to is called the *label space*.

Typically, $\mathcal{X}$ is a subset of $\mathbb{R}^d$ and $\mathcal{Y}$ is binary label space either as $\{0, 1\}$ or $\{-1, 1\}$. For example, $\mathcal{X}$ can be $\{0, 1\}^d$ to indicate the presence or absence of something in the instance space, say a specific set of words in an email if the task is to classify emails with labels 0 and 1 for non-spam or spam.

Remark 8.1. *To connect back to our previous terminology, we see that the data space $\mathbb{X} = (\mathcal{X}, \mathcal{Y})$ (we will later write $\mathbb{X} \times \mathbb{Y}$ to avoid X being used for both feature and label) is split into the featue space and the label space. The random variable we are observing is a pair (X, Y) and a collection of n samples is the training dataset.*

8.1 Linear Classifiers

Let us say that we are trying to device a classification rule based on instance space $\mathcal{X} = \mathbb{R}^d$ and label space $\mathcal{Y} = \{-1, 1\}$.

One of the simplest such rule involves taking weighted sums of the x_i's until it exceeds a threshold to determine if it should be labelled by $+1$ or not. Such rules involve finding a hyperplane of dimension $d - 1$ to separate

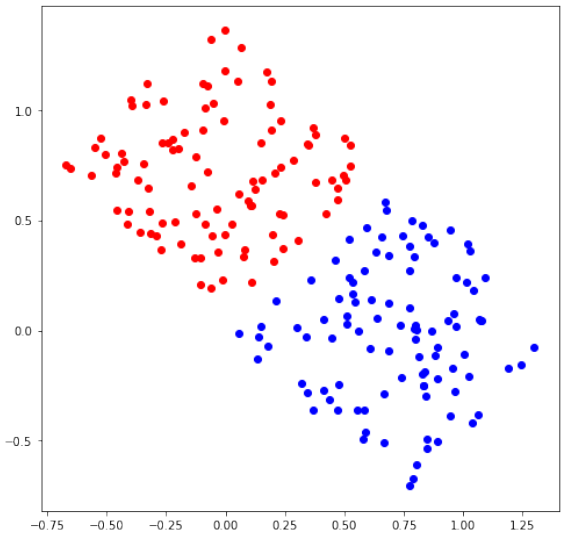


Figure 8.1: Linearly separable data with labels $+1$ or red and -1 or blue.

out the data with the same labels on each side of the hyperplane. Such a classification rule is called a *linear separator*.

Fig. 8.1 shows an example of data that is *linearly separable* and thus ideal for linear separators.

8.1.1 Linearly Separable Dataset

Consider the following linearly separable dataset, where we can draw a line (hyper-plane in $\mathbb{R}^2$) to separate the data points with different labels on either side of the line.

8.1.2 The perceptron algorithm

In the history of artificial intelligence and neural network research, linear classifiers of this type were called **perceptrons.** (Fisher's linear discrimination analysis (LDA, 1936) is also given by a linear classifier). In this section we present a training algorithm for a perceptron, invented in 1957 at the Cornell Aeronautical Laboratory by Frank Rosenblatt. The algorithm created a great deal of interest when it was first introduced. As we will see, it is guaranteed to converge if there exists a hyperplane that correctly classifies the training data.

The perceptron algorithm tries to find a linear separator, i.e. a hyperplane in $\mathbb{R}^d$ that separates the two classes. The task is thus to find w and t such that for the training data S, the data consists of pairs (x_i, l_i) the x_i represents our features and the l_i our labels or target.

$$w \cdot x_i > t \quad \text{for each } x_i \text{ labeled } +1$$
$$w \cdot x_i < t \quad \text{for each } x_i \text{ labeled } -1$$

Adding a new coordinate to our space allows us to consider $\hat{x}_i = (x_i, 1)$ and $\hat{w} = (w, t)$, this allows us to rewrite the inequalities above as

$$(\hat{w} \cdot \hat{x}_i) l_i > 0.$$

The algorithm

1. $w = 0$
2. while there exists x_i with $x_i y_i \cdot w \leq 0$, update $w := w + x_i y_i$

[12]:
```
@interact
def _(n_steps=(0,(0..63))):
        # X = (n_points,3)
        # W = (n_points,3)
        n_points = X.shape[0]
        W = np.array([0,0,0])
        P=points(zip(X1[:,0],X1[:,1]),color='blue')
        P+=points(zip(X2[:,0],X2[:,1]),color='red')

        k = 0
        max_iter=10000
        j = 0
        while ((k < n_steps) and (j < max_iter)):
                i = j % n_points
                j+=1
                if (X[i,:]@W * yall[i] <= 0):
                        W = W + X[i,:]*yall[i]
                        P+=points(X[i,:2],color='yellow')
                        k+=1
        print(W)
```

Theorem 8.2. *If there exists w^* such that $w^* \cdot x_i y_i \geq 1$ for all i. Then the perceptron algorithm finds a w satisfying $w \cdot x_i y_i \geq 0$ for all i in at most $r^2 |w^*|^2$ updates, where $r = \max_i |x_i|$.*

Proof. Lets say we are given a sequence of points $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Let w_i, for $i = 0, \ldots$, denote the weight at update i. That is, $w_0 = 0$ and for every update i, there is a corresponding index $I(i) \in \{1, \ldots, n\}$ such that

$$w_i \cdot x_{I(i)} y_{I(i)} \leq 0.$$

Since we then update the weights, we also have

$$w_{i+1} = w_i + x_{I(i)}y_{I(i)}.$$

Assuming the statement is true, you expect that $w_i \to w^*$ in some capacity, at least the direction should be close. Secondly, we also know that the weights most likely grow since we are always adding vectors to it. With this in mind, let us track two quantities and see how they react to an update: First let us compute

$$w_{i+1} \cdot w^* = (w_i + x_{I(i)}y_{I(i)}) \cdot w^* w_i \cdot w^* + w^* \cdot x_{I(i)}y_{I(i)} \geq w_i \cdot w^* + 1. \tag{8.1}$$

Then let us compute

$$\begin{aligned} |w_{i+1}|^2 = w_{i+1} \cdot w_{i+1} &= (w_i + x_{I(i)}y_{I(i)}) \cdot (w_i + x_{I(i)}y_{I(i)}) \\ &= |w_i|^2 + 2w_i \cdot x_{I(i)}y_{I(i)} + |x_{I(i)}|^2 \\ &\leq |w_i|^2 + r^2. \end{aligned} \tag{8.2}$$

Since $w_0 = 0$ we get from iterating (8.1) and (8.2) for $i = 0, \ldots, m$ that

$$\begin{aligned} w_m \cdot w^* &\geq m \\ |w_m|^2 &\leq mr^2. \end{aligned}$$

If we use Cauchy-Schwartz for the dot-product we get from the above that

$$m \leq w_m \cdot w^* \leq |w_m||w^*| \leq |w^*|\sqrt{m}r.$$

Now, dividing the left hand side and right hand side with $\sqrt{m}$ and skipping the middle we get

$$\sqrt{m} \leq |w^*|r,$$

which when squared gives the result. □

So this theorem guarantees that if the two classes can be separated then the preceptron will also find a separator in finite time. This is interesting since finding the plane that minimizes the error is NP-hard, so this tells us that if there is separation the problem is "easy".

What about non-linearly separable data. Let $B_r = \{x \in \mathbb{R}^2 : |x| < r\}$, then for instance

$$X = (B_4 \setminus B_3) \cup B_1$$

and let $g^* = \mathbb{1}_{B_1}$. We cannot separate these sets using a linear classifier, see simulation in notebooks.

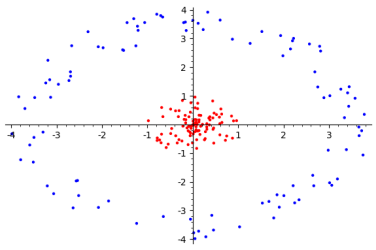


Figure 8.2: Linearly non-separable data in two dimensions.

8.2 Kernelization

What about non-linearly separable data. Take for instance the data formed by the following operations with points at different distances from the origin in two dimensions, as shown in Figure 8.2.

$$X = (B_4 \setminus B_3) \cup B_1$$

and let $c^* = B_1$. We cannot separate these sets using a linear classifier

[25]:
```
A = np.random.normal(size=(100,2))
A_unit = A/(np.linalg.norm(A,axis=1).reshape(-1,1))
radial_A = 3+np.random.uniform(size=(100,1))
P=points(A_unit*radial_A,color='blue')

B = np.random.normal(size=(100,2))
B_unit = B/(np.linalg.norm(B,axis=1).reshape(-1,1))
radial_B = np.random.uniform(size=(100,1))
P+=points(B_unit*radial_B,color='red')
P.show()
```

[25]: we can however separate the following mapping of X. Namely in $\mathbb{R}^2$ we can do

$$\phi(x) = (x_1, x_2, x_1^2 + x_2^2) \in \mathbb{R}^3$$

This is clearly linearly separable as we can see from the following 3d plot shown in Figure 8.3.

[27]:
```
A_2d = A_unit*radial_A
A_3d = np.concatenate([A_2d,np.linalg.norm(A_2d,axis=1).reshape(-1,1)^2],axis=1)
B_2d = B_unit*radial_B
B_3d = np.concatenate([B_2d,np.linalg.norm(B_2d,axis=1).reshape(-1,1)^2],axis=1)

P=points(A_3d,size=20,color='blue')
P+=points(B_3d,size=20,color='red')
P.show()
```

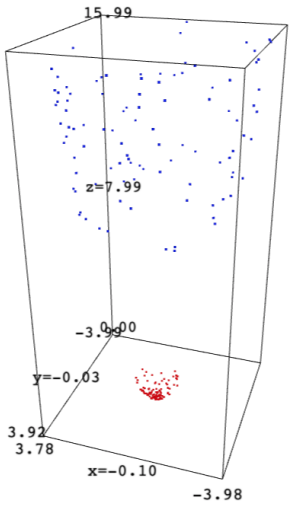


Figure 8.3: Linearly separable in three dimensions after $(x_1, x_2) \mapsto (x_1, x_2, x_1^2 + x_2^2)$.

> Remember the extra dimension that we always add to simplify notation. Therefore the full ϕ in the above examples is $\hat{\phi}(x) = (x_1, x_2, x_1^2 + x_2^2, 1)$.

So if we transform the $x \to \phi(x)$ for some good transformation ϕ then our perceptron will try to solve

$$w \cdot \phi(x_i) l_i > 0$$

furthermore, remember how we constructed w using the perceptron algorithms, i.e. using additions of $x_i l_i$, which transforms into $\phi(x_i) l_i$, and we start with $w = 0$, this gives that the weight has the form

$$w = \sum_{i=1}^{n} c_i \phi(x_i)$$

for numbers c_i. The perceptron algorithm becomes just addition and subtraction of certain c_i's by 1.

Furthermore

$$w \cdot \phi(x_i) = \sum_{i=1}^{n} c_j \phi(x_j) \cdot \phi(x_i) = \sum_{i=1}^{n} c_i k_{ij}$$

where $k_{ij} = \phi(x_i) \cdot \phi(x_j)$.

Is it easy to find such a mapping ϕ? No, it is actually quite difficult. Furthermore, if the mapping ϕ is high dimensional we might need to do alot

of computation, which is not so efficient. What if we had a function $k(x, y)$ that could be written as

$$k(x, y) = \phi(x) \cdot \phi(y)$$

for some ϕ and k is easier to compute, then our life would be simpler. Also, what if we are given a function $k(x, y)$ and we would like to know if it is a "kernel function".

Lemma 8.3. *Given a sequence of points $\{x_i\}_{i=1}^n$, $x_i \in \mathbb{R}^d$, and given an $n \times n$ matrix K, which is symmetric and positive semidefinite. Then, there is a mapping $\phi : \mathbb{R}^d \to \mathbb{R}^m$ (for some m) such that $K_{ij} = \phi(x_i) \cdot \phi(x_j)$.*

Exercise 8.4. *Prove the above Lemma using the following outline for it:*

1. *$K = Q\Lambda Q^T$ (eigendecomposition)*
2. *K is positive definite, all eigenvalues ≥ 0, so we can define $B = Q\Lambda^{1/2}$.*
3. *$K = BB^T$*
4. *define $\phi(x_i) = B_{i\cdot}$, i.e. the i:th row of B, then $K_{ij} = \phi(x_i) \cdot \phi(x_j)$.*
5. *What is the size of m?*

We now have a way to identify whenever a matrix K is a kernel matrix. There are some standard choices of kernel functions one could try, that produces positive semi-definite matrices whenever all points x_i are distinct.

Definition 8.5. *We call a function $k(x, y) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a kernel function if there is a mapping $\phi : \mathbb{R}^d \to \mathbb{R}^m$ (for some m) such that $k(x, y) = \phi(x) \cdot \phi(y)$.*

Theorem 8.6. *Suppose $k_1, k_2 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ are kernel functions. Then*

1. *For any constant $c \geq 0$, ck_1 is a kernel function.*
2. *For any scalar function f, $k(x, y) = f(x)f(y)k_1(x, y)$ is a kernel function.*
3. *$k_1 + k_2$ is a kernel function. 4. k_1k_2 is a kernel function.*

Exercise 8.7. *Prove Theorem 8.6.*

Corollary 8.8. *The following functions are kernel function*

- *$k(x, y) = (\gamma x \cdot y + r)^k$, ($k \in \mathbb{N}$) polynomial*
- *$k(x, y) = x \cdot y$, linear*

Exercise 8.9. *Prove the above corollary by multiple applications of Theorem 8.6.*

8.2.1 Other types of Kernels

The above concept of a kernel function is a simplification, we can also have the following kernels for which ϕ maps to "infinite dimensions":

1. $k(x, y) = e^{-\gamma|x-y|}$, called Radial Basis Function
2. $k(x, y) = \tanh(\gamma x \cdot y + r)$, sigmoidal

8.3 Theoretical guarantees

Recall that in the pattern recognition model (Section 4.1.3) we assume that the supervisors conditional distribution $F(y|x)$ is discrete, and can take k different values, $y = 0, \ldots, k-1$.

Recall the $0-1$ loss function for $z = (x, y)$

$$L(z, u) = \begin{cases} 0 & \text{if } y = u \\ 1 & \text{if } y \neq u \end{cases}$$

that is, the loss is 1 if u is the wrong value and 0 if it is correct. The pattern recognition problem is the problem of minimizing the functional

$$R(g) = \int L(y, g(x)) dF(x, y) = \mathbb{E}\left[L(Y, g(X))\right]$$

where $(X, Y) \sim F(x, y)$.

Also recall that,

$$\mathbb{E}\left[L(Y, g_\lambda(X))\right] = \mathbb{P}(\{Y \neq g_\lambda(X)\}).$$

As we have alluded to in Section 5.3 is that we want to minimize the empirical version of the risk and hope that we can get reasonable concentration estimates, as in Theorem 5.27.

Definition 8.10. *Given a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, and assume that $Z = ((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)) \overset{\text{IID}}{\sim} F(x, y)$ is a sequence of $\mathbb{R}^{m+1}$ valued random variables taking values in the data space $\mathbb{X} \times \mathbb{Y}$. We define the empirical risk for a function $g : \mathbb{X} \to \mathbb{Y}$ as*

$$\hat{R}_n(g) = \hat{R}_n(Z; g) = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, g(X_i)).$$

Note that the empirical risk is a statistic evaluated on Z.

We would like to minimize the risk, but we only have access to $Z = ((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)) \overset{\text{IID}}{\sim} F(x, y)$, so we can in practice only try to minimize the empirical risk. So given a model space $\mathcal{M}$ we consider

$$\hat{g}_n^* := \hat{g}_n^*(Z) := \operatorname*{argmin}_{g \in \mathcal{M}} \hat{R}_n(g).$$

The first realization here is now that $\hat{g}_n^*$ is a random variable that depends on Z, which means that it is not immediate that

$$\hat{R}_n(Z; g_n^*(Z))$$

is unbiased, in fact, since we are minimizing the risk it is quite possibly downward biased. However, even if it is downward biased one could hope that in some cases the bias is small when n is large.

8.3.1 Guarantees with a held out testing set

The problem with using the value of the empirical risk above is that we are evaluating on the "training-data", if we however have access to a testing dataset, then we can give some guarantees in the pattern recognition problem.

Consider a data set $T_{n+m} := \{(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})\}$ sampled i.i.d from $F_{X,Y}$. We consider $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ (which we dub the training data) and $\{(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})\}$ (which we dub the testing data). Define $\hat{\phi}$ the **empirical risk minimizer** on the **training dataset**, namely

$$\hat{R}_n(\hat{\phi}) = \min_{\phi \in \mathcal{M}} \hat{R}_n(\phi)$$

then since the **testing dataset** is independent of the training dataset and hence $\hat{\phi}$ is independent of the testing data, we deduce using Theorem 3.6 that if $\hat{R}_m(\phi)$ denotes the empirical risk over the testing dataset we have (Provided R is nice enough)

$$\mathbb{P}(|\hat{R}_m(\hat{\phi}) - R(\hat{\phi})| > \epsilon \mid T_n) < 2e^{-C\epsilon^2 n}. \tag{8.3}$$

Now again using the tower property we can get

$$\mathbb{P}(|\hat{R}_m(\hat{\phi}) - R(\hat{\phi})| > \epsilon) < 2e^{-C\epsilon^2 n}. \tag{8.4}$$

This is a procedure which always works when the loss is bounded, like $0-1$ loss. The risk on the testing data-set can even be exchanged for another loss, i.e. different than the training loss.

Exercise 8.11. *In the above we are mentioning that R needs to be nice enough, why is that? Does $0-1$ loss work? Why?*

Furthermore, we used the tower property to derive (8.4) from (8.3), how does this work?

Remark 8.12. *In many machine learning text-books that are practically oriented, you will see the recommendation that the training/testing split should be 70/30. In the pattern recognition problem this doesn't make much sense,*

it is better to determine with how high probability you want the bound to hold and use that to choose the number of samples to reach a tight enough interval.

Remark 8.13. *Note that the testing estimate above is only valid if done once, as the probability is over the testing set. Since we only have one, it can only be used once.*

It should be noted that during other courses you will encounter what seems to be a violation to the above, i.e. with the introduction of a test set and a validation set. In this setup the test set is used several times to select the best out of a set of different models, the best model is then chosen and the performance is evaluated on the **validation data**. In this setup one is not too concerned with the fact that the test set is used several times, as it is only used to select the model. The final performance evaluation, done once will satisfy the concentration bound (8.3).

Exercise 8.14. *In kaggle competitions and other online data-science competitions, several teams try to produce a model on a training data-set. When the team submits their proposed model it is validated on a hidden test-set (same for all teams). The teams are then ranked according to the score on the hidden test set. Does this mean that the best team had the best model? Do you think it would look the same if we repeated this with other hidden test-sets?*

8.3.2 Other test metrics

In the practical usage of the pattern recognition problem, one often sees the use of the test-metrics Precision and Recall. For class 1 they are defined as the following conditional probabilities

$$\begin{aligned} \text{Precision:} \quad & \mathbb{P}(Y = 1 \mid g(X) = 1) \\ \text{Recall:} \quad & \mathbb{P}(g(X) = 1 \mid Y = 1). \end{aligned}$$

Recall in particular is often used in medical testing and are then called sensitivity.

Exercise 8.15. *Lets say we have trained an ML model and gotten g as output, and lets say we want to use the test-set to estimate precision and recall. Lets say you wish to give a guarantee using for instance "concentration of measure" in the following scenarios*

1. *The function g is always* 1
2. *The function g is always* 0

3. $P(Y = 1)$ *is close to* 0

4. $P(Y = 1)$ *is close to* 1.

Which of these problems are easier and which are harder? What if we switch to estimating precision and recall for class 0*?*

8.4 Empirical Risk Minimization for Linear Classifiers

If we restrict the **"complexity"** of the decision functions, complexity will be defined. Then, we can give some guarantees without having to resort to a test-set that can only be used once. A form of a-priori estimate. These estimates are very very strong, but a bit restrictive.

Consider the data space $\mathbb{X} \times \mathbb{Y} = \mathbb{R}^m \times \{0, 1\}$, then a linear decision function is given as

$$\phi_a(x) = \begin{cases} 1 & \text{if } (a')^T x + a_0 > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $a = (a', a_0) \in \mathbb{R}^{m+1}$. The model space corresponding to linear decision rules is

$$\mathcal{M} = \left\{\phi_a : a \in \mathbb{R}^{m+1}\right\} = \{ \text{ all linear decision rules } \}.$$

We can index $\mathcal{M}$ using $a \in \mathbb{R}^{m+1}$.

8.4.1 A classifier with finitely many hyperplanes (without testing)

Instead of choosing $\mathcal{M}$ to be indexed by $a \in \mathbb{R}^{m+1}$ with uncountably infinitely many possibilities for ϕ, we will limit ourself to minimizing the empirical risk over finitely many linear decision rules – exactly $2\binom{n}{m}$ that are defined by each of the m choices from the n training points in $D_n = \{X_1, \ldots, X_n\}$.

Consider choosing m arbitrary points $(X_{i_1}, X_{i_2}, \ldots, X_{i_m})$ from the training data $\{X_1, X_2, \ldots, X_n\}$, and let $(a')^T x + a_0 = 0$ be the hyper-plane containing these m points, i.e., $x = (x_{i_1}, x_{i_2}, \ldots, x_{i_m})$. If we assume that X_i are continuous random variables, the m points are in *general position*[1] with probability 1 and this hyperplane is unique determining two decision rules:

$$\phi_+(x) = \begin{cases} 1 & \text{if } a^T x + a_0 > 0 \\ 0 & \text{otherwise} \end{cases}$$

[1] https://en.wikipedia.org/wiki/General_position and Exercise 8.17

and

$$\phi_-(x) = \begin{cases} 1 & \text{if } a^T x + a_0 < 0 \\ 0 & \text{otherwise} \end{cases}$$

with empirical risks or misclassification training errors $\hat{R}_n(\phi_+)$ and $\hat{R}_n(\phi_-)$. Thus to each m-tuple of data points we can associate two decision rules to yield a total of $2\binom{n}{m}$ such decision rules. Let us denote these decision rules by $\mathcal{M}_n = \{\phi_1, \ldots, \phi_{2\binom{n}{m}}\}$.

Now let $\hat{\phi}$ be a decision rule that minimizes $\hat{R}_n(\phi_i)$ over all $i \in \{1, 2, \ldots, 2\binom{n}{m}\}$, i.e.

$$\hat{R}_n(\hat{\phi}) = \min_{\phi \in \mathcal{M}_n} \hat{R}_n(\phi). \tag{8.5}$$

Theorem 8.16. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and consider (X_1, Y_1), …, (X_n, Y_n) be an i.i.d. sequence of random variables, X_i being continuous and taking values in $\mathbb{R}^m$ and $Y_i \in \{0, 1\}$ is discrete. If $\hat{\phi}$ is defined as in (8.5), then, if $2m/n \leq \epsilon \leq 1$, we have*

$$\mathbb{P}\{R(\hat{\phi}) > \inf_{\mathcal{M}} R(\phi) + \epsilon\} \leq e^{2m\epsilon} \left(2\binom{n}{m} + 1\right) e^{-n\epsilon^2/2}.$$

Proof. Denote $\phi^* \in \mathcal{M}$ a decision rule that satisfies

$$R(\phi^*) = \inf_{\mathcal{M}} R(\phi).$$

Furthermore, note that if we take a $\phi \in \mathcal{M}$ then at most it can make m better predictions than $\hat{\phi}$, as it could be that the plane that defines ϕ does not touch any X_i,

$$\hat{R}_n(\hat{\phi}) \leq \hat{R}_n(\phi) + \frac{m}{n}.$$

Now elementary considerations give together with the above,

$$\begin{aligned} I_0 := R(\hat{\phi}) - \inf R(\phi) &= R(\hat{\phi}) - \hat{R}_n(\hat{\phi}) + \hat{R}_n(\hat{\phi}) - \inf R(\phi) \\ &\leq R(\hat{\phi}) - \hat{R}_n(\hat{\phi}) + \hat{R}_n(\phi^*) - R(\phi^*) + \frac{m}{n} \\ &\leq \max_{1 \leq i \leq n} (R(\phi_i) - \hat{R}_n(\phi_i)) + \hat{R}_n(\phi^*) - R(\phi^*) + \frac{m}{n} \\ &=: \max_{1 \leq i \leq n} I_{1,i} + I_2 + \frac{m}{n} \end{aligned}$$

The reason we want to bound the difference this way is that we wish to apply concentration inequalities to both of these terms, i.e. we want to use

the fact that if n is large, there is a high probability that the empirical risk is close to the true risk.

From the properties of probability (monotonicity and Boole's inequality) we get (denoting $N = 2\binom{m}{n}$)

$$\begin{aligned}\mathbb{P}(I_0 > \epsilon) &\leq \mathbb{P}(\max_{1\leq i\leq N} I_{1,i} + I_2 + \frac{m}{n} > \epsilon) \\ &\leq \mathbb{P}(\max_{1\leq i\leq N} I_{1,i} > \epsilon/2) + \mathbb{P}(I_2 > \epsilon/2 - \frac{m}{n}) \\ &\leq \sum_{i=1}^{N} \mathbb{P}(I_{1,i} > \epsilon/2) + \mathbb{P}(I_2 > \epsilon/2 - \frac{m}{n}) \end{aligned} \tag{8.6}$$

Now, let us first bound $\mathbb{P}(I_2 > \epsilon/2 - m/n)$. This is easy, because $L(Y_i, \phi^*(X_i))$ is a sequence of independent bounded random variables, i.e. we can apply Hoeffdings inequality Theorem 3.6 to get (if $\epsilon/2 - m/n > 0$)

$$\mathbb{P}(I_2 > \epsilon/2 - m/n) = \mathbb{P}(\hat{R}_n(\phi^*) - R(\phi^*) > \epsilon/2 - \frac{m}{n}) \leq e^{-2n(\epsilon/2 - \frac{m}{n})^2}. \tag{8.7}$$

To bound $\mathbb{P}(I_{1,i} > \epsilon/2)$ we will make use of the tower property Theorem 2.60 and note that

$$\mathbb{P}(I_{1,i} > \epsilon/2) = \mathbb{E}\left[\mathbb{P}(I_{1,i} > \epsilon/2 \mid X_{i_1}, \ldots, X_{i_m})\right] \tag{8.8}$$

Let $K_i = \{i_1, \ldots, i_m\}$ be the set of indices of points used to construct ϕ_i, then

$$\begin{aligned}&\mathbb{P}(I_{1,i} > \epsilon/2 \mid \{X_k, k \in K_i\}) \\ &\qquad = \mathbb{P}\left(R(\phi_i) - \frac{1}{n}\sum_{j=1}^{n} L(Y_j, \phi_i(X_j)) > \epsilon/2 \;\middle|\; \{X_k, k \in K_i\}\right) \\ &\qquad \leq \mathbb{P}\left(R(\phi_i) - \frac{1}{n}\sum_{j=1, j\notin K_i}^{n} L(Y_j, \phi_i(X_j)) > \epsilon/2 \;\middle|\; \{X_k, k \in K_i\}\right).\end{aligned}$$

Now, from this point we can go different paths, but we will use the observation that all $L(Y_j, \phi_i(X_j))$ are Bernoulli$(R(\phi_i))$ for $j \notin K_i$, if we add $Z_1, \ldots, Z_m$ i.i.d. from Bernoulli$(R(\phi_i))$, also independent of $L(Y_j, \phi_i(X_j))$

for $j \notin K_i$, then we can apply Hoeffding to the following

$$\begin{aligned}
&\mathbb{P}\left(R(\phi_i) - \frac{1}{n}\sum_{j=1, j\notin K}^{n} L(Y_j, \phi_i(X_j)) > \epsilon/2 \;\middle|\; \{X_k, k \in K\} \right) \\
&\leq \mathbb{P}\left(R(\phi_i) - \frac{1}{n}\sum_{j=1, j\notin J}^{n} L(Y_j, \phi_i(X_j)) - \frac{1}{n}\sum_{j=1}^{m} Z_i > \epsilon/2 - \frac{m}{n} \;\middle|\; \{X_k, k \in K\} \right) \\
&\leq e^{-2n(\epsilon/2 - m/n)^2}.
\end{aligned}$$

Recalling (8.6)–(8.8) we obtain

$$\begin{aligned}
\mathbb{P}(I_0 > \epsilon) &\leq 2\binom{n}{m} e^{-2n(\epsilon/2 - m/n)^2} + e^{-2n(\epsilon/2 - \frac{m}{n})^2} \\
&= \left(2\binom{n}{m} + 1 \right) e^{-2n(\epsilon/2 - m/n)^2}.
\end{aligned}$$

The final step is to realize that

$$2n(\epsilon/2 - m/n)^2 > \frac{n\epsilon^2}{2} - 2m\epsilon.$$

□

Exercise 8.17. *Consider $X_1, \ldots, X_{m+1}$ be i.i.d. $\mathbb{R}^m$ valued continuous random variables. Use the tower property to prove that the probability of these points to lie in the same hyperplane is zero. Using this, show the assumption of general position needed to construct the decision rules we worked with above.*

8.5 Preliminaries for VC theory

Definition 8.18 (empirical measure)**.**

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \in A}$$

We can deduce also that

Lemma 8.19.

$$R(\phi_n^*) - \inf_{\phi \in \mathcal{C}} R(\phi) \leq 2 \sup_{\phi \in \mathcal{C}} |\hat{R}_n(\phi) - R(\phi)|$$

8.6 VC theory

We have just studied a specific class of decision rules that where constructed using the observations, and we see that the rule selected by (8.5) is indeed very good. It performs very closely to the best possible!! However our method of proof relies on the way that the rule is constructed and in particular does not allow us to minimize the empirical risk over $\mathcal{M}$ instead of $\mathcal{M}_n$. To cope with this, we need to develop some theory that stems from the works of Vapnik and Chervnonenkis, [SLT]. For this we will transition from viewing the decision rules as functions to viewing them as sets, as a decision rule from $\mathcal{M}$ splits $\mathbb{R}^m \times \{0,1\}$ into two pieces, one which gives zero loss and one part which gives loss 1. Define,

$$\mathcal{A} := \{\{(x,y) \in \mathbb{R}^m \times \{0,1\} : L(y,\phi(x)) = 1\} : \phi \in \mathcal{M}\}, \tag{8.9}$$

and denote by $\nu = dF(x,y)$ and ν_n the empirical measure with respect to the data set $D_n = \{(X_1,Y_1),\dots,(X_n,Y_n)\}$. The above definition allows us to rephrase

$$\mathbb{P}(\sup_{\mathcal{M}} |R_n(\phi) - R(\phi)| > \epsilon) = \mathbb{P}(\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| > \epsilon). \tag{8.10}$$

Exercise 8.20. *Derive (8.10). Think about the following, given a decision function $\phi \in \mathcal{M}$, then for this function ϕ there is a corresponding set $A \in \mathcal{A}$ as in (8.9). The measure $\nu(A)$ is simply*

$$\nu(A) = \mathbb{P}((X,Y) \in A) = \mathbb{P}(L(Y,\phi(X)) = 1) = \mathbb{P}(Y \neq \phi(X))$$

and the empirical measure is

$$\nu_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(X_i,Y_i)\in A} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{L(Y_i,\phi(X_i))=1} = \frac{1}{n}\sum_{i=1}^n L(Y_i,\phi(X_i)).$$

This puts our problem in the framework of uniform convergence of empirical measures **UCEM**. The uniform is because we have the supremum inside the probability. Recall, we have already seen an example of this!! If we consider the sets $A = (-\infty, a)$ then we are in the setting of Theorem 5.27.

If $|\mathcal{A}| < \infty$, then we could simply use Hoeffdings inequality, Theorem 3.6, to obtain

$$P(\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| > \epsilon) \leq 2|\mathcal{A}|e^{-2n\epsilon^2}.$$

Exercise 8.21. *Use the Union bound and Theorem 3.6 to prove the above inequality.*

However, even in the simple case of linear decision functions, $\mathcal{M}$, we have an uncountably infinite set.

Lemma 8.22. *Consider a sequence of i.i.d. random variables $Z_1, \dots, Z_{2n} \sim \nu$, split this into two datasets $D_n = \{Z_1, \dots, Z_n\}$ and $D'_n = \{Z_n+1, \dots, Z_2 n\}$. Let $n\epsilon^2 \geq 2$, then if we denote ν_n, ν'_n the empirical measure over D_n and D'_n respectively, we have*

$$\mathbb{P}\left[\sup_{A\in\mathcal{A}} |\nu(A) - \nu_n(A)| > \epsilon\right] \leq 2\,\mathbb{P}\left[\sup_{A\in\mathcal{A}} \left|\nu_n(A) - \nu'_n(A)\right| > \frac{\epsilon}{2}\right]$$

Proof. Let $A^* \in \mathcal{A}$ be a set for which $|\nu_n(A^*) - \nu(A^*)| > \epsilon$, if such a set exists, otherwise we can let A^* be a fixed set in $\mathcal{A}$. **NOTE:** A^* is a random set that depends on D_n. Now

$$\begin{aligned}
&\mathbb{P}(\sup_{A\in\mathcal{A}} \left|\nu_n(A) - \nu'_n(A)\right| > \epsilon/2) \\
&\geq \mathbb{P}(\left|\nu_n(A^*) - \nu'_n(A^*)\right| > \epsilon/2) \\
&\geq \mathbb{P}(|\nu(A^*) - \nu_n(A^*)| > \epsilon, \left|\nu(A^*) - \nu'_n(A^*)\right| < \epsilon/2) \\
&= \mathbb{E}\left[\mathbb{1}_{|\nu(A^*)-\nu_n(A^*)|>\epsilon}\, \mathbb{P}(\left|\nu(A^*) - \nu'_n(A^*)\right| < \epsilon/2 \mid D_n)\right]
\end{aligned}$$

Now, conditioned on D_n, the set A^* is fixed and $\nu'_n(A^*)$ is the mean of n independent Bernoulli($\nu(A^*)$) r.v.s., hence using Chebyshev's inequality (Proposition 3.2)

$$\mathbb{P}(\left|\nu(A^*) - \nu'_n(A^*)\right| < \epsilon/2 \mid D_n) \geq 1 - \frac{\frac{1}{n}\nu(A^*)(1-\nu(A^*))}{(\epsilon/2)^2} \geq 1/2,$$

in the last step we used the assumption $n\epsilon^2 \geq 2$. Putting it all together we now get

$$\begin{aligned}
\mathbb{P}(\sup_{A\in\mathcal{A}} \left|\nu_n(A) - \nu'_n(A)\right| > \epsilon/2) &\geq \frac{1}{2}\,\mathbb{E}\left[\mathbb{1}_{|\nu(A^*)-\nu_n(A^*)|>\epsilon}\right] \\
&= \frac{1}{2}\,\mathbb{P}(|\nu(A^*) - \nu_n(A^*)| > \epsilon) \\
&= \frac{1}{2}\,\mathbb{P}(\sup_{A\in\mathcal{A}} |\nu(A) - \nu_n(A)| > \epsilon)
\end{aligned}$$

where in the last step we used the definition of A^*. □

Now, this lemma gives us precisely what we want, namely to reduce the size of $\mathcal{A}$ from infinite to finite. This we do as follows

- Given a dataset $D_k = \{Z_1, \dots, Z_k\}$ we say that $A, B \in \mathcal{A}$ are equivalent if $A \cap D_k = B \cap D_k$.
- This equivalence relation defines equivalence classes on $\mathcal{A}$ given D_k, let us denote this set $\mathcal{A}_{D_k}$.

- It is clear that
  1. $\mathcal{A}_{D_k}$ is finite,
  2. it satisfies $|\mathcal{A}_{D_k}| \leq |2^{D_k}|$,
  3. non-decreasing with k. (In most interesting cases it grows with k).

Example 8.23. *Consider $\mathbb{R}^2$ and consider the set $\mathcal{M}$ linear decision rules and construct the corresponding $\mathcal{A}$. Let us now take say two points and labels $(x_0, y_0), (x_1, y_1) \in \mathbb{R}^2 \times \{0, 1\}$. Given two different linear functions $\phi_1, \phi_2 \in \mathcal{M}$ construct $A, B \in \mathcal{A}$, as follows*

$$A = \{(x, y) \in \mathbb{R}^2 \times \{0, 1\}, L(y, \phi_1(x)) = 1\}$$
$$B = \{(x, y) \in \mathbb{R}^2 \times \{0, 1\}, L(y, \phi_2(x)) = 1\}$$

we would then say that A, B are equivalent if $A \cap \{(x_0, y_0), (x_1, y_1)\} = B \cap \{(x_0, y_0), (x_1, y_1)\}$. What does this mean? It means that $L(y_0, \phi_1(x_0)) = L(y_0, \phi_2(x_0))$ and $L(y_1, \phi_1(x_1)) = L(y_1, \phi_2(x_1))$ which is the same as saying that $\phi_1(x_0) = \phi_2(x_0)$ and $\phi_1(x_1) = \phi_2(x_1)$. In short, the two decision functions ϕ_1, ϕ_2 assigns the same values to x_0, x_1 and are thus undistinguishable on these points.

Again, let us repeat. The concept of grouping together several decision functions is to say that on our dataset each labeling has with it an equivalence class of decision functions which produce said labeling.

Definition 8.24. *The largest size of $\mathcal{A}_{D_n}$ for a given n is called the shattering number for $\mathcal{A}$ given n*

$$s(\mathcal{A}, n) = \sup_{x_1, \ldots, x_n} |\mathcal{A}_{\{x_1, \ldots, x_n\}}|$$

Example 8.25. *Consider the decision function as being an inequality i.e. $\phi(x) = \mathbb{1}_{x > x_0}$. Then the corresponding sets $\mathcal{A}$ is as follows*

$$\{(x, y), x \in \mathbb{R}, y \in \{0, 1\} : \mathbb{1}_{x > x_0} \neq y\}$$
$$= \{(x, y) : x \in (0, x_0], y = 1\} \cup \{(x, y) : x \in (x_0, \infty), y = 0\}.$$

Consider now two points $(x_1, 1), (x_2, 0)$ with $x_1 < x_2$, then we are counting the number of sets of the form $\{(x_1, 1), (x_2, 0)\} \cap A$. For any decision function there is a threshold x_0, either $x_0 < x_1, x_2$ and we get the set

$$\{(x_2, 0)\}$$

if $x_1 < x_0 < x_2$ *then we get*

$$\{(x_1, 1), (x_2, 0)\}$$

and if $x_1 < x_2 < x_0$

$$\{(x_1, 1)\}$$

Thus all in all we have three sets. Note that we did not actually create the empty set.

This is all equivalent to saying that the number of distinct labelings that the decision function can produce for two points is 3.

Definition 8.26. *For a set of points* $G_n = \{z_1, \ldots, z_n\}$ *we say that* $\mathcal{A}$ *shatters* G_n *if*

$$|\mathcal{A}_{\{z_1,\ldots,z_n\}}| = 2^n$$

What does the above actually mean in terms of the decision function? It means that the decision function can produce all possible labelings of the corresponding set $\{x_1, \ldots, x_n\}$, with $z_i = (x_i, y_i)$.

Lemma 8.27. *Let* $\sigma_1, \ldots, \sigma_n$ *be a i.i.d. sequence of Rademacher random variables, (i.e. is equal to* 1 *or* -1 *with equal probability), then*

$$\mathbb{P}\left[\sup_{A \in \mathcal{A}_{D_n \cup D'_n}} \left|\nu_n(A) - \nu'_n(A)\right| > \frac{\epsilon}{2}\right] \leq 4s(\mathcal{A}, n) \sup_{A \in \mathcal{A}} \mathbb{P}\left[\frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i \mathbb{1}_A(Z_i)\right| > \frac{\epsilon}{4}\right]$$

Proof. The above definition of equivalence classes suggests that we should be able to use the union bound to prove this lemma. We however need to circumvent the technical hurdle that the equivalence classes depend on $D_n \cup D'_n$. This can however be done by again performing a symmetrization with respect to the sign of $\nu_n(A) - \nu'_n(A)$ (this is the right hand side of the above), for details see [PTPR, Thm 12.4]. □

Lemma 8.28.

$$\mathbb{P}\left[\frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i \mathbb{1}_A(Z_i)\right| > \frac{\epsilon}{4}\right] \leq 2e^{-n\epsilon^2/32}$$

Proof.

$$\mathbb{P}\left[\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\mathbb{1}_A(Z_i)\right| > \frac{\epsilon}{4}\right] = \mathbb{E}\left[\mathbb{P}\left[\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\mathbb{1}_A(Z_i)\right| > \frac{\epsilon}{4}\;\middle|\; D_n\right]\right]$$

Now the conditional probability is easy to bound, as given D_n $\sum_{i=1}^{n}\sigma_i\mathbb{1}_A(Z_i)$ is the sum of n independent and bounded (in $\{-1,1\}$) random variables, we can use Hoeffdings bound (Theorem 3.6) to get

$$\mathbb{P}\left[\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\mathbb{1}_A(Z_i)\right| > \frac{\epsilon}{4}\;\middle|\; D_n\right] \leq 2e^{-n\epsilon^2/32}$$

which proves the lemma. □

We are now ready to prove the VC generalization bound

Theorem 8.29. *Consider a sequence $Z_1, \ldots, Z_n \sim \nu$ of i.i.d. random variables and let $\mathcal{A}$ be a set of ν-measurable sets, then if $n\epsilon^2 \geq 2$ we have*

$$\mathbb{P}(\sup_{A\in\mathcal{A}}|\nu_n(A) - \nu(A)| > \epsilon) \leq 8s(\mathcal{A}, n)e^{-n\epsilon^2/32}$$

Exercise 8.30. *Prove the above theorem using Lemmas 8.22, 8.27 and 8.28.*

We also have this immediate corollary

Corollary 8.31. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and consider (X_1, Y_1), ..., (X_n, Y_n) be an i.i.d. sequence of random variables, X_i being continuous and taking values in $\mathbb{R}^m$ and $Y_i \in \{0, 1\}$ is discrete. Then if $n(\epsilon)^2 \geq 8$ the following holds*

$$\mathbb{P}(\sup_{\mathcal{M}}|R_n(\phi) - R(\phi)| > \epsilon) \leq 8s(\mathcal{A}, n)e^{-n\epsilon^2/64}.$$

In the above, $\mathcal{A}$ is derived from $\mathcal{M}$ as in (8.9).

8.7 Vapnik Chervonenkis dimension

Definition 8.32. *The VC-dimension of $\mathcal{A}$, denoted by $\mathcal{V}_\mathcal{A}$, equals the largest integer $n \geq 1$ such that*

$$s(\mathcal{A}, n) = 2^n.$$

If the above equality holds for all n we say that $\mathcal{V}_\mathcal{A} = \infty$.

Example 8.33. *Consider sets of the form*

$$\mathcal{A} = \{(0, x_0] \times \{1\}\} \cup \{(x_0, \infty) \times \{0\} : x_0 \in \mathbb{R}\}$$

which corresponds to a decision rule as in Example 8.25. What is $s(\mathcal{A}, 1)$? Consider a single points (x_1, y_1), $x_1 \in \mathbb{R}$ and $y_1 \in \{0, 1\}$. Then for $y_1 = 0$ we get

$$\mathcal{A}_{(x_1,y_1)} = \{\{(x_1, y_1)\}, \emptyset\}$$

and for $y_1 = 1$ we get also two sets. So $s(\mathcal{A}, 1) = 2$ which is 2^1 so $\mathcal{V}_A \geq 1$. Now consider two points, we already saw this in Example 8.25 where we only got 3 sets which is less than $2^2 = 4$ (check that this is so for any other labeling as well). Conclusion is that $\mathcal{V}_A = 1$.

Example 8.34. *Lets consider the sets*

$$\mathcal{A} = \{\{(a, b) \times (c, d)\} \times \{0\} \cup \{(a, b) \times (c, d)\}^C \times \{1\} \subset \mathbb{R}^2 \times \{0, 1\} : a < b, c < d\}$$

that is, this corresponds to a classifier which classifies everything within an axis parallel rectangle as 1 and outside as 0. Now, the realization should be that it is enough to check how many different labelings we can create using rectangles. Consider a diamond pattern set of points in $\mathbb{R}^2$, i.e. $\{(1, 0), (0, 1), (-1, 0), (0, -1)\}$, using axis-parallel rectangles it should be clear that we can create all possible labelings.

Now consider 5 points, how do we realize that no matter how we do this can we create all labelings. To see this, assume for contradiction we can produce any labeling. Find the minimum enclosing rectangle for the five points. Let us now think that the points which touch the edges will be given as class 1 and the last point as class 0, but this point is inside the set and as such cannot be labeled as that. This gives us that the VC dimension of these axis parallel rectangles is 4.

Example 8.35. *Let us now consider the following class of sets consisting of polygons as in the example above. Then for any number of points placed along a unit circle we can produce any labeling, why? because if we select a subset of the points and then construct a polygon with those points as corners then this will label them as we wish. This means that any number of points can be labeled and we thus get that the VC dimension is infinite.*

Lemma 8.36 (Sauer–Shelah lemma (1972))**.** *For any positive integer N we have*

$$s(\mathcal{H}, N) \leq \sum_{i=0}^{\mathcal{V}_{\mathcal{H}}-1} \binom{N}{i}.$$

Proof. Out of scope. □

Lemma 8.37. *The Sauer–Shelah lemma (Lemma 8.36) is a polynomial upper bound, i.e.*

$$\sum_{i=0}^{k-1} \binom{N}{i} \leq \left(\frac{Ne}{k}\right)^k.$$

Proof. Let $\lambda \in (0,1)$ then

$$\begin{aligned} 1 &= (\lambda + (1-\lambda))^N \\ &\geq \sum_{i=1}^{\lambda N} \binom{N}{i} \lambda^i (1-\lambda)^{n-1} \\ &\geq \sum_{i=1}^{\lambda N} \binom{N}{i} \left(\frac{\lambda}{1-\lambda}\right)^{\lambda n} (1-\lambda)^n \end{aligned}$$

Thus

$$\begin{aligned} \sum_{i=1}^{\lambda N} \binom{N}{i} &\leq e^{N((\lambda-1)\log(1-\lambda) - \lambda\log(1-\lambda))} \\ &\leq e^{N(\lambda - \lambda\log(1-\lambda))} \\ &= \left(\frac{eN}{\lambda N}\right)^{\lambda N} \end{aligned}$$

Then for $k = \lambda N$ we have our result. □

Let us now turn our focus back towards the problem of the linear classifier.

Lemma 8.38. *(informal) A linear classifier in $\mathbb{R}^m$ has VC-dimension $m+1$.*

Exercise 8.39. *State and prove exactly what the above lemma means.*

Exercise 8.40. *Continuing on from the above exercise, apply Corollary 8.31 and Lemmas 8.36 and 8.37 to obtain a generalization of Theorem 8.16 which you state and prove.*

8.8 What if you don't care about $\inf R(\phi)$?

We begin with an extension of Theorem 3.6 (Hoeffdings theorem) to a sequence of dependent variables, specifically we consider

Definition 8.41. *A sequence of random variables $V_1, \ldots$ is a martingale difference sequence if*

$$\mathbb{E}\left[V_{i+1} \mid V_1, \ldots, V_i\right] = 0, \quad a.s.$$

for every $i > 0$. A sequence of random variables $V_1, V_2, \ldots$ is called a martingale difference sequence with respect to a sequence of random variables $X_1, X_2, \ldots$, if for every $u > 0$, V_i is a function of $X_1, X_2, \ldots$, and

$$\mathbb{E}\left[V_{i+1} \mid X_1, \ldots, X_i\right] = 0, \quad a.s.$$

This extended version of Hoeffdings theorem will be proved in the theoretical foundations course, if you are interested in the proof, see [PTPR, Thm 9.1].

Theorem 8.42. *Let $X_1, \ldots$ be a sequence of RV's and assume that $V_1, \ldots$ is a martingale difference sequence with respect to $X_1, \ldots$. Assume that there exists random variables $Z_1, \ldots$ and nonnegative constants $c_1, \ldots$ such that for every $i > 0$, Z_i is a function of $X_1, \ldots, X_{i-1}$ and*

$$Z_i \leq V_i \leq Z_i + c_i, \quad a.s.$$

Then for any $\epsilon > 0$ and n

$$\mathbb{P}\left(\sum_{i=1}^{n} V_i \geq \epsilon\right) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n} c_i^2}}$$

and

$$\mathbb{P}\left(\sum_{i=1}^{n} V_i \leq -\epsilon\right) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n} c_i^2}}$$

Interestingly, the empirical error probability of the empirically optimal classifier is with high probability close to its expected value. This is very interesting, and in the case that if the empirical minimizer of the risk is a consistent estimator, we get the best one could hope for. However it should be mentioned again that $R_n(\phi_n^*)$ is quite often a downward biased estimator of $R(\phi^*)$, that is, $\mathbb{E}\left[R_n(\phi_n^*)\right] < \inf R(\phi^*)$. Therefore, this whole deal with complexity tells us that $R_n(\phi_n^*)$ is asymptotically consistent if we have bounded VC-dimension.

Theorem 8.43. *Consider a sequence $(X_1, Y_1), \ldots, (X_n, Y_n) \sim \nu$ of i.i.d. random variables, let $\mathcal{C}$ be an arbitrary set of classification rules. Let $\phi_n^* \in \mathcal{C}$ be the rule that minimizes the empirical risk given $(X_1, Y_1), \ldots, (X_n, Y_n)$, i.e.*

$$R_n(\phi_n^*) = \min_{\phi \in \mathcal{C}} R_n(\phi).$$

Then for every n and $\epsilon > 0$,

$$\mathbb{P}\left(\left|\hat{R}_n(\phi_n^*) - \mathbb{E}\left[\hat{R}_n(\phi_n^*)\right]\right| > \epsilon\right) < 2e^{-n\epsilon^2/2}$$

Proof. Begin by defining

$$\begin{aligned} Q_i &:= \mathbb{E}\left[R_n(\phi_n^*) \mid X_1, \ldots, X_i\right], \quad i = 1, \ldots, n \\ Q_0 &:= \mathbb{E}\left[R_n(\phi_n^*)\right] \end{aligned}$$

furthermore, note that $Q_n = R_n(\phi_n^*)$ and that

$$Q_{i-1} - \frac{1}{n} \leq Q_i \leq Q_{i-1} + \frac{1}{n}$$

since changing the value of one pair (X_i, Y_i) can only change the risk by $1/n$ (1 extra or one less error in classification). If we now denote

$$V_i = Q_i - Q_{i-1}, \quad i = 1, \ldots, n$$

then V_i is a martingale difference sequence w.r.t. (X_i, Y_i) for $i = 1, \ldots, n$. Furthermore, if we define $Z_i = -1/n$ then

$$Z_i \leq V_i \leq Z_i + \frac{2}{n}.$$

Applying Theorem 8.42 we obtain

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} V_i\right| > \epsilon\right) \leq 2e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}(2/n)^2}} = 2e^{-n\epsilon^2/2}$$

□

8.9 Bibliography

The perceptron and kernel part is mostly from [BlHo]. Section 8.4 is mostly covered in [PTPR] Chapter 4. Section 8.6 and the rest is scattered around in [PTPR] and other sources.

Chapter 9

Regression

The main difference between regression and pattern recognition is that the loss function $l(y, g(x))$ is real valued instead of being discrete. What we will do in this chapter is to outline how to modify the ideas for the pattern recognition problem to obtain generalization estimates for real valued loss functions, which includes regression.

It however turns out that dealing with unbounded loss functions is technically difficult and we will not cover it here, if you want more information, take a look at [SLT].

As in the pattern recognition problem, we want to bound

$$\mathbb{P}(\sup |R_n(\phi) - R(\phi)| > \epsilon) \tag{9.1}$$

we did that by rephrasing this as a problem of estimating empirical measures on certain classes of sets. How do we do the same for the problem where $l(y, g(x))$ can take any value between $[0, 1]$ for instance?

Consider a function $0 \leq \Phi(z) \leq 1$, and consider a sequence of i.i.d. random variables $Z, Z_i \sim \nu$, then

$$\begin{aligned}
\mathbb{E}\left[\phi(Z)\right] - \frac{1}{n}\sum_{i=1}^{n}\phi(Z_i) & \\
&= \int_0^1 \left(\nu(\Phi(z) > t) - \nu_n(\phi(z) > t)\right) dt \\
&\leq \sup_{\beta\in[0,1]} \left(\nu(\Phi(z) > \beta) - \nu_n(\phi(z) > \beta)\right) \int_0^1 dt \\
&= \sup_{\beta\in[0,1]} \left(\nu(\Phi(z) > \beta) - \nu_n(\phi(z) > \beta)\right)
\end{aligned}$$

where ν_n is the empirical measure based on $Z_1, \ldots, Z_n$.

That is, if we have a model space $\mathcal{M}$ of decision functions and a loss $l(y, g(x))$, for $g \in \mathcal{M}$ taking values in $[0, 1]$, then if we construct the class of

sets as follows

$$\mathcal{A} = \big\{\{(x,y) : l(y,g(x)) - \beta > 0\} : \beta \in (0,1)\big\}, \tag{9.2}$$

we can rewrite (9.1) in the following way

$$\mathbb{P}(\sup_{\mathcal{M}} |R_n(\phi) - R(\phi)| > \epsilon) \leq \mathbb{P}(\sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)| > \epsilon).$$

Now we are exactly the same situation as in (8.10) and we can apply Theorem 8.29 to get

Corollary 9.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let (X_1, Y_1), ..., (X_n, Y_n) be an i.i.d. sequence of random variables, X_i being continuous and taking values in $\mathbb{R}^m$ and $Y_i \in \mathbb{R}$. Then if $n\epsilon^2 \geq 8$ the following holds*

$$\mathbb{P}(\sup_{\mathcal{M}} |R_n(\phi) - R(\phi)| > \epsilon) \leq 8s(\mathcal{A}, n)e^{-n\epsilon^2/64}.$$

In the above, $\mathcal{A}$ is derived from $\mathcal{M}$ as in (9.2).

Example 9.2. *Assume that $g(x) \in [0,1]$ and that $l(y,x) = (y-x)^2$, then*

$$\{(x,y) : l(y,g(x)) - \beta > 0\} = \{(x,y) : |y - g(x)| > \sqrt{\beta}\}$$

that is, for every fixed value of x, the set is all y which are at distance greater than $\sqrt{\beta}$ from $g(x)$. That is, this is the complement to a tubular region around the graph $g(x)$.

A similar observation can be made for any convex loss function, it is thus clear that the growth function will often only depend on the complexity of $\mathcal{M}$ instead of depending on the choice of loss.

9.1 Guarantees with a held out testing set

Consider as in Section 8.3.1 the following notation. Consider a data set $T_{n+m} := \{(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})\}$ sampled i.i.d from $F_{X,Y}$. We consider $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ (which we dub the training data) and $\{(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})\}$ (which we dub the testing data). Define $\hat{\phi}$ the **empirical risk minimizer** on the **training dataset**, namely

$$\hat{R}_n(\hat{\phi}) = \min_{\phi \in \mathcal{M}} \hat{R}_n(\phi)$$

then since the **testing dataset** is independent of the training dataset and hence $\hat{\phi}$ is independent of the testing data, we want to use Theorem 3.6 but now we would like to consider the loss function to be unbounded. This

happens in the case of mean square error for instance. Consider the most usual quantity (mean squared error)

$$R(\phi) = \mathbb{E}\left[(Y - g(X))^2\right]$$

Previously we had $0-1$ loss and we thus knew that the random variable $L(Y, \phi(X))$ was bounded and we could immediately apply Theorem 3.6. Since we do not know that we have to make some assumptions to move forward, but this is getting into advanced topics and is outside the scope of this course, since how do we know that the assumptions make sense? Anyways, we will for simplicity make the assumption that $(Y - \phi(X))^2$ is sub-Gaussian with parameter $\lambda(\phi)$, i.e. where the parameter depends on the function ϕ. If you are up to it, you should spend some time thinking about why this is so. Hint: think of $\phi(x) = ax + b$ being a linear function and let $Y = 0$ and $X \sim \text{Bernoulli}(1/2)$, that is the sub-Gaussian parameter of $(\phi(X))^2$?

Anyways, we can get the following bound for the Risk using Theorem 3.13 that if $\hat{R}_m(\phi)$ denotes the empirical risk over the testing dataset we have

$$\mathbb{P}(|\hat{R}_m(\hat{\phi}) - R(\hat{\phi})| > \epsilon \mid T_n) < 2e^{-\frac{\epsilon^2 n}{2\lambda(\hat{\phi})^2}}. \tag{9.3}$$

We now run into some trouble, since we have a random variable on the right hand side of the bound, namely $\lambda(\hat{\phi})$ so we cannot use the tower property as we did in deriving (8.4) since we would need to be able to compute

$$\mathbb{E}\left[e^{-\frac{\epsilon^2 n}{2\lambda(\hat{\phi})^2}}\right].$$

However, this is usually not a problem. Since, if we adhere to the Train-Test philosophy we are actually only interested in (9.3), as at the point of having trained and found $\hat{\phi}$, the value $\lambda(\hat{\phi})$ is computable given only some assumptions on X.

Remark 9.3. *It should be noted that it is more often the case that* $(Y - \phi(X))^2$ *is sub-Exponential. This happens for instance if* Y *is Gaussian, since its squared, see Lemma 3.15.*

9.1.1 R^2

A common metric used in evaluating regression models is the so called R^2. The reason for the name comes from the theory of linear regression, where R^2 is actually the correlation squared. The metric is an empirical one. First consider what is called the fraction of variance unexplained

$$\widehat{FVU}(\hat{\phi}; T_{n+m} \setminus T_n) = \frac{\frac{1}{m}\sum_{i=n+1}^{m}(Y_i - \hat{\phi}(X_i))^2}{\frac{1}{m-1}\sum_{i=n+1}^{m}(Y_i - \frac{1}{m}\sum_{i=n+1}^{m} Y_i)^2} = \frac{\hat{R}_m(\hat{\phi})}{\hat{V}_m[Y]}$$

Lets explain the terms in the above expression as it looks quite crowded. The left hand side is the expression that we define as the Fraction of Variance Explained (FVU) and it takes the proposed (Trained) function $\hat{\phi}$ and the test set, i.e., $T_{n+m} \setminus T_n = \{(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})\}$. The ratio on the right hand side: the numerator is the empirical test risk $\hat{R}_m(\hat{\phi})$ and the denominator is the empirical variance of Y over the testing set.

Remark 9.4. *Due to the fact that in the definition of FVU the function $\hat{\phi}$ can be anything, we can have FVU take any non-negative value. Specifically, it can be greater than* 1.

Now, we define a version of R^2 that comes from the idea of variance explained and it is just $1 - FVU$, but beware, it can be negative so the term R^2 does not make sense, as it is not a square (only in the case of linear regression). You will encounter this confusion as you go out into industry, so make sure you get it right!

Can we make a concentration statement about FVU? Well, easiest is to realize that the true FVU given $\hat{\phi}$ is just

$$FVU(\hat{\phi}) = \frac{R(\hat{\phi})}{\mathbb{V}[Y]}$$

we can given the above discussion actually get intervals for both the numerator and denominator separately, i.e. if we assume that Y^2 is sub-Gaussian with parameter σ we get

$$\mathbb{P}(|\hat{R}_m(\hat{\phi}) - R(\hat{\phi})| > \epsilon \mid T_n) < 2e^{-\frac{\epsilon^2 n}{2\lambda(\hat{\phi})^2}}$$
$$\mathbb{P}\left(\left|\hat{V}_m[Y] - \mathbb{V}[Y]\right| > \epsilon \mid T_n\right) < 2e^{-\frac{\epsilon^2 n}{2\sigma^2}}$$

Using the union bound Lemma 1.12 we get that

$$\mathbb{P}\left(|\hat{R}_m(\hat{\phi}) - R(\hat{\phi})| \leq \epsilon \text{ and } \left|\hat{V}_m[Y] - \mathbb{V}[Y]\right| \leq \epsilon \mid T_n\right) \geq 1 - 2e^{-\frac{\epsilon^2 n}{2\lambda(\hat{\phi})^2}} - 2e^{-\frac{\epsilon^2 n}{2\sigma^2}}$$

Thus we can write a bound for the ratio by rearranging a bit and assuming that all the quantities are non-negative, i.e. $\hat{R}_m(\hat{\phi}) - \epsilon \geq 0$ and $\hat{V}_m[Y] - \epsilon \geq 0$.

$$\mathbb{P}\left(\frac{\hat{R}_m(\hat{\phi}) - \epsilon}{\hat{V}_m[Y] + \epsilon} \leq \frac{R(\hat{\phi})}{\mathbb{V}[Y]} \leq \frac{\hat{R}_m(\hat{\phi}) + \epsilon}{\hat{V}_m[Y] - \epsilon} \mid T_n\right) \geq 1 - 2e^{-\frac{\epsilon^2 n}{2\lambda(\hat{\phi})^2}} - 2e^{-\frac{\epsilon^2 n}{2\sigma^2}}$$

However, the problem is that often we do not know much about what is sub-Gaussian etc. but we might be in a situation where things are bounded and one could hope to apply Theorem 3.6. The problem is that in this case,

often the sum of square of the residual is quite small (good model) and perhaps also the variance of Y is small, in this case the Hoeffding inequality is too rough and we will employ a stronger inequality called **Bennett's inequality**. It looks quite complicated, but all the things are computable and we do it in the notebooks:

Theorem 9.5 (Bennett's inequality)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables with finite variance such that $\mathbb{P}(X_i \leq b) = 1$ with mean zero. Let and $\sigma^2 = \mathbb{V}[X_i]$. Then for any $\epsilon > 0$,*

$$\mathbb{P}(\overline{X}_n \geq \epsilon) \leq \exp\left(-\frac{n\sigma^2}{b^2} h\left(\frac{b\epsilon}{\sigma^2}\right)\right)$$

where $h(u) = (1+u)\log(1+u) - u$ for $u > 0$.

In what case can we apply this theorem. Lets make the assumption that $|Y| \leq 1$ (can be done via scaling of the data). Consider again that we found our model $\hat{\phi}$ by training on T_n and let $b = \max_X \hat{\phi} - \min_X \hat{\phi}$. The constant b can be a bit tricky to get, but we can guess its value by taking the max and the min over the data.

$$\mathbb{P}(|\hat{R}_m(\hat{\phi}) - R(\hat{\phi})| > \epsilon_1 \mid T_n) \leq \exp\left(-\frac{n\sigma^2}{b^2} h\left(\frac{b\epsilon_1}{\sigma^2}\right)\right)$$

where $\sigma^2 = \mathbb{V}[L(Y, \hat{\phi}(X))]$ (can also be estimated from data). For the Y part we can do

$$\mathbb{P}\left(\left|\hat{V}_m[Y] - \mathbb{V}[Y]\right| > \epsilon_2 \mid T_n\right) < \exp\left(-n\sigma^2 h\left(\frac{\epsilon_2}{\sigma^2}\right)\right)$$

where $\sigma^2 = \mathbb{V}[(Y - \mathbb{E}[Y])^2]$. One option is to find ϵ_1, ϵ_2 such that

$$\frac{\alpha}{2} = \exp\left(-\frac{n\sigma^2}{b^2} h\left(\frac{b\epsilon_1}{\sigma^2}\right)\right)$$
$$\frac{\alpha}{2} = \exp\left(-n\sigma^2 h\left(\frac{\epsilon_2}{\sigma^2}\right)\right)$$

which then gives a final bound of

$$\mathbb{P}\left(\frac{\hat{R}_m(\hat{\phi}) - \epsilon_1}{\hat{V}_m[Y] + \epsilon_2} \leq \frac{R(\hat{\phi})}{\mathbb{V}[Y]} \leq \frac{\hat{R}_m(\hat{\phi}) + \epsilon_1}{\hat{V}_m[Y] - \epsilon_2} \mid T_n\right) \geq 1 - \alpha$$

Remark 9.6. *There is an example in the Regression notebook where this inequality is used in practice.*

9.2 Bibliography

The generalization bounds and definition of the VC dimension for a more general loss than in the pattern recognition case is a very natural extension. Some parts of the above can be found in [SLT], but the connection to empirical measures was not done. As we know, we should expect some of these results to hold for unbounded loss functions, provided that we can bound the tail of the empirical risk. As far as I know, the most general results can be found in [SLT], Chapter 5. For Bennett's inequality, see for instance [BLM].

Chapter 10

High dimension

10.1 Introduction: Volume of the unit ball in d dimensions

In this chapter we will deal with the geometry of high dimensions, specifically we will touch upon the volume of unit balls and how to sample from them and from unit spheres. Generating points from a unit ball or sphere is very useful thing in applications and simulation.

Definition 10.1. *Given a radius $r > 0$ we define the d-**dimensional ball** as the set*

$$B_r(x) := \{y \in \mathbb{R}^d : |x - y| < r\}.$$

*We also denote the d-**dimensional sphere** as the set*

$$S_r(x) := \{y \in \mathbb{R}^d : |x - y| = r\}.$$

*Whenever $r = 1$ we call $B_1(x), S_1(x)$ **unit ball** and **unit sphere** respectively. If $x = 0$ we omit it from the notation, and use $B_r = B_r(0)$ and $S_r = S_r(0)$.*

Remark 10.2. *In this chapter we will be using the volume of sets in $\mathbb{R}^d$, but how do we define the volume? It is simply as follows*

$$|E| = \int_E dx = \int \mathbb{1}_E dx = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbb{1}_E(x_1, \ldots, x_d) dx_1 \ldots dx_d.$$

But how does the volume of a unit ball change with dimension? Intuitively you would perhaps say that it does not change, so lets use the law of large numbers to convince you otherwise.

To start, let us define the spherical Gaussian

Model 10.3. *A continuous $\mathbb{R}^d$ valued random variable Z with density function*

$$f(x) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}|x|^2\right), \quad x \in \mathbb{R}^d$$

is called a **spherical Gaussian**. *In short, each coordinate is a standard Gaussian and are independent of eachother.*

Let us also define the normalized Gaussian

Model 10.4. *Let Z be a spherical Gaussian in $\mathbb{R}^d$ and consider $Y = (2\pi)^{-1/2}Z$ then Y is called a* **normalized Gaussian**, *the density is*

$$f(x) = \exp\left(-\pi|x|^2\right), \quad x \in \mathbb{R}^d.$$

Lemma 10.5. *Let $d > 4\pi$ then for $B_1 \subset \mathbb{R}^d$ there exists a constant $C > 1$ that does not depend on dimension such that*

$$|B_1| \leq \frac{C}{d}.$$

Proof. Let $Z \in \mathbb{R}^d$ be a normalized Gaussian, then we have that the density of Z, f, satisfies

$$\begin{aligned} f(0) &= 1 \\ f(z) &\geq e^{-\pi}, \quad z \in B_1. \end{aligned}$$

Now for the normalized Gaussian each component of $Z = (Z_1, \ldots, Z_d)$ are i.i.d. and they all are Gaussians with variance $(2\pi)^{-1/2}$, this means that $|Z|^2 = \sum_i |Z_i|^2$ is now a sum of independent r.v.s and if we use Proposition 3.2

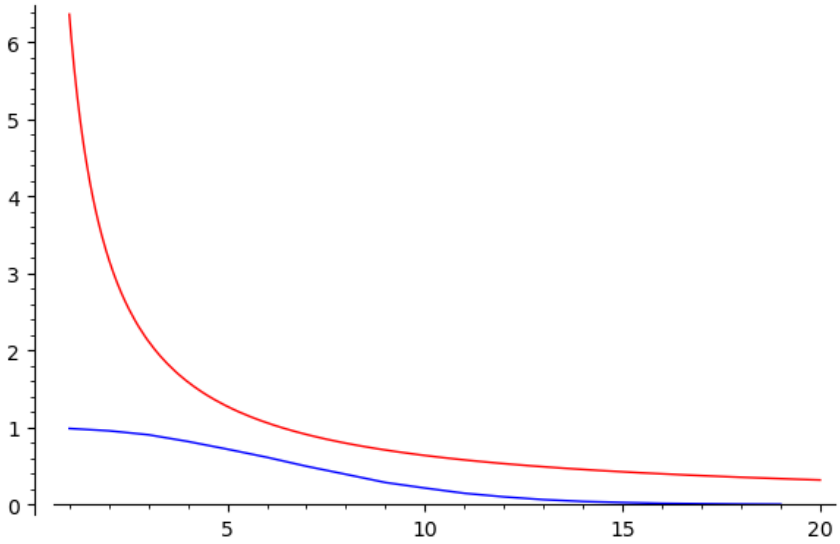
$$\mathbb{P}\left(\left||Z|^2 - \mathbb{E}\left[|Z|^2\right]\right| \geq \epsilon\right) \leq \frac{\operatorname{Var}(|Z|^2)}{\epsilon^2} = \frac{d\operatorname{Var}(|Z_1|^2)}{\epsilon^2} = \frac{cd}{\epsilon^2} \tag{10.1}$$

in the second to last step we used independence and in the last step we used that $\operatorname{Var}(|Z_1|^2) = c$ is a number that we can compute but we will skip it here. However, we also know that $\mathbb{E}\left[|Z|^2\right] = \frac{d}{2\pi}$, so if we choose $\epsilon = (d - 2\pi)/2\pi$ we get (f is the density for Z)

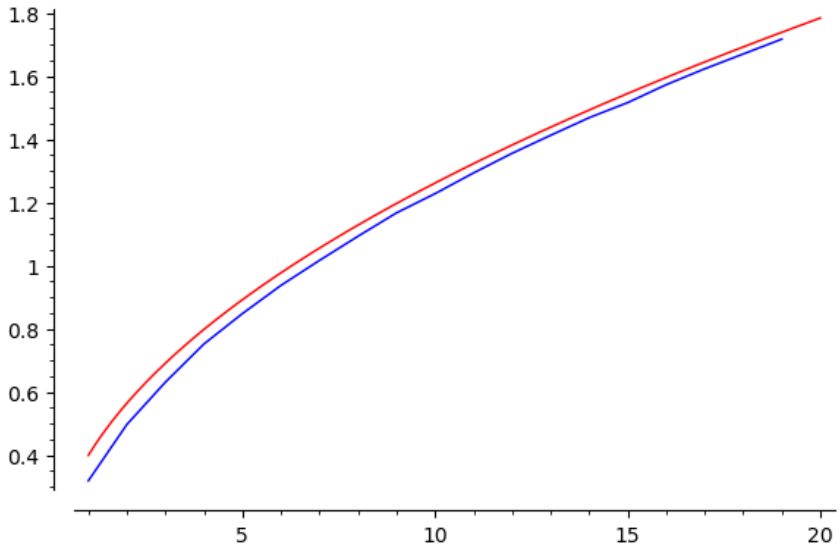
$$\begin{aligned} \frac{|B_1|}{e^\pi} &\leq \int_{B_1} f(z)dz = \mathbb{P}(Z \in B_1) \\ &\leq \mathbb{P}\left(|Z|^2 \geq \epsilon + \mathbb{E}\left[|Z|^2\right] \text{ or } \mathbb{E}\left[|Z|^2\right] - \epsilon \geq |Z|^2\right) \\ &= \mathbb{P}\left(\left||Z|^2 - \mathbb{E}\left[|Z|^2\right]\right| \geq \epsilon\right) \\ &\leq \frac{cd}{((d-2\pi)/(2\pi))^2} \leq \frac{C}{d} \end{aligned}$$

for some constant $C > 1$. □

Thus the volume of the unit ball will decrease with dimension.



In the picture you can see the blue curve being the estimated probability of a Gaussian landing inside the unit ball for different dimensions, while the red curve denotes the upper bound given by Lemma 10.5. We did a fairly poor job at capturing the behavior, the actual volume seems to be much smaller than our estimate. However, the estimate (10.1) also suggests that $|Z|$ should concentrate around $\sqrt{\frac{d}{2\pi}}$, below you can see the plots of the estimated $|Z|$ vs the expected.



This seems fairly spot on, interesting!

Exercise 10.6. *The proof above used the concentration inequality (Chebyshev), which is a fairly weak one. Can you improve on the estimate above using another concentration inequality? Do this before you read on...*

10.2 The geometry of high dimension

The scaling property of volume. Lets say we have a cube centered at the origin, namely the cube can be written as $Q = [-l, l]^d$ where d is the dimension, the volume is the product of the side-lengths and thus $|Q| = (2l)^d$. Scaling each side of the cube by $(1-\epsilon)$ where ϵ is a small number gives us that the volume also scales with $(1-\epsilon)^d$, this gives us the formula

$$|(1-\epsilon)Q| = (1-\epsilon)^d |Q|$$

Lets divide this equation by the volume of Q, we get

$$\frac{|(1-\epsilon)Q|}{|Q|} = (1-\epsilon)^d \to 0$$

as $d \to \infty$. The conclusion is that most of the volume is located close to the surface of the cube.

Lemma 10.7. *Let $E \subset \mathbb{R}^d$ and let $\epsilon \in (0, 1]$, then*

$$(1-\epsilon)^d |E| = |(1-\epsilon)E|$$

where $(1-\epsilon)E := \{(1-\epsilon)x : x \in E\}$.

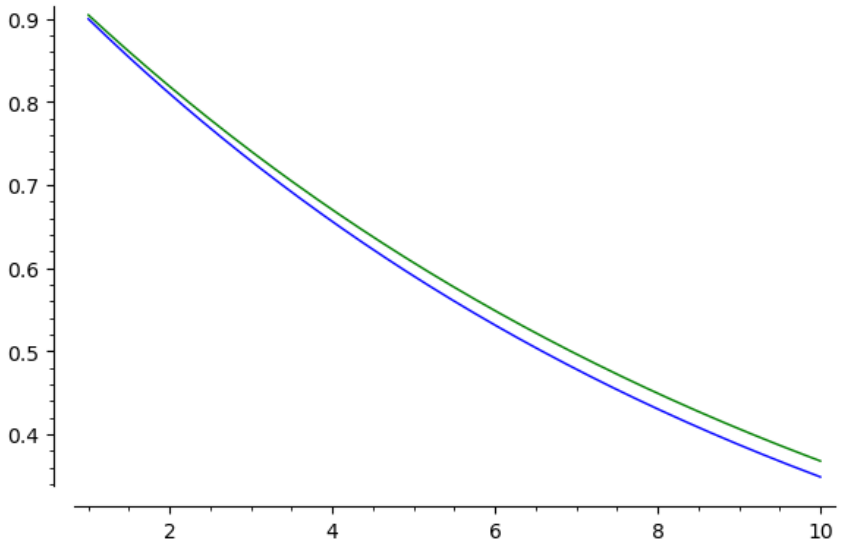
Proof. By the change of variables formula ($y_1 = (1-\epsilon)x_1$ we get $dy_1 = (1-\epsilon)dx_1$) and our area becomes

$$\begin{aligned}
|E| &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbb{1}_E(x_1, \ldots, x_d) dx_1 \ldots dx_d \\
&= \frac{1}{(1-\epsilon)^d} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbb{1}_E(y_1/(1-\epsilon), \ldots, y_d/(1-\epsilon)) dy_1 \ldots dy_d \\
&= \frac{1}{(1-\epsilon)^d} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbb{1}_{(1-\epsilon)E}(y_1, \ldots, y_d) dy_1 \ldots dy_d \\
&= \frac{1}{(1-\epsilon)^d} |(1-\epsilon)E|
\end{aligned}$$

which is what we wanted to prove. □

[269]:
```
P=plot((1-0.1)^x,1,10)
P+=plot(exp(-0.1*x),1,10,color='green')
P.show()
```

[269]:



Based on the above, we can choose $\epsilon = 1/d$ which gives us that most of the volume is contained in the annulus below.

10.3 Properties of the unit ball

Let us now do one of the main computations of this chapter, namely the volume of the unit ball is computed exactly. The computation is a bit complicated but follows the following simple structure:

1. Write the volume of the unit ball as an integral of the constant function 1 over the unit ball
2. Use a radial coordinate system to rewrite that integral so that we get the integral over the surface of a unit ball instead.
3. Compute the integral of the Gaussian kernel in two ways, one using the fact that $\exp(|x|^2) = \exp(|x_1|^2)\exp(|x_2|^2)\ldots\exp(|x_d|^2$, the second one using radial coordinates
4. The radial part of the Gaussian integral gives rise to the Gamma function, which is a generalization of the factorial, the spherical part is just the area of the unit sphere (which is the one we are after).

Theorem 10.8. *The volume of the unit ball in d dimensions is*

$$|B_1| = \frac{2\pi^{\frac{d}{2}}}{d\Gamma(\frac{d}{2})}$$

where Γ *is the aptly named Gamma-function. If* k *is an integer then* $\Gamma(k) = (k-1)!$*, which gives us for even dimensions that*

$$|B_1| = \frac{2\pi^{\frac{d}{2}}}{d(\frac{d}{2}-1)!}.$$

Proof. We begin by first writing down what we want to compute

$$|B_1| = \int_{B_1} dx$$

Step 2: Surface integral

$$\int_{B_1} dx = \int_{S^d} \int_0^1 \left|\frac{dx}{dr}\right| dr d\Omega$$

where $\left|\frac{dx}{dr}\right|$ is the Jacobian of the change of variables and $d\Omega$ is the surface element on the unit sphere S_1 (think of the area of a tiny square on the surface of a ball, in 3d we can think of the longitude and latitue coordinates).

$$\left|\frac{dx}{dr}\right| = r^{d-1}$$

The conclusion is that

$$\int_{B_1} dx = \int_{S_1} \int_0^1 \left|\frac{dx}{dr}\right| dr d\Omega = \frac{|S_1|}{d}.$$

In the above we used $|S_1| := \int_{S_1} d\Omega$.

Step 3a: Gaussian kernel trick (repeated integrals)

This is where we use the Gaussian kernel trick, first note that

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

The standard normal random variable has a normalizing factor which is $1/\sqrt{\pi}$.

[273]:
```
x = var('x')
integrate(exp(-x^2),x,-infinity,infinity)
```

[273]:
```
sqrt(pi)
```

$$\int_{\mathbb{R}^d} e^{-|x|^2} dx = \int_{\mathbb{R}^d} \prod_{i=1}^{d} e^{-x_i^2} dx = \prod_{i=1}^{d} \int_{\mathbb{R}^d} e^{-x_i^2} dx_i = \pi^{d/2}$$

Step 3b: Gaussian kernel trick (spherical coordinates)

Let us compute the same integral again, but this time using spherical coordinates

$$\int_{\mathbb{R}^d} e^{-|x|^2} dx = \int_{S_1} \int_0^\infty e^{-r^2} r^{d-1} dr d\Omega = |S_1| \int_0^\infty e^{-r^2} r^{d-1} dr$$

now doing the change of variables $t = r^2$ we get $dt = 2rdr$ and thus

$$\int_0^\infty e^{-r^2} r^{d-1} dr = \int_0^\infty e^{-t} t^{\frac{d-1}{2}} \frac{1}{2\sqrt{t}} dt = \frac{1}{2} \int_0^\infty e^{-t} t^{\frac{d}{2}-1} dt = \frac{1}{2} \Gamma\left(\frac{d}{2}\right)$$

In conclusion assembling the previous step with this

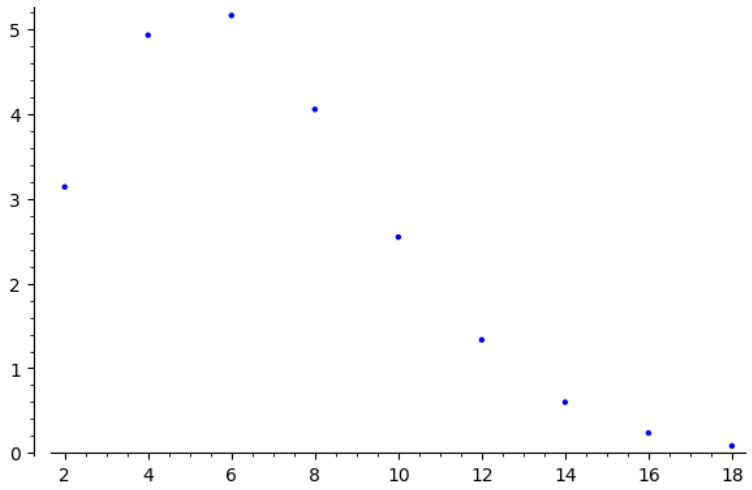
$$\pi^{d/2} = \int e^{-|x|^2} dx = |S_1| \frac{1}{2} \Gamma\left(\frac{d}{2}\right).$$

Step 4: Assembly time

Assemling step 1 and 2 together with the line above we get

$$|B_1| = |S_1|/d = \frac{\pi^{d/2}}{\frac{1}{2}\Gamma\left(\frac{d}{2}\right) d}$$

which completes the proof. □



10.4 Uniform at random from a ball and sphere

Oftentimes we want to work with what is denoted as uniform at random from the unit sphere or the unit ball. Let us define these

Model 10.9. *We say that an $\mathbb{R}^d$ valued random variable Z is* **uniform at random from the unit sphere** *if $Z \in S_1$ and for any A we have*

$$\mathbb{P}(Z \in A) = \frac{1}{|S_1|} \int_{S_1} \mathbb{1}_A(\theta) d\Omega(\theta)$$

where the integral above is the surface integral on the sphere, here $d\Omega$ is the surface element on S_1. We denote this as $Z \sim \text{Uniform}(S_1)$.

Remark 10.10. *We havent really defined what a surface element is, but the heuristic understanding is enough for now.*

This definition is easier to grasp

Model 10.11. *We say that an $\mathbb{R}^d$ valued random variable Z is* **uniform at random from the unit ball** *if $Z \in B_1$ and for any A we have*

$$\mathbb{P}(Z \in A) = \frac{1}{|B_1|} \int_{B_1} \mathbb{1}_A(z) dz = \frac{|A \cap B_1|}{|B_1|}.$$

In short, the probability of landing inside $A \cap B_1$ is given by the proportion of the volume it makes up out of B_1. We say $Z \sim \text{Uniform}(B_1)$.

How about generation? Let us start with 2 dimensions.

10.4.1 Generating points uniformly at random from a circle

Lets say that we want to generate a uniformly at random variable on the unit circle. One suggestion would be to generate two coordinates X and Y i.i.d. from Uniform$(-1, 1)$ and then projecting (X, Y) onto the unit circle. However, this results in the picture below, when we plot the distribution of angles.

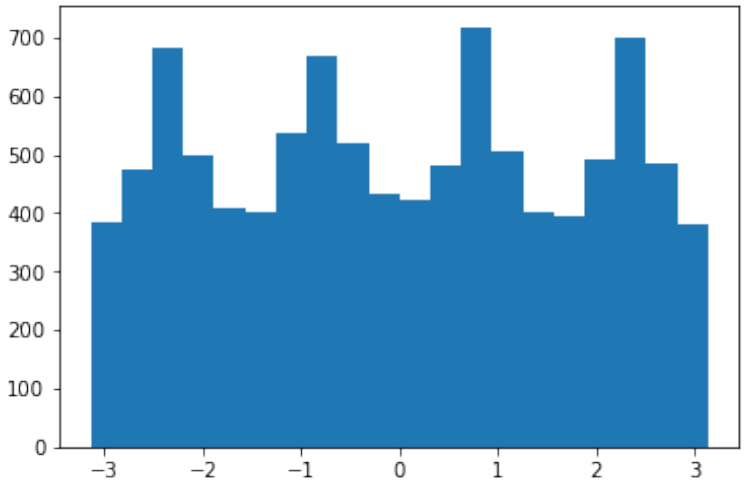
Lets take a look, denote the projection π, i.e.

$$\pi(x, y) = (X/\sqrt{X^2 + Y^2}, Y/\sqrt{X^2 + Y^2}),$$

then the density of the angle

$$f_{\angle \pi(X,Y)}(\theta) = c_0 \int_0^\infty p_{(X,Y)}(t\cos(\theta), t\sin(\theta)) dt \tag{10.2}$$

for some constant c_0. Since $p_{(X,Y)}(x, y)$ is constant (uniform distribution) the above basically measures the length of the line starting from the origin $(0, 0)$ and stretches out in direction $(\cos(\theta), \sin(\theta))$ and reaches the edge of the unit square $[-1, 1]^2$. This is why we see four peaks in the plot below, one for each corner.



Remark 10.12. *From (10.2) we see that if $p_{(X,Y)}$ is rotationally symmetric, i.e. $p_{(X,Y)}(t\cos(\theta), t\sin(\theta)) = p(t)$ for some p then the distribution over angles become uniform.*

This warrants the following definition

Definition 10.13. *We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is* **rotationally symmetric** *if there exists a function $g : \mathbb{R} \to \mathbb{R}$ such that*

$$f(x) = g(|x|)$$

for all $x \in \mathbb{R}^d$.

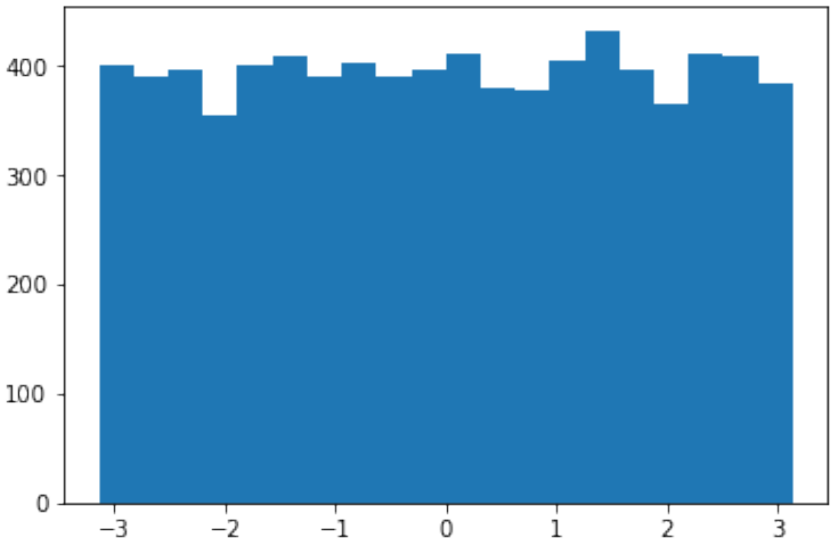
From Remark 10.12 we see that if we could sample from the unit disk, then the projection trick will produce samples uniform on the unit circle. How do we sample from the unit disk? We can use the concept of rejection sampling Algorithm 1, i.e. our sampling density is Uniform$([-1,1]^d)$ and our target density is $\frac{1}{|B_1|}\mathbb{1}_{B_1}(x)$.

Exercise 10.14. *The rejection sampling suggested above, is that equivalent to sampling from* Uniform$([-1,1]^d)$ *and accepting only those samples that lie inside the unit disk?*

[281]:
```
XY = np.random.uniform(-1,1,size=(10000,2))
XY_inCircle = XY[np.linalg.norm(XY,axis=1) < 1]
```

```
XY_inCircle = XY_inCircle / np.linalg.
 ↪norm(XY_inCircle,axis=1).reshape(-1,1)

import pylab
_=pylab.hist(np.arctan2(XY_inCircle[:,1],XY_inCircle[:
 ↪,0]),bins=20)
```



We know from Remark 10.12 that this is uniform on the unit circle, however is it a reasonable approach in higher dimension? We already showed that the volume of the unit ball decreases rapidly with dimension while the volume of the cube is 2^d, so the probability of being inside the unit ball is decreasing very rapidly,

Exercise 10.15. *What happens with the rejection sampling algorithm above when d is large?*

10.4.2 Uniform at random on the unit sphere in high dimension

What was the problem that we had, well basically if we sample from the unit square that distribution is not rotationally symmetric, thus if we sample from a rotationally symmetric distribution then it does not matter, we can just scale any sample to be on the unit circle. A prime example of a rotationally symmetric random variable is the multidimensional Gaussian.

Lemma 10.16. *Let Z be a d dimensional spherical Gaussian (see Model 10.3) then the density for f is rotationally symmetric (see Definition 10.13), as such the density for $\pi(Z)$ is uniform on the unit sphere S_1, where π is*

$$\pi(z) = \frac{z}{|z|}.$$

Exercise 10.17. *Prove the above lemma using radial coordinates.*

10.4.3 Uniform at random from the unit ball B_1?

In the above we learned how to generate uniform at random from the unit sphere. How can we use this to fill out the entire ball? Perhaps we think that if we take $r \sim \text{Uniform}([0,1])$ and $\theta \sim \text{Uniform}(S_1)$, do we then get $r\theta \sim \text{Uniform}(B_1)$? We know that $r\theta \mid r$ is uniform on S_r but $r\theta$ is not uniform, it turns out that we need to scale r as seen in the following theorem.

Theorem 10.18. *Let the dimension $d > 1$ be fixed, let $r \sim \text{Uniform}([0,1])$ and let $\theta \sim \text{Uniform}(S_1)$. Then*

$$r^{\frac{1}{d}}\theta \sim \text{Uniform}(B_1).$$

Proof. Assume that $X \sim \text{Uniform}(B_1)$, represent X in polar coordinates, i.e. $r_X\theta_X$, where $r_X \in [0,1]$ and $\theta_X \in S_1$. We know that given r_X the distribution for θ_X is uniform on the unit sphere. Secondly we know that r_X and θ_X are independent. The goal is to compute the density of r_X:

$$F_{r_X}(r) = \mathbb{P}(r_X \leq r) = \mathbb{P}(r_X\theta_X \in B_r) = \frac{|B_r|}{|B_1|} = \frac{r^d|B_1|}{|B_1|} = r^d,$$

where we just used the definition of the uniform distribution Model 10.11 and Lemma 10.7. If we now wish to sample from F_{r_X} we can use the inversion sampling technique (Theorem 5.38) and note that if $r \sim \text{Uniform}([0,1])$, then $F_{r_X}^{-1}(r) \sim F_{r_X}$. This proves our theorem. □

Remark 10.19. *The scaling of $r^{1/d}$ where $r \sim \text{Uniform}([0,1])$ tells us that we are more likely to get points with radius close to 1, than we are getting points with radius close to 0. This points to the fact that the most interesting things happen close to the unit sphere.*

10.5 High dimensional annulus theorem

The interesting thing is that in high dimension, random variables tend to concentrate on a spherical shell. Remember that for a d-dimensional spherical Gaussian X with standard deviation 1 in each dimension satisfies

$$\mathbb{E}\left[|X|^2\right] = d.$$

Thus one could expect that $|X|$ concentrates around $\sqrt{d}$, as hinted to in Lemma 10.5. The next theorem makes this sharper.

Theorem 10.20. *For a d-dimensional RV X with mean 0, with each component, sub-Gaussian with parameter 1 and variance a^2, then for any $\beta \leq \sqrt{d}$ we have*

$$\mathbb{P}\left(\sqrt{d}|a| - \beta \leq |X| \leq \sqrt{d}|a| + \beta\right) < 2e^{-\frac{\beta^2}{128}}.$$

Proof. First note that $|X|^2 = \sum_{i=1}^{d} |X_i|^2$ is the sum of independent random variables, and since X_i is sub-Gaussian with parameter 1, $|X_i|^2$ is sub-exponential with parameter 8, Lemma 3.15. Hence a simple application of Theorem 3.14 tells us that

$$\mathbb{P}\left(\frac{1}{d}|X|^2 - \frac{1}{d}\mathbb{E}\left[|X|^2\right] > \epsilon\right) < e^{-\frac{\epsilon^2 d}{128}} \wedge e^{-\frac{(\epsilon+1)d}{16}}$$

however

$$\mathbb{P}(|X|^2 > a^2 d + d\epsilon) > \mathbb{P}(|X| > |a|\sqrt{d} + \sqrt{d\epsilon})$$

so, denoting $\beta = \sqrt{d\epsilon}$ and since

$$\frac{(\epsilon+1)d}{16} = \frac{(\beta^2/d + 1)d}{16} = \frac{\beta^2 + d}{16} > \frac{\beta^2}{16}$$

and

$$\frac{\epsilon^2}{128} = \frac{\beta^4/d}{128} < \frac{\beta^4/\beta^2}{128} = \frac{\beta^2}{128} < \frac{\beta^2}{16}$$

hence we get

$$P(|X| > |a|\sqrt{d} + \beta) < e^{-\frac{\beta^2}{128}}.$$

The other side of the inequality is obtained in a similar way, and, together with the union bound gives the result. □

10.6 Bibliography

This section is losely built on [BlHo].

Chapter 11

Dimensionality reduction

11.1 Random Projection and Johnson – Lindenstrauss Lemma

We saw in the previous chapter that there is a concentration effect happening in high dimension, namely that length of vectors with sub-Gaussian components tend to concentrate on an annuli. This can be leveraged as an algorithm, i.e. the random projection algorithm. This works because i.i.d. vectors with i.i.d. components are essentially orthogonal, so choosing k random vectors i.i.d. we should expect to get k decently close basis vectors of the space. Here we are relying on the independence to get orthogonality but we dont normalize length, if we did that we would loose the almost orthogonality, instead we rely on the high dimension to give us vectors that have a certain length with high probability.

Theorem 11.1 (**Random Projection**). *Let v be a fixed vector in $\mathbb{R}^d$ of length 1, fix $\epsilon \in (0,1)$ and let $U_1, \dots, U_k \in \mathbb{R}^d$ be i.i.d., mean 0, being sub-Gaussian with parameter 1 in each component and having variance a^2. Consider the projection onto $(U_1, \dots, U_k)$*

$$f(v) = (U_1 \cdot v, \dots, U_k \cdot v) : \mathbb{R}^d \to \mathbb{R}^k,$$

then

$$\mathbb{P}\left(\left||f(v)| - \sqrt{k}|a||v|\right| \geq \epsilon\sqrt{k}|a||v|\right) \leq 2e^{-\frac{k\epsilon^2}{128}}.$$

Proof. Let us first assume that $|v| = 1$. Now, since each component of U_i is sub-Gaussian with parameter 1, and the components are independent, we get

$$\mathbb{E}\left[e^{sU_i \cdot v}\right] = \prod_{j=1}^{d} \mathbb{E}\left[e^{s(U_i)_j v_j}\right] \leq \prod_{j=1}^{d} e^{s^2 (v_j)^2/2} = e^{s^2 \sum_{j=1}^{d} v_j^2/2} = e^{s^2/2}.$$

That is, $f(v)$ satisfies the prerequisites for Theorem 10.20 and we get for $\beta \leq \sqrt{k}$

$$\mathbb{P}\left(\sqrt{k}|a| - \beta \leq |f(v)| \leq \sqrt{k}|a| + \beta\right) < 2e^{-\frac{\beta^2}{128}}.$$

Setting $\epsilon = \beta/\sqrt{k} \in (0,1)$ and scaling back the length of v we get the statement of the theorem. □

Perhaps not overly exciting as we only have a single data-point, however we can use the union bound to extend this theorem to multiple points

Theorem 11.2 (**Johnson Lindenstrauss**). *For any $0 < \epsilon < 1$ and any integer n, let $k > \frac{384\ln(n)}{\epsilon^2}$. For any set of n points $\{v_1, \ldots, v_n\} \in \mathbb{R}^d$ then the random projection defined in Theorem 11.1 satisfies*

$$\mathbb{P}\left((1-\epsilon)\sqrt{k}|v_i - v_j| \leq f(v_i - v_j) \leq (1+\epsilon)\sqrt{k}|v_i - v_j|\right) \geq 1 - \frac{3}{2n}$$

Proof. Since the random projection f is linear we could for each pair $v_i - v_j$ apply Theorem 11.1 and get for $a = 1$

$$\mathbb{P}\left(\left||f(v_i - v_j)| - \sqrt{k}|v_i - v_j|\right| \geq \epsilon\sqrt{k}|v_i - v_j|\right) \leq 2e^{-\frac{k\epsilon^2}{128}}.$$

There are $\binom{n}{2} < n^2/2$ pairs, so by the union bound we get

$$\mathbb{P}\left(\exists i, j : \left||f(v_i - v_j)| - \sqrt{k}|v_i - v_j|\right| \geq \epsilon\sqrt{k}|v_i - v_j|\right) \leq n^2 e^{-\frac{k\epsilon^2}{128}}.$$

Now, if we choose k such that

$$n^2 e^{-\frac{k\epsilon^2}{128}} = 1/n$$

this becomes

$$k = \frac{384\ln(n)}{\epsilon^2}.$$

□

Remark 11.3. *Note that this usually requires k to be quite large, however we are proving the probability that all distances are preserved. It is usually better if we can allow more error, see Fig. 11.1. The reason for this is that the data itself is IID and as such we can think of Theorem 11.1 as providing a p for a Bernoulli trial, but this is of course not rigorous.*

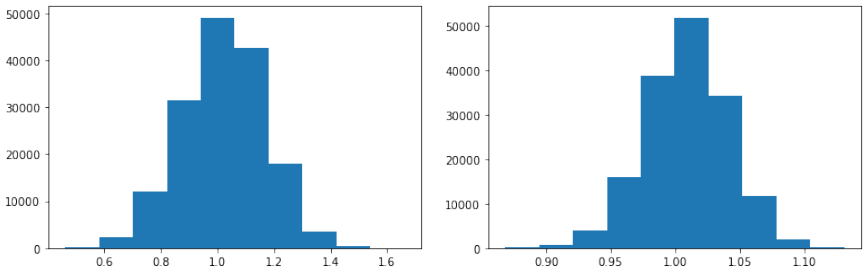


Figure 11.1: The distribution of relative error on the Olivetti faces dataset using only $k = 20$ and $k = 400$ respectively.

11.2 SVD (Singular Value Decomposition)

Something that works "better" in medium high dimension (whatever that means) is **SVD** or **Singular Value Decomposition**.

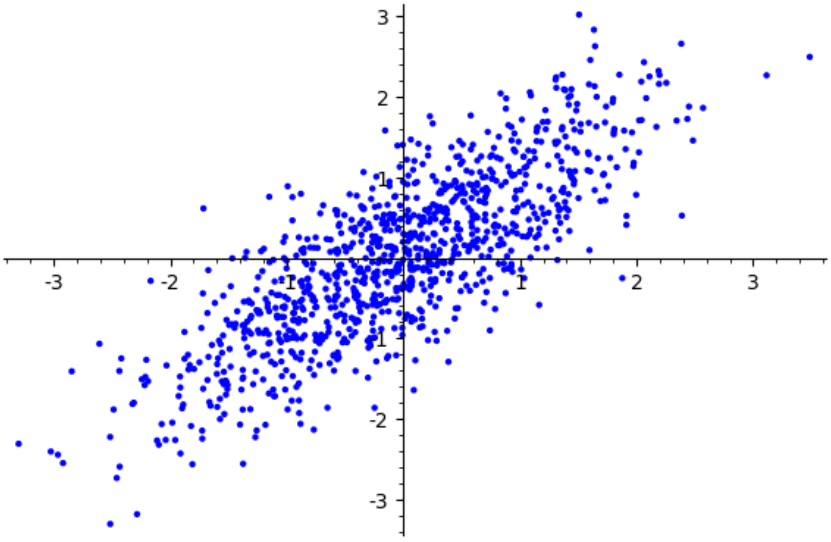


Figure 11.2: Sample data for SVD

Lets say that we wish to represent the data using a low-dimensional subspace, think of a low-dimensional plane. In the case of 2d there is only 1d planes (lines), but if you have, say 100 dimensions we could consider the best fitting 10 dimensional plane. What we mean with best fitting is that the distance from the point to its projection onto our subspace is as small as possible. Think of our 2d example above, then we would like to find the line such that orthogonal projection gives the smallest error. Just looking at the plot we would take the line $y = x$.

But how do we formulate this rigorously? Well we will solve another problem, and later see that it is the same

Remark 11.4. *Consider a line given by the unit vector v, and consider a point x then the projection of x onto v is as above given by*

$$(v \cdot x)v$$

We will now use these ideas applied to IID samples of points $\{X_1, \ldots, X_n\} \in \mathbb{R}^m$ with zero empirical mean (we have centered them). Let $v \in \mathbb{R}^m$ be a unit vector. Consider the projection of each X_i onto v but only consider the proportion i.e. $X_i \cdot v$, then define

$$Y_i = (X_i \cdot v)$$

The line with maximal empirical variance can be written as ($\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i = \frac{1}{n}\sum_{i=1}^n X \cdot v = 0$ since we assumed zero empirical mean)

$$\begin{aligned} v_1 &:= \arg\max_{\|v\|=1} \frac{1}{n}\sum_i (Y_i - \overline{Y}_n)^2 \\ &= \arg\max_{\|v\|=1} \sum_{j=1}^n |X_i \cdot v|^2. \end{aligned}$$

If we construct a matrix A of size $n \times m$ with rows X_i then we can rewrite

$$\sum_{j=1}^n |X_i \cdot v|^2 = |Av|^2$$

and our problem reduces to the linear algebra problem of given an $n \times m$ matrix A to find the direction that is most "expanded/least contracted" by A, in the following sense

$$\arg\max_{\|w\|=1} |Av|.$$

Remark 11.5. *Note, the singular vectors are not necessarily unique, in fact if v is a singular vector, then so is $-v$. We can also have ties, in that case we arbitrarily pick one. We assume that the singular vectors can be picked uniquely, for instance by requiring no ties and that we fix the sign as to make the vector unique.*

Definition 11.6. *The vector $v_1 \in \mathbb{R}^m$ of the $(n \times m)$ matrix A, defined as*

$$v_1 := \arg\max_{\|v\|=1} |Av|$$

is called the **first singular vector** *of A. The value $\sigma_1(A)$ defined as*

$$\sigma_1(A) := |Av_1|$$

is called the **first singular value** *of A.*

Now that we have defined the first singular vector, we can define the second singular vector. This is simply a vector that is orthogonal to v_1 again solving our maximum problem, i.e.

$$v_2 := \arg \max_{\|v\|=1, v \perp v_1} |Av|.$$

We can interpret this as follows, consider the plane given by the first singular vector v_1 as the normal, then we can consider our new problem by finding the vector v that maximizes $|(P_1 A)v|$ where

$$PA = \begin{bmatrix} P_1 X_1 \\ P_1 X_2 \\ \vdots \\ P_1 X_n \end{bmatrix}, \quad P_1 x = x - (x \cdot v_1) v_1.$$

where P_1 is the projection of a vector onto the plane $v_1 \cdot x = 0$. See Fig. 11.3 for the result of the projection.

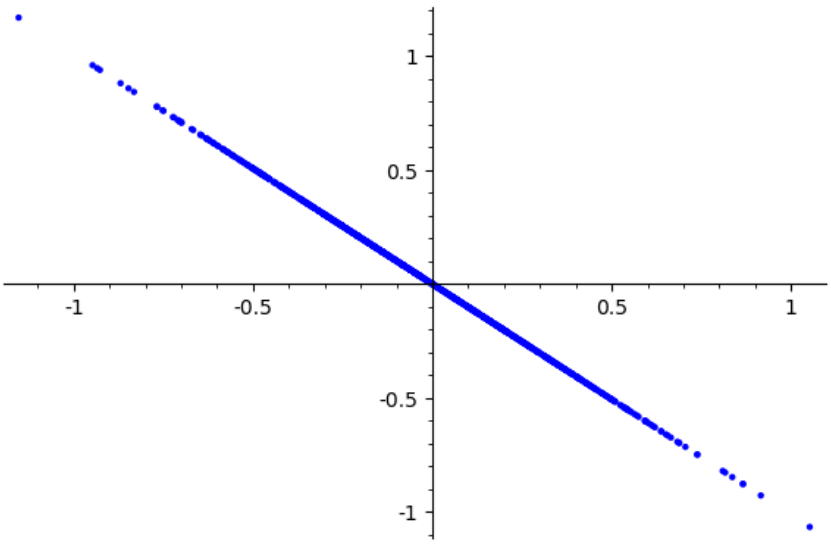


Figure 11.3: The data from Fig. 11.2 projected onto the normal of the plane defined by v_1.

This can be extended, all the way until we have m vectors. That is, there are m singular vectors. To connect this to something which you have already seen in linear algebra, note that

$$\arg \max_{\|v\|=1} |Av| = \arg \max_{\|v\|=1} |Av|^2 = \arg \max_{\|v\|=1} \langle Av, Av \rangle = \arg \max_{\|v\|=1} \langle A^T Av, v \rangle$$

If we let $(v_1, \ldots, v_m)$ be the eigenvectors (all orthogonal) and $\lambda_1, \ldots, \lambda_m$ be the eigenvalues of $A^T A$ (all positive and ordered decreasingly) then we can write

$$v = \sum_{i=1}^{m} a_i v_i$$

which allows us to write

$$\begin{aligned}\langle A^T A v, v\rangle &= \left\langle A^T A \left(\sum_{i=1}^{m} a_i v_i\right), \sum_{i=1}^{m} a_i v_i \right\rangle \\ &= \sum_{i=1}^{m} \left\langle \lambda_i a_i v_i, \sum_{i=1}^{m} a_i v_i \right\rangle = \sum_{i=1}^{m} \lambda_i a_i^2\end{aligned}$$

now, since $1 = \|v\| = \sqrt{a_1^2 + \ldots + a_m^2}$, the above is maximized if $a_1 = 1$ and all other $a_i = 0$, since λ_1 is the largest eigenvalue. We have now proved

Lemma 11.7. *Let A be an $(n \times m)$ matrix, and let v_1 be the first singular vector of A and let $\sigma_1(A)$ be the first singular value (with $\sigma_2(A) < \sigma_1(A)$), then v_1 is an eigenvector of the $m \times m$ matrix $A^T A$, and*

$$\max_{\|v\|=1} |Av| = |Av_1| = \sqrt{\lambda_1} = \sigma_1(A)$$

where λ_1 is the first eigenvalue of $A^T A$.

Remark 11.8. *The problem which can occur in the above is that for multiple eigenvalues we have to choose one of them and identify that with the singular vector, but this is just up to a permutation of indices. Secondly, the above works for any singular value. I.e. every singular vector is an eigenvector and every singular value is the square root of an eigen-value.*

Remark 11.9. *In the context where A is constructed from our IID vectors $X_1, \ldots, X_n$, we see that σ_1 is the standard deviation in the direction of the first singular vector. Furthermore, the matrix $A^T A$ will then be the empirical covariance matrix. I.e. we are looking at the eigenvectors and eigenvalues of the empirical covariance matrix.*

Theorem 11.10 (**Greedy Algorithm**). *Let A be an $n \times d$ matrix with singular vectors $v_1, ..., v_r$. For $1 \leq k \leq r$, let V_k be the subspace spanned by $v_1, \ldots, v_k$. For each k, V_k is the best fit k-dimensional subspace for A.*

Here, V_k is defined as

$$V_k = \{\alpha_1 v_1 + \ldots + \alpha_k v_k : (\alpha_1, \ldots, \alpha_k) \in \mathbb{R}^k\} =: span(\{v_1, \ldots, v_k\}).$$

What do we mean by best fit? Let $\tilde{V}_k$ be another k-dimensional subspace consider the distance of a point p to the k-dimensional subspace $\tilde{V}_k$, such a space is spanned by an orthonormal basis $\tilde{v}_1, \ldots, \tilde{v}_k$, the distance from p to $\tilde{V}_k$ can be seen to be

$$\|p - proj_{\tilde{V}_k} p\| = \|p - \sum_{i=1}^{k} (\tilde{v}_i \cdot p)\tilde{v}_i\|$$

We mean that V_k is the k-dimensional subspace that minimizes

$$\sum_{i=1}^{n} \|X_i - proj_{\tilde{V}_k} X_i\|^2$$

But we can use the Pythagorean theorem to get

$$\sum_{i=1}^{n} \left(\|proj_{\tilde{V}_k} X_i\|^2 + \|X_i - proj_{\tilde{V}_k} X_i\|^2 \right) = \sum_{i=1}^{n} \|X_i\|^2$$

and thus we can get

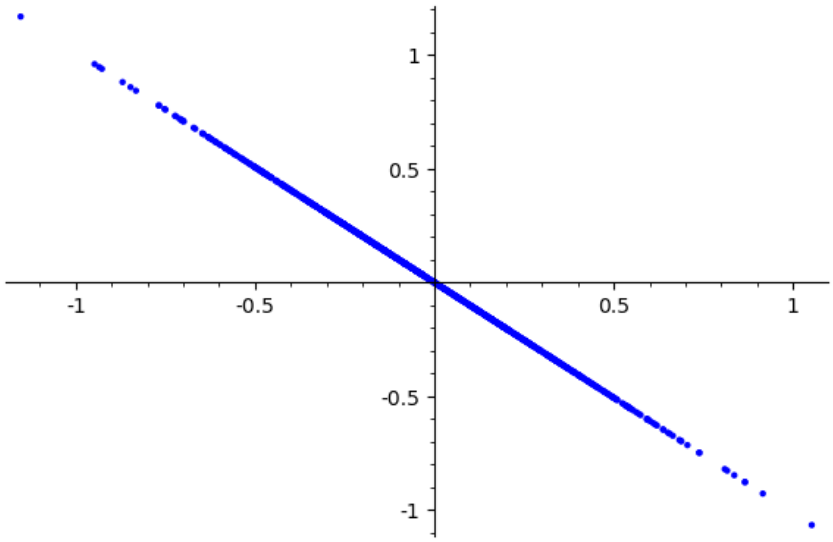
$$\sum_{i=1}^{n} \left(\|X_i\|^2 - \|proj_{\tilde{V}_k} X_i\|^2 \right) = \sum_{i=1}^{n} \|X_i - proj_{\tilde{V}_k} X_i\|^2$$

From the above we see that the best fitting subspace is the subspace that maximizes the "variance" in the sense that we have seen. The point I am making is that we can rephrase the theorem as saying that finding $v_1, \ldots, v_k$ in a greedy way by maximizing the variance is the same as directly minimizing the variance of the deviation from the subspace. This thus answers our question in the beginning of the section.

If we run the greedy algorithm we get the following on the data plotted above.

```
[-0.71191709 -0.70226352] 43.587923587503624
```

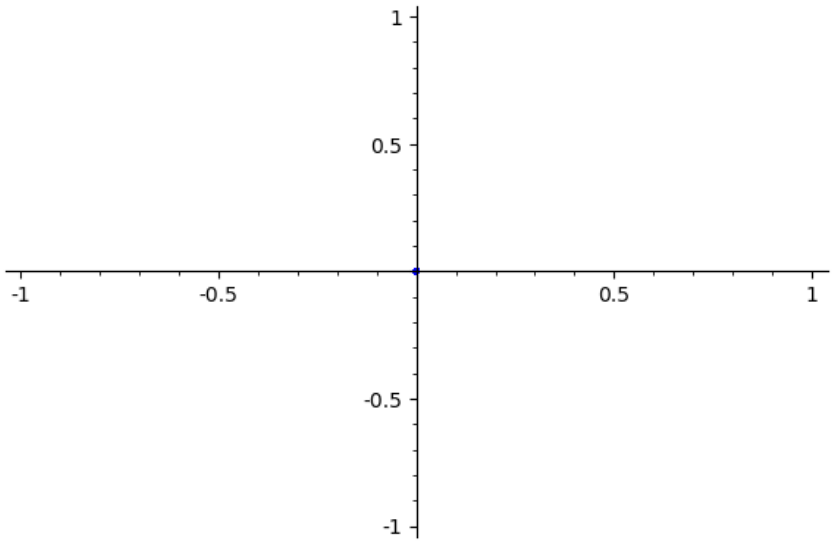
As we can see this is pretty much identical to the vector in the 45 degree direction. To find the second singular vector, we simply project our data onto the plane spanned by having v_1 as the normal.



We then get the following

```
[ 0.70226351 -0.7119171 ] 14.365618434588386
```

Its clear which direction this is headed. Let us also look at what happens when we project the data onto the plane with normal v_2.



Exercise 11.11. *What have we done? Two projections in a row? What is the projection of a projection?*

It should be clear from the definition and Theorem 11.10 that $proj_{V_m} A = A$. That is, if we use all possible singular vectors, then we can represent the data from A as points in V_m. That is any row of A can be written as a linear combination of all the singular vectors.

Singular Value Decomposition of a Matrix

Remember that we said that if we compute m singular vectors of the $n \times m$ dimensional matrix A, then

$$proj_{V_m} A = A$$

this implies that we can write each row in A as $X_i = \sum_{j=1}^{m} (X_i \cdot v_j) v_j$ which we can now rewrite as

$$A = \sum_{j=1}^{m} A v_j v_j^T$$

denoting $u_i := \frac{Av_i}{\sigma_i}$ we see that the above expression becomes

$$A = \sum_{j=1}^{m} \sigma_j u_j v_j^T \tag{11.1}$$

This is the singular value decomposition. I.e. we have decomposed A into a sum of matrices, that is $u_j v_j^T$ is $n \times m$ matrices.

Rewriting the above equation in matrix format we get

$$A = UDV^T$$

where U is the matrix with $u_1, \ldots$ as the columns, D is a diagonal matrix with σ_i as the diagonal and V is the matrix with v_i as columns of V.

Definition 11.12. *The vectors u_i are called the* **left singular vectors**.

11.2.1 The power method

Another way to prove Lemma 11.7 is to consider the matrix $A^T A$ using our decomposition above to get

$$A^T A = (UDV^T)^T (UDV^T) = (VDU^T UDV^T) = VD^2 V^T$$

since $U^T U = I$ which comes from the fact that the columns are orthonormal.

Exercise 11.13. *Prove that the left singular vectors are orthonormal.*

This means that for any column v_i in V

$$A^T A v_i = V D^2 V^T v_i = \sigma_i^2 v_i$$

so we see that v_i is the i:th eigenvector of $A^T A$ with eigenvalue σ_i^2. We can thus find the singular vectors by trying to find the eigenvectors of $A^T A$.

How do we find the eigenvectors of $B = A^T A$? Well first note that

$$B^k = (V D^2 V^T)^k = (V D^{2k} V^T)$$

by the same argument as above, i.e. $V^T V = I$. Thus we see that if $\sigma_1 > \sigma_2$ then if we let k be large enough then

$$B^k \approx (\sigma_1)^{2k} v_1 v_1^T.$$

11.3 PCA

What is PCA, well basically it is a coordinate transformation from the original coordinates to the coordinate system given by the singular vectors. Since V is orthonormal it is as simple as a product, i.e.

$$A = UDV^T$$

Recall that each row in A is a data point i.e. an m dimensional vector and that V is an orthonormal basis, as such we project each point in A onto each basis vector from V by using dot products, as in $(X_i \cdot v_i)v_i$, the coordinate in the basis is just $X_i \cdot v_i$, and as such we get

$$PCA(A) = AV = UDV^T V = UD$$

Remark 11.14. *Warning: In the beginning of this section we assumed that our data had empirical mean zero. Thus in order to use this we first have to center the data.*

11.4 SVD in Action

This is all cool and such, but what can you do with it?

Singular value decomposition can be used in the following ways

11.4.1 Factor Analysis

- Studying underlying factors. The famous `g factor`: proposed by Spearman (Spearman correlation), to describe "general intelligence" as a singular vector based on data about IQ, Math ability and other cognitive tests. This is also called Factor analysis.
- Compressing a representation of data, as a dimensional reduction technique. This is similar to the rank k approximation idea.

11.4.2 Example on compressing data

Lets consider the Mnist dataset, which is handwritten digits from 0 to 9. These are represented as 8×8 pixel images and will be put together as a single array of length 64. As such we have points in $\mathbb{R}^d$ with $d = 64$. The number of points is 1797. If we assemble all these images into a matrix as before, where each row is a datapoint (image) we get a matrix A of shape 1797×64. Recall from (11.1) we have that A is the sum of m matrices of shape $n \times m$, we will now sum this from 1 to 10 instead and consider

$$A_k := \sum_{i=1}^{k} \sigma_j u_j v_j^T$$

for $k = 10$. That is, we are using 10 singular vectors to represent the digits, in Fig. 11.4 you can see 10 uncompressed sample images and in Fig. 11.5 you can see the same 10 samples but compressed. What do we mean, we mean that if X_i is an image, it will be row i of A, the compressed image will be row i of A_{10}.

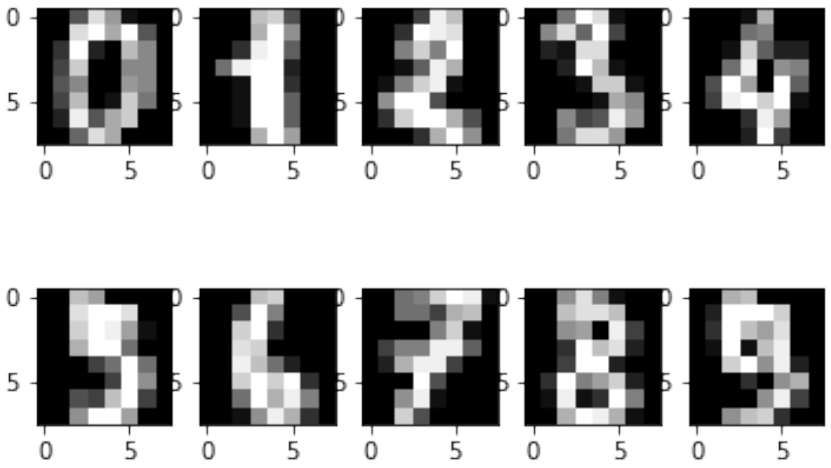


Figure 11.4: 10 sample images from Mnist

```
Number of data_points: 1797, number of features: 64,␣
↪Number of components: 10
```

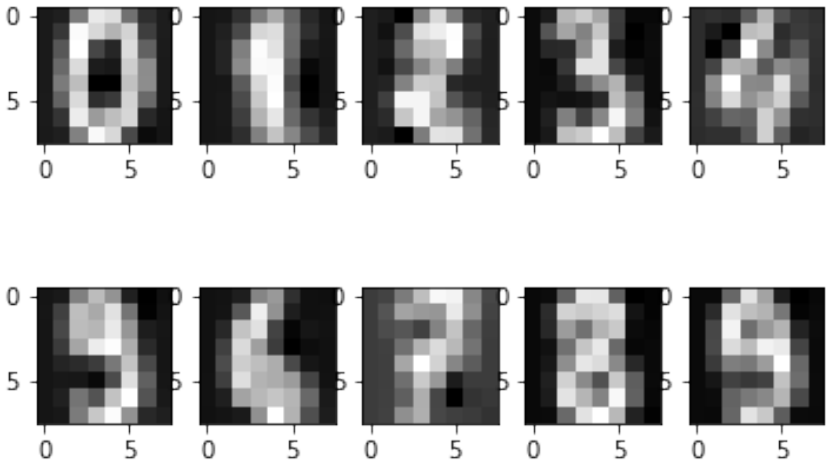


Figure 11.5: The data from Fig. 11.4 projected onto the plane defined by the first 10 singular vectors.

What we can see is that even with only 10 components we were able to fairly well represent the digits, although it is clear that some are not so easy.

Reconstruction error

The reconstruction error is defined as the error we make in the compression, i.e. the square distance between the real image and the target image,

$$\text{Reconst} := \sqrt{\sum_{i=1}^{n} \|(A)_i - (A_k)_i\|^2}$$

For those who know linear algebra this is nothing else than the Frobenious norm of $A - A_k$. Using the decomposition (11.1) we get that

$$A - A_k = \sum_{j=k+1}^{m} \sigma_j u_j v_j^T$$

and the norm of this is simply $\sqrt{\sum_{j=k+1}^{m} \sigma_j^2}$. As such the sum of squares of the remaining singular values are giving us the reconstruction error.

Explained variance

Explained variance is how much percentage of the total variance is captured by our singular vectors. Remember the interpretation of the singular values as the standard deviation, as such the variance explained of the first k components is just the sum of the singular values squared and divided by the total variance.

11.4.3 Anomaly detection and reconstruction error

The approach taken in Section 11.4.2 can be used for a rudimentary form of anomaly detection, which incidentally works really well.

The point is here is that we compress data into the matrix A_k, we estimate the distribution function for $\|(A)_i - (A_k)_i\|$ using the samples and then use this to select quantiles that we will use for detection of an anomaly.

11.5 Theoretical analysis

The PCA components are eigenvector of the empirical covariance matrix. Namely, let $Z = (X_1, \dots, X_d) \sim F_Z$ and consider an i.i.d. sequence of $Z_1, \dots, Z_n$. Covariance matrix is

$$\mathbb{E}\left[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^T\right]$$

assuming that Z has mean zero, lets consider

$$\mathbb{E}[ZZ^T] = (\mathbb{E}[X_i X_j])_{i,j}$$

there are $d^2/2$ such values. Now the empirical covariance matrix is that we use the empirical mean to estimates each component of the matrix, i.e.

$$\hat{\Sigma}_{i,j} = \frac{1}{n}\sum_{k=1}^{n}(Z_k)_i(Z_k)_j$$

if now each component of Z_k is sub-Gaussian then we can use concentration to get something like

$$\mathbb{P}(|\hat{\Sigma}_{i,j} - (\mathbb{E}[X_i X_j])_{i,j}| > \epsilon) < e^{-c\epsilon n}$$

using the union bound we can thus get

$$\mathbb{P}(\max_{i,j}|\hat{\Sigma}_{i,j} - (\mathbb{E}[X_i X_j])_{i,j}| > \epsilon) < \frac{d^2}{2}e^{-c\epsilon n}$$

The d^2 in the estimate is however quite suboptimal and there is an improvement over the above, which follows from the so-called **Matrix Bernstein inequality**.

Theorem 11.15. *Let $X_1, \dots, X_n$ be centred i.i.d, random vectors in $\mathbb{R}^d$. Suppose that for all i, $\mathrm{Var}(X_i) = \Sigma$ and $\mathbb{P}(\|X_i\|_2 \leq \sqrt{C}) = 1$ for some C. Then for all $\epsilon > 0$*

$$\mathbb{P}\left[\left\|\hat{\Sigma}_n - \Sigma\right\| > \epsilon\right] \leq 2de^{-\frac{n\epsilon^2}{2C(C+2\epsilon/3)}}$$

Remark 11.16. *The notation $\|\Sigma\|$ for matrices, denotes the operator norm.*

This theorem tells us that with high probability the estimated covariance matrix will be close to the true covariance Σ if we have many observations. However, the PCA method relied on computing the eigen-values and eigen-vectors of $\hat{\Sigma}_n$. Closeness in the matrix norm allows us to say something about the closeness of the eigen-values

Theorem 11.17 (Weyls theorem)**.** *Let $\hat{\Sigma} = \Sigma + E$, where Σ and E are symmetric matrices. Let λ_i and $\hat{\lambda}_i$ be the i:th eigen-values of Σ and $\hat{\Sigma}$ respectively. Then*

$$\max_{i=1,\dots,d} |\hat{\lambda}_i - \lambda_i| \leq \|E\|.$$

This is all well and good, but estimating the eigen-vectors is a difficult problem.

Example 11.18. *Consider $\Sigma = \begin{bmatrix} 1, & 0 \\ 0, & 1 \end{bmatrix}$ and $E = \begin{bmatrix} 0, & \epsilon \\ \epsilon, & 0 \end{bmatrix}$. The eigenvalues of Σ are 1 and 1. The eigen-vectors of Σ are $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$. The eigen-values of $\hat{\Sigma} := \Sigma + E$ is $1 + \epsilon$ and $1 - \epsilon$. However, for any ϵ, the eigenvectors of $\hat{\Sigma}$ are $\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.*

The problem in the above example is the closeness of the eigen-values for $\hat{\Sigma}$, since Σ has a double eigen-value. This poses problems as it is an unstable problem and we have no hope. What we can say though, is that if the eigen-values are only simple, then we can expect stability. We will not cover that in this course, but if you want to dig deeper, check-out [WW].

11.6 Reconstruction error

Introduce the class

$$\mathcal{P}_k = \{\Pi : \mathbb{R}^d \to \mathbb{R}^d \mid \Pi \text{ is an orthogonal projection of rank } k\}.$$

Consider a $\mathbb{R}^d$ valued random variable $X \in L^2(\mathbb{P})$. Define the loss function $L(X, \Pi(X)) = \|X - \Pi(X)\|_2^2$, for $X \in \mathbb{R}^d$, then define the reconstruction error of the projection operator Π as

$$\mathcal{R}(\Pi) = \mathbb{E}\left[L(Z, \Pi(Z))\right].$$

The minimizer of the risk Π_k^* is defined as

$$\Pi_k^* = \arg\min_{\Pi \in \mathcal{P}_k} \mathcal{R}(\Pi).$$

As we have seen above with singular value decomposition etc. we have that Π_k^* is the projection onto the first k eigen-vectors of the covariance matrix Σ.

The empirical minimization problem is

$$\hat{\Pi}_k^* = \arg\min_{\Pi \in \mathcal{P}_k} \frac{1}{n} \sum_{i=1}^{n} L(X_i, \Pi(X_i)) = \arg\min_{\Pi \in \mathcal{P}_k} \frac{1}{n} \sum_{i=1}^{n} \|X_i - \Pi(X_i)\|_2^2.$$

The excess risk is defined as

$$\mathcal{E}_k := \mathcal{R}(\hat{\Pi}_k^*) - \mathcal{R}(\Pi_k^*)$$

as in Chapter 8 the goal is to bound the excess risk with high probability.

We have the following estimate

Lemma 11.19. *In the setting above, if we define $\hat{\Sigma}$ the empirical covariance matrix, then*

$$\mathcal{E}_k \leq \sqrt{2k} \|\Sigma - \hat{\Sigma}\|_2$$

Proof. See [ReWa, Proposition 2.2]. □

Thus, we have from Theorem 11.15

Theorem 11.20. *Let $X_1, \ldots, X_n$ be centred i.i.d, random vectors in $\mathbb{R}^d$. Suppose that for all i, $\mathrm{Var}(X_i) = \Sigma$ and $\mathbb{P}(\|X_i\|_2 \leq \sqrt{C}) = 1$ for some C. Then for all $\epsilon > 0$*

$$\mathbb{P}\left[\mathcal{E}_k > \epsilon\right] \leq \mathbb{P}\left[\left\|\hat{\Sigma}_n - \Sigma\right\| > \epsilon/\sqrt{2k}\right] \leq 2de^{-\frac{n\epsilon^2}{4C(C+2\epsilon/3)k}}$$

11.7 Bibliography

The first part concerning SVD is losely built on [BlHo]. For the Bernstein inequality Theorem 11.15 see [WW, Corollary 6.20]. If you want to dig deeper into reconstruction errors, see [ReWa].

Chapter 12

Group Assignments

12.1 Group Assignment 1

- Prove Lemma 1.14
- Prove Lemma 2.8
- Prove property 4 of Theorem 2.18
- Solve Exercise 2.59
- Prove the "tower property" (Theorem 2.60) for a discrete random variable.

12.2 Group Assignment 2

- Prove Corollary 3.7.
- Prove Lemma 3.15, properties 1-4.
- Solve Exercise 3.16
- Solve Exercise 4.7
- Prove Theorem 4.9 with all details, basically referring to all the properties of the indicator function used, the monotonicity of measures etc.

12.3 Group Assignment 3

- Solve Exercise 5.20
- Solve Exercise 6.11

- Solve Exercise 6.18
- Solve Exercise 7.12
- Solve Exercise 7.16

Bibliography

[BlHo] Blum, A., Hopcroft, J., and Kannan, R. *Foundations of data science.* Cambridge University Press, 2020.

[BLM] Boucheron, S., Lugosi, G., Massart, P. *Concentration inequalities: a nonasymptotic theory of independence*, Oxford University Press, 2013.

[PTPR] Devroye, L., Gyorfi, L. and Lugosi, G., *A Probabilistic Theory of Pattern Recognition*, Springer 1997

[HD] T. E. Hull and A. R. Dobell, *Random Number Generators*, SIAM Review 1962 4:3, 230-254

[K] Knuth, Donald (1997). Seminumerical Algorithms. The Art of Computer Programming. 2 (3rd ed.). Reading, MA: Addison-Wesley Professional. pp. 10–26.

[ReWa] Reiß, M., and Wahl, M. *Nonasymptotic upper bounds for the reconstruction error of PCA.* The Annals of Statistics 48.2 (2020): 1098-1123.

[SLT] Vapnik, V. , *Statistical Learning Theory*, Wiley, 1998.

[WW] M. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, 2019.

[W] Wasserman, L. All of Statistics. Springer Texts in Statistics, Springer, 2004.

Index