Lecture Notes for Data and Uncertainty (first part)

Jochen Bröcker

December 29, 2020

1 Introduction

Sections 1 to 10 will cover probability theory, integration, and a bit of statistics (roughly in that order). The second part (Sect. 11-13) –**not part of these notes**– will explain an important technique in statistics called Monte Carlo simulations. This introduction will motivate probability theory and statistics a little bit, and in particular why we need concepts from measure theory and integration, which is often perceived as abstract and complicated.

Probability

It is not easy to explain what probability theory is about without sounding tautological. One might say that it allows to quantify uncertainty or chance, but then what do we mean by "uncertainty" or "chance"?

De Moivre's seminal textbook "The Doctrine of Chances" [dM67] is widely considered as the first textbook on probability theory, and the theory has undergone enormous developments since then. In particular, there is an axiomatic framework which has been universally adopted and which we will discuss in this course. But even though De Moivre's book was first published in 1718, there is still some debate as to the interpretation of probability theory, or in other words, to what this nice axiomatic framework actually pertains. Different interpretations of probability have been put forward, but somewhat fortunately to the student, the differences matter little as far as the mathematics is concerned. Nonetheless, we will briefly mention the most prominent interpretations of probability (see the Wikipedia page on "Probability Interpretations" for more information).

The classical definition of probability put forward by Laplace ([Lap95]), "consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence." These cases equally possible might be termed *elementary events*, and the probability of any event A is the number of elementary events it contains, divided by the number of all elementary events. Evidently, this definition assumes that for any given problem, (i) the number of elementary events is finite, and (ii) every event can be expressed as the union of elementary events. We will see below that this theory is not powerful enough to deal with certain questions we are interested in.

Frequentism Frequentism considers experiments which can, at least in principle, be repeated as often as desired under constant external conditions. Internal phenomena however may lead to variable outcomes under repetition. Whether these phenomena are fundamentally deterministic (like throwing a coin) or inherently random (like radioactive decay) is irrelevant. The probability of an outcome is then defined to be the limiting observed frequency of this outcome. But the probability is only well defined if the observed frequencies actually converge, and it is very difficult to provide useful criteria as to when this is the case (other than just saying that they have to converge). We see that this theory struggles to explain exactly what experiments it applies to.

Subjectivism Subjectivism maintains that probabilities express the degree of belief a certain individual has as to whether a certain statement about the real world is true or not. For example, I think it is likely, but not certain, that the Riemann hypothesis is correct. It is not necessary that the individual actually uses probabilities; as Savage [Sav71] and others have shown, these probabilities can be inferred or "elicited" from the individual's behaviour. In other words, any possible action an individual may take in the face of an uncertain event can be explained by means of a number, interpreted as that individual's subjective probability of the event. The proof of this fact however assumes that the individual acts rationally in a sense specified by Savage's axioms, and these axioms are not altogether unobjectionable. In fact, there is strong evidence that people's behaviour under uncertainty can deviate considerably from these axioms.

Just for illustrative purposes, here is a little example showing how the elicitation of (subjective) probabilities might work in practice. You can play this game with a friend of yours (if you have a few pounds to spare).

Example 1.1. The statement is: "At the time of his death, Isaac Newton had lost all but one of his teeth." This statement is known to be false; Apart from his primary teeth, Newton lost only one tooth during his lifetime, but you don't reveal this to your friend at this point. You merely say that you know the answer. You offer your friend to pay her p pounds if the statement is correct, q pounds if it is not correct, and *she* can choose the numbers p and q, *but* she will have to pay $\frac{1}{2}(p^2+q^2)$ pounds back to you in order to play the game. Once your friend has chosen p and q, you reveal the answer and exchange the money. You can convince yourself of the following facts, which support the interpretation of p and q as your friend's subjective probabilities for and against the statement:

1. There is an incentive for your friend to play the game, as there is a way to make at least 25 pence for sure.

2. There is an incentive for your friend to state two numbers p, q which are nonnegative and so that $p+q=1$, any deviation from this will incur a certain loss.

Why the classical interpretation of probability is insufficient

For the remainder of the introduction, we will discuss why the classical interpretation of probability, which has a finite number of events with equal probability as a basis, is insufficient. In fact, any theory with a finite number of events is insufficient, whatever their probability. We will see that measure theoretic probability is the right answer to this problem, so the following is really a motivation to study measure theory and integration.

Consider the following model of a *fair coin*: A *binary sequence* is a sequence $x_1, x_2, \ldots$ where $x_k = 0$ or 1 only for all k. We identify 1 with "Head" and 0 with "Tail". Let $\Omega :=$ the set of all binary sequences. A generic element is written as $\omega = (\omega_1, \omega 2, \ldots)$. An *event* is a subset of Ω. A probability $\mathbf{P}$ is a function which assigns a number between 0 and 1 to events so that

1. $\mathbf{P}(\Omega) = 1$,

2. $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$ whenever the events A and B are disjoint.

These assumptions are reasonable for any probability. In the present example, we assume

1. $\mathbf{P}(\{\omega \in \Omega; \omega_k = 1\}) = p$ for any k (and this implies $\mathbf{P}(\{\omega \in \Omega; \omega_k = 0\}) = 1-p$),

2. Different throws are independent.

The second condition means the following: Whenever $x_1, \ldots, x_n$ is a binary sequence of finite length, then

$$\mathbf{P}(\{\omega \in \Omega; \omega_1 = x_1, \ldots, \omega_n = x_n\}) = p^{\sum_{k=1}^n x_k}(1-p)^{n-\sum_{k=1}^n x_k}.$$

Based on these assumptions, we can now work out things like

$$\mathbf{P}(\text{No. of "Heads" in 10 throws}) = p^{10}$$

etc. In other words, we can deal with events that depend on finitely many throws, only. But now let $S_n = \sum_{k=1}^n \omega_k$ = No. of "Heads" in n throws, and consider the statement

$$\frac{S_n}{n} \to p \qquad \text{as } n \to \infty. \tag{1.1}$$

This statement, known as the *Law of Large Numbers*, "should hold", but cannot be analysed in a "finite" framework, since the very existence of the limit is a random event, depending on more than just finitely many throws (in fact, neither the value of the limit nor its existence depends on the first n throws, however large n is).

To see where the problem lies, let us delve into this a little further. We assume $p = \frac{1}{2}$ for simplicity. Fix some $\epsilon > 0$ and consider the events $A_k := \{\text{all } \omega \text{ so that } |\frac{S_k}{k} - \frac{1}{2}| > \epsilon\}$. You might know from your previous studies that

$$\mathbf{P}(A_k) \leq \frac{C}{k}$$

for some constant C depending on ϵ. In fact, with a little more work you can show that

$$\mathbf{P}(A_k) \leq C \cdot \lambda^k \tag{1.2}$$

for some $C > 0$ and $\lambda < 1$, both depending on ϵ. Either estimate shows that $\frac{S_k}{k}$ will concentrate about $1/2$ with increasing probability. But this does not imply that the statement (1.1) holds for any ω. In fact, if ω is so that the statement (1.1) holds, then ω can be a member of only finitely many A_k! So we need to investigate the event $B := \{\text{all } \omega \text{ which are in infinitely many } A_k\}$. If we can show that this event has probability zero no matter how we pick ϵ, then the statement (1.1) is true for all ω, except perhaps for some ω's in a set of probability zero.

The most challenging bit in this discussion is the construction of the event B. Consider first $B_n := A_n \cup A_{n+1} \cup \ldots$. This is the event that $\{|\frac{S_k}{k} - \frac{1}{2}| > \epsilon\}$

for *some* $k \geq n$. Now if there is an ω which is a member of A_k for infinitely many k's, then it must be in *all* B_n's; hence $B \subset B_n$ for any n, and therefore

$$\mathbf{P}(B) \leq \mathbf{P}(B_n) \qquad \text{for all } n = 1, 2, \ldots.$$

We aim to show that the probability of B_n goes to zero. Now by Equation (1.2)

$$\mathbf{P}(A_n \cup A_{n+1} \cup \ldots \cup A_{n+m}) \leq \sum_{k=n}^{n+m} \lambda^k = \frac{\lambda^n - \lambda^{n+m+1}}{1-\lambda} < \frac{\lambda^n}{1-\lambda}.$$

The left hand side is increasing in m and bounded (by one or in fact by the right hand side) and therefore convergent, hence

$$\lim_{m\to\infty} \mathbf{P}(A_n \cup A_{n+1} \cup \ldots \cup A_{n+m}) \leq \frac{\lambda^n}{1-\lambda}. \tag{1.3}$$

And here is the point: We would like to use that the left hand side is in fact $\mathbf{P}(B_n)$, that is

$$\lim_{m\to\infty} \mathbf{P}(A_n \cup A_{n+1} \cup \ldots \cup A_{n+m}) \stackrel{?}{=} \mathbf{P}(A_n \cup A_{n+1} \cup \ldots) = \mathbf{P}(B_n). \tag{1.4}$$

Assuming this is correct for the moment and using it in Equation (1.3), we obtain

$$\mathbf{P}(B_n) \leq \frac{\lambda^n}{1-\lambda} \to 0 \qquad \text{if } n \to \infty,$$

and we can conclude that the statement (1.1) holds for all sequence of heads and tails which are not in B, but this happens probability zero.

So what is the problem with Equation (1.4)? We clearly have

$$\mathbf{P}(A_n \cup A_{n+1} \cup \ldots \cup A_{n+m}) \leq \mathbf{P}(B_n).$$

so all we can say is that

$$\lim_{m\to\infty} \mathbf{P}(A_n \cup A_{n+1} \cup \ldots \cup A_{n+m}) \leq \mathbf{P}(B_n),$$

but there is no apriori reason why there should be equality here. If we want relations like (1.4) to be correct, we have to add this as an assumption. In other words, we will only work with probabilities where this is correct. But there is then another question: Is this consistent with our assumptions (a) and (b) made at the beginning of this section, and the other properties we would like a probability to have? This is a nontrivial question which we will address in this chapter (the answer is "yes").

Statistics

Statistics works with real data, that is, quantitative observations from real world experiments. The aim is to explain the data by means of models, mostly probabilistic models. So when compared to probability theory, in a sense we go in the opposite direction. Rather than investigating a given probability model (e.g. a fair coin) we ask the question: is there a probability model consistent with given data? For example, suppose we observe the two data sets

$$\{H, H, T, H, T, H, T, T, H\} \tag{1.5}$$
$$\{H, H, H, H, T, T, T, T, H\} \tag{1.6}$$

alledgedly coming from tossing a fair coin. Is this data consistent with the model of a fair coin? You might observe that both data sets contain 5 heads and 4 tails. In that sense, the data sets are not entirely atypical for nine tosses of a fair coin. But you might also observe that the H's and T's are sprinkled quite evenly across the first data set while in the second, there seem to be unusually long runs of heads and tails. So the first data set might plausibly come from a fair coin but not the second. Statistics is about making these conclusions more quantitative.

2 Sigma algebras and probability measures

In this section, we discuss probabilities and events, that is the sets we want to assign probabilities to. We start with some fundamental definitions. Let Ω, A, B be sets. Familiarity with the notations $A \subset \Omega$, $A \cup B$, $A \cap B$, $\varnothing$ is assumed. Further

$$A \setminus B := \{x \in A; x \notin B\}, \qquad \text{read "}A \text{ without } B\text{"}$$
$$A^{\mathsf{c}} := \Omega \setminus A, \qquad \text{read "Complement of } A \text{ in } \Omega\text{"}.$$

The notation A^{c} is used if Ω is clear from the context. If the elements of a set A are again sets, we call A a system or family of sets.

Definition 2.1. Let Ω be a set. A system $\mathcal{A}$ of subsets of Ω is called an *algebra* if

1. $\varnothing \in \mathcal{A}$
2. $A \in \mathcal{A} \Rightarrow A^{\mathsf{c}} \in \mathcal{A}$.
3. $A_1, \ldots, A_n \in \mathcal{A} \Rightarrow \bigcup_{k=1}^{n} A_k \in \mathcal{A}$.

Further, $\mathcal{A}$ is a *sigma algebra* if

4. $A_1, A_2, \cdots \in \mathcal{A} \Rightarrow \bigcup_{k=1}^{\infty} A_k \in \mathcal{A}$.

An algebra formalises the intuition behind "events". Considering sigma algebras rather than just algebras, that is where 3 holds for countably many A_n rather than just finitely many, is important as we have seen in the introduction. Members of $\mathcal{A}$ are called *events* or *measurable sets*.

Definition 2.2. Let $\mathcal{A}$ be an algebra. A function $\mathbf{P} : \mathcal{A} \longrightarrow [0,1]$ is a probability if it satisfies

1. *Normalisation:* $\mathbf{P}(\Omega) = 1$
2. *Additivity:* If $A_1, \ldots, A_n \in \mathcal{A}$, with $A_i \cap A_j = \varnothing$ for $i \neq j$, then $\sum_{k=1}^{n} \mathbf{P}(A_k) = \mathbf{P}(\bigcup_{k=1}^{n} A_k)$.
3. *Continuity at $\varnothing$:* If $A_1, A_2, \ldots \in \mathcal{A}$, with $A_1 \supset A_2 \supset \ldots$ and $\cap A_j = \varnothing$, then $\mathbf{P}(A_k) \to 0$ for $k \to \infty$.

Again, the intuition is clear. The continuity at $\varnothing$ is important for technical reasons, as we have seen in the introduction (the connection will be made clear in Exercise 2.2). It is possible to construct examples of a probability on an algebra that is not continuous at $\varnothing$. Note that a probability satisfies $\mathbf{P}(\varnothing) = 0$ (Exercise 2.2).

Definition 2.3. A pair $(\Omega, \mathcal{A})$ with Ω a set and $\mathcal{A}$ a sigma algebra is called a *measurable space*. A triple $(\Omega, \mathcal{A}, \mathbf{P})$ with Ω a set, $\mathcal{A}$ a sigma algebra, and $\mathbf{P}$ a probability is called a *probability space*.

Note that algebras are very much smaller than sigma algebras, so it should be much easier to define $\mathbf{P}$ just on an algebra (examples later).

Definition 2.4. Let $\mathcal{A}$ be an arbitrary family of subsets of Ω. Then $\sigma(\mathcal{A})$ is defined as the smallest σ-algebra containing $\mathcal{A}$.

In Exercise 2.1 you will show that this concept is well defined.

Theorem 2.5 (The Measure Extension Theorem, also known as MET or Hahn-Carathéodory theorem)**.** *Let $\mathcal{A}$ be an algebra and $\mathbf{P}$ a probability on $\mathcal{A}$. Then there exists a unique probability $\tilde{\mathbf{P}}$ on $\sigma(\mathcal{A})$ with $\tilde{\mathbf{P}}|_{\mathcal{A}} = \mathbf{P}|_{\mathcal{A}}$. Further, if $A \in \sigma(\mathcal{A})$, then for any $\epsilon > 0$ there exist disjoint sets $A_1, \ldots, A_n \in \mathcal{A}$ with $\tilde{\mathbf{P}}(A \triangle \bigcup_{k=1}^{n} A_k) \leq \epsilon$.*

Sketch of a proof, see e.g. [Hal74]. For any $Y \subset \Omega$, put $\mathbf{P}^*(Y) = \inf \sum_{k=1}^{\infty} \mathbf{P}(A_k)$, inf taken over $A_1, A_2, \cdots \in \mathcal{A}$, with $Y \subset \bigcup_k A_k$. Now

1. $\mathbf{P}^*|_{\mathcal{A}} = \mathbf{P}|_{\mathcal{A}}$ ("$\leq$" is trivial).

2. Consider the family of sets $\mathcal{M}$: a set $A \subset \Omega$ is a member of $\mathcal{M}$ if $\forall E \subset \Omega$ it holds that $\mathbf{P}^*(E) \geq \mathbf{P}^*(E \cap A) + \mathbf{P}^*(E \setminus A)$. One then proves that $\mathcal{M}$ is a $\sigma-$algebra with $\mathcal{M} \supset \mathcal{A}$.

3. $\mathbf{P}^*$ is a measure on $\mathcal{M}$.

4. The approximation result is relatively straightforward from the definition of $\mathbf{P}^*$.

□

We fix the uniqueness part, which is true under weaker conditions:

Theorem 2.6 (Uniqueness of probabilities)**.** *Let $\mathcal{A}$ be a family of sets so that for any two sets $A_1 \in \mathcal{A}$, $A_2 \in \mathcal{A}$, also $A_1 \cap A_2 \in \mathcal{A}$. (This is true for instance if $\mathcal{A}$ is an algebra.) Further, let $\mathbf{P}, \mathbf{Q}$ be two probabilities on $\sigma(\mathcal{A})$, the sigma algebra generated by $\mathcal{A}$. Then if $\mathbf{P}(A) = \mathbf{Q}(A)$ for any set $A \in \mathcal{A}$, they agree on $\sigma(\mathcal{A})$.*

For a proof see [Bre73], Proposition 2.23. The following theorem ensures that there exists a probability on the unit interval which on any subinterval is given by the length of that subinterval. For a proof, see for instance [JP00], Chapter 7.

Theorem 2.7 (The Lebesgue measure)**.** *A halfopen interval on $[0, 1]$ is a set of the form $[a, b[$, where $0 \leq a < b \leq 1$. Let $\mathcal{A}$ be the family of sets which are unions of finitely many disjoint halfopen intervals. Then $\mathcal{A}$ is an algebra. To each $A \in \mathcal{A}$ we assign $\lambda(A) :=$ the total length of A. This is a probability on $\mathcal{A}$ (the continuity at $\varnothing$ requires proof, see for instance [JP00] for a somewhat more general statement). It now follows from Theorem 2.5 that λ can be extended to a probability on $\sigma(\mathcal{A})$, which is the Borel algebra (see Definition 3.1).*

Exercises for Section 2

Exercise 2.1. Let Ω be a set.

1. Show that the power set 2^Ω is a sigma algebra.

2. Show that $\mathcal{S}_1 \cap \mathcal{S}_2$ is a sigma algebra for any two sigma algebras $\mathcal{S}_1, \mathcal{S}_2$.

3. Use the previous two items to show that $\sigma(\mathcal{A})$ in Definition 2.4 makes sense, i.e. there exist sigma algebras containing $\mathcal{A}$, and among these there exists a smallest possible one.

Exercise 2.2. Let Ω be a set, $\mathcal{A}$ an algebra, $\mathbf{P} : \mathcal{A} \to [0,1]$ a set function satisfying properties 1 and 2 in Definition 2.2.

1. Show that $\mathbf{P}(\varnothing) = 0$.
2. Show that property 3 in Definition 2.2 is equivalent to *sigma additivity*: If $A_1, A_2, \ldots$ is a sequence of sets in $\mathcal{A}$ with $A_i \cap A_j = \varnothing$ for any $i \neq j$, and if $\bigcup_k A_k \in \mathcal{A}$ as well, then $\sum_{k=1}^{\infty} \mathbf{P}(A_k) = \mathbf{P}(\bigcup_k A_k)$.
3. Show that property 3 in Definition 2.2 is equivalent to *continuity from above*: If $A_1, A_2, \ldots \in \mathcal{A}$, with $A_1 \supset A_2 \supset \ldots$ and $\cap A_j = A$ with $A \in \mathcal{A}$, then $\mathbf{P}(A_k) \to \mathbf{P}(A)$ for $k \to \infty$.
4. Show that property 3 in Definition 2.2 is equivalent to *continuity from below*: If $A_1, A_2, \ldots \in \mathcal{A}$, with $A_1 \subset A_2 \subset \ldots$ and $\cup A_j = A$ with $A \in \mathcal{A}$, then $\mathbf{P}(A_k) \to \mathbf{P}(A)$ for $k \to \infty$.
5. Show that for any series $A_1, A_2, \ldots$ of disjoint sets in $\mathcal{A}$, we have $\mathbf{P}(A_n) \to 0$ (in fact, $\mathbf{P}(A_n)$ must be summable).

3 Measurable Functions and Integration

A probability can be seen as a generalised form of volume. As with the standard volume in Euclidean space, it is possible to integrate functions against probabilities. We want to define an integral which, to some extent, can be interchanged with pointwise limits of functions. Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space.

Definition 3.1.
1. On $\mathbb{R}$ we define the Borel-algebra $\mathcal{B}$ as the smallest $\sigma-$algebra containing all open sets (see 2.4).
2. A function $f : \Omega \longrightarrow \mathbb{R}$ is *measurable* or a *random variable* if $f^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}$.

The definition of a random variable guarantees that sets such as $\{\omega \in \Omega; a < f(\omega) < b\} = f^{-1}(]a, b[)$ can be assigned a probability to. To prove that a function is measurable, it is enough to check that $\{\omega; f(\omega) > a\} \in \mathcal{A}$ for any $a \in \mathbb{R}$ (see Exercise 3.2).

Theorem 3.2.
1. If f_n, $n \in \mathbb{N}$ are random variables, so are the pointwise $\limsup f_n$, $\liminf f_n$, $\lim f_n$ *(if the last exists).*

2. If $f^{(k)}$, $k = 1, \ldots, d$ are random variables and $\phi : \mathbb{R}^d \to \mathbb{R}$ is a continuous function, then the function $\psi : \omega \to \phi(f^{(1)}(\omega), \ldots, f^{(d)}(\omega))$ is a random variable.

Proof. To prove item 1, pick $a \in \mathbb{R}$. Then $\{\omega; \sup_k f_{n+k}(\omega) > a\} = \bigcup_k \{\omega; f_{n+k}(\omega) > a\} \in \mathcal{A}$, so $\sup_k f_{n+k}$ is measurable for every n by the remark after Definition 3.1. $\inf_k f_{n+k}$ is similar (take $\bigcap_k \{\dots\}$). But

$$\liminf_n f_n = \sup_n \inf_k f_{n+k},$$
$$\limsup_n f_n = \inf_n \sup_k f_{n+k}.$$

So they are measurable. If $\lim_n f_n$ exists, it is equal to lim sup and lim inf. To prove the second item, we note that the statement is true if $f^{(1)}, \dots, f^{(d)}$ are simple functions. Further, we will show later on that every nonnegative random variable is the pointwise limit of simple functions, and this is easily seen to extend to general (not necessarily nonnegative) random variables. We can conclude that ψ is the pointwise limit of simple functions and thus a random variable by item 1. □

The integral

We want to define an integral $\int f \mathrm{d}\mathbf{P}$ for random variables, which we will also write as $\mathbb{E}(f)$, generalising the expectation value.

But first a remark about limits and increasing sequences. A sequence $\{x_n, n \in \mathbb{N}\}$ of real numbers is called *increasing* if $x_1 \leq x_2 \leq \dots$. If $\{x_n\}$ is increasing, then $x_n \to x$ and $x_n \uparrow x$ have the same meaning, namely that $x = \sup_n x_n$. Note that x might be infinite, but if it is not, we have $x = \lim_{n\to\infty} x_n$. (We stress that per definition, the limit of a sequence is *always* finite.) If $\{x_n\}$ is not increasing though, then $x_n \uparrow x$ is meaningless, while $x_n \to x$ means that $x = \lim_n x_n$.

For a sequence $\{f_n, n \in \mathbb{N}\}$ of real valued functions on some set Ω, the limits $\lim_n f_n = f$ and $f_n \to f$ are understood pointwise (unless otherwise stated), that is $\lim_n \{f_n(\omega)\} = f(\omega)$ and also $f_n(\omega) \to f(\omega)$ for every $\omega \in \Omega$. The sequence $\{f_n\}$ is called *increasing* if $\{f_n(\omega), n \in \mathbb{N}\}$ is an increasing sequence for every $\omega \in \Omega$, and we write $f_n \uparrow f$ if $f_n(\omega) \uparrow f(\omega)$ for every $\omega \in \Omega$.

The integral of a random variable can be constructed along the following steps. See [Kle14, Hal74, Doo94, Dud89] for details.

1. For $A \in \mathcal{A}$, define the *indicator function*

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{else.} \end{cases}$$

2. A random variable $f : \Omega \to \mathbb{R}$ is *simple* if it assumes finitely many values, say $\{f_1, \ldots, f_n\} \subset \mathbb{R}$. We can write

$$f = \sum_{l=1}^{k} f_l \cdot \mathbf{1}_{B_l}$$

with $B_l = f^{-1}(\{f_l\})$ for all $l = 1, \ldots k$. Note that $B_l \in \mathcal{A}$ for all $l = 1, \ldots k$ because f is assumed measurable.

3. With f, g simple, so are $f \cdot g$, $\alpha f + \beta g$, $\alpha, \beta \in \mathbb{R}$, $\max\{f, g\}$ and $|f|$ (these operations are understood pointwise).

4. Every nonnegative random variable $f : \Omega \to \mathbb{R}_{\geq 0}$ is the pointwise monotone increasing limit of simple functions.

Proof. Define $g_n : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$

$$g_n(x) = \begin{cases} k + \frac{l}{2^n} & \text{if } k + \frac{l}{2^n} < x \leq k + \frac{l+1}{2^n} \\ & \text{for } k = 1 \ldots n-1,\ l = 0 \ldots 2^n - 1 \\ n & \text{if } x > n. \end{cases}$$

Clearly $g_n(x) \uparrow x$, $\forall\ x \in \mathbb{R}_{\geq 0}$. Now put $f_n := g_n \circ f$, then clearly f_n is simple and $f_n \uparrow f$. □

5. For f simple, define

$$\int f \, \mathrm{d}\mathbf{P} = \sum_{k=1}^{n} f_k \mathbf{P}(B_k).$$

6. Prove that the integral is linear, monotone (i.e. $f \leq g \Rightarrow \int f \, \mathrm{d}\mathbf{P} \leq \int g \, \mathrm{d}\mathbf{P}$) and $|\int f \, \mathrm{d}\mathbf{P}| \leq \int |f| \, \mathrm{d}\mathbf{P}$.

7. If f is a nonnegative random variable and $\{f_n\}$ is a sequence of simple functions and $f_n \uparrow f$ (e.g. as in setp 4), then $\{\int f_n \, \mathrm{d}\mathbf{P}\}$ is an increasing sequence of real numbers and

$$\sup_n \int f_n \, \mathrm{d}\mathbf{P} = \sup_g \int g \, \mathrm{d}\mathbf{P}, \tag{3.1}$$

where "$\sup_g$" is over all simple g with $f \geq g$. This will be proved in exercise 3.3. We define $\int f \, \mathrm{d}\mathbf{P}$ as either side of Equation (3.1). This might be a nonnegative real number or ∞. But if $\int f \, \mathrm{d}\mathbf{P} < \infty$, then $\int f_n \, \mathrm{d}\mathbf{P} \to \int f \, \mathrm{d}\mathbf{P}$ for $n \to \infty$.

8. For a general random variable $f : \Omega \to \mathbb{R}$, put $f_+ := \max\{f, 0\}$, $f_- := f_+ - f$ (now f_+, f_- are nonnegative) and set

$$\int f \, \mathrm{d}\mathbf{P} := \int f_+ \, \mathrm{d}\mathbf{P} - \int f_- \, \mathrm{d}\mathbf{P}$$

if at least one them is finite. If both are finite, f is called *integrable*.

We stress that the integral of a nonegative random variable is always well defined (but maybe infinite). In particular $\int |f| \, \mathrm{d}\mathbf{P}$ is always well defined for any random variable f, and f is integrable if and only if $\int |f| \, \mathrm{d}\mathbf{P} < \infty$.

Lemma 3.3 (Properties of the integral)**.** *The integral enjoys the properties in step (6) if both $\int |f| d\mathbf{P} < \infty$ and $\int |g| \mathrm{d}\mathbf{P} < \infty$.*

Proof. The linearity and the monotonicity for integrals over nonnegative simple functions is assumed proved in step 6. The additivity for integrals over nonnegative functions f, g is shown by observing that if $f_n, g_n, n \in \mathbb{N}$ are nonnegative simple functions with $f_n \uparrow f$ and $g_n \uparrow g$, then $f_n + g_n \uparrow f + g$ with $f_n + g_n$ nonnegative and simple. The additivity of the integral in this case then follows from the additivity of the integral for nonnegative simple functions and step 7 above. To show the monotonicity for integrals over nonnegative functions $f \leq g$, we take nonnegative simple functions $f_n, g_n, n \in \mathbb{N}$ with $f_n \uparrow f$ and $g_n \uparrow g$. Now note that $h_n = \max\{f_n, g_n\}$ is also nonnegative and simple with $h_n \uparrow g$, and further $f_n \leq h_n$. It follows from step 7 that $\int f \mathrm{d}\mathbf{P} \leq \int g \mathrm{d}\mathbf{P}$. To prove the additivity in the general case, observe first that $|f + g| \leq |f| + |g|$ and hence $\int |f + g| \mathrm{d}\mathbf{P} < \infty$ by the monotonicity for nonnegative functions. From the identity

$$(f + g)_+ + f_- + g_- = (f + g)_- + f_+ + g_+$$

we obtain by the additivity for nonnegative random variables that

$$\int (f + g)_+ \mathrm{d}\mathbf{P} + \int f_- \mathrm{d}\mathbf{P} + \int g_- \mathrm{d}\mathbf{P} = \int (f + g)_- \mathrm{d}\mathbf{P} + \int f_+ \mathrm{d}\mathbf{P} + \int g_+ \mathrm{d}\mathbf{P}.$$

Note that by integrability, all the terms in this identity are finite. Rearranging and using the definition of the integral for general f and g gives the result. To prove the monotonicity in the general case, we use the linearity

(in the line marked with $(*)$) to obtain

$$\begin{aligned}\int f\mathrm{d}\mathbf{P} &= \int f_+\mathrm{d}\mathbf{P} - \int f_-\mathrm{d}\mathbf{P} \\ &\leq \int f_+\mathrm{d}\mathbf{P} + \int f_-\mathrm{d}\mathbf{P} \\ &= \int (f_+ + f_-)\mathrm{d}\mathbf{P} \qquad (*) \\ &= \int |f|\mathrm{d}\mathbf{P}.\end{aligned}$$

Similarly, one proves that $-\int f\mathrm{d}\mathbf{P} \leq \int |f|\mathrm{d}\mathbf{P}$ which gives the result. □

Interchange of integral with a.s. limits

The most important reason for introducing this integral (as opposed to using the Riemann integral) is the nice behaviour of the integral under pointwise limits.

Theorem 3.4 (Monotone Convergence)**.** *Suppose* $\{f_n, n \in \mathbb{N}\}$ *is an increasing sequence of nonnegative random variable, and* $f_n \uparrow f$*. Then*

$$\int f_n \, \mathrm{d}\mathbf{P} \uparrow \int f \, \mathrm{d}\mathbf{P}. \tag{3.2}$$

Proof. According to step 4, for every $n \in \mathbb{N}$ there exists a sequence $\{f_{n,m}, m \in \mathbb{N}\}$ of simple nonegative random variable with $\lim_{m\to\infty} f_{n,m} = f_n$. Let $g_n = \max\{f_{k,l}, k, l \leq n\}$. This is a increasing sequence of simple functions. On the one hand,

$$g_n \leq f_n \leq f \qquad \text{for all } n. \tag{3.3}$$

On the other hand if we fix $\epsilon > 0$ and $\omega \in \Omega$ we can find n and $m \geq n$ so that

$$\begin{aligned}f(\omega) &\leq f_n(\omega) + \epsilon/2 \\ f_n(\omega) &\leq f_{n,m}(\omega) + \epsilon/2\end{aligned}$$

and since $m \geq n$ we have

$$f_{n,m} \leq g_m.$$

Taking these three estimates together gives

$$f \leq g_m + \epsilon.$$

This fact together with the estimate (3.3) proves

$$g_n \uparrow f.$$

The result now follows from the definition of the integral in Step 7. □

Note that the right hand side in Equation (3.2) might be infinity. Further, the theorem remains true if the function f assumes the value ∞, but we haven't quite defined the integral for such functions (the extension is not difficult). Also, it actually suffices that $\int f_n d\mathbf{P} \geq 0$ rather than $f_n \geq 0$ for the theorem to hold, see [Dud89].

Theorem 3.5 (Fatou Lemma)**.** *If $\{f_n\}$ is a sequence of nonnegative random variable then*

$$\int \liminf f_n \, \mathrm{d}\mathbf{P} \leq \liminf \int f_n \, \mathrm{d}\mathbf{P}. \tag{3.4}$$

Before proving this, a little example for illustration.

Example 3.6. We will later see that on $\Omega = [0, 1]$ equipped with the Borel algebra (i.e. the sigma algebra generated by all open sets on $[0, 1]$) one can define a probability by the formula $\mathbf{P}(A) = \int_A \mathrm{d}x$. The integral with respect to $\mathbf{P}$ is of course the standard Lebesgue integral on the unit interval (or the Riemann integral if the integrand is continuous). Define

$$f_n(x) = n \cdot \mathbf{1}_{[0,\frac{1}{n}]}(x).$$

Now $\liminf f_n = \lim f_n = 0$, and hence the left hand side of Equation (3.4) is zero. But $\int f_n(x) \, \mathrm{d}x = 1$ and therefore $\liminf \int f_n(x) \, \mathrm{d}x = 1$, hence the right hand side is one. This helps me to remember which direction the inequality goes in Fatou's lemma. Further, the example demonstrates that the integral is in general *not* exchangeable with pointwise limits. Some additional condition (like monotonicity in Theorem 3.4) is necessary. A different but still sufficient condition will be discussed presently.

Proof of Fatou's Lemma. Since

$$\inf_k f_{n+k} \leq f_{n+l} \qquad \text{for all } l \in \mathbb{N},$$

we get by integrating that

$$\int \inf_k f_{n+k} \, \mathrm{d}\mathbf{P} \leq \int f_{n+l} \, \mathrm{d}\mathbf{P} \qquad \text{for all } l \in \mathbb{N},$$

so we take the inf over l and obtain

$$\int \inf_k f_{n+k} \, \mathrm{d}\mathbf{P} \leq \inf_k \int f_{n+k} \, \mathrm{d}\mathbf{P}. \tag{3.5}$$

We now want to take the limit $n \to \infty$ on both sides of this inequality. Note that $\inf_k f_{n+k}$ is a monotone sequence in n of nonegative functions, and hence

$$\lim_n \int \inf_k f_{n+k} \, \mathrm{d}\mathbf{P} = \int \lim_n \inf_k f_{n+k} \, \mathrm{d}\mathbf{P} = \int \liminf_n f_{n+k} \, \mathrm{d}\mathbf{P}$$

by monotone convergence and the definition of lim inf. On the right hand side, taking the limit simply gives $\liminf_n \int f_n \, \mathrm{d}\mathbf{P}$. □

The next theorem shows that the integral can be interchanged with pointwise limits provided the sequence of functions is bounded. The boundedness condition replaces the monotonicity condition in the Monotone Convergence Theorem (note that the sequence in Example 3.6 is neither bounded nor monotone).

Theorem 3.7 (Bounded Convergence)**.** *Let $\{f_n, n \in \mathbb{N}\}$ be a sequence of random variable with $|f_n| \leq C$, and $f_n \to f$ for $n \to \infty$. Then f is integrable and $\int f_n \, \mathrm{d}\mathbf{P} \longrightarrow \int f \, \mathrm{d}\mathbf{P}$.*

A more general version of this theorem goes under the name *Dominated Convergence Theorem*, in which the condition $|f_n| \leq C$ is replaced with $|f_n| \leq g$ where g is an integrable function. The conclusions are the same.

Proof. Clearly $|f| \leq C$ as well so we get $\int |f| \, \mathrm{d}\mathbf{P} \leq C$, proving that f is integrable. Since $f_n + C$, and $f + C$ are nonnegative, we can apply Fatou and get (after subtracting the constant again from both sides)

$$\int f \, \mathrm{d}\mathbf{P} \leq \liminf_n \int f_n \, \mathrm{d}\mathbf{P}.$$

The same can be done with $-f_n$ and $-f$; we get

$$\int -f \, \mathrm{d}\mathbf{P} \leq \liminf_n \int -f_n \, \mathrm{d}\mathbf{P} = -\limsup_n \int f_n \mathrm{d}\mathbf{P},$$

or after multiplying with -1:

$$\int f \, \mathrm{d}\mathbf{P} \geq \limsup_n \int f_n \, \mathrm{d}\mathbf{P}.$$

In summary, we have shown that

$$\liminf_n \int f_n \, \mathrm{d}\mathbf{P} \geq \int f \, \mathrm{d}\mathbf{P} \geq \limsup_n \int f_n \, \mathrm{d}\mathbf{P},$$

completing the proof. □

Definition 3.8 (Equivalence of random variables)**.**

1. Let $f_1, f_2 : \Omega \to \mathbb{R}$ functions (not necessarily measurable). We say

$$f_1 = f_2 \qquad \text{almost surely (a.s.)}$$

or f_1 and f_2 are equivalent if $f_1(\omega) = f_2(\omega)$ for all ω in a measurable set Ω_1 with $\mathbf{P}(\Omega_1) = 1$. (One can check that this is indeed an equivalence relation.)

2. If f is an integrable random variable, we can put

$$\int \hat{f} \, \mathrm{d}\mathbf{P} := \int f \, \mathrm{d}\mathbf{P},$$

for any $\hat{f}$ which is equivalent to f.

3. For integrable random variable f we define the L_1*–norm* by $\|f\|_1 = \int f \, \mathrm{d}\mathbf{P}$.

The L_1–norm is in fact not a norm on functions, only a pseudo–norm: $\|f\|_1 = 0$ does not quite imply $f = 0$. But by Exercise 3.4, $f = 0$ almost surely, and therefore $\|f - g\| = 0$ means that f and g are equivalent. So strictly speaking, $\|.\|_1$ is a norm on equivalence classes of functions.

Definition 3.9 (The space L_1)**.**

1. The space of integrable functions (or strictly speaking, their equivalence classes) with the norm $\|.\|_1$ is denoted as $L_1(\Omega, \mathcal{A}, \mathbf{P})$ or just L_1 if the probability space is clear from the context.

2. If $\{f_n\}$ is a sequence of integrable random variables and f is another random variable so that $\|f_n - f\|_1 \to 0$ as $n \to \infty$, we will say that $\{f_n\}$ converges to f in L_1 or write $f_n \overset{L_1}{\to} f$.

Theorem 3.10 (Completeness of L_1)**.** *Suppose $\{f_n\}$ is a sequence of random variable which is Cauchy with respect to $\|\cdot\|_1$. Then there exists an integrable random variable f with $f_n \to f$ in L_1. Further, if f' is another random variable with this property, then $f = f'$ a.s.*

This result is one of the main drivers behind the development of measure and integration. With regards to Theorem 3.10 and also Definition 3.9,2, it has to be kept in mind that L_1 limits need not be unique; a sequence $\{f_n\}$ of random variables can converge in L_1 against two different functions f and f' at the same time, however, f and f' will be equivalent.

Exercises for Section 3

Exercise 3.1. In this exercise, we fill in some details to Section 3. Let $(\Omega, \mathcal{A})$ be a measurable space (i.e. a set Ω with a sigma algebra $\mathcal{A}$). Consider a function $f : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the Borel algebra.

1. Consider the family $\mathcal{A}_0$ of all sets of the form $f^{-1}(B)$ where $B \in \mathcal{B}$. Show that $\mathcal{A}_0$ is a sigma algebra on Ω. ($\mathcal{A}_0$ is referred to as the sigma algebra *generated* by f.)

2. Consider the family $\mathcal{B}_0$ of all sets $B \subset \mathbb{R}$ so that $f^{-1}(B) \in \mathcal{A}$. Show that $\mathcal{B}_0$ is a sigma algebra on $\mathbb{R}$.

3. Conclude that f is a random variable if $\mathcal{B}_0$ from the previous item contains $\mathcal{B}$.

4. Use the previous item and Exercise 3.2 to prove the remark after Definition 3.1: f is a random variable if $\{\omega \in \Omega; f(\omega) > a\} \in \mathcal{A}$ for any $a \in \mathbb{R}$.

Exercise 3.2. In this exercise[1], we learn more about the Borel algebra $\mathcal{B}$ on $\mathbb{R}$. (Recall that $\mathcal{B}$ is the smallest sigma algebra which contains all open sets.) Show that $\mathcal{B}$ is actually the smallest sigma algebra which contains all sets of the form $]a, \infty]$ for any $a \in \mathbb{R}$. You need to prove that if $\tilde{\mathcal{B}}$ is a sigma algebra containing all sets of the form $]a, \infty]$ for any $a \in \mathbb{R}$, then $\tilde{\mathcal{B}}$ must contain all open sets. Proceed along the following steps:

1. Show that $\tilde{\mathcal{B}}$ contains all left open right closed intervals, i.e. sets of the form $]a, b]$ with $a < b$.

2. Show that $\tilde{\mathcal{B}}$ contains all open intervals (Hint: $]a, b[= \cup_{n=1}^{\infty}]a, b - \frac{1}{n}]$).

3. Show that $\tilde{\mathcal{B}}$ contains countable unions of open intervals.

4. Show that every open set in $\mathbb{R}$ is the union of countably many open intervals (this is difficult, so skip if you want), and conclude that $\tilde{\mathcal{B}}$ contains every open set.

Exercise 3.3. In this exercise, we will prove item (7) in the construction of the integral.

[1] This exercise might require bookwork. Check for example [Dud89]

1. Because the f_n are an increasing sequence of functions, the same is true for the real numbers $\int f_n \mathrm{d}\mathbf{P}$. Therefore $c = \lim_n \int f_n \mathrm{d}\mathbf{P}$ exists. Show that the following statement implies item (7): If g is simple and $g \leq f$, then
$$\int g \mathrm{d}\mathbf{P} \leq c. \tag{3.6}$$
The following steps will establish this statement.

2. Set $\epsilon > 0$ and define the sets $M_n = \{\omega \in \Omega; f_n(\omega) > g(\omega) - \epsilon\}$. Show that these sets are measurable, that $M_1 \subset M_2 \subset \ldots$, and that $\cup_{n=1}^{\infty} M_n = \Omega$.

3. Justify all "$\geq$" signs in the following:
$$\int f_n \mathrm{d}\mathbf{P} \geq \int f_n \cdot \mathbf{1}_{M_n} \mathrm{d}\mathbf{P} \geq \int g \cdot \mathbf{1}_{M_n} \mathrm{d}\mathbf{P} - \epsilon \mathbf{P}(M_n) \tag{3.7}$$

4. Use sigma additivity to establish that $\mathbf{P}(M_n) \to 1$, and that $\int g \cdot \mathbf{1}_{M_n} \mathrm{d}\mathbf{P} \to \int g \mathrm{d}\mathbf{P}$ (remember that g is simple). Using this in Equation (3.7) gives
$$c = \lim_n \int f_n \mathrm{d}\mathbf{P} \geq \int g \mathrm{d}\mathbf{P} - \epsilon$$
for any ϵ, establishing (3.6).

Exercise 3.4. Show that if f is a nonnegative random variable with $\int f \mathrm{d}\mathbf{P} = 0$, then $f = 0$ *almost surely*, that is $f(\omega) = 0$ for all ω in a set Ω_1 with $\mathbf{P}(\Omega_1) = 1$. Hint: Consider the sets $A_n = \{\omega : f(\omega) > 1/n\}$ and show that $n \cdot f \geq \mathbf{1}_{A_n}$ to get an upper bound on $\mathbf{P}(A_n)$. What can you now say about $\cup_{n=1}^{\infty} A_n$?

Exercise 3.5. In this exercise, we will introduce the concept of densities. Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space. Let f be a nonnegative random variable, and suppose that $\int f \mathrm{d}\mathbf{P} = 1$. On $\mathcal{A}$, define the set function F by
$$F(A) = \int \mathbf{1}_A \cdot f \, \mathrm{d}\mathbf{P}.$$

1. Show that F is a probability on $(\Omega, \mathcal{A})$. To prove that F is sigma additive, you need to invoke the Monotone Convergence Theorem.

2. Show that $\mathbf{P}(A) = 0$ implies $F(A) = 0$. (Attention: this is not immediately obvious; assume first that f is simple, then use Monotone Convergence).

We will say that f is a *density* for F. The next item will show that densities are (essentially) unique.

3. Using Exercise 3.4, show that if two densities f and g give rise to the same probability F, then $f = g$ almost everywhere. Hint: let $h = f - g$ and consider h_+, h_-.

4 Transformations

This short chapter is devoted to transformations, the push–forward of probabilities and the transformation formula. The material is important for later parts of this chapter but also for dynamical systems.

Let $(\Omega_k, \mathcal{A}_k)$, $k = 1, 2$ be two measurable spaces. In this context, a mapping $T : \Omega_1 \to \Omega_2$ is defined as *measurable* if $T^{-1}(A) \in \mathcal{A}_1$ for all $A \in \mathcal{A}_2$. Note that random variables as defined 3.1 are just a special case of this, namely with $(\Omega_2, \mathcal{A}_2) = (\mathbb{R}, \mathcal{B})$. Let $\mathbf{P}$ be a measure on $(\Omega_1, \mathcal{A}_1)$. Then the formula $T_*\mathbf{P}(A) := \mathbf{P}(T^{-1}(A))$ for all $A \in \mathcal{A}_2$ defines a probability $T_*\mathbf{P}$ on $(\Omega_2, \mathcal{A}_2)$ called the *pushforward* of $\mathbf{P}$ under T. That the pushforward is a probability will be proved in Exercise 4.1.

Theorem 4.1 (Transformation formula)**.** *If $f : (\Omega_2, \mathcal{A}_2) \to (\mathbb{R}, \mathcal{B})$ random variable, either positive or integrable w.r.t. $T_*\mathbf{P}$, then*

$$\int_{\Omega_2} f \mathrm{d}(T_*\mathbf{P}) = \int_{\Omega_1} f \circ T \mathrm{d}\mathbf{P}.$$

Proof. We prove this for simple functions first. If $f = \sum_{k=1}^n f_k \cdot \mathbf{1}_{A_k}$, we have on the left hand side

$$\int_{\Omega_2} f \ \mathrm{d}(T_*\mathbf{P}) = \sum_{k=1}^n f_k \cdot \mathbf{P}(T^{-1}(A_k)).$$

On the right hand side we obtain

$$\begin{aligned}\int_{\Omega_2} f \circ T \ \mathrm{d}\mathbf{P} &= \sum_{k=1}^n f_k \cdot \int \mathbf{1}_{A_k} \circ T \ \mathrm{d}\mathbf{P} \\ &= \sum_{k=1}^n f_k \cdot \int \mathbf{1}_{T^{-1}(A_k)} \ \mathrm{d}\mathbf{P} = \sum_{k=1}^n f_k \cdot \mathbf{P}(T^{-1}(A_k)),\end{aligned}$$

establishing the transformation formula for simple functions. The rest of the proof is covered in Exercise 4.2 □

Exercises for Section 4

Exercise 4.1. This exercise fills in several details to the beginning of Section 4 in preparation of the transformation formula 4.1. Let $(\Omega_k, \mathcal{A}_k), k = 1, 2$ be measurable spaces, $\mathbf{P}$ is a measure on $(\Omega_1, \mathcal{A}_1)$. Further, $T : (\Omega_1, \mathcal{A}_1) \to (\Omega_2, \mathcal{A}_2)$ is a measurable mapping and $f : (\Omega_2, \mathcal{A}_2) \to (\mathbb{R}, \mathcal{B})$ a random variable.

1. Show that the pushforward $T_*\mathbf{P}$ defined by $T_*\mathbf{P}(A) := \mathbf{P}(T^{-1}(A))$ for all $A \in \mathcal{A}_2$ is a probability on the sigma algebra $\mathcal{A}_2$.

2. Show that $f \circ T : (\Omega_1, \mathcal{A}_1) \to (\mathbb{R}, \mathcal{B})$ is a random variable.

3. If $S : (\Omega_0, \mathcal{A}_0) \to (\Omega_1, \mathcal{A}_1)$ is another measurable mapping, show that $T \circ S : (\Omega_0, \mathcal{A}_0) \to (\Omega_2, \mathcal{A}_2)$ is measurable. (Hint: the previous item is a special case of this statement.)

Exercise 4.2. In this exercise, we actually prove the transformation formla 4.1. The same setup is as in theorem 4.1, and we assume it has been proved for simple functions.

1. Use the Monotone Convergence Theorem and the fact that the pushforward is a probability to prove theorem 4.1 in the case that $f \geq 0$.

2. For integrable f prove theorem 4.1 by considering f_+ and f_- and using the previous item.

5 Products spaces and product measures, Fubini-Theorem

A rectangle in $\mathbb{R}^2$ is the carthesian product of two intervals in $\mathbb{R}$, and the volume of the rectangle is the product of the volumes (i.e. lengths) of these two intervals. Rather than using the standard volume, basically the same can be done with probabilities, resulting in product probabilities. We will also look at the integral of functions against product probabilities and prove that such integrals can be computed as iterated integrals.

Consider a sequence $(\Omega_k, \mathcal{A}_k)$, $k \in \mathbb{N}$ of measurable spaces. We define the *Cartesian Product*

$$\Omega := \prod_{k \in \mathbb{N}} \Omega_k := \text{ sequences } (\omega_1, \omega_2, \dots) \text{ with } \omega_k \in \Omega_k \text{ for all } k \in \mathbb{N}. \quad (5.1)$$

A sigma algebra can be introduced on Ω as follows. A *rectangular cylinder* is a set of the form

$$\{\omega \in \Omega; \omega_k \in A_k, k \in \mathbb{N}\},$$

where $A_k \in \mathcal{A}_k$ for all $k \in \mathbb{N}$, and $A_k \neq \Omega_k$ for only finitely many k. Now let $\mathcal{A} :=$ smallest *sigma* algebra on Ω containing all rectangular cylinders. Notation $\mathcal{A} := \bigotimes_{k\in\mathbb{N}} \mathcal{A}_k$. The measurable space $(\Omega, \mathcal{A})$ is called the measurable product of $(\Omega_k, \mathcal{A}_k)$, $k \in \mathbb{N}$. A carthesian product over finitely many factors $(\Omega_k, \mathcal{A}_k)$, $k = 1 \dots K$ is defined in the same way (the requirement that $\mathcal{A}_k \neq \Omega_k$ for only finitely many k in the definition of rectangular cylinders is of course not needed then).

Example 5.1. We define

$$\mathbb{R}^\infty := \prod_{k\in\mathbb{N}} \mathbb{R}, \qquad \mathcal{B}_\infty := \bigotimes_{k\in\mathbb{N}} \mathcal{B}(\mathbb{R}),$$

using $(\mathbb{R}, \mathcal{B})$ for all factors. Let $(\Omega, \mathcal{A})$ be another measurable space. A mapping

$$\begin{aligned} f : (\Omega, \mathcal{A}) &\longrightarrow (\mathbb{R}^\infty, \mathcal{B}_\infty) \\ \omega &\longrightarrow f(\omega) = (f_1(\omega), f_2(\omega), \dots) \end{aligned}$$

is measurable if and only if each component f_k is a random variable

Proof. Exercise 5.1. □

Lemma 5.2. *The set of finite unions of all rectangular cylinders is an algebra.*

Proof. Exercise 5.2. □

Definition 5.3. 1. For any finite $I \subset \mathbb{N}$, we define the *projections*

$$\begin{aligned} \pi_I : \prod_{k\in\mathbb{N}} \Omega_k &\longrightarrow \prod_{k\in I} \Omega_k \\ (\omega_1, \omega_2, \dots) &\longrightarrow (\omega_{k_1}, \dots, \omega_{k_N}), \end{aligned}$$

where $k_1 < \dots < k_N \in I$.

2. If $\mathbf{P}$ is a probability on $(\prod_{k\in I} \Omega_k, \bigotimes_{k\in I} \mathcal{A}_k)$ we define the I-marginal as $\mathbf{P}_I := \pi_I * \mathbf{P}$, which is a probability on $(\prod_{k\in I} \Omega_k, \bigotimes_{k\in I} \mathcal{A}_k)$.

3. $\mathbf{P}$ is called a *product probability* if for every rectangular cylinder

$$A = \{\omega \in \Omega; \omega_k \in A_k; k \in \mathbb{N}\}$$

we have

$$\mathbf{P}(A) = \prod_{k\in\mathbb{N}} \mathbf{P}_{\{k\}}(A_k), \tag{5.2}$$

where $\mathbf{P}_{\{k\}}$ is the marginal for $I = \{k\}$. Note that in Equation 5.2, only finitely many factors are $\neq 1$. In particular for finite products

$$\Omega = \Omega_1 \times \cdots \times \Omega_N, \qquad \mathcal{A} = \mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_N$$

we have that for $A_1 \in \mathcal{A}_1, \ldots, A_N \in \mathcal{A}_N$

$$\mathbf{P}(A_1 \times \ldots A_N) = \mathbf{P}_{\{1\}}(A_1) \cdot \cdots \cdot \mathbf{P}_{\{N\}}(A_N).$$

Theorem 5.4. *Let $\mathbf{P}, \mathbf{Q}$ two probabilities on $(\Omega, \mathcal{A}) = (\prod_{k\in\mathbb{N}} \Omega_k, \bigotimes_{k\in\mathbb{N}} \mathcal{A}_k)$ with all marginals being the same. Then*

$$\mathbf{P} = \mathbf{Q}.$$

Proof. The condition just means that $\mathbf{P} = \mathbf{Q}$ on rectangular cylinders. The rest of the proof is Exercise 5.3. □

Theorem 5.5 (Fubini-Tonelli theorem)**.** *Consider $(\Omega, \mathcal{A}) = (\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ with product measure $\mathbf{P} = \mathbf{P}_1 \otimes \mathbf{P}_2$. Then for every random variable $f : \Omega \to \mathbb{R}$,*

1. *For all $\omega_1 \in \Omega_1$ the function $\omega_2 \to f(\omega_1, \omega_2)$ is measurable.*
2. *If $f \geq 0$ or if for all $\omega_1 \in \Omega_1$ the function $\omega_2 \to f(\omega_1, \omega_2)$ is $\mathbf{P}_2$-integrable, then the function $\omega_1 \to \int f(\omega_1, \omega_2) \cdot d\mathbf{P}_2(\omega_2)$ is measurable.*
3. *If $f \geq 0$ then,*

$$\int f \mathrm{d}\mathbf{P} = \int [\int f(\omega_1, \omega_2) \cdot d\mathbf{P}_2(\omega_2)] d\mathbf{P}_1(\omega_1).$$

4. *If f is $\mathbf{P}$ integrable, then the function $\omega_1 \to \int f(\omega_1, \omega_2) \cdot d\mathbf{P}_2(\omega_2)$ is $\mathbf{P}_1$-integrable and*

$$\int f \mathrm{d}\mathbf{P} = \int [\int f(\omega_1, \omega_2) \cdot d\mathbf{P}_2(\omega_2)] d\mathbf{P}_1(\omega_1).$$

We will use two lemmata.

Lemma 5.6. *Items (1,2) hold for indicators* $\mathbf{1}_A$, $A \in \mathcal{A}$.

Lemma 5.7.

$$\mathbf{P}(A) = \int \mathbf{1}_A \mathrm{d}\mathbf{P} = \int [\int \mathbf{1}_A(\omega_1, \omega_2) \cdot d\mathbf{P}_1] d\mathbf{P}_2.$$

Proof of Lemma 5.6. Put $\mathcal{D} :=$ set of all $A \subset \Omega$ so that (a),(b) hold for indicators $\mathbf{1}_A$. If $A = A_1 \times A_2$, $A_1 \in \mathcal{A}_1$, $A_2 \in \mathcal{A}_2$, then $\mathbf{1}_A = \mathbf{1}_{A_1}(\omega_1) \cdot \mathbf{1}_{A_2}(\omega_2)$ and (a, b) hold trivially. Thus $\mathcal{D} \supset$ all cyclinders.

Now let $A_1 \subset A_2 \subset \cdots \in \mathcal{D}$. Then

$$\mathbf{1}_{A_k}(\omega_1, \omega_2) \uparrow \mathbf{1}_{\bigcup_{k=1}^{\infty} A_k}(\omega_1, \omega_2) \quad \forall (\omega_1, \omega_2) \in \Omega,$$

so in particular for ω_1 fixed. Hence

$$\omega_2 \longrightarrow \mathbf{1}_{\bigcup_{k=1}^{\infty} A_k}(\omega_1, \omega_2) \quad \text{is measurable.}$$

Further

$$\omega_1 \to \int \mathbf{1}_{\bigcup_{k=1}^{\infty}}(\omega_1, \omega_2) d\mathbf{P}_2 = \int \lim_n \mathbf{1}_{A_k}(\omega_1, \omega_2) \mathrm{d}\mathbf{P}_2$$
$$\overset{\text{Monot. conv.}}{=} \lim_n \int \mathbf{1}_{A_k}(\omega_1, \omega_2) \mathrm{d}\mathbf{P}(\omega_2) \qquad \text{(measurable!)}$$

So $\bigcup_{k=1}^{\infty} A_k \in \mathcal{D}$. If $A_1 \supset A_2 \supset \cdots \in \mathcal{D}$, prove that $\bigcap_{k=1}^{\infty} A_k \in \mathcal{D}$ along similar lines, using dominated convergence. We have shown that $\mathcal{D}$ contains all rectangular cylinders and is a *monotone class*. A family $\mathcal{D}$ is a monotone class if

1. $A_1 \subset A_2 \subset \cdots \in \mathcal{D} \Rightarrow \bigcup_k A_k \in \mathcal{D}$
2. $A_1 \supset A_2 \supset \cdots \in \mathcal{D} \Rightarrow \bigcap_k A_k \in \mathcal{D}$.

This implies that $\mathcal{D} \supset \mathcal{A}$. □

Proof of Lemma 5.7. It is trivial to verify 5.7 for rectangular cylinders. However, the right hand side makes sense for any $A \in \mathcal{A}$ and forms a probability (σ-additivity comes from monotone convergence). We thus get 5.7 by the MET. □

Completion of proof of 5.5. We have shown that 5.5 holds for indicators and clearly also for simple functions. If $f \geq 0$ random variable, take $f_n \uparrow f$, f_n simple. Now the functions

$$\omega_2 \longrightarrow f_n(\omega_1, \omega_2) \qquad \text{and} \qquad \omega_1 \longrightarrow \int f_n(\omega_1, \omega_2) d\mathbf{P}_2,$$

respectively, are measurable and converge monotonically to

$$\omega_2 \longrightarrow f(\omega_1, \omega_2) \quad \forall \omega_1 \qquad \text{and} \qquad \omega_1 \longrightarrow \int f(\omega_1, \omega_2) \mathrm{d}\mathbf{P}_2 \quad \forall \omega_2,$$

respectively (the second by monotone convergence). Finally, because $\int f_n \mathrm{d}\mathbf{P} = \int [f_n(\omega_1, \omega_2) \mathrm{d}\mathbf{P}_2] \mathrm{d}\mathbf{P}_1$ and both sides converge by monotone convergence to $\int f \mathrm{d}\mathbf{P}$ and $\int [\int f(\omega_1, \omega_2) \mathrm{d}\mathbf{P}_2] \mathrm{d}\mathbf{P}_1$, respectively, Theorem 5.5 is proved for random variable $f \geq 0$.

If f is integrable, 5.5 holds for f_+ and f_-. The only thing that needs proving is that

$$\omega_1 \longrightarrow \int f(\omega_1, \omega_2) \mathrm{d}\mathbf{P}_2 = \int f_+(\omega_1, \omega_2) \mathrm{d}\mathbf{P}_2 - \int f_-(\omega_1, \omega_2) \mathrm{d}\mathbf{P}_2 \qquad (5.3)$$

is well defined (no $\infty - \infty$ situation occurs). Let $N_+ = \{\omega_1 \in \Omega_1; \int f_+(\omega_1, \cdot) \mathrm{d}\mathbf{P}_2 = \infty\}$. We must have $\mathbf{P}_1(N_+) = 0$, because

$$\infty > \int f_+ \mathrm{d}\mathbf{P} = \int [\int f_+(\omega_1, \omega_2) \mathrm{d}\mathbf{P}_2] \mathrm{d}\mathbf{P}_1.$$

Similarly for $N_- = \{\omega_1 \in \Omega_1; \int f_-(\omega_1, \cdot) \mathrm{d}\mathbf{P}_2 = \infty\}$. So 5.3 is well defined apart from $\omega_1 \in N_+ \cap N_-$. which has $\mathbf{P}(N_+ \cap N_-) = 0$. □

Remark 5.8.

1. The integrability condition cannot be omitted. It's not hard to find cases where $\int |f| \mathrm{d}\mathbf{P} = \infty$ and then both sides of (c) are well defined but fail to be equal.

2. To verify that f is integrable, one might use item 3 of the Fubini–Tonelli theorem which says that $\int |f| \mathrm{d}\mathbf{P} = \int [\int |f|(\omega_1, \omega_2) \cdot d\mathbf{P}_2(\omega_2)] d\mathbf{P}_1(\omega_1)$.

3. One can use the r.h.s. of 5.7 to define $\mathbf{P}$ from the marginals. We have shown that r.h.s. is well defined for $A \in \mathcal{A}$ and is $\sigma-$additive. But one then need to invoke MET to show that

$$\int [\int \mathbf{1}_A(\omega_1, \omega_2) \mathrm{d}\mathbf{P}_1] \mathrm{d}\mathbf{P}_2 = \int [\int \mathbf{1}_A(\omega_1, \omega_2) \mathrm{d}\mathbf{P}_2] \mathrm{d}\mathbf{P}_1.$$

4. Fubini-Tonelli extends to finite products.

Exercises for Section 5

Exercise 5.1. Prove the statement in Example 5.1.

Exercise 5.2. Show Lemma 5.2: The family of sets which are finite unions of cylinders is an algebra of subsets of $\prod_{k\in\mathbb{N}} \Omega_k$. (Hint: The complement is the tricky bit. Start with assuming (and later showing) that if A and B are cylinders, then $A \setminus B$ is a finite union of cylinders.)

Exercise 5.3. Demonstrate Theorem 5.4. You can use without proof the Theorem 2.6.

Exercise 5.4. Setup is as in Section 5

1. Consider a set of the form

$$B = \{\omega; \omega_k \in A_k \text{ for all } k \in \mathbb{N}\},$$

with $A_k \in \mathcal{A}_k$ for all $k \in \mathbb{N}$. (B is *not necessarily* a cylinder!) Show that B is nonetheless measurable.

2. Demonstrate that for a product probability $\mathbb{Q}$ (see Definition 5.3, item 3) and with B as in the previous item, $\mathbb{Q}(B) = \lim_{n\to\infty} \prod_{k=1}^{n} \mathbf{P}_k(A_k)$.

6 Distributions and independence

In this section, we will change notation somewhat, bringing it closer to standard notation in probability theory. Further, we introduce the important concept of independence.

Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space. Random variables are measurable functions with values in $(\mathbb{R}^d, \mathcal{B}_d)$ (where $d = \infty$ possible) and are denoted by capital letters:

$$X : (\Omega, \mathcal{A}) \to (\mathbb{R}^d, \mathcal{B}_d).$$

If $d = 1$, we put $\mathbb{E}(X) := \int X \mathrm{d}\mathbf{P}$ ("expectation value"). If $d < \infty$, $\mathbb{E}(X)$ is taken component wise. Throughout this section, the symbol d stands for a finite integer or for ∞, unless otherwise stated.

Definition 6.1. 1. The *distribution* of a random variable $X : (\Omega, \mathcal{A}) \to (\mathbb{R}^d, \mathcal{B}_d)$ is defined as $P_X := X_*\mathbf{P}$.

2. An I-marginal of X is the distribution $X_I := (X_{k_1}, \ldots, X_{k_N})$, where $I = \{k_1 < \cdots < k_N\}$.

Note that an I-marginal of X according to Definition 6.1 is the same as an I-marginal of $\mathbf{P}_X$ according to Definition 5.3 (Exercise 6.2).

Lemma 6.2. *Suppose that $X : \Omega \to \mathbb{R}^d$ is a random variable and $f : \mathbb{R}^d \to \mathbb{R}$ a measurable function (with respect to the Borel algebra on both the domain and range). Further, suppose that $f \circ X$ is integrable. Then*

$$\mathbb{E}(f \circ X) = \int_{\mathbb{R}^d} f(x) \mathrm{d}\mathbf{P}_X(x).$$

Proof. This is essentially the transformation formula, see Exercise 6.1. □

Lemma 6.3. *Two random variables*

$$X, Y : \Omega \longleftrightarrow \mathbb{R}^d$$

have the same distribution iff they have the same I-marginals.

Proof. If d is finite, then the I-marginal for $I = \{1, \ldots, d\}$ is actually the distribution. If d is infinite, let $A = \{x \in \mathbb{R}^\infty; x_{n_k} \in A_k, n_k \in I\}$ be a rectangular cylinder for some $I = \{n_1, \ldots, n_k\} \subset \mathbb{N}$ and some $A_1, \ldots, A_k \in \mathcal{B}_1$. Then

$$\begin{aligned} &P_{X_I}(A_1 \times \cdots \times A_k) = \mathbf{P}(\{\omega; X_{n_k} \in A_k, n_k \in I\}) \\ &= \mathbf{P}(X \in A) = P_X(A). \end{aligned} \tag{6.1}$$

and the same for Y. If the I-marginals agree, then Equation (6.1) shows that P_X and P_Y agree on rectangular cylinders, so Theorem 5.4 gives $P_X = P_Y$. If on the other hand $P_X = P_Y$, then Equation (6.1) (read from right to left) shows that the I-marginals agree. □

Independence

This paragraph has only two definitions. Some facts about independent random variables will be explored in the exercises.

Definition 6.4. Let $X_1, X_2, \ldots$ random variables with values in $\mathbb{R}$. They are called independent if any I marginal is a product probability. This means that for any $N \in \mathbb{N}$, any index set $I = \{k_1 < \ldots < k_N\}$ and any selection of sets $B_1, \ldots, B_N$ in $\mathcal{B}(\mathbb{R})$ the relation

$$\mathbf{P}(X_{k_1} \in B_1, \ldots, X_{k_N} \in B_N) = \mathbf{P}(X_{k_1} \in B_1) \cdot \ldots \cdot \mathbf{P}(X_{k_N} \in B_N)$$

holds.

Definition 6.5. For any random variable $X : \Omega \to \mathbb{R}^d$, $d < \infty$ define the Covariance matrix

$$\mathrm{Cov}(X) = \mathbb{E}([X_i - \mathbb{E}X_i][X_j - \mathbb{E}X_j])_{i,j}$$

and the variance

$$\mathbb{V}(X) = \mathrm{tr}[\mathrm{Cov}(X)] = \mathbb{E}([X - \mathbb{E}X]^2)$$

(both are finite if $\sum_{k=1}^n X_k^2$ is integrable). Finally, if $Y : \Omega \longrightarrow \mathbb{R}^{d'}$ random variable ($d' < \infty$), then

$$\mathrm{Cov}(X, Y) = \mathbb{E}([X_i - \mathbb{E}X_i][Y_j - \mathbb{E}Y_j]) \qquad \in \mathbb{R}^{d \times d'}.$$

It is easy to see that $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)^T$. Note that $\mathrm{Cov}(X)$ is symmetric and nonnegative definite, because

$$v^T \mathrm{Cov}(X) v = \mathbb{E}\left(\left(v^T \left(X - \mathbb{E}\left(X\right)\right)\right)^2\right) \geq 0.$$

We write $A \geq 0$ if $A \in \mathbb{R}^{d \times d}$ symmetric nonnegative definite. Also, $A \geq B$ means A, B symmetric and $A - B \geq 0$. Similarly, "$>$" means positive definite.

Modelling with random variables

The following lemma might sound abstract, but its interpretation is very simple. Suppose we observe data from the real world, and we want to model them as random variables, say $X_1, \ldots, X_d$, each of them real valued. But "modelling" almost always means to merely specify the distribution of those random variables. The probability space $(\Omega, \mathcal{A}, \mathbf{P})$ on which these random variables live, and in fact the variables themselves are usually not specified. The following lemma simply says that this is not a problem, and in the proof you will find a canonical choice for these missing ingredients:

Lemma 6.6. *If μ is a probability on $(\mathbb{R}^d, \mathcal{B}_d)$, then there exists a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ and a measurable random variable $X : \Omega \longrightarrow: \mathbb{R}^d$ so that $\mu = P_X$.*

Proof. Take $\Omega = \mathbb{R}^d$, $\mathcal{A} = \mathcal{B}_d$, $\mathbf{P} = \mu$ and $X(\omega) = \omega$. □

Exercises for Section 6

Exercise 6.1. In the setup of Lemma 6.2, show that $f \circ X$ is a random variable and prove the formula.

Exercise 6.2. For P_X the distribution of some $X : (\Omega, \mathcal{A}, \mathbf{P}) \to (\mathbb{R}^d, \mathcal{B}_d)$, we have defined the concept of I–marginals in Definition 6.1. Show that Definition 5.3 however is also applicable and gives the same concept of I–marginals. (Hint: this is used in the proof of Lemma 6.3).

Exercise 6.3. In this exercise, d is finite. Consider a random variable $X : \Omega \to \mathbb{R}^d$ with distribution P_X which has a density $p : (\mathbb{R}^d, \mathcal{B}_d) \to (\mathbb{R}_{\geq 0}, \mathcal{B})$ with respect to the n–dimensional Lebesgue measure.

1. Let $f : (\mathbb{R}^d, \mathcal{B}_d) \to (\mathbb{R}, \mathcal{B})$ be integrable with respect to P_X. Show that
$$\int f(x) \mathrm{d}P_X(x) = \int f(x) p(x) \mathrm{d}x$$
Start with f being a simple function and proceed as usual. (Note that this extends Lemma 6.2)

2. Show that the marginals of μ have densities as well. Hint: For example $\mu_{\{1\}}$ has the density
$$p_1(x_1) = \int_{\mathbb{R}^{d-1}} p(x_1, x_2, \ldots, x_d) \mathrm{d}x_2 \ldots \mathrm{d}x_d.$$

Exercise 6.4. A little bit about independence.

1. Show that random variables $X_1, X_2, \ldots$ with values in $\mathbb{R}$ are independent if and only if for any $n \in \mathbb{N}$ and any selection $f_1, \ldots, f_n$ of bounded and measurable functions the relation
$$\mathbb{E}(f_1(X_1) \cdot \ldots \cdot f_n(X_n)) = \mathbb{E}(f_1(X_1)) \cdot \ldots \cdot \mathbb{E}(f_n(X_n))$$
holds.

2. Suppose that random variables $X_1, \ldots, X_d$ with values in $\mathbb{R}$ are independent, and their distribution has a density p as in exercise 6.3. Show that
$$p(x) = p_1(x_1) \cdot \ldots \cdot p_d(x_d)$$
where p_k is the density of the distribution of X_k for each $k = 1, \ldots, d$.

3. Suppose that random variables X_1, X_2 with values in $\mathbb{R}$are independent, and there are sets B_1, B_2 in $\mathcal{B}(\mathbb{R})$ so that
$$\{\omega; X_1(\omega) \in B_1\} = \{\omega; X_2(\omega) \in B_2\},$$
Then $\mathbf{P}(\{\omega; X_1(\omega) \in B_1\}) = 0$ or 1.

Exercise 6.5. Let $X = (X_1, \ldots, X_d)$ random variables (d is finite). The distribution of X is said to be *normal* or *Gaussian* if it has a density $p : \mathbb{R}^d \to \mathbb{R}_{>0}$ with respect to Lebesgue measure given by the formula

$$p(x; \mu, \Gamma) = \frac{1}{\sqrt{\det(2\pi\Gamma)}} \exp\big(-\frac{1}{2}(x-\mu)^T\Gamma^{-1}(x-\mu)\big)$$

where $\mu \in \mathbb{R}^d$ and Γ is a positive definite $d \times d$–matrix.

1. Show that $\mathbb{E}(X_k) = \mu_k$ and $\mathrm{Cov}(X) = \Gamma$.
2. Show that the marginals of the distribution of X are normal as well, and determine the expectation value and covariance matrix.
3. Let A be a surjective $m \times d$–matrix ($m \leq d$) and $b \in \mathbb{R}^m$. Show that $AX + b$ has again a normal distribution, and determine the expectation value and covariance matrix.
4. Show that $X_1, \ldots, X_d$ are independent if and only if the covariance matrix is diagonal.

7 Introduction to Statistics

The main point of statistics is to "identify" P_X for some random variable $X : \Omega \longrightarrow \mathbb{R}^d$ given X assumes certain values $(x_1, \ldots, x_d)$. Here, d is finite, but one is also interested what happens if d becomes large, e.g. are we able to reconstruct P_X if $d \to \infty$?

Definition 7.1 (Parametric estimation problem)**.** Let $\Theta \subset \mathbb{R}^p$ open ($p < \infty$). This is the *parameter space*. Let $\mathcal{H} := \{\mathbf{P}_\theta : \theta \in \Theta\}$ be a set of probability measures on $(\mathbb{R}^n, \mathcal{B}_n)$ (i.e. distributions). We call $\mathcal{H}$ the *hypothesis*. The quadruple $(\mathbb{R}^n, \mathcal{B}_n, \Theta, \mathcal{H})$ we will call a *parametric estimation problem*.

Since the parametric estimation problem is about identifying the *distribution* of $X = (X_1, \ldots, X_n)$, we are free to chose Ω, $\mathcal{A}$, and X as in Lemma 6.6 to ensure that X has the desired distribution. This is the choice we made in Definition 7.1. We will write $\mathbb{E}_\theta(\phi) := \int_{\mathbb{R}^n} \phi(x) \mathrm{d}\mathbf{P}_\theta$ for any function ϕ integrable with respect to $\mathbf{P}_\theta$. Note that $\mathbf{P}_\theta$ is *not* the distribution of some random variable θ, but that θ is a parameter of that distribution. So our notation in the present section differs slightly from that of Section 6.

The aim in the parametric estimation problem is to find measurable functions ("estimators")

$$t : (\mathbb{R}^n, \mathcal{B}_n) \to (\Theta, \mathcal{B}_p)$$

so that the distribution of t under $\mathbf{P}_\theta$, that is $t_*\mathbf{P}_\theta$ "concentrates" around θ for all $\theta \in \Theta$. There are many ways to understand the word "concentrates". We will use the concept of *mean square error* to quantify this:

Definition 7.2. Fix $\mathcal{H}$ and estimator t.

1. The bias of t is the function
$$\begin{aligned} b :&\Theta \longrightarrow \mathbb{R}^p, \\ &b(\theta) = \mathbb{E}_\theta(t) - \theta. \end{aligned}$$
2. t is unbiased if $b(\theta) = 0$ for all $\theta \in \Theta$.
3. The mean-square-error of t is
$$\mathrm{mse}(\theta) = \mathbb{E}_\theta([t-\theta]^2).$$

Lemma 7.3. *If $\mathbb{E}_\theta(t^2) < \infty$ for all $\theta \in \Theta$, then*
$$\mathrm{mse}(\theta) = \underbrace{\mathbb{E}_\theta([t - \mathbb{E}_\theta(t)]^2)}_{=:V_\theta(t)} + b(\theta)^2.$$

Proof.
$$\begin{aligned} \mathrm{mse}(\theta) = \mathbb{E}_\theta([t-\mathbb{E}_\theta(t)]^2) &= \mathbb{E}_\theta\left([t - \mathbb{E}_\theta(t) + \mathbb{E}_\theta - \theta]^2\right) \\ &= V_\theta(t) + (b(\theta))^2 - 2\underbrace{\mathbb{E}_\theta\left([t-\mathbb{E}_\theta(t)][\underbrace{\theta - \mathbb{E}_\theta(t)}_{\text{constant}}]\right)}_{0}. \end{aligned}$$

□

Definition 7.4 (The standard setup)**.** Let $(\mathbb{R}^n, \mathcal{B}_n, \Theta, \mathcal{H})$ be a parametric estimation problem. We will say that the parametric estimation problem has the *standard setup* if $\mathcal{H}$, the set of candidate distributions, is given by a family of product densities, that is
$$\mathbf{P}_\theta(A) = \int_{\mathbb{R}^n} \mathbf{1}_A \cdot p(x,\theta)\mathrm{d}x,$$
where $p(x;\theta) = \prod_{k=1}^n f(x_k,\theta)$ and $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a density (i.e. $\int_{\mathbb{R}} f(x,\theta)\mathrm{d}x = 1$).

It follows that if the parametric estimation problem is in the standard setup, the $X_1, \ldots, X_n$ are independent, with all X_k having the same distribution given by the density $f(\cdot, \theta)$.

Example 7.5. Assume the standard setup 7.4 with $f(x, \theta) = g(x-\theta)$, where $g : \mathbb{R} \longrightarrow \mathbb{R}_{\geq 0}$ has properties

1. $\int g(x)dx = 1$.
2. $\int xg(x)dx = 0$.
3. $\int x^2 g(x)dx = 1$.

Now item 1 ensures f is a density. From item 2 we get

$$\mathbb{E}_\theta(x_k) = \int x \cdot f(x, \theta)dx = \int xg(x-\theta)dx \overset{(2)}{=} \theta,$$

and $\mathrm{Cov}_\theta(x) = 1$. We will now try to estimate θ. We use $t : \mathbb{R}^n \to \mathbb{R}, t(x) = \frac{1}{n}\sum_{k=1}^n x_k$. First we calculate the bias:

$$\mathbb{E}_\theta(t) = \mathbb{E}_\theta\left(\frac{1}{n}\sum_{k=1}^n x_k\right) = \frac{1}{n}\sum_{k=1}^n \mathbb{E}_\theta x_k \overset{(d)}{=} \theta \Rightarrow \text{Bias} = 0.$$

Variance:

$$\begin{aligned}
\mathbb{V}_\theta(t) &= \mathbb{E}_\theta([t-\theta]^2) \\
&= \mathbb{E}_\theta\left([\frac{1}{n}\sum_{k=1}^n x_k - \theta]^2\right) \\
&= \mathbb{E}_\theta\left([\frac{1}{n}\sum_{k=1}^n (x_k - \theta)]^2\right) \\
&= \frac{1}{n^2}\sum_{k,j=1}^n \mathbb{E}\left([x_k-\theta][x_j-\theta]\right) \\
&= \frac{n}{n^2} = \frac{1}{n},
\end{aligned}$$

so $\mathrm{mse}(\theta) = \mathbb{V}_\theta(t) = \frac{1}{n} \longrightarrow 0$.

Definition 7.6 (The Fisher Information regularity conditions)**.** A parametric estimation problem $(\mathbb{R}^n, \mathcal{B}_n, \Theta, \mathcal{H})$ with d finite is said to satisfy the *Fisher Information (FI) regularity conditions* if the following is true:

1. There is a measure ν on $(\mathbb{R}^d, \mathcal{B}_d)$ and a function $p : (\mathbb{R}^d, \Theta) \longrightarrow \mathbb{R}_{\geq 0}$ so that $p(\cdot, \theta)$ is a random variable $\forall \theta \in \Theta$, and

$$\mathbf{P}_\theta(A) = \int_A p(x, \theta)\, d\nu(x).$$

   So p is a density w.r.t. ν (see Exercise 3.5).

2. $\forall x \in \mathbb{R}^d$, $\theta \in \Theta : D_\theta p(x, \theta)$ exists.

3.

$$\int D_\theta p(x, \theta) d\nu(x) \underset{\text{Assume this is permitted}}{=} D_\theta \int p(x, \theta) d\nu(x) = D_\theta 1 = 0.$$

4. $C(\theta) = \{x \in \mathbb{R}^d; p(x, \theta) > 0\}$ does not depend on θ.

Theorem 7.7 (The Cramér–Rao lower bound)**.** *Suppose the FI-conditions 7.6 are met. Further, suppose that t is an unbiased estimator of θ. Then*

$$\mathrm{Cov}_\theta(t) = \mathbb{E}_\theta\left[(t - \theta)_i (t - \theta)_j\right] \geq \mathcal{I}^{-1},$$

where $\mathcal{I} = \mathrm{Cov}_\theta\left(D_\theta \log p(x, \theta)\right)$ is the Fisher *information, provided that $\mathcal{I}$ is invertible.*

We will need the following

Lemma 7.8. *Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space. Suppose $X : \Omega \to \mathbb{R}^{d_1}$, $Y : \Omega \to \mathbb{R}^{d_2}$, finite variances. If $\mathrm{Cov}(Y) > 0$, then*

$$\mathrm{Cov}(X) \geq \mathrm{Cov}(X, Y)\mathrm{Cov}(Y)^{-1}\mathrm{Cov}(Y, X).$$

Further, equality holds here if and only if there is $M \in \mathbb{R}^{d_1 \times d_2}$ so that $X = MY$.

Proof. Put

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix}$$

and consider

$$\mathrm{Cov}(Z) = \begin{bmatrix} \mathrm{Cov}(X) & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(Y, X) & \mathrm{Cov}(Y) \end{bmatrix}.$$

Put

$$W = \begin{bmatrix} \mathbf{1} \\ -\mathrm{Cov}(Y)^{-1} \cdot \mathrm{Cov}(Y, X) \end{bmatrix} \in \mathbb{R}^{(d_1 + d_2) \times d_1}.$$

Now

$$0 \leq W^T \mathrm{Cov}(Z) W = \mathrm{Cov}(X) - \mathrm{Cov}(X,Y)\mathrm{Cov}(Y)^{-1}\mathrm{Cov}(Y,X).$$

If equality holds, then $W^T\mathrm{Cov}(Z)W = \mathrm{Cov}(W^T Z) = 0$. Checking the diagonal elements, we get $W^T Z = 0$ or $X = \underbrace{\mathrm{Cov}(X,Y)\mathrm{Cov}(Y)^{-1}}_{M} \cdot Y$. □

Proof of the Cramér–Rao bound. By taking the derivative of the identity $1 = \int p(x,\theta)\mathrm{d}\nu(x)$ we get

$$0 = \int \partial_{\theta_j} \log p(x,\theta) \cdot p(x,\theta)\mathrm{d}\nu(x). \tag{7.1}$$

From unbiasedness, we get

$$0 = \int [t(x) - \theta]_i p(x,\theta)\mathrm{d}\nu(x) = \mathbb{E}_\theta(t) - \theta,$$

and taking the derivative of this identity and using (7.1) we obtain

$$0 = -\delta_{ij} + \int [t(x) - \theta_i] \cdot \partial_{\theta_j} \log p(x,\theta) \cdot p(x,\theta)\mathrm{d}\nu. \tag{7.2}$$

Equation (7.2) can be written as $\mathbf{1} = \mathrm{Cov}\left(t - \theta, D_\theta \log p(X,\theta)\right)$. Now the lemma gives the result. □

Lemma 7.9. *Under appropriate regularity conditions*

$$\mathcal{I} = -\mathbb{E}_\theta \left(\partial^2_{\theta_i\theta_j} \log p(X,\theta) \right).$$

Proof. Exercise 7.1. □

Definition 7.10. Let $(\mathbb{R}^n, \mathcal{B}_n, \Theta, \mathcal{H})$ be a parametric estimation problem, and suppose that t is an unbiased estimator. If

$$\mathrm{Cov}_\theta(t) \leq \mathrm{Cov}_\theta(t') \quad \forall \theta \in \Theta$$

for any other unbiased estimator t', then t is called Uniformly Minimum Variance Unbiased (UMVU) estimator.

Example 7.11 (Another example)**.** Standard setup 7.4, with $f(x,\theta) = \frac{1}{\theta}\exp(-\frac{x}{\theta})$, $\theta \in \mathbb{R}_+$, $x \in \mathbb{R}_{\geq 0}$. Consider the estimator $t = \frac{1}{n}\sum_{k=1}^n x_k$. For this estimator, we get $\mathbb{E}_\theta(t) = \theta$, $\mathbb{V}_\theta(t) = \frac{\theta^2}{n}$. The FI-regularity conditions are easily checked, and

$$\mathcal{I} = -\mathbb{E}_\theta \left(\partial^2_\theta \log p(x,\theta) \right) = \frac{n}{\theta^2}$$

as the following calculations show:

$$\begin{aligned}
-\log p(x,\theta) &= \sum_{k=1}^{n} \log(\theta) + \frac{x_k}{\theta} = n\log(\theta) + \sum_{k=1}^{n} \frac{x_k}{\theta}, \\
\partial_\theta^2(-\log p(x,\theta)) &= -\frac{n}{\theta^2} + 2\frac{\sum_k x_k}{\theta^3}, \\
\mathbb{E}_\theta\left(\partial_\theta^2(-\log p(X,\theta))\right) &= -\frac{n}{\theta^2} + 2\frac{n\cdot\theta}{\theta^3} = \frac{n}{\theta^2}.
\end{aligned}$$

Hence t has minimum variance!

Definition 7.12 (The maximum likelihood estimator)**.** Suppose a parametric estimation problem is in the standard setup and satisfies the FI-regularity conditions 7.6. The *Likelihood* $L : \Theta \times \mathbb{R}^n \to \mathbb{R}$ is given by

$$L(\theta, x) = p(x, \theta),$$

Consider the set $C = \{x \in \mathbb{R}^d; f(x,\theta) > 0\}$ and suppose that for any $x \in C$ the likelihood as a function of θ, that is $\theta \to L(\theta, x)$, has a unique maximiser $t(x)$. In other words, for any $x \in C$ and $\theta \in \Theta$ it holds that

$$L(t(x), x) \geq L(\theta, x),$$

with equality here if and only if $\theta = t(x)$. Then t is called the *Maximum Likelihood estimator* (MLE).

Remark 7.13.

1. It often pays off to maximise $l := \log L(\cdot, x)$, the $log-likelihood$.
2. In the standard setup, we have

$$p(x,\theta) = \prod_{k=1}^{n} f(x_k, \theta),$$

so $l = \log L(\theta, x) = \sum_{k=1}^{n} \log f(x_k, \theta)$.

Example 7.14.

1. Assume the standard setup with $f(x,\theta) = \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right)$, $\theta > 0, x \geq 0$. Then

$$l(\theta, x) = -\frac{1}{\theta}\sum_{k=1}^{n} x_k - n\log\theta.$$

To find the MLE, we solve the normal equations:

$$\frac{\partial}{\partial\theta} l(\theta, x) = \frac{1}{\theta^2}\sum_{k=1}^{n} x_k - \frac{n}{\theta}$$

$$\frac{\partial}{\partial\theta} l(t, x) = 0 \iff 0 = \frac{1}{t^2}\sum_{k=1}^{n} x_k - \frac{n}{t}$$

$$\iff t(x_1, \ldots, x_n) = \frac{1}{n}\sum_{k=1}^{n} x_k \quad \text{as in Example 7.11.}$$

2. Assume the standard setup with

$$f(x, \theta) = \frac{x^m \cdot \exp(-\theta x)\theta^{m+1}}{m!},$$

where $m \in \mathbb{N}$ is a fixed.

$$l(\theta, x) = \sum_{k=1}^{n} \{\log \frac{1}{m!} + m \log x_k - \theta x_k + (m+1)\log\theta\}$$

$$\frac{\partial}{\partial\theta} l(\theta, x) = -\sum_{k=1}^{n} x_k + n \cdot (m+1)/\theta$$

$$\frac{\partial}{\partial\theta} l(t, x) = 0 \iff t = \frac{(m+1)}{\frac{1}{n}\sum_{k=1}^{n} x_k}.$$

Remark 7.15. It is not always true that the MLE is UMVU. But if the parametric estimation problem satisfies the standard setup, the FI-regularity conditions, and some more conditions, then t has, asymptotically for large n, a normal distribution with mean θ and variance $\mathcal{I}(\theta)^{-1}$. See [vdV00], chapter 5, for precise statements and proofs of this result.

Exercises for Section 7

Exercise 7.1. Demonstrate the formula in Lemma 7.9 for the Fisher information and state the conditions under which this formula is correct.

Exercise 7.2. Assume the $\mathbb{R}$–valued random variables $X_1, \ldots, X_n$ (n is finite) are independent, nonegative, and for all $k = 1, \ldots, n$, the distribution of X_k has a density with respect to Lebesgue measure, given by

$$p(x, \theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \le x \le \theta \\ 0 & \text{else.} \end{cases}$$

where $\theta \in \mathbb{R}_{>0}$. This is called a *uniform distribution*. Given that X_k can never be larger than θ, we put $T(x) = \max\{x_1, \ldots, x_n\}$ as an estimator for θ.

1. Show that
$$\mathbf{P}(T \leq z) = \frac{z^n}{\theta^n}, \qquad 0 < z < \theta.$$
2. Conclude from the previous item that T has a density given by
$$q(z) = \frac{nz^{n-1}}{\theta^n}, \qquad 0 < z < \theta.$$
3. Show that T is biased, and suggest an estimator for θ based on T which would be unbiased.

Exercise 7.3. Let $x_1, \ldots, x_n$ be independent and identically distributed random variables, and the distribution of each x_k has a density $f(x;\theta)$ with respect to Lebesgue measure of the form
$$f(x;\theta) = a(x)\exp(\theta h(x) - \phi(\theta)),$$
where $a \geq 0$ and h are given random variables, and $x \in I$ with I some interval not depending on θ. Densities of this kind are referred to as *exponential families* (not to be confused with the exponential distribution, which is a special case). You can assume that $f(x;\theta)$ is a well defined density for all θ in some open interval Θ. Further, you can assume that differentiation under the integral sign is permitted in the following, that the function $\theta \to \mathbb{E}_\theta(h)$ is one-to-one and continuous in both directions, and that $0 < \mathbb{V}_\theta(h) < \infty$ for all $\theta \in \Theta$.

1. By considering the condition $\int_{x\in I} f(x;\theta)\mathrm{d}x = 1$, provide an expression for the function ϕ.
2. By differentiating $\int_{x\in I} f(x;\theta)\mathrm{d}x = 1$ with respect to θ, demonstrate that $\mathbb{E}_\theta(h) = \frac{\mathrm{d}\phi}{\mathrm{d}\theta}$. Differentiate again to show that $\mathbb{V}_\theta(h) = \frac{\mathrm{d}^2\phi}{\mathrm{d}\theta^2}$.
3. Show that the maximum likelihood estimator $\hat{\theta}$ for θ is a solution of the equation $\frac{1}{n}\sum_{k=1}^n h(x_k) = \frac{\mathrm{d}\phi}{\mathrm{d}\theta}(\hat{\theta})$.
4. Show that the maximum derived in the last item is actually a maximum at least locally (Hint: You might want to use results from item 2 for this proof).

5. Demonstrate that the Fisher information (of the joint density) is given by $n\frac{\mathrm{d}^2\phi}{\mathrm{d}\theta^2}$.

6. For this last item, you need to use the Law of large numbers: If $Y_1, \ldots, Y_n$ are independent and identically distributed random variables with finite variance, then $\frac{1}{n}\sum_{k=1}^{n} Y_k \to \mathbb{E}(Y_1)$ almost surely for $n \to \infty$. Using this, prove that if θ_0 corresponds to the true distribution of $x_1, \ldots, x_n$, then $\hat{\theta} \to \theta_0$.

8 Conditional probabilities and Conditional expectations

Let $(\Omega, \mathcal{A}, \mathbf{P})$ probability space. Consider L_2, the space of all random variables $f : \Omega \to \mathbb{R}$ so that $\int f^2 \mathrm{d}\mathbf{P} < \infty$. This is a Hilbert space with scalar product $\langle f, g\rangle := \int fg\mathrm{d}\mathbf{P}$. Let f be an element of this Hilbert space and S be a closed subspace. Then there exists $\hat{f} \in S$ which is the "best approximation" f, which means

$$\left\| f - \hat{f} \right\|^2 = \langle f - \hat{f}, f - \hat{f}\rangle \leq \|f - g\|^2, \quad \forall g \in S,$$

and equality occurs here if and only if $g = \hat{f}$. We now claim that $f - \hat{f}$ (i.e. the approximation error) is perpendicular to S, that is $\langle f - \hat{f}, g\rangle = 0$ for any $g \in S$. To see this, note that for any $g \in S$ we have

$$\left\| f - (\hat{f} + g) \right\|^2 = \left\| f - \hat{f} \right\|^2 + \|g\|^2 - 2\langle f - \hat{f}, g\rangle.$$

Suppose $\exists g \in S$ with $\langle f - \hat{f}, g\rangle = m \neq 0$, then replace g in the relation above with $g' = \frac{m}{\|g\|^2} g$, which gives

$$\begin{aligned}\left\| f - (\hat{f} + g') \right\|^2 &= \left\| f - \hat{f} \right\|^2 + \frac{m^2}{\|g\|^2} - 2\frac{m^2}{\|g\|^2} \\ &= \left\| f - \hat{f} \right\|^2 - \frac{m^2}{\|g\|^2} < \left\| f - \hat{f} \right\|^2,\end{aligned}$$

which means that $\hat{f} + g'$ is a better approximation than $\hat{f}$, which is a contradiction. Hence $\langle f - \hat{f}, g\rangle = 0$, or

$$\int f \cdot g\mathrm{d}\mathbf{P} = \int \hat{f} \cdot g\mathrm{d}\mathbf{P} \quad \text{for all}\, g \in S. \tag{8.1}$$

We want to use this with S being the space of all random variables g so that $\int g^2 \mathrm{d}\mathbf{P} < \infty$ and g is $\mathcal{F}$-measurable, where $\mathcal{F}$ is some σ-algebra on Ω with $\mathcal{F} \subset \mathcal{A}$. It is clear that S is a subspace of L_2 and is in itself a Hilbert space. This implies that S is closed in L_2, hence we can find $\hat{f}$ so that (8.1) is correct. Note that for the special S we have chosen here, (8.1) would be true if

$$\int f \cdot \mathbf{1}_A \mathrm{d}\mathbf{P} = \int \hat{f} \cdot \mathbf{1}_A \mathrm{d}\mathbf{P}, \, \forall A \in \mathcal{F} \tag{8.2}$$

by approximation. But (8.2) makes sense even if $\int |f| \mathrm{d}\mathbf{P} < \infty$, which is weaker that $\int f^2 \mathrm{d}\mathbf{P} < \infty$. This leads us to the following definition.

Definition 8.1. Let $\int |f| d\mathbf{P} < \infty$, $\mathcal{F} \subset \mathcal{A}$, $\mathcal{F}$ a sigma-algebra. Then the *conditional expectation of f given $\mathcal{F}$*, written as $\mathbb{E}(f|\mathcal{F})$, is any $\mathcal{F}$-measurable function $\hat{f}$ satisfying (8.2).

Theorem 8.2. *Let $\int |f| d\mathbf{P} < \infty$, $\mathcal{F} \subset \mathcal{A}$, $\mathcal{F}$ a sigma algebra.*

1. *There exists conditional expectation $\mathbb{E}(f|\mathcal{F})$.*
2. *Suppose $f^{(1)}, f^{(2)}$ are $\mathcal{F}$-measurable and satisfy (8.2), then*

$$f^{(1)}(\omega) = f^{(2)}(\omega)$$

for $\omega \in \Omega_1$, with $\mathbf{P}(\Omega_1) = 1$.

Proof. Suppose first $f \geq 0$. Put $c = \int f d\mathbf{P}$ and define

$$F(A) := \frac{1}{c} \int f \cdot \mathbf{1}_A \mathrm{d}\mathbf{P} \tag{8.3}$$

for any $A \in \mathcal{F}$. Note that this is a probability on $\mathcal{F}$ (see Exercise 8.1). Using the Radon Nykodym theorem, it can be shown that $\exists \mathcal{F}$-measurable random variable $\hat{f} \geq 0$ with

$$F(A) = \frac{1}{c} \int \hat{f} \cdot \mathbf{1}_A \mathrm{d}\mathbf{P} \tag{8.4}$$

for any $A \in \mathcal{F}$. Combining (8.3,8.4) is (8.2). The uniqueness is like the uniqueness for densities. For general f, consider f_+ and f_-. □

Remark 8.3 (Defining properties of the conditional expectation)**.** Let's say you have some $\hat{f}$ and you suspect that

$$\hat{f} = \mathbb{E}(f|\mathcal{F}).$$

To verify this, you have to check that

1. $\hat{f}$ is $\mathcal{F}$-measurable

2. $$\int f \cdot g \mathrm{d}\mathbf{P} = \int \hat{f} \cdot g \mathrm{d}\mathbf{P}$$

   for any function g which is $\mathcal{F}-$measurable and bounded (in fact, it suffices to check this for all g of the form $\mathbf{1}_A$ with $A \in \mathcal{F}$).

Lemma 8.4 (Properties of $\mathbb{E}(f|\mathcal{F})$)**.**

1. *Linear in f.*
2. $f \geq 0 \Rightarrow \mathbb{E}(f|\mathcal{F}) \geq 0$ *a.s.*
3. *If $\mathcal{Y} \subset \mathcal{F} \subset \mathcal{A}$ are sigma-algebras, then*

   $$\mathbb{E}(\mathbb{E}(f|\mathcal{F})|\mathcal{Y}) = \mathbb{E}(f|\mathcal{Y})$$

   (Law of the Iterated Expectations).

Proof. Exercise 8.2. □

Definition 8.5.

1. Let $g : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ measurable. The family of sets
   $$\sigma(g) = \{g^{-1}(A) : A \in \mathcal{A}'\}$$
   is a sigma algebra, called the *sigma algebra generated by g* ($\mathcal{A}'$ is fixed). Measurability implies $\sigma(g) \subset \mathcal{A}$.
2. $\mathbb{E}(f|g) := \mathbb{E}(f|\sigma(g))$. Note that this is a random variable on $(\Omega, \mathcal{A})$.
3. The following is a slightly different concept of conditional expectation. Let $X : (\Omega, \mathcal{A}) \to (\mathbb{R}^d, \mathcal{B}_d)$. Then $\mathbb{E}(f|X = x)$ is any random variable $\hat{f}$ satisfying
   $$\int \mathbf{1}_B(x) \cdot \hat{f}(x) \mathrm{d}\mathbf{P}_X(x) = \int \mathbf{1}_B \circ X(\omega) \cdot f(\omega) \mathrm{d}\mathbf{P}(\omega)$$
   for all $B \in \mathcal{B}_d$. Note that $\mathbb{E}(f|X = x)$ is a random variable on $(\mathbb{R}^d, \mathcal{B}_d)$.

Lemma 8.6. $\hat{f}(x) = \mathbb{E}(f|X = x) \iff \hat{f}(X(\omega)) = \mathbb{E}(f|\sigma(X))(\omega)$.

Proof. Exercise 8.3. □

Definition 8.7 (Conditional Probability)**.** The conditional expectation of an indicator function has a special interpretation. Let $A \in \mathcal{A}$.

1. $\mathbf{P}(A|\mathcal{F}) := \mathbb{E}(\mathbf{1}_A|\mathcal{F})(\omega)$ is called the *conditional probability* of A given $\mathcal{F}$.

2. If X is a random variable, we define $\mathbf{P}(A|X) := \mathbb{E}(\mathbf{1}_A|X)$ and call it conditional probability of A given X.

3. Using the alternative concept of conditional expectation in Definition 8.5, item 3 we define $\mathbf{P}(A|X = x) := \mathbb{E}(\mathbf{1}_A|X = x)$.

Note that for $B \in \mathcal{F}$ we have the formula

$$\int \mathbf{P}(A|\mathcal{F}) \cdot \mathbf{1}_B \, \mathrm{d}\mathbf{P} = \int \mathbf{1}_A \cdot \mathbf{1}_B \cdot \, \mathrm{d}\mathbf{P} = \mathbf{P}(A \cap B).$$

Lemma 8.8 (Bayes–Rule)**.**

$$X : (\Omega, \mathcal{A}) \longrightarrow (\mathbb{R}^{d_1}, \mathcal{B}_{d_1})$$
$$Y : (\Omega, \mathcal{A}) \longrightarrow (\mathbb{R}^{d_2}, \mathcal{B}_{d_2}),$$

$d_1 + d_2 < \infty$. *Suppose that* $Z = (X, Y)$ *has density* $p(z) = p(x, y)$. *Then*

$$\mathbf{P}(X \in B|Y = y) = \frac{\int_B p(x, y)dx}{\int_{\mathbb{R}^{d_1}} p(x, y)dx}$$

for all $B \in \mathcal{B}_{d_1}$.

Proof. Exercise 8.4. □

Regular conditional probabilities

Consider sigma-algebra $\mathcal{F} \subset \mathcal{A}$ and conditional probability $\mathbf{P}(A|\mathcal{F})(\omega)$. We have a mapping

$$\mu : (\mathcal{A} \times \Omega) \longrightarrow [0, 1],$$
$$\mu(A, \omega) = \mathbf{P}(A|\mathcal{F})(\omega),$$

so that

1. $\forall A \in \mathcal{A}, \omega \to \mu(A, \omega)$ is $\mathcal{F}-$measurable random variable indexed by ω.

We would also like to have

2. $\forall \omega \in \Omega$; $A \longrightarrow \mu(A, \omega)$ is a probability on A.

But there is a problem: Note that the relation

$$\lim_{n\to\infty} \sum_{k=1}^{n} \mu(A_k, \omega) = \mu(\bigcup_{k=1}^{\infty} A_k, \omega) \tag{8.5}$$

for pairwise disjoint $A_1, A_2, \cdots \in \mathcal{A}$ merely holds for $\omega \in \Omega_0$ with $\mathbf{P}(\Omega_0) = 1$. Although these are "almost all ω", the set Ω_0 where Equation (8.5) holds depends on $A_1, A_2 \ldots$. Now μ would have to be modified on $\Omega_0^{\complement}$ in order to render Equation (8.5) correct for *all* ω. We then have to repeat this for any sequence $(A_1, A_2, \ldots)$ of measurable and pairws. disjoint sets. There are uncountably many such sequences, hence uncountably many "problem sets" $\Omega_0^{\complement}$, and their union might have nonzero measure.

Theorem 8.9. *Let $X : (\Omega, \mathcal{A}) \to (\mathbb{R}^d, \mathcal{B}_d)$, $d = \infty$ permitted, $\mathcal{F} \subset \mathcal{A}$ sigma algebra. Then the conditional distribution*

$$P_X(B|\mathcal{F}) := \mathbf{P}(\{\omega; X(\omega) \in B|\mathcal{F})$$

has a regular version *$\mu : (\mathcal{B}_d \times \Omega) \to [0, 1]$, that is for any $B \in \mathcal{B}_d$ the equation $P_X(B|\mathcal{F})(\omega) = \mu(B, \omega)$ holds, provided $\omega \in \Omega_B$, where $\mathbf{P}(\Omega_B) = 1$, and μ satisfies conditions (1,2) at the beginning of this paragraph.*

Proof. See [Bre73], theorem 4.34. The structure of $\mathcal{B}_d$ enters in an essential way. □

Exercises for Section 8

Exercise 8.1. Prove that the mapping F in the proof of Theorem 8.2 is a probability on $\mathcal{F}$.

Exercise 8.2. Prove Lemma 8.4.

Exercise 8.3. Prove Lemma 8.6.

Exercise 8.4. Prove Lemma 8.8.

9 The Central Limit Theorem

Let $X_1, X_2, \ldots$ random variable, $X_k : \Omega \longrightarrow \mathbb{R}$, independent with identical distribution (i.i.d.), $\mathbb{E}(X_k) = m$, $\mathbb{E}((X_k - m)^2) = \sigma^2 < \infty$. Then $\mathbb{E}\left(\sum_{k=1}^{n}(X_k - m)\right) = 0$, $\mathbb{E}\left((\sum_{k=1}^{n}(X_k - m))^2\right) = n \cdot \sigma^2$, so $Z_n := \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{X_k - m}{\sigma}$ has mean zero and unit variance. Use your computer to play around with specific examples to convince yourselves that Z_n does not converge pointwise

(or any strong sense). On the other hand, if for example $X_k = 1$ or -1, both with probability $1/2$, it can be shown that

$$\begin{aligned}\mathbf{P}(Z_n \leq z) &= P_{Z_n}((-\infty, z]) \\ &\longrightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{1}{2}x^2\right) dx \qquad \text{(standard normal).}\end{aligned}$$

Note that this does *not* involve the convergence of random variables but rather of *distributions*, that is probabilities on $(\mathbb{R}, \mathcal{B})$; write $\mathcal{D}$ for the family of all distributions on $(\mathbb{R}, \mathcal{B})$. For $\mu_1, \mu_2, \cdots \in \mathcal{D}$ we could define

$$\mu_n \xrightarrow{\mathcal{D}} \mu \in \mathcal{D} \quad \text{if} \quad \mu_n(A) \longrightarrow \mu(A) \forall A \in \mathcal{B},$$

but this is too strong. Let for example μ_n = uniform measure of $[0, 1/n]$, then $\mu_n(\{0\}) = 0 \ \forall n$ but clearly we would like

$$\mu_n \xrightarrow{\mathcal{D}} \delta_0,$$

which according to our too strong definition would necessitate $\mu_n(\{0\}) \to 1$. The following definition of "$\xrightarrow{\mathcal{D}}$" is weaker.

Definition 9.1.

1. If $\mu \in \mathcal{D}$, call $F_\mu(z) := \mu((-\infty, z])$ the *Cumulative Distribution Function* (CDF) of μ.

2. We will say that $\mu_n \xrightarrow{\mathcal{D}} \mu$ (read "μ_n converges in distribution to μ") if one of the equivalent condition holds (we will not prove this):

   (a) $\int \phi d\mu_n \longrightarrow \int \phi d\mu$ for any bounded and continuous function ϕ on $\mathbb{R}$.

   (b) $F_{\mu_n}(z) \longrightarrow F_\mu(z)$ at every z where F_μ continuous.

Note that the integral in item 2a is always well defined because ϕ is assumed bounded. Further, it can be shown that two distributions μ_1 and μ_2 agree if

$$\int \phi d\mu_1 = \int \phi d\mu_2$$

for any bounded and continuous function ϕ on $\mathbb{R}$. This implies that the limit in distribution is well defined.

Definition 9.2. The characteristic function of $\mu \in \mathcal{D}$ is

$$f(u) := \int e^{iux} d\mu(x) \qquad u \in \mathbb{R}.$$

This is well defined since $|e^{iux}| = 1$.

Lemma 9.3. *The characteristic function f of some $\mu \in \mathcal{D}$ has the following properties:*

1. *f is uniformly continuous.*
2. *$f(0) = 1$.*
3. *$|f(u)| \leq 1$.*
4. *$f(-u) = \overline{f}(u)$.*

Proof. Only item 1 is nontrivial

$$\begin{aligned} |f(u+h) - f(u)| &= |\int e^{i(u+h)x} - e^{iux}) d\mu(x)| \\ &= |\int (e^{ihx} - 1) e^{iux} d\mu(x)| \\ &\leq \int |(e^{ihx} - 1)| d\mu(x) =: \delta(h), \end{aligned}$$

and $\delta(h) \to 0$ if $h \to 0$ by the bounded convergence theorem. □

Theorem 9.4. *Let f_k be the characteristic function of μ_k, $k = 1, 2$. Then $f_1 = f_2$ implies $\mu_1 = \mu_2$.*

Proof. We will show that $\int \phi d\mu_1 = \int \phi d\mu_2$ for any bounded and continuous function ϕ on $\mathbb{R}$. If

$$\int \exp(iux) d\mu_1 = \int \exp(iux) d\mu_2,$$

then $\int g d\mu_1 = \int g d\mu_2$ for

$$g(x) = \sum_{\text{finite } k} \alpha_k e^{iu_k x}. \tag{9.1}$$

Now let $\epsilon_n \leq 1$, $\epsilon_n \to 0$. Any bounded and continuous function ϕ can be approximated uniformly on $[-n, n]$ by functions g_n of the form (5.3), so

$$|\phi(x) - g_n(x)| \leq \epsilon_n$$

for $x \in [-n, n]$. But since $\phi(x) \leq M$ and $\epsilon_n \leq 1$, and since g is periodic, we have

$$|g_n| \leq M + 1.$$

So (9.1) gives

$$\int g_n d\mu_1 = \int g_n d\mu_2$$

and hence

$$\begin{aligned}\int \phi d\mu_1 &= \int g_n d\mu_1 + \int \phi - g_n d\mu_1 \\ &= \int g_n d\mu_2 + \int \phi - g_n d\mu_1 \\ &= \int \phi d\mu_2 + \int \phi - g_n d\mu_1 + \int g_n - \phi d\mu_2.\end{aligned}$$

It remains to prove that by choosing n large enough, $\int \phi - g_n d\mu_1$ and $\int \phi - g_n d\mu_2$ can be made as small as we want. Fix $\epsilon > 0$. Note that for any $z > 0$ we have

$$\begin{aligned}\int |\phi - g_n| d\mu_1 &= \int_{[-z,z]} |\phi - g_n| d\mu_1 + \int_{[-z,z]^c} |\phi - g_n| d\mu_1 \\ &\leq \int_{[-z,z]} |\phi - g_n| d\mu_1 + (M+1)\mu_1([-z,z]^c).\end{aligned}$$

Since $\mu_1([-z,z]^c) \to 0$ for $z \to \infty$, we can pick z so large that $(M + 1)\mu_1([-z,z]^c) \leq \frac{\epsilon}{2}$. Next, pick n_0 so large that $z \leq n$ and further $\epsilon_n \cdot 2z \leq \frac{\epsilon}{2}$ whenever $n \geq n_0$. Using this in the estimate above gives

$$\int |\phi - g_n| d\mu_1 \leq \epsilon$$

whenever $n \geq n_0$. The reasoning for the integral $\int |\phi - g_n| d\mu_2$ is the same $\square$

Theorem 9.4 demonstrates that no two distributions can have the same characteristic function. The following theorem strengthens this and demonstrates why characteristic functions are so important.

Theorem 9.5. *Assume that $f_n, n \in \mathbb{N}$ are characteristic functions of some distributions μ_n, and further that $f_n \to f$ pointwise if $n \to \infty$, where f is also the characteristic function of some distribution μ. Then $\mu_n \xrightarrow{\mathcal{D}} \mu$.*

Proof. (The proof will be incomplete.) We know that $\int \phi \mathrm{d}\mu_n \to \int \phi \mathrm{d}\mu$ for every function ϕ of the form $\phi(x) = \exp(iux)$ for some u, so we can assume that this convergence takes place if ϕ is of the form (9.1) in the proof of Theorem 9.4. We have to show that this implies $\int \phi \mathrm{d}\mu_n \to \int \phi \mathrm{d}\mu$ for every continuous and bounded function ϕ on $\mathbb{R}$. We will do this under the *wrong* assumption that all μ_n are concentrated on a compact interval I. So let ϕ be a continuous and bounded function on I, and let $\epsilon > 0$. (All integrals that follow will be over I.) As in the proof of Theorem 9.4 we can find a function g of the form (9.1) in the proof of Theorem 9.4 so that

$$\sup_{x \in I} |g(x) - \phi(x)| \leq \frac{\epsilon}{3}.$$

With g chosen, take n_0 so large that

$$|\int g \mathrm{d}\mu_n - \int g \mathrm{d}\mu| \leq \frac{\epsilon}{3}$$

for any $n \geq n_0$. This implies

$$\begin{aligned}
&|\int \phi \mathrm{d}\mu_n - \int \phi \mathrm{d}\mu| \\
&\leq |\int (\phi - g) \mathrm{d}\mu_n + \int g \mathrm{d}\mu_n - \int g \mathrm{d}\mu + \int (g - \phi) \mathrm{d}\mu| \\
&\leq |\int (\phi - g) \mathrm{d}\mu_n| + |\int g \mathrm{d}\mu_n - \int g \mathrm{d}\mu| + |\int (g - \phi) \mathrm{d}\mu| \\
&\leq \epsilon.
\end{aligned}$$

□

Two remarks about this:

1. The wrong assumption of compactness can be removed, because from the fact that $f_n \to f$ it is possible to show (with some further work) that for any $\epsilon > 0$ there exists a compact interval I_ϵ so that $\mu_n(I_\epsilon^{\mathsf{c}}) < \epsilon$ for all n.

2. Note that in our version of the theorem, we do not only assume that $f_n \to f$ but also that f is a characteristic function. In general, a pointwise limit of characteristic functions *need not be* a characteristic function! If we don't assume that f is a characteristic function, a compensating assumption has to be made.

We will use Theorem 9.5 to prove the Central Limit Theorem. The following Taylor expansion is the key:

Lemma 9.6. *Let $\mathbb{E}(X^2) < \infty$, then*

$$f(u) = 1 + iu \cdot \mathbb{E}(X) - \frac{1}{2}u^2\mathbb{E}(X^2) + \delta(u)u^2,$$

with $\delta(u) \to 0$ as $u \to 0$.

Proof.

$$\exp(iux) = 1 + iux - \frac{1}{2}(ux)^2 + \frac{1}{2}(ux)^2 \cdot \varphi(u,x),$$

with $\varphi(u,x) = \cos(\theta_1(u,x)) + i\sin(\theta_2(u,x)) - 1$, $\theta_1, \theta_2 \in \mathbb{R}$, $|\theta_n| \leq 1$. Further $\mathbb{E}\left((uX)^2\varphi(u,X)\right) = u^2 \cdot \mathbb{E}(X^2 \cdot \varphi(u,X))$ so we get that $(uX)^2\varphi(u,X)$ is integrable for every u. Further $|X^2 \cdot \varphi(u,X)| \leq 3X^2$, and $\varphi(u,X) \to 0$ if $u \to 0$, so we get the statement by the dominated convergence theorem. □

The Central Limit Theorem

Consider the situation at the beginning of this section: $X_1, X_2, \ldots$ are random variable, $X_k : \Omega \longrightarrow \mathbb{R}$, independent with identical distribution μ, and $\mathbb{E}(X_k) = m$, $\mathbb{E}((X_k - m)^2) = \sigma^2 < \infty$. We consider the characteristic function of $Z_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{X_k - m}{\sigma}$:

$$\begin{aligned}
f_n(u) &= \mathbb{E}\left(\exp\{i \cdot u \cdot \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{X_k - m}{\sigma}\}\right) \\
&= \mathbb{E}\left(\prod_{k=1}^{n} \exp\{i \cdot u \cdot \frac{X_k - m}{\sqrt{n} \cdot \sigma}\}\right) \\
&= \left[\mathbb{E}\left(\exp\{iu\frac{X_k - m}{\sqrt{n}\sigma}\}\right)\right]^n \\
&\overset{9.6}{=} [1 - \frac{1}{n}\left(\frac{1}{2} \cdot u^2 + \delta(\frac{u}{\sqrt{n}} \cdot u^2)\right)]^n \\
&\longrightarrow e^{-\frac{1}{2}u^2}
\end{aligned}$$

for $n \to \infty$. A simple calculation gives that this is the characteristic function of the standard normal distribution, i.e.

$$\frac{1}{\sqrt{2\pi}} \int \exp(-iux) e^{-\frac{1}{2}x^2} \, \mathrm{d}x = e^{-\frac{1}{2}u^2}.$$

Using Theorem 9.5, we arrive at the following

Theorem 9.7. *If $X_1, X_2, \ldots$ are random variable, $X_k : \Omega \longrightarrow \mathbb{R}$, independent with identical distribution μ, $\mathbb{E}(X_k) = m$, and $\mathbb{E}((X_k - m)^2) = \sigma^2 < \infty$. Then the characteristic function of $Z_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{X_k - m}{\sigma}$ converges pointwise to the characteristic function of the standard normal distribution.*

10 Literature on measure theory and integration

The following books cover measure theory and integration, mostly somewhat more general than in this chapter. [JP00] is nice and brief, strongly recommended. Some proofs are ommitted. [Dud89] is unusual in that it covers analysis and probability alongside each other, including aspects of functional analysis, measure theory, and advanced aspects of probability theory. The presentation is superb. [Hal74] is an absolute classic. Halmos' fame as a mathematical expositor began with this book. Focusses on measures on locally compact spaces which is somewhat outdated. [Doo94] a very consise text which nonetheless covers everything that is important.

Concerning probability theory, I recommend the following. [Kle14] a modern accout of measure theory which touches upon many aspects of probability theory as well. For an introductory text it is often somewhat too concise. [Bre73] A classic in probability theory. Written in Breiman's very personal but highly readable style, it gives a wonderful introduction to the subject, and whoever thinks it "too theoretical" should look at Breiman's later career. This book does not cover measure theory and integration in detail though. [Fel66, Fel70] Feller's two books on probability theory are even more classic in probability theory than [Bre73]. Again, does not cover measure theory and integration in detail.

For data assimilation, I believe that [Jaz70] is a good introduction, albeit not a rigorous account, and written by an engineer rather than an atmospheric scientist. It's a must–have though for everyone working in data assimilation.

Finally, there is a growing amount of very decent lecture notes available on the internet, for instance

Daniel Ocone's homepage:
`http://www.math.rutgers.edu/~ocone`

Stefan Grossinsky's homepage:
`http://homepages.warwick.ac.uk/~masgav`

Pavel Chigansky's homepage:
`http://pluto.huji.ac.il/~pchiga/teaching.html`

11 Solutions to selected exercises

Exercise 2.1

1. For a set Ω the power set 2^Ω of Ω is the set of all of its subsets. Following Definition 3, we need to verify three properties for the power set to be a sigma algebra. First, since $\varnothing$ is a subset of Ω, $\varnothing \in 2^\Omega$. Second, if $A \in 2^\Omega$ then $A \subset \Omega$ which implies that $A^\mathsf{c} \subset \Omega$. This means that $A^\mathsf{c} \in 2^\Omega$. Finally for countably many elements $A_1, A_2, \cdots \in 2^\Omega$, we have $A_1, A_2, \cdots \subset \Omega$, hence $\cup_{k=1}^\infty A_k \subset \Omega$. This means that $\cup_{k=1}^\infty A_k \in 2^\Omega$ as required.

2. Suppose $\mathcal{S}_1, \mathcal{S}_2$ are sigma algebras. Following Definition 3, we need to verify three properties for $\mathcal{S}_1 \cap \mathcal{S}_2$ to be a sigma algebra. First, since $\varnothing \in \mathcal{S}_1$ and $\varnothing \in \mathcal{S}_2$ we have that $\varnothing \in \mathcal{S}_1 \cap \mathcal{S}_2$. Second, if $A \in \mathcal{S}_1 \cap \mathcal{S}_2$ then $A \in \mathcal{S}_k$, $k = 1, 2$. Since $\mathcal{S}_k$, $k = 1, 2$ are sigma algebras, we deduce that $A^c \in \mathcal{S}_k$, $k = 1, 2$ which further implies that $A^c \in \mathcal{S}_1 \cap \mathcal{S}_2$. Finally, let $A_1, A_2, \cdots \in \mathcal{S}_1 \cap \mathcal{S}_2$. Then, $A_1, A_2, \cdots \in \mathcal{S}_k$, $k = 1, 2$. Since $\mathcal{S}_k$, $k = 1, 2$ are sigma algebras we deduce that $\cup_{j=1}^\infty A_j \in \mathcal{S}_k$, $k = 1, 2$ which further implies that $\cup_{j=1}^\infty A_j \in \mathcal{S}_1 \cap \mathcal{S}_2$. In conclusion the three required properties for a sigma algebra are satisfied.

3. For a set Ω, let $\mathcal{A}$ be an arbitrary family of subsets of Ω. We define $\mathfrak{F}$ to be the family of all sigma algebras on Ω that contain the family $\mathcal{A}$ of subsets of Ω. The power set 2^Ω by definition contains $\mathcal{A}$ and from the previous item it is a sigma algebra on Ω. Hence, $2^\Omega \in \mathfrak{F}$. So $\mathfrak{F}$ contains at least one sigma algebra (and maybe more). We take the intersection of all these sigma algebras and call the result $\bar{\mathcal{A}}$. For sure, $\bar{\mathcal{A}} \supset \mathcal{A}$. But since the intersection of sigma algebras is also a sigma algebra we deduce that $\bar{\mathcal{A}}$ is a sigma algebra on Ω containing $\mathcal{A}$. Further, $\bar{\mathcal{A}}$ is contained in any other sigma algebra in $\mathfrak{F}$ and is therefore the smallest sigma algebra containing $\mathcal{A}$.

Exercise 2.2

1. Since $\varnothing$ and Ω are disjoint, the additivity property implies:

$$P(\varnothing) + P(\Omega) = P(\varnothing \cup \Omega) = P(\Omega),$$

and since $P(\Omega) = 1 < \infty$ we deduce that $P(\varnothing) = 0$.

2. ($\Rightarrow$) Consider countably many pairwise disjoint sets $A_n, n = 1, 2, \ldots$ in $\mathcal{A}$ so that $\cup A_n$ is in $\mathcal{A}$ as well. We want to show that $\sum_{n=1}^\infty P(A_n) =$

$P(\cup A_n)$. We define:

$$B_n = \cup A_k \setminus (A_1 \cup \cdots \cup A_n)$$

for $n \in \mathbb{N}$. We then have that $B_1 \supset B_2 \supset \ldots$ and $\cap B_n = \varnothing$. Hence, since we assume that continuity at the empty set holds we deduce:

$$\lim_{n\to\infty} P(B_n) = P(\cap B_n) = 0.$$

Furthermore, due to the disjointness of the A_i's, we have

$$0 = \lim_{n\to\infty} P(B_n) = \lim_{n\to\infty} \Big(P(\cup A_k) - \sum_{k=1}^{n} P(A_k)\Big) = P(\cup A_k) - \sum_{k=1}^{\infty} P(A_k),$$

which gives the required equality.

($\Leftarrow$) For countably many sets $A_n, n = 1, 2, \ldots$ in $\mathcal{A}$ such that $A_1 \supset A_2 \supset \ldots$ and $\cap A_k = \varnothing$, we need to show that $\lim_{n\to\infty} P(A_n) = 0$. We define:

$$B_n = A_n \cap A_{n+1}^c$$

for $n \in \mathbb{N}$. We then have that for $i \neq j$, $B_i \cap B_j = \varnothing$. Moreover, $\cup B_k = \cup A_k = A_1 \in \mathcal{A}$ and from the sigma additivity of the family of subsets $\{B_n\}_{n\in\mathbb{N}}$ we also have that $\sum_{k=1}^{\infty} P(B_k) = P(\cup B_k) = P(A_1)$. So we have:

$$\begin{aligned} 0 &= \lim_{n\to\infty} \Big(P(\cup B_k) - \sum_{k=1}^{n} P(B_k)\Big) \\ &= \lim_{n\to\infty} P(\cup B_k \setminus (B_1 \cup \cdots \cup B_n)) \\ &= \lim_{n\to\infty} P(A_1 \setminus (A_1 \cap A_{n+1}^c)) \\ &= \lim_{n\to\infty} P(A_{n+1}), \end{aligned}$$

namely, $\lim_{n\to\infty} P(A_n) = 0$ as required.

3. Using the previous item we will show that sigma additivity is equivalent to continuity from above.

   ($\Rightarrow$) Consider countably many sets A_n, $n \in \mathbb{N}$ with $A_1 \supset A_2 \supset \ldots$. We define

$$B_n = A_n \setminus A_{n+1}$$

for all $n \in \mathbb{N}$. We have then that for $i \neq j$, $B_i \cap B_j = \varnothing$ and $\cup B_k = \cup A_k = A_1 \in \mathcal{A}$. Moreover, using the sigma additivity of the family $\{B_n\}_{n\in\mathbb{N}}$, we have $\sum P(B_k) = P(\cup B_k)$. Furthermore,

$$\begin{aligned} 0 &= \lim_{n\to\infty} \Big(P(\cup B_k) - \sum_{k=1}^{n} P(B_k) \Big) \\ &= \lim_{n\to\infty} P(\cup B_k \setminus (B_1 \cup \cdots \cup B_n)) \\ &= \lim_{n\to\infty} P(A_1 \setminus (A_1 \cap A_{n+1}^c)) \\ &= \lim_{n\to\infty} P(A_{n+1}). \end{aligned}$$

($\Leftarrow$) Consider sets $\{A_n\}_{n\in\mathbb{N}}$ with $A_i \cap A_j = \varnothing$ for $i \neq j$. We define:

$$B_n = \cup_{k=n}^{\infty} A_k$$

for all $n \in \mathbb{N}$. Then we have that $B_1 \supset B_2 \supset \ldots$ and $\cup B_k = \cup A_k$. Moreover, it is not difficult to verify that

$$\cap B_n = \cap_{n=1}^{\infty} (\cup_{k=n}^{\infty} A_k) = \varnothing.$$

From the continuity from above property on the family $\{B_n\}_{n\in\mathbb{N}}$ we have that $\lim_{n\to\infty} P(B_n) = P(\cap B_n) = 0$. Furthermore,

$$\cup A_k = \Big(\cup_{k=1}^{n} A_k \Big) \cup \Big(\cup_{k=n+1}^{\infty} A_k \Big) = \cup_{k=1}^{n} \cup B_{n+1},$$

and therefore

$$P(\cup A_k) = \sum_{k=1}^{n} P(A_k) + P(B_{n+1}), \forall n \in \mathbb{N}.$$

By taking the limit $n \to \infty$, we get

$$P(\cup A_k) = \sum_{k=1}^{\infty} P(A_k) + P(\cap B_k) = \sum_{k=1}^{\infty} P(A_k).$$

4. Using a previous item we will show that sigma additivity is equivalent to continuity from below.

   ($\Rightarrow$) Consider countably many sets with $A_1 \subset A_2 \subset \ldots$ and $\cup A_k \in \mathcal{A}$. We define:

$$B_n = A_n \setminus (A_1 \cup \cdots \cup A_{n-1})$$

for all $n \in \mathbb{N}$. Then we have that for $i \neq j$, $B_i \cap B_j = \varnothing$ and $\cup B_k = \cup A_k$ and $\cup_{j=1}^{k} B_j = A_k$. So we have that $P(A_k) = \sum_{j=1}^{k} P(B_j)$ which implies that:

$$\lim_{k\to\infty} P(A_k) = \sum_{k=1}^{\infty} P(B_k) = P(\cup B_k) = P(\cup A_k),$$

where for the second equality we have used the sigma additivity for the family of sets $\{B_n\}_{n\in\mathbb{N}}$.

($\Leftarrow$) Consider countably many pairwise disjoint sets A_n, $n \in \mathbb{N}$. We define:

$$B_n = A_1 \cup \cdots \cup A_n$$

for all $n \in \mathbb{N}$. Then we have that $B_1 \subset B_2 \subset \ldots$ and $\cup B_k = \cup A_k$. Furthermore:

$$\begin{aligned}\sum_{n=1}^{\infty} P(A_n) &= \lim_{n\to\infty} \sum_{k=1}^{n} P(A_k) = \lim_{n\to\infty} P(A_1 \cup \cdots \cup A_n) \\ &= \lim_{n\to\infty} P(B_n) = P(\cup B_k) = P(\cup A_k),\end{aligned}$$

where the fourth equality follows from the continuity from below property of the family $\{B_i\}_{i\in\mathbb{N}}$ of subsets.

5. For $A_1, A_2, \ldots$ countably many disjoint sets in $\mathcal{A}$. Then from the additivity property we have for all $n \in \mathbb{N}$:

$$\sum_{k=1}^{n} P(A_k) = P(\cup_{k=1}^{n} A_k) \leq P(\Omega) = 1.$$

Since for all $n \in \mathbb{N}$, $\sum_{k=1}^{n} P(A_k) \leq 1$ by taking the limit as $n \to \infty$ we have:

$$\sum_{n=1}^{\infty} P(A_n) < \infty,$$

i.e. the series converges. This implies that $P(A_n) \to 0$ as $n \to \infty$.

Exercise 3.1

1. We check Definition 3.
   (a) $\varnothing \in \mathcal{B}$, and $f^{-1}(\varnothing) = \varnothing$, so $\varnothing \in \mathcal{A}_0$.
   (b) Let $A \in \mathcal{A}_0$. Then $\exists B \in \mathcal{B} : f^{-1}(B) = A$. Now $B^{\mathsf{c}} \in \mathcal{B}$ and $f^{-1}(B^{\mathsf{c}}) = f^{-1}(B)^{\mathsf{c}} = A^{\mathsf{c}}$, hence $A^{\mathsf{c}} \in \mathcal{A}_0$.

(c) If $A_1, A_2, ... \in \mathcal{A}_0$, then $\exists B_1, B_2, ... \in \mathcal{B}$ so that $f^{-1}(B_k) = A_k$. Therefore $\bigcup_k A_k = \bigcup_k f^{-1}(B_k) = f^{-1}(\underbrace{\bigcup_k B_k}_{\in \mathcal{B}}) \in \mathcal{A}_0$.

2. We check Definition 3.

   (a) Since $f^{-1}(\varnothing) = \varnothing \in \mathcal{A}$, $\varnothing \in \mathcal{B}_0$.

   (b) If $f^{-1}(B) \in \mathcal{A}$, then $f^{-1}(B^{\mathsf{c}}) = f^{-1}(B)^{\mathsf{c}} \in \mathcal{A}$. This shows $B \in \mathcal{B}_0 \implies B^{\mathsf{c}} \in \mathcal{B}_0$.

   (c) If $f^{-1}(B_k) \in \mathcal{A}\ \forall k \in \mathbb{N}$, then $f^{-1}(\bigcup_k B_k) = \bigcup_k f^{-1}(B_k) \in \mathcal{A}$. This shows $B_1, B_2, ... \in \mathcal{B}_0 \implies \bigcup_k B_k \in \mathcal{B}_0$.

3. If $\mathcal{B}_0$ contains $\mathcal{B}$, then $f^{-1}(B) \in \mathcal{A}$ for all sets $B \in \mathcal{B}$. Hence f is a random variable.

4. Let $\mathcal{D}$ be the sets of the form $\{x \in \mathbb{R}, x > a\}$, and $\mathcal{B}_0$ as in item 2. We know by assumption $\mathcal{D} \subset \mathcal{B}_0$. Since $\mathcal{B}_0$ is a σ-algebra by (2), we have $\sigma(\mathcal{D}) \subset \mathcal{B}_0$. But by (4.2), $\sigma(\mathcal{D}) = \mathcal{B}$. Hence $\mathcal{B} \subset \mathcal{B}_0$. It follows from (3) that f is a random variable.

Exercise 3.3

1. You can find simple $\tilde{g} \leq f$ so that $\int \tilde{g} \mathrm{d}\mathbb{P}$ is arbitrarily close to $\sup \int g \mathrm{d}\mathbb{P}$ in the theorem. Hence if $c < \sup \int g \mathrm{d}\mathbb{P}$, you could find g so that $c < \int g \mathrm{d}\mathbb{P}$, violating the statement, i.e. the statement implies $c \geq \sup \int g \mathrm{d}\mathbb{P}$. On the other hand, since all f_n are simple and no greater than f, we must have $c \leq \sup \int g \mathrm{d}\mathbb{P}$.

2. $f_n - g$ is measurable, so $M_n = \{f_n - g > -\epsilon\}$ are measurable sets. $M_1 \subset M_2 \subset ...$ follows because f_n is monotone increasing. Suppose that ω were in none of the M_n, then

$$f_n(\omega) \leq g(\omega) - \epsilon \leq f(\omega) - \epsilon$$

   $\forall n$, so $f_n(\omega) \nrightarrow f(\omega)$, which is a contradiction. Hence $\bigcap_n M_n = \varnothing \implies \bigcup_n M_n = \Omega$.

3. We know that f_n, g and $\mathbf{1}_{M_n}$ are simple, so $f_n \cdot \mathbf{1}_{M_n}, g \cdot \mathbf{1}_{M_n}, \sum \mathbf{1}_{M_n}$ are too. Now $f_n \geq f_n \cdot \mathbf{1}_{M_n} \geq (g - \epsilon) \cdot \mathbf{1}_{M_n}$ due to the definition of M_n. Now the relation (3.7) follows from monotonicity.

4. $\mathbb{P}(M_n) = \mathbb{P}(\bigcup_{k=1}^{n} M_k) \to \mathbb{P}(\bigcup_{k}^{\infty} M_k) = 1$. Now let $g\mathbf{1}_{M_n} = \sum_{k=1}^{m} g_k \cdot \mathbf{1}_{B_k \cap M_n}$. By the same argument as above we get $\mathbb{P}(B_k \cap M_n) = \mathbb{P}(\bigcup_{l=1}^{n}(B_k \cap M_k)) \xrightarrow{n\to\infty} \mathbb{P}(B_k)$. Hence $\int \mathbf{1}_{M_n} g \mathrm{d}\mathbb{P} = \sum_{k=1}^{m} g_k \cdot \mathbb{P}(B_k \cap M_n) \xrightarrow{n\to\infty} \sum_{k=1}^{m} g_k \cdot \mathbb{P}(B_k) = \int g \mathrm{d}\mathbb{P}$.

Exercise 3.4

Put $A_n = \{\omega; f(\omega) > \frac{1}{n}\}$, then for $\omega \in A_n$; $n \cdot f(\omega) \geq 1$, and if $\omega \notin A_n$, $n \cdot f(\omega) \geq 0$, so $n \cdot f(\omega) \geq \mathbf{1}_{A_n}(\omega)\ \forall \omega \in \Omega$. This gives $0 = n \int f \mathrm{d}\mathbb{P} \geq n \cdot \int \mathbf{1}_{A_n} \mathrm{d}\mathbb{P} = \mathbb{P}(A_n)$, so

$$\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \lim_{m\to\infty} \mathbb{P}(\bigcup_{n=1}^{m} A_n) \leq \sum_{n=1}^{m} \mathbb{P}(A_n) = 0.$$

But if $f(\omega) > 0$ for some ω, then $f(\omega) > \frac{1}{n}$ for some n, hence $\omega \in A_n$ for some n, hence $\omega \in \bigcup_{n\in\mathbb{N}} A_n$, but this set has probability zero.

Exercise 4.1

1. We shall check the three defining properties of probability one by one.

   (a) Note that $T^{-1}(\Omega_2) = \Omega_1$. Then $T_\star\mathbb{P}(\Omega_2) = \mathbb{P}(T^{-1}(\Omega_2)) = \mathbb{P}(\Omega_1) = 1$, using the assumption that $\mathbb{P}$ is a probability on $(\Omega_1, \mathcal{A}_1)$.

   (b) If $A, B \in \mathcal{A}_2$ and $A \cap B = \phi$ then $T^{-1}(A) \cap T^{-1}(B) = T^{-1}(A \cap B) = T^{-1}(\phi) = \phi$, hence $T^{-1}(A)$ and $T^{-1}(B)$ are disjoint. Therefore by additivity of $\mathbb{P}$, $T_\star\mathbb{P}(A \cup B) = \mathbb{P}(T^{-1}(A \cup B)) = \mathbb{P}(T^{-1}(A) \cup T^{-1}(B)) = \mathbb{P}(T^{-1}(A)) + \mathbb{P}(T^{-1}(B)) = T_\star\mathbb{P}(A) + T_\star\mathbb{P}(B)$, i.e. $T_\star\mathbb{P}$ is additive.

   (c) Suppose $A_k \in \mathcal{A}_2$ for all $k \in \mathbb{N}$ and $A_1 \supseteq A_2 \supseteq \cdots$ with $\cap_{k\in\mathbb{N}} A_k = \phi$. It follows that $T^{-1}(A_1) \supseteq T^{-1}(A_2) \supseteq \cdots$ and $\cap_{k\in\mathbb{N}} T^{-1}(A_k) = T^{-1}(\cap_{k\in\mathbb{N}} A_k) = \phi$. Hence by continuity of $\mathbb{P}$ at ϕ, $T_\star\mathbb{P}(A_k) = \mathbb{P}(T^{-1}(A_k)) \to 0$, so $T_\star\mathbb{P}$ is continuous at ϕ as well.

2. If $f : (\Omega_2, \mathcal{A}_2) \to (\mathbb{R}, \mathcal{B})$ is a random variable, then $f^{-1}(B) \in \mathcal{A}_2$ for all $B \in \mathcal{B}$. Further, if $T : (\Omega_1, \mathcal{A}_1) \to (\Omega_2, \mathcal{A}_2)$ is measurable, then $T^{-1}(A_2) \in \mathcal{A}_1$ for all $A_2 \in \mathcal{A}_2$, in particular if we take $A_2 = f^{-1}(B)$.

Hence $(f \circ T)^{-1}(B) = T^{-1}(f^{-1}(B)) \in \mathcal{A}_1$ for all $B \in \mathcal{B}$, implying that $f \circ T$ is a random variable.

Exercise 4.2

1. Suppose $f : (\Omega_2, \mathcal{A}_2) \to (\mathbb{R}, \mathcal{B})$ is measurable and non-negative. Take $(f_n)_{n\in\mathbb{N}}$ a sequence of simple functions with $f_n \uparrow f$ (e.g. as in step 4 of the integral construction). We have, from theorem 5.1,

$$\int_{\Omega_2} f_n \, \mathrm{d}(T_\star \mathbb{P}) = \int_{\Omega_1} f_n \circ T \, \mathrm{d}\mathbb{P}.$$

   By monotone convergence, the left-hand-side converges to $\int_{\Omega_2} f \mathrm{d}(T_\star \mathbb{P})$. Further, since $f_n \circ T \uparrow f \circ T$, the right-hand-side converges to $\int_{\Omega_1} f \circ T \mathrm{d}\mathbb{P}$, again by monotone convergence. By uniqueness of limits we get

$$\int_{\Omega_2} f \, \mathrm{d}(T_\star \mathbb{P}) = \int_{\Omega_1} f \circ T \, \mathrm{d}\mathbb{P}.$$

2. If $f = f_+ - f_-$ is integrable with respect to $T_\star \mathbb{P}$, then

$$\infty > \int_{\Omega_2} f_+ \, \mathrm{d}(T_\star \mathbb{P}) = \int_{\Omega_1} f_+ \circ T \, \mathrm{d}\mathbb{P}$$

   and

$$\infty > \int_{\Omega_2} f_- \, \mathrm{d}(T_\star \mathbb{P}) = \int_{\Omega_1} f_- \circ T \, \mathrm{d}\mathbb{P}$$

   using part (1). Subtracting the second expression from the first and observing $f_+ \circ T - f_- \circ T = (f_+ - f_-) \circ T = f \circ T$ gives the result.

Exercise 5.1

We first prove the statement "f is a random variable implies f_k are random variables for all $k \in \mathbb{N}$." Fix $k \in \mathbb{N}$ and $B \in \mathcal{B}$, and consider the rectangular cylinder $C := \{x \in \mathbb{R}^\infty : x_k \in B\}$. Then $C \in \mathcal{B}_\infty$ and hence by our assumption $f^{-1}(C) \in \mathcal{A}$. But

$$\begin{aligned} f^{-1}(C) &= \{\omega \in \Omega : f_k(\omega) \in B\} \\ &= f_k^{-1}(B). \end{aligned}$$

Hence $f_k^{-1}(B) \in \mathcal{A}$, implying that f_k is a random variable.

For the converse statement, fix a non-negative integer L, the indices $k_1, \cdots, k_L \in \mathbb{N}$, the Borel sets $B_1, \cdots, B_L \in \mathcal{B}$, and the rectangular cylinder

$$C = \{x \in \mathbb{R}^\infty : x_{k_1} \in B_1, \cdots, x_{k_L} \in B_L\}.$$

Then

$$\begin{aligned} f^{-1}(C) &= \{\omega \in \Omega : f_{k_1}(\omega) \in B_1, \cdots, f_{k_L}(\omega) \in B_L\} \\ &= \cap_{m=1}^L \{\omega \in \Omega : f_{k_m}(\omega) \in B_m\} \\ &= \cap_{m=1}^L f_{k_m}^{-1}(B_m). \end{aligned}$$

Since we have assumed that f_k are random variables for all $k \in \mathbb{N}$ and $\mathcal{A}$ is a sigma-algebra (which is closed under finite and countable intersections), the right-hand-side is in $\mathcal{A}$. So we have shown $f^{-1}(C) \in \mathcal{A}$ for any rectangular cylinder. The conclusion now follows as in Exercise 4.1: $\mathcal{B}_0$, the family of all sets $B \subseteq \mathbb{R}^\infty$ such that $f^{-1}(B) \in \mathcal{A}$, is a sigma-algebra. Since we have shown that $\mathcal{B}_0$ contains all rectangular cylinders, we have $\mathcal{B}_\infty = \sigma(\{C : C \text{ is a rectangular cylinder}\}) \subseteq \mathcal{B}_0$ and in particular the pre-image of any rectangular cylinder is in $\mathcal{A}$.

Exercise 7.2

Following the assumptions we have:

1. For $z \in [0, \theta]$ we have:

$$P(T \leq z) = \prod_{k=1}^n P(x_k \leq z) = \prod_{k=1}^n \int_0^z \frac{1}{\theta} \, \mathrm{d}x = \prod_{k=1}^n \frac{z}{\theta} = \frac{z^n}{\theta^n}.$$

2. From the previous item, we have that:

$$P(T \leq z) = \frac{z^n}{\theta^n} = \int_0^z \frac{n x^{n-1}}{\theta^n} \, \mathrm{d}x$$

for all $z \in [0, \theta]$. Hence, we have that the distribution of T has a density function $q(z) = \frac{n z^{n-1}}{\theta^n}$, $z \in [0, \theta]$.

3. Using the density, we find:

$$E_\theta(T) = \int_0^\theta x \frac{n x^{n-1}}{\theta^n} \, \mathrm{d}x \frac{n}{n+1} \theta$$

so T is biased. From the linearity of E_θ we deduce that putting $\hat{T} = \frac{n+1}{n} T$, we have that $E_\theta(\hat{T}) = \frac{n+1}{n} E_\theta(T) = \theta$, so $\hat{T}$ is unbiased.

Exercise 7.3

1. We have

$$1 = \int_I f(x;\theta)\,\mathrm{d}x = \int_I a(x)e^{\theta h(x)-\phi(\theta)}\,\mathrm{d}x = e^{-\phi(\theta)}\int_I a(x)e^{\theta h(x)}\,\mathrm{d}x.$$

   Hence

$$e^{\phi(\theta)} = \int_I a(x)e^{\theta h(x)}\,\mathrm{d}x \Leftrightarrow \phi(\theta) = \log\Big(\int_I a(x)e^{\theta h(x)}\,\mathrm{d}x\Big).$$

2. We differentiate w.r.t. θ and then use differentiation under the integral sign which is permitted by assumption. We have

$$0 = \int_I a(x)(h(x)-\phi'(\theta))e^{\theta h(x)-\phi(\theta)}\,\mathrm{d}x = \int_I (h(x)-\phi(\theta))f(x;\theta)\,\mathrm{d}x,$$

   which then give us:

$$\int_I h(x)f(x;\theta)\,\mathrm{d}x = \phi'(\theta)\int_I f(x;\theta)\,\mathrm{d}x = \phi'(\theta) = \frac{d\phi}{d\theta}(\theta).$$

   But the left-hand-side of the above equation is by definition equal to $E_\theta(h)$ i.e. $E_\theta(h) = \frac{d\phi}{d\theta}$ as required. Now we differentiate again w.r.t. θ and by using again differentiation under the integral sign we obtain:

$$\begin{aligned}
\frac{d^2\phi}{d\theta^2} &= \int_I h(x)\frac{d}{d\theta}f(x;\theta)\,\mathrm{d}x \\
&= \int_I h(x)(h(x)-\frac{d\phi}{d\theta})f(x;\theta)\,\mathrm{d}x \\
&= \int_I h(x)^2 f(x;\theta)\,\mathrm{d}x - \frac{d\phi}{d\theta}\int_I h(x)f(x;\theta)\,\mathrm{d}x \\
&= E_\theta(h^2) - E_\theta(h)E_\theta(h) \\
&= V_\theta(h).
\end{aligned}$$

   Here we have used the definition of variance and the properties of E_θ.

3. Let $\hat{\theta}$ the MLE for θ. Then $\hat{\theta}$ maximizes the log-likelihood

$$l(\theta) = \log L(\theta, x) = \sum_{k=1}^{n} \log f(x_k;\theta).$$

(Here we have the standard setup 7.4 so we use the log-likelihood.) Hence, we have that

$$\begin{aligned}
0 = \frac{dl}{d\theta}(\hat{\theta}) &= \frac{d}{d\theta}\Big|_{\theta=\hat{\theta}} \sum_{k=1}^{n} \log f(x_k;\theta) \\
&= \sum_{k=1}^{n} \frac{d}{d\theta}\Big|_{\theta=\hat{\theta}} \Big(\log a(x_k) + \theta h(x_k) - \phi(\theta) \Big) \\
&= \sum_{k=1}^{n} h(x_k) - n\frac{d\phi}{d\theta}(\hat{\theta}) \\
\Rightarrow \frac{1}{n}\sum_{k=1}^{n} h(x_k) &= \frac{d\phi}{d\theta}(\hat{\theta}).
\end{aligned}$$

4. From the previous item we have that the derivative of log-likelihood w.r.t. θ is:

$$\frac{dl}{d\theta} = \sum_{k=1}^{n} h(x_k) - n\frac{d\phi}{d\theta}.$$

We differentiate again w.r.t to θ and have:

$$\frac{d^2 l}{d\theta^2} = -n\frac{d^2\phi}{d\theta^2} = -nV_\theta(h),$$

where the last equality follows from the second item. Now $\hat{\theta}$ derived in the previous item (i.e. as a critical point of the log-likelihood) satisfies:

$$\frac{dl}{d\theta}(\hat{\theta}) = 0, \quad \frac{d^2 l}{d\theta^2}(\hat{\theta}) = -nV_{\hat{\theta}}(h) < 0.$$

This implies that the critical point of the log-likelihood is at least a local maximum for the log-likelihood.

5. We calculate:

$$\begin{aligned}
\log f(x_k,\theta) &= \sum_{k=1}^{n} \{\log a(x_k) + \theta h(x_k) - \phi(\theta)\} \\
\partial_\theta(\log f(x_k,\theta)) &= \sum_{k=1}^{n} \{h(x_k) - \phi'(\theta)\} \\
\frac{\partial^2}{\partial\theta^2}(\log f(x_k,\theta)) &= -\sum_{k=1}^{n} \phi''(\theta) = -n\phi''(\theta).
\end{aligned}$$

Then using Lemma 7.9 we have that the Fisher information is given by:

$$I = -E_\theta\Big(- n\phi''(\theta)\Big) = n\phi''(\theta).$$

6. We will be using the law of large numbers which says that if $Y_1, \ldots, Y_n$ independent, identically distributed RV with finite variance then

$$\frac{1}{n}\sum_{k=1}^{n} Y_k \to E(Y_1)$$

almost surely for $n \to \infty$. We note that $\hat{\theta}$ is uniquely determined due to the 1-1 property of $E_\theta(h)$. By assumption $x_1, \ldots, x_n$ are independent and identically distributed and it is easy to see that the same holds for $Y_k = h(x_k)$. Since by assumption $V_\theta(h)$ is finite the law of large numbers implies $\frac{1}{n}\sum_{k=1}^{n} h(x_k) \to E_\theta(h(x_1))$ almost surely for $n \to \infty$. Using items two and three we then deduce that $\frac{\mathrm{d}\phi}{\mathrm{d}\theta}(\hat{\theta}) \to E_\theta(h(x_1))$ almost surely as n goes to infinity. Moreover, by assumption $\frac{\mathrm{d}\phi}{\mathrm{d}\theta}$ is 1-1 and continuous, thus $\hat{\theta}$ converges to θ as $n \to \infty$.

Exercise 8.2

We can assume that f is integrable (otherwise $\mathbb{E}[f|\mathcal{F}]$ is not well defined). We use Remark 8.3 throughout.

1. Let f, g be integrable, $\lambda, \mu \in \mathbb{R}$. Then $\lambda f + \mu g$ is integrable. Put $h := \lambda\mathbb{E}[f|\mathcal{F}] + \mu\mathbb{E}[g|\mathcal{F}]$ then h is $\mathcal{F}$-measurable, and for any $A \in \mathcal{F}$ we have

$$\begin{aligned}\int \mathbf{1}_A \cdot h \mathrm{d}\mathbb{P} &= \int \mathbf{1}_A\{\lambda\mathbb{E}[f|\mathcal{F}] + \mu\mathbb{E}[g|\mathcal{F}]\}\mathrm{d}\mathbf{P} \\ &= \lambda\int \mathbf{1}_A\mathbb{E}[f|\mathcal{F}]\mathrm{d}\mathbb{P} + \mu\int \mathbf{1}_A\mathbb{E}[g|f]\mathrm{d}\mathbf{P} \\ &\overset{(8.3,2)}{=} \lambda\int \mathbf{1}_A f\mathrm{d}\mathbf{P} + \mu\int \mathbf{1}_A g\mathrm{d}\mathbf{P} \\ &= \int \mathbf{1}_A\{\lambda f + \mu g\}\mathrm{d}\mathbf{P}.\end{aligned}$$

This implies $\mathbb{E}[\lambda f + \mu g|\mathcal{F}] = \lambda\mathbb{E}[f|\mathcal{F}] + \mu\mathbb{E}[g|\mathcal{F}]$.

2. Suppose not. Then there is $\epsilon > 0$ so that $A := \{\omega \in \Omega, \mathbb{E}[f|\mathcal{F}] \leq -\epsilon\}$ has positive probability. Further, $A \in \mathcal{F}$ by (8.3,1). By (8.3,2) and

since $f \geq 0$ we get:

$$0 = \int f \cdot \mathbf{1}_A \mathrm{d}\mathbb{P} = \int \mathbb{E}[f|\mathcal{F}]\mathbf{1}_A \mathrm{d}\mathbb{P} \leq -\epsilon\mathbb{P}(A).$$

which is a contraditction.

3. Put $h := \mathbb{E}[\mathbb{E}[f|\mathcal{F}]|\mathcal{Y}]$. Then by Remark 8.3, h is $\mathcal{Y}$-measurable and $\forall A \in \mathcal{Y}$,

$$\int \mathbf{1}_A \mathbb{E}[f|\mathcal{F}]\mathrm{d}\mathbb{P} = \int \mathbf{1}_A h \mathrm{d}\mathbb{P}.$$

But since $A \in \mathcal{F}$ as well, the left hand side is equal to $\int \mathbf{1}_A f \mathrm{d}\mathbb{P}$. By Remark 8.3, $h = \mathbb{E}[f|\mathcal{Y}]$. □.

References

[Alb07] J. Albert. Bayesian computation with r. *Springer*, 2007.

[Bar65] A. A. Barker. Monte carlo calculations of the radial distribution functions for a proton-electron plasma. *Australian Journal of Physics*, 18:119–133, 1965.

[Bol10] W. M. Bolstad. Understanding computational bayesian statistics. *Wiley*, 2010.

[Bre73] Leo Breiman. *Probability*. Addison-Wesley, Reading, Mass, 1973.

[BS00] I. Beichl and F. Sullivan. The metropolis algorithm. *Computing in Science and Engineering*, 2(1):65–69, 2000.

[CSI00] M.-H. Chen, Q.-M. Shao, and J. G. Ibrahim. Monte carlo methods in bayesian computation. *Springer*, 2000.

[Dev86] L. Devroye. Non-uniform random variate generation. *Springer-Verlag*, 1986.

[dM67] A. de Moivre. *The doctrine of chances or, a method of calculating the probabilities of events in play*. Frank Cass & Co., Ltd., London, 1967. Exact reproduction of the 1738 (second) edition.

[Doo94] J.L. Doob. *Measure Theory*. Graduate Texts in Mathematics. Springer New York, 1994.

[Dud89] R.M. Dudley. *Real Analysis and Probability.* Chapman & Hall, 1989.

[Fel66] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. John WIley & Sons, Inc., New York, 1966.

[Fel70] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. John WIley & Sons, Inc., New York, 1970.

[GCSR04] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. Bayesian data analysis. *Chapman & Hall, CRC*, 2004.

[Gey92] C. Geyer. Practical markov chain monte carlo. *Statistical Science*, 7(4):473–483, 1992.

[GL06] D. Gamerman and H. F. Lopes. Markov chain monte carlo: Stochastic simulation for bayesian inference. *Chapman & Hall, CRC*, 2006.

[Hal74] P.R. Halmos. *Measure theory.* Number 18 in Graduate Texts in Mathematics. Springer, 1974.

[Has70] W. Hastings. Monte carlo sampling methods using markov chains and their application. *Biometrika*, 57:97–109, 1970.

[Jaz70] Andrew H. Jazwinski. *Stochastic Processes and Filtering Theory*, volume 64 of *Mathematics in Science and Engineering.* Academic Press, 1970.

[JP00] J. Jacod and P.E. Protter. *Probability Essentials.* Hochschultext / Universitext. Springer, 2000.

[Kle14] Achim Klenke. *probability theory.* Springer, 2014.

[Lap95] Pierre-Simon Laplace. *Philosophical essay on probabilities*, volume 13 of *Sources in the History of Mathematics and Physical Sciences.* Springer-Verlag, New York, 1995. Translated from the fifth (1825) French edition.

[Liu04] J. S. Liu. Monte carlo strategies in scientific computing. *Springer*, 2004.

[MRR+53] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[Owe13] Art B. Owen. *Monte Carlo theory, methods and examples.* 2013.

[Pes73] P. H. Peskun. Optimum monte-carlo sampling using markov chains. *Biometrika*, 60:607–612, 1973.

[RC04] C. P. Robert and G. Casella. Monte carlo statistical methods. *Springer*, 2004.

[RC10] C. P. Robert and G. Casella. Introducing monte carlo methods with r. *Springer*, 2010.

[RK08] R. Y. Rubinstein and D. P. Kroese. Simulation and the monte carlo method. *Wiley*, 2008.

[Sav71] Leonard J. Savage. Elicitation of personal probabilities and expectation. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

[Sch95] M. J. Schervish. Theory of statistics. *Springer Series in Statistics, Springer-Verlag*, 1995.

[SE96] Chib S. and Greenberg E. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1996.

[SS06] D. S. Sivia and J. Skilling. Data analysis: a bayesian tutorial. *Oxford University Press*, 2006.

[vdV00] A.W. van der Vaart. *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[vN51] J. von Neumann. Various techniques used in connection with random digits. monte carlo methods. *Nat. Bureau Standards*, pages 36–38, 1951.